ABSTRACT

Title of dissertation: THE ROLE OF STRUCTURAL INFORMATION IN THE RESOLUTION OF LONG-DISTANCE DEPENDENCIES

Anton Malko
Doctor of Philosophy, 2019

Dissertation directed by: Professor Colin Phillips
Department of Linguistics

The main question that this thesis addresses is: in what way does structural information enter into the processing of long-distance dependencies? Does it constrain the computations, and if so, to what degree? Available experimental evidence suggests that sometimes structurally illicit but otherwise suitable constituents are accessed during dependency resolution. Subject-verb agreement is a prime example (Wagers et al., 2009; Dillon et al., 2013), and similar effects were reported for negative polarity items (NPIs) licensing (Vasishth et al., 2008) and reflexive pronouns resolution (Parker and Phillips, 2017; Sloggett, 2017). Prima facie this evidence suggests that structural information fails to perfectly constrain real-time language processing to be in line with grammatical constraints. This conclusion would fall neatly in line with an assumption that human sentence processing relies on cue-based memory (e.g McElree et al., 2003; Lewis and Vasishth, 2005; Van Dyke and Johns, 2012; Wagers et al., 2009, a.m.o.), the key property of which is the fragility of memory search, which can return irrelevant results if they look similar enough to the

relevant ones. The attractiveness of such an approach lies in its parsimony: there is independent evidence that general purpose working memory is cue-based (Jonides et al., 2008), so we do not need to postulate any language specific mechanisms. Additionally, the processing of multiple linguistic dependencies can be analyzed within the same theoretical framework.

Cue-based approach has also been argued to be the best one in terms of its empirical coverage: some of the experimental evidence was assumed to only be explainable within it (the absence of ungrammaticality illusions in subject-verb agreement is the main example, to which we will return in more detail later). However, recently several other approaches have been suggested which would be able to account for these cases (Eberhard et al., 2005; Xiang et al., 2013; Sloggett, 2017; Hammerly et al., draft.april.2018). These approaches usually assume separate processing mechanisms for different linguistic dependencies, and thus lose the parsimonious attractiveness of cue-based memory models. They also take a different stance on the role of structural information in real-time language processing, assuming that structural cues do accurately guide the dependency resolution. A priori there is no reason why they could not turn out to be true. But given the theoretical attractiveness of cue-based models in which structural information does not categorically restrain processing, it is important to critically evaluate these recent claims. In this thesis, we focus on reflexive pronouns and on the novel pattern reported in Parker and Phillips (2017) and Sloggett (2017): the finding that reflexive pronouns are sensitive to the properties of structurally inaccessible antecedents in some specific conditions (interference effect). The two works report consistent findings, but the accounts

they give take opposite perspectives on the role of structural information in reflexive resolution. Our aim in this thesis is to assess the reliability of these findings and to experimentally investigate cases which would hopefully provide clearer evidence on how the structure guides reflexives processing.

To this aim, we conduct two direct replications of Parker and Phillips (2017) and four novel experiments further investigating the properties of the interference effect. None of the six experiments provided strong statistical support for the previous findings. After ruling out several possible confounds and analyzing numerical patterns (which go in the expected direction and are consistent with previous results), we conclude that interference effect is likely real, but may be less strong than the previous studies would lead to believe. These results can be used for setting more realistic expectations for future studies regarding the size of the effect and statistical power necessary to detect it. With respect to our main goal of distinguishing between cue-based and alternative accounts of the interference effect, we tentatively conclude that cue-based approaches are preferred; however, one has to assume that some structural features are able to categorically rule out illicit antecedents. Further highly powered studies are necessary to verify and confirm these conclusions.

The role of structural information in the resolution of long-distance dependencies

by

Anton Malko

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2019

Advisory Committee:
Professor Colin Phillips, Chair
Assistant Professor Ellen Lau
Professor Jeffrey Lidz
Associate Professor Omer Preminger
Associate Professor Robert Slevc

Acknowledgments

I have finally reached the stage of writing this part of my dissertation. The path to this point has been long and not always straightforward, and I would like to express my gratitude to people who supported me during the process and who made this journey even remotely possible (the list is far from being exhaustive, of course!).

My greatest thank you goes to my advisor, Colin Phillips. Colin patiently guided me through my sometimes confused thoughts and continuously pushed me to re-think my assumptions, to shape my ideas better and to communicate them extremely clearly - all of this without explicitly holding my hand and giving me enough room for trying things out on my own. Without Colin's guidance and support this thesis would truly have been impossible, and I owe him my deepest gratitude.

I greatly enjoyed our discussions and conversations with Ellen Lau. She was always able to provide encouraging advice and also point out to me some unexpected connections between my ideas and existing research. Ellen somehow always managed to make things brighter and clearer than what they had appeared to me - it was such a comforting and stimulating experience.

The research I discuss in this thesis would have been much harder without the logistical support I received. First, I would like to thank those who helped by providing native speaker judgments, helping to prepare experimental stimuli and collecting the data: Julia Buffinton, Hanna Muller, Maggie Kandel, Lalitha Balachandran, Cassidy Wyatt. I express my gratitude to all of them. Thank you also to Kim Kwok, who has been tremendously helpful and supportive in all the administrative questions, especially considering enormous amounts of work she was (and is!) doing. Finally, I want to acknowledge the financial support I received from the Department of Linguistics, Language Science Center and NSF (the research reported in this thesis is based upon work supported by the National Science Foundation under Grant No.1449815).

Thank you to the people who sparked and supported my interest in statistics and modeling. To Michael Dougherty and Kevin O'Grady for teaching my first ever statistic courses in a fun way and making me excited about them - I don't think I would have been able to write a lot of this thesis (especially Chapter 4!) without the foundations they gave to me. To Naomi Feldman, who introduced me to computational psycholinguistics, and to Philip Resnik for his exciting and extremely lucid discussion of the NLP ideas in the computational linguistics class. If I am now interested in computational modeling and am not afraid of MCMC methods, it is because of Naomi and Philip. Also thank you to my fellow grad students, Phoebe Gaston, Hanna Muller and Adam Liter for sharing my interest in statistics, for frequent related discussions we had and for the little quest for understanding the Bayesian approaches we pursued together.

I would have never learned and known the things above if I had not come to Maryland, and I would like to express my gratitude to several people who made this possible. First and foremost, to Natalia Slioussar, who I was very lucky to have as my Master's thesis advisor. My PhD dissertation would not have been even considered if I have not met her in my MA years. Natalia introduced me to (psycho)linguistics in general, being very generous with her time and advice. The current thesis would have probably been very different, if not for our work on agreement attraction in Russian, which has quite directly connected me to the topics I consider here. It was also her who greatly influenced my decision to come to Maryland. I am deeply grateful for all the guidance, support and advice Natalia gave me in the years I have known her. People from NYI 2012 in Saint-Petersburg have shaped my idea of coming to a graduate school abroad at all - most prominently, Sabine Iatridou. And thank you to Nikos Angelopoulos and Christine Boucher for supporting me in this idea and for the great fun we had together in Saint-Petersburg.

My gratitudes list would be incomplete without mentioning my friends from inside and outside of the department, who added a lot of fun and enjoyment to the graduate school experience. Thank you to the Hyattsville gang - Allyson Ettinger, Lara Ehrenhofer, Kasia Hitczenko, Phoebe Gaston, Christian Brodbeck, Paulina Lyskawa, Miloš Nikolić, Suddhasattwa Das, Amit Nag - for all the great experiences we had, including, to name just a few, conversations about everything and anything, kayaking, movie nights, Easter Eggs hunt, playing Settlers of Catan, and several kidnappings. I remember these fondly. Ilia Kurenkov and Natalia Lapinskaya were among the very first people I met having come to Maryland and they remained good friends throughout my time here. Thanks to them I have not forgotten my Russian completely! Ilia was also the person who introduced me to LaTeXand git, and thus indirectly contributed to the creation of this document. Thank you to Sol Lago, Shota Momma, William Matchin, Eric Pelzl, Nick Huang for having many invariably enjoyable conversations throughout these years.

I cannot even start expressing my gratitude towards Lena. I would have hardly survived the final stages of writing without her continuous support; but so much more would have been impossible without her. Thank you for all the lessons you taught me and the love and joy you shared with me.

Finally, thank you to my family for giving me the foundations which allowed me to make it this far, and for supporting me and believing in me all the way through.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

The main question that this thesis addresses is: in what way does structural information enter into the processing of long-distance dependencies? Does it constrain the computations, and if so, to what degree? Available experimental evidence suggests that sometimes structurally illicit but otherwise suitable constituents are accessed during dependency resolution. Subject-verb agreement is a prime example (Wagers et al., 2009; Dillon et al., 2013), and similar effects were reported for negative polarity items (NPIs) licensing (Vasishth et al., 2008) and reflexive pronouns resolution (Parker and Phillips, 2017; Sloggett, 2017). Prima facie this evidence suggests that structural information fails to perfectly constrain real-time language processing to be in line with grammatical constraints. This conclusion would fall neatly in line with an assumption that human sentence processing relies on cue-based memory (e.g. McElree et al., 2003; Lewis and Vasishth, 2005; Van Dyke and Johns, 2012; Wagers et al., 2009, among many others), which has similarity-based interference as a key property. The attractiveness of such an approach lies in its parsimony: there is independent evidence that general purpose working memory is cue-based (Jonides et al., 2008), so we do not need to postulate any language specific mechanisms. Additionally, the processing of multiple linguistic dependencies can be analyzed within the same theoretical framework.

Empirically, some of the experimental evidence was assumed to only be explainable within cue-based approach (the absence of ungrammaticality illusions in subject-verb agreement is the main example, to which I will return in more detail later). However, recently several other approaches have been suggested which would be able to account for these cases, even if we assume that structural information accurately guides the parsing (Eberhard et al., 2005; Hammerly et al., 2018; Xiang et al., 2013; Sloggett, 2017). These approaches usually assume separate mechanisms for different linguistic dependencies, and thus lose the parsimonious attractiveness of cue-based memory models. A priori there is no reason why they could not turn out to be true. But given the theoretical attractiveness of cue-based models in which structural information does not categorically restrain processing, it is important to critically evaluate these recent claims. In this thesis, I focus on reflexive pronouns and on the novel pattern reported in Parker and Phillips (2017) and Sloggett (2017). The two works report consistent findings, but the accounts they give take opposite perspectives on the role of structural information in reflexive resolution. Our aim in this thesis is to experimentally investigate cases which could provide clearer evidence on how structure guides reflexives processing. The rest of this chapter is structured as follows. First, I survey available experimental evidence in subject-verb agreement and reflexive resolution. Then, I describe a prominent cue-based model used to account for these data and show how it can explain them. As we will see, adopting this explanation for the data will make us assume that structural information is not able to categorically rule out illicit dependency formation. I proceed to discuss the recent alternative approaches which assume that structure restricts the processing to be compliant with the grammar. Finally, I describe the cases which would distinguish between the two classes of approaches and outline the experiments I carry out to investigate these cases.

1.1 Empirical evidence

In this section I discuss the evidence for grammatically illicit constituents being considered during real-time processing. I start the discussion with the most popular experimental paradigm used in this line of research, and then I turn to empirical data on subject-verb agreement, NPI licensing and reflexive resolution.

Feature-mismatch paradigm is very widely used in the investigations of long-distance dependencies formation. It is based on the observation that people quickly notice if feature composition of the two dependency elements does not match (for instance, the subject is singular and the verb is plural) (e.g Osterhout and Mobley, 1995; Osterhout et al., 1997). We can use this reaction to assess whether feature mismatch with a grammatically inaccessible, but otherwise plausible, dependency component elicits similar complications. Thus, the paradigm is usually structured as follows. People are presented with sentences containing the dependency of interest and two possible ways to resolve it - one grammatically licensed and the other one not. For example, people may be presented with a sentence like "The key to the cabinet is rusty". This sentence contains two NPs, both of which could in principle control verbal agreement, but only one of them - "the key" - is in the correct structural position to do so, i.e. is a grammatically licensed agreement controller. In what follows, I will call the grammatically licit option “target”, and the grammatically illicit - “lure”. The features of the target and the lure are factorially manipulated, so that they either match or mismatch the other dependency element. In the example above, the number of both NPs could be manipulated so that they match or mismatch the number of the verb. We do expect to observe some indication of processing difficulty in target mismatch conditions as compared to target match. But of critical interest is whether we observe any reaction to feature mismatch with the lure. If we do, we can conclude that at some point in time feature information on the lure had entered the processing system in a way in which it was able to affect dependency resolution process. In what follows, I will refer to any lure-related effects as “lure-match effects” or “interference”. I will talk about “facilitation” (“facilitatory interference”), if the lure-match conditions are read faster than lure-mismatch conditions, and about “inhibition” (“inhibitory interference”) otherwise.

It is important to keep in mind that the interpretation of the results obtained with this paradigm is not always straightforward. Its goal is to assess whether the presence of the lure affects processing in any measurable way: it may be differences in interpretations, reaction times, electrophysiological responses... However, just the fact that a certain constituent affects processing does not mean that it is actively accessed by the parser; the differences we observe may well stem from other sources. Therefore, additional steps must be made in order to ensure that the evidence we obtain is indeed the evidence about access mechanisms. I will return to this discussion after we have surveyed available experimental evidence.

Subject-verb agreement Subject-verb agreement presents a prime example of a dependency for which lure-match effects are observed: people tend to mis-identify the agreement controller in sentences like (1b) - (1a), a phenomenon known as "agreement attraction". Both of the sentences are ungrammatical, since the head of the subject noun phrase ("key") mismatches the verb in number. However, this violation is perceived as milder in the second sentence. This effect is apparent in acceptability judgments and reaction time (RT) profile (with (1b) eliciting higher ratings and faster RTs than (1a)) (Wagers et al., 2009; Dillon et al., 2013; Hammerly et al., 2018), as well as in electrophysiological measures (Xiang et al., 2013; Tanner et al.). I will refer to this effect as "grammaticality illusion": an ungrammatical sentence may have been fleetingly considered grammatical by the readers. Grammaticality illusions have been observed in a number of languages and with multiple grammatical features (e.g., see Lago et al. (2015) for Spanish, Tucker et al. (submitted) for Arabic and Slioussar and Malko (2016) for Russian).

(1) a. *The key to the cabinet are on the table.

b. *The key to the cabinets are on the table.

c. The key to the cabinets is on the table.

d. The key to the cabinet is on the table.

While the presence of lure-match effect in ungrammatical sentences is uncontroversial, the patterns arising in grammatical sentences are less clear. There are three logical possibilities ((1c) is exactly as difficult as (1d); (1c) is easier than (1d); (1c) is harder than (1d)). Each possibility would support a different class of models, as I will discuss in section 1.2.2.

The first possibility - that the processing difficulty of grammatical sentences does not depend on the number-marking of the lure - is commonly assumed to be an actuality, based on a number of studies (e.g. Wagers et al., 2009; Dillon et al., 2013; Tanner et al.). This pattern is known as the “absence of ungrammaticality illusions” (i.e. there are no cases when people perceive grammatical sentences as degraded). However, this observation may be overly general. Two recent meta-analytic reviews suggest that the literature contains a fair amount of evidence for people being faster in (1c) as compared to (1d). Jäger et al. (2017) review reaction time (self-paced, eye-tracking) and electrophysiological studies and find that 12 out of 25 report statistically significant effects in this direction (12 studies report no significant effects and one reports the opposite pattern)[1]. These qualitative conclusions are supported by quantitative Bayesian meta-analysis of reaction time data: it estimates the speed-up in (1c) to be of 6.6ms with the 95% of the posterior probability mass in the interval [-16.2, 3.7] (91% of the probability mass below 0)[2]. Hammerly et al. (draft.april.2018) reach similar conclusions in an informal review of the literature with somewhat less detailed comparison. They review a partially overlapping subset of reaction time studies of agreement attraction; in addition, they also look

[1] I only report the data for configurations with singular subjects, since there is almost no data for configurations with plural subjects.

[2] One has to keep in mind that for eye-tracking, only first-pass reading times were entered into the analysis.

at interpretation studies. Out of 26 tested instances of agreement attraction configurations[3], in 11 cases the authors reported significant effects in both grammatical and ungrammatical sentences. This evidence will be of crucial importance when we choose between possible accounts of agreement attraction in section 1.2.2.

NPIs Negative Polarity Items (NPIs) licensing is another dependency whose formation is arguably prone to interference (Vasishth et al., 2008; Xiang et al., 2008; Parker and Phillips, 2016). NPIs are items like "any", "ever", "lift a finger" and others which can only occur in a special linguistic environment. The exact nature of the environment is still debated , but one generalization is that NPIs have to occur in a scope of a downward entailing element, e.g. negation. It is possible to conceptualize this relation as an item-to-item dependency[4]: when people encounter an NPI, they look for a downward entailing element in structurally higher position, and if they find it, the NPI is licensed. Interestingly for our purposes, sentences with an NPI licensor in an inappropriate structural position (2c) are perceived as better than sentences without a licensor altogether (2b) (Vasishth et al., 2008; Parker and Phillips, 2016).

(2) a. No diplomats have ever supported a drone strike.

[3] Most of them correspond to whole studies, but in a few cases, Hammerly et al. (draft.april.2018) count subsets of stimuli in the same study separately.

[4] Although this conceptualization is contested, and as we will see in the discussion of the evidence, this is the reason why NPI data does not provide strong evidence for or against the use of structural information

b. *The diplomats have ever supported a drone strike.

c. *The diplomats that no congressman could trust have ever supported a drone strike.

Reflexive pronouns I now turn to the evidence which will be most relevant for my thesis - influence of structurally inappropriate antecedents on the resolution of reflexive pronouns. The evidence is rather mixed. On the one hand, a number of experiments in a variety of methodologies (cross-modal priming, eye-tracking, ERPs) failed to provide robust evidence that lures affect reflexives resolution (cross-modal priming: Nicol and Swinney (1989); self-paced reading: Badecker and Straub (2002) (Exp.5,6); eye-tracking: Sturt (2003); Dillon et al. (2013); Cunnings and Sturt (2014); ERPs: Xiang et al. (2008) among others). On the other hand, some studies do report interference effects (Badecker and Straub (2002) (Exp.4), King et al. (2012a); Cunnings and Felser (2013); Patil et al. (2016); Parker and Phillips (2017); Sloggett (2017)). I will discuss these groups of studies in order.

The first study to consider the influence of lures on reflexive resolution was Nicol and Swinney (1989). They report the results of an experiment relying on cross-modal priming paradigm. The participants were auditorily presented with sentences like (3). At the moment when people heard the pronominal, a word appeared on the screen, and people were asked to decide whether it was a real English word. The word could be related to one of the potential antecedents or unrelated to any of them. The logic underlying the experiment is that if during the course of anaphora resolution people reactivate an antecedent, words related to that antecedent would be primed, and thus people reaction times would be faster as compared to the unrelated word conditions. If binding principles are accurately applied, we expect that only words related to "boxer" and "skier" would be primed when people encounter "him" and only words related to "doctor" will be primed when people encounter "himself". This is exactly what Nicol and Swinney report.

(3) The boxer told the skier that the doctor for the team would blame *him* / *himself* for the injury.

These results suggest that people were successfully using structural information to restrict antecedent selection, and a number of subsequent studies fell in line with this conclusion. Sturt (2003) examined sentences like (4) (the examples come from Exp.1 and 2, correspondingly).

(4) a. Jonathan was pretty worried at the City Hospital. The surgeon who treated Jonathan / Jennifer had pricked himself / herself with a used syringe needle.

b. Jonathan / Jennifer was pretty worried at the City Hospital. He / She remembered that the surgeon had pricked himself / herself with a used syringe needle.

Neither Experiment 1 or 2 provided evidence for any detectable effects of the lure in early measures[5]. Dillon et al. (2013) considered sentences like (5). Agreement

[5] Experiment 1 did provide evidence for some late effects. First, in re-read times on the reflexive

conditions (5a) showed a profile indicative of agreement attraction: lure match conditions being read faster than lure mismatch conditions within target mismatch sentences only[6]. However, no evidence for the impact of the lure was found in the reflexive conditions (5b). King et al. (2012b) report similar results[7].

(5) a. The new executive who oversaw the middle manager / managers apparently was / were dishonest about the companys profits.

b. The new executive who oversaw the middle manager / managers apparently doubted himself / themselves on most major decisions.

Similarly to Dillon et al. (2013), Xiang et al. (2008) compared NPI licensing to re-

the target match conditions were read faster than target mismatch conditions within lure mismatch conditions only. Or, to put it another way, target mismatch-lure match conditions were the slowest of all. Second, in re-read times on the pre-final region, lure match conditions were faster than lure mismatch conditions within target match conditions only. A follow-up sentence interpretation experiment suggested that in lure match conditions people provide more answers compatible with the resolution of the dependency to the lure. I have to note, however, that a replication of this experiment by Cunnings and Sturt (2014, Exp.1) failed to find support for the late effects of the lure. It is also worth noting that in both studies the number of participants was rather low, 24 for Sturt (2003) and 28 for Cunnings and Sturt (2014, Exp.1). Thus, the statistical power that these studies had was likely low . Given this, I do not consider this evidence against the absence of lure-match effects in Sturt (2003) as very strong.

[6] In this case, “target” refers to the subject of the main clause, and “lure” - to the subject of the embedded clause.

[7] This study did report lure match effects in a different configuration, but I will return to them later

flexive resolution in sentences like (6) using EEG. NPI conditions did show effects of structurally inappropriate licensor: the amplitude of P600[8]was reduced as compared to the sentences with no licensor at all. However, no similar evidence was found for the reflexive sentences: regardless of whether the lure matched or mismatch the reflexive in gender, target mismatch conditions elicited smaller P600.

(6) a. **NPI, grammatical**: No restaurants [that the local newspapers have recommended in their dining reviews] have ever gone out of business[9].

b. **NPI, illicit licensor**: The restaurants [that no local newspapers have recommended in their dining reviews] have ever gone out of business

c. **NPI, no licensor**: Most restaurants that the local newspapers have recommended in their dining reviews have ever gone out of business

d. **Reflexive, target match:** The tough soldier that Fred treated in the military hospital introduced himself to all the nurse.

e. **Reflexive, target mismatch, lure match:** The tough soldier that Katie treated in the military hospital introduced herself to all the nurses.

f. **Reflexive, target mismatch, lure mismatch:** The tough soldier that Fred treated in the military hospital introduced herself to all the nurses.

[8] An event-related potential component often associated with the processing of syntactic violations. Bigger amplitude is taken to be associated with more severe reaction to the violation.

[9] I omit an additional manipulation of NPI licensor type, since it is not relevant for interpreting these results.

While the studies above report consistent evidence for reflexive resolution being accurately guided by structural information, other studies have reached different conclusions. Badecker and Straub (2002) used feature-mismatch paradigm in a self-paced reading study with sentences like (7). They reported a *slow-down* in the second region after the reflexive when both the lure and the target matched the reflexive in features.

(7) Jane / John thought that Bill owed himself another opportunity to solve the problem.

King et al. (2012a) compare conditions with verb-adjacent and non-verb-adjacent reflexives to test the following conjecture. The verbs likely re-activate their arguments, including subject, regardless of the presence of the reflexive. Reflexive pronouns are often found in direct object position, so if the information about the subject persists in the focus of attention, it might be too prominent to allow any effects of the lure be detected. If this conjecture is right, we will only observe the effects of the lure in the latter case. This is what King et al. report[10]. Only for non-adjacent reflexives lure match conditions were read faster than lure-mismatch conditions within target mismatch conditions.

[10] A caveat is in order: as far as I know, these results have never been published as a research paper. The information is available as a poster and a conference abstract, but the amount of statistical and numerical details is obviously much smaller than what would be possible in a paper. Thus, the interpretation presented here is correct to the degree that I interpret this limited information correctly.

(8) a. **Verb adjacent:** The mechanic who spoke to John / Mary sent himself / herself a package.

b. **Non verb adjacent:** The mechanic who spoke to John / Mary sent a package to himself / herself.

Cunnings and Felser (2013) investigated whether reflexive resolution is affected by participants' working memory characteristics. They used eye-tracking with feature mismatch paradigm, and additionally median-splitted their participants into two groups according to their working memory capacity (as assessed by Daneman-Carpenter test (Daneman and Carpenter, 1980)). In two experiments they found limited evidence for the influence of the lure. In Experiment 1, it was observed at the pre-final region for both memory groups in first fixation duration; in Experiment 2, only participants in the low-working-memory group were affected by the lure (but the effect was noticeable earlier, at the reflexive itself). An eye-tracking study by Patil et al. (2016) reports the influence of lure match on first-pass regression probability[11], observing more regressions in lure match conditions as compared to lure mismatch conditions within target match only.

Finally, Parker and Phillips (2017) (furthermore, PP2017) and Sloggett (2017) (furthermore, S2017) provide perhaps the clearest evidence of lure-match effects in reflexive resolution. PP2017 rely on the same feature mismatch manipulation as most of the previous studies, with an additional twist: they manipulate the degree to

[11] I.e. the probability that the eyes will move to earlier regions during the first pass on the current region

which the target mismatches the reflexives. In all of the previous studies, the target either fully matched the reflexive in morphological features, or mismatched it in one morphological feature. Parker and Phillips additionally considered conditions where the target mismatches the reflexive in two morphological features, as exemplified in (9). They found that people were much faster to read the reflexive if it matched the features of the lure, but only in the 2-feature target mismatch conditions. The authors considered target mismatch in all possible two-features combinations of gender, number and animacy and found similar effects for all of them.

(9) The talented actor/actress mentioned that the attractive spokeswomen praised himself for a good job. .

I will slightly delay the discussion of the interpretation that PP2017 give. I will only note that according to their account, having mismatch in two features between the reflexive and the target is crucial for observing lure-match effects. Therefore it is not surprising that a big proportion previous studies failed to do so: all of them used 1-feature target mismatch configurations.

S2017 replicates most of PP2017 findings. Moreover, he demonstrates that the lure match effect in PP2017 configurations depends not only on the degree of feature match between the reflexive and the potential antecedents, but also on discourse factors. First, he shows that the identity of the embedding verb is important - in sentences like (10a) lure-match effects were only observed with report verbs like "say", but not with perception verbs like "hear" (Experiment 1b). Second, he shows that the identity of the target is important - lure-match effects do not appear if the target is an indexical pronoun like “I” or “you” (10b) (Experiments 3b,4b). Finally, in an interpretation study (Experiment 1c), he demonstrates that lure-match effects are potentially detectable even in grammatical sentences. In this experiment people were presented with sentences like (10c) in a self-paced reading manner, and at the end of each sentence had to answer a comprehension question. For grammatical sentences like those in (10c), the question was probing the interpretation of the reflexive, e.g.: “Who was misrepresented at the meeting? The librarian / the schoolgirl”. People chose answers compatible with non-local resolution (e.g. “The librarian”) in roughly 30% of the time despite the sentence being fully grammatical[12].

(10) a. The librarian / janitor said / heard that the schoolboys misrepresented herself at the meeting.

b. The actor/actress said that Joanna I horribly misrepresented herself in the article.

c. The librarian / janitor said / heard that the schoolgirl misrepresented herself at the meeting.

[12] While this may, indeed, indicate that target mismatch is not a necessary pre-condition for lure-match effects to arise, we have concerns about how the data fromS2017 Experiment 1c should be interpreted. We discuss these concerns later in the thesis.

1.2 Evidence interpretation

The evidence reviewed above suggest that at least sometimes grammatically irrelevant constituents do affect real-time dependency formation. This information in itself, however, is not sufficient to claim that grammatical constraints fail to uniquely guide the parser. In this section we will review three types of accounts of how lure-match effects come to be. We will call an account "structure-strict" if it suggests that structural information can categorically rule out illicit dependency formation, and "structure-defeasible" otherwise. The evidence above could be used to support "structure-defeasible" accounts only if we manage to argue that "structure-strict" accounts fit the evidence less well or should be dispreferred on theoretical grounds. We will show that such an argument could be made. We will then discuss potential counter-arguments, and this will lead us to the main goal of the thesis - evaluating these counter-arguments in order to understand whether our original argument for "structure-defeasible" models should or should not be abandoned.

1.2.1 Accounts of lure-match effects

The first class of accounts can be called **"representational"**. They come in different flavors (e.g. Nicol et al., 1997; Vigliocco and Nicol, 1998; Franck et al., 2002; Eberhard et al., 2005), but they all share the assumption that lure-match effects arise because syntactic representation of the target is distorted: e.g. the whole phrase receives its number marking not from its head but from the complement. Interference effects arise when a correctly functioning access mechanism contacts an invalid linguistic representation. Thus, representational accounts have to be "structure-strict" - they assume that structure perfectly guides the search for the target constituent.

A second type of accounts could be called **"grammatical"** (e.g. Xiang et al., 2013; Sloggett, 2017). These accounts basically suggest that lure match effects are not a result of a processing error, rather, they are a manifestation of grammatical strategies licensed by the language. Xiang et al. (2013) suggest an account of this sort for the NPI data. They argue that constraints on the dependency licensor should not always be analyzed as "be in the c-command domain of a downward-entailing expression". Rather, the grammar also provides an alternative licensing path, via pragmatic inferences from the context. In certain cases this strategy could lead to spurious inferences, which would lead to illusory NPI licensing. Sloggett (2017) suggests this type of account for lure-match effects in reflexives, arguing that they should be analyzed as instances of logophoric behavior of the anaphors. We will discuss this account in much more detail in a later section. These accounts are also "structure-strict": if one wants to use grammatical principles as the explanation, one has to assume that they are applied accurately.

Finally, a third class of accounts could be called **"memory"** accounts: they assume that the errors arise at the level of memory operations underlying language comprehension (Lewis and Vasishth, 2005; McElree, 2006; Van Dyke and McElree, 2006; Engelmann et al., draft4; Jäger et al., 2017; Parker et al., 2017). The most common assumption is that memory *access* is faulty and sometimes the parser may access constituents not licensed by the grammar. This amounts to "structure-defeasibility" in my terms. However, memory models are not inherently "structure-defeasible" - one could modify model parameters so that the parser would be predicted to only access grammatically licit constituents. Such models would not predict lure-match effects arising from faulty memory access. One could hope that they can be easily ruled simply by the fact that lure-match effects are empirically attested. Unfortunately, lure-match effects can arise even in "structure-strict" memory models, since memory access is not the only place where things could go wrong.

Another such place is *encoding*. It might be the case that when one item needs to be entered into memory and another item with similar features is already present there, the encoding for one or both items may get degraded, perhaps because it is difficult to keep distinct bindings for similar features (Nairne, 1990; Vasishth et al., 2017). The degraded representation will be more difficult to access during retrieval, causing a slow-down. Thus, similarity between the target and the lure might affect RTs even if retrieval operates completely faithfully[13]. Originally, this argument has only been applied to cases of inhibitory interference: slow-down when the lure matches the target in features (Dillon et al., 2013; Chow et al., 2014). However, recently Patil et al. (2016, p.17) and Parker et al. (2017, p.129) pointed out that facilitatory interference (speed-up when the lure matches the features on the other end of the dependency, but not on the target) does not have to stem from retrieval stage either. For example, if representations can be degraded during the encoding stage, they will be degraded more in (1a) than in (1b), since in the former case the

[13] Notice that in this case the dependency formation is completely licensed by the grammar, but is made more difficult by extra-linguistic factors.

two nouns have a greater overlap in features. Thus, even if memory access is always accurate and the target is reliably retrieved in all cases, it will take longer time in (1a), creating a profile of reaction times consistent with facilitatory interference.

1 a. *The key to the cabinet are on the table.
b. *The key to the cabinets are on the table.
c. The key to the cabinets is on the table.
d. The key to the cabinet is on the table.

Another possible locus of lure-match effects in memory models is in post-retrieval operations, e.g. repair. Different types of models assume different memory dynamics. Lewis and Vasishth (2005) model, which we will discuss in detail later, assumes that an item's prominence in memory determines both how likely it is to be faithfully recovered and how quickly this item can be made available for processing. Thus this model can account for reaction times differences directly; if needed, repair processes could be assumed, but without having a good reason for doing so, a parsimonious Lewis and Vasishth-type model could make do without them. On the other hand, in McElree (2006) framework access times are always constant: an item's prominence only determines the probability of successfully accessing the item. If this model is closer to the truth, RT differences would have to be produced by some other mechanism. One possibility would be that longer reaction times stem from re-analysis/repair processes. E.g. if retrieval fails, a second retrieval could be initiated to compensate for that. The slow-down due to additional retrievals could potentially account for lure-match effects[14]. Notice that in this case, RTs again will not necessarily tell us anything useful about how structural information is used to constrain real-time processing: repair processes might involve a different set of processing routines/constraints[15]. Another possibility is that the strength of the recovered representation might matter: even if any item can be retrieved equally fast, some will be recovered more faithfully, and the processes downstream from memory retrieval may be sensitive to these differences.

[14] Consider the following possibility. As we discuss in section 1.3, cue-based models assume that when a memory item has to be accessed, a set of *retrieval cues* is specified; all items are matched in parallel to these cues, and receive a boost in prominence for each matching feature; the most prominent item is then accessed with some probability of success. Now, the lure would be more prominent in (1b) than in (1a) because it would match the number of the verb, which presumably would be a part of the set of retrieval cues. Fewer re-analysis operations would be initiated in the first case, potentially leading to a speed up (but see Nicenboim and Vasishth (2018) who suggest it is not clear how repair processes would even work in ungrammatical sentences and how the RTs profiles in those would be accounted by McElree model).

[15] Of course, we might argue that repair is a normal part of everyday language use and thus identifying the constraints involved is useful. Still, one needs to be careful: if we are talking about "first-pass" processing that just directly leads to interpretation, the set of memory mechanism we can "blame" for errors is narrower and better defined. If we are talking about repair, the spectrum is much wider: it may be that repair relies on the "usual" memory access procedures with some minor tweaks. Or maybe the "usual" memory access is used but with completely different set of parameters. Or maybe qualitatively different mechanisms are involved. Without clearly specifying our theory of repair, it would be hard to figure out what exact conclusions should we draw out of our RT data.

1.2.2 Choosing an account

We have looked at three groups of accounts: representational, memory, and grammatical. As we have seen, the first two groups of accounts have to assume that structural information can successfully constrain dependency resolution. Memory accounts can be either "structure-strict" or "structure-defeasible", depending on the assumptions one makes about the access properties and the source of behavioral evidence: encoding, retrieval, repair. Only if we can unambiguously attribute the observed lure-match effects to the retrieval processes, will we be able to make a claim about "structure-defeasibility" of normal real-time sentence processing. In this section we will discuss why one might, indeed, decide that memory retrieval processes stand behind lure-match effects, and why, therefore, one would want to believe the "structure-defeasibility" thesis.

Ruling out representational accounts The biggest argument against representational models is empirical: it is the absence of illusions of ungrammaticality in agreement attraction (representational accounts have been historically developed to deal with agreement data) (Wagers et al., 2009). In these accounts, the distortion of the target NP happens phrase-internally due to morphosyntactic mechanisms (e.g. erroneous feature percolation from the head level to the phrase level) and independently of other factors such as sentence well-formedness. Thus, it is as likely to happen in grammatical as in ungrammatical sentences, and we should commonly observe cases in which a perfectly well-formed sentence is perceived as ungrammatical.

However, such cases are rarely reported.

Another potential argument against representational accounts is that they are not easily expandable to cover other dependencies which exhibit similar empirical behavior (NPI licensing or reflexive resolution). In contrast, memory models can explain multiple phenomena in terms of the same underlying mechanisms, and thus may be preferred, if they are comparable in terms of empirical coverage. A priori this may seem counter-intuitive: since we know that different linguistic dependencies have qualitatively different constraints on them, wouldn't we expect them to behave qualitatively different during the processing as well? I argue that while this is possible, it is not a logical necessity.

It is inevitable that at some point during language processing linguistic constraints have to be translated into processing instructions. I argue that it may be advantageous to assume that this translation process is very straightforward at the interface: every linguistic constraint is turned into a retrieval cue[16]. However, from this point onwards what happens with these cues is determined only by the properties of the memory system: i.e. it is agnostic about where the cues came from and just knows how to perform operations with them. Consider anaphora as an example. Under Chomskyan Binding theory, an anaphor should be bound by a local (roughly, a clause-mate) and c-commanding antecedent. According to our hypothesis, these specifications would be converted to retrieval cues like "+ local" and a "+ c-commands current NP" cue[17]. Another set of constraints - semantic or

[16] See section 1.3.

[17] Real constraints will have to be more complicated. The actual definition of locality domain

perhaps pragmatic - require that the anaphor matches its antecedent in ϕ-features - correspondingly, ϕ-features of the anaphor would be translated to retrieval cues as well (e.g + sg, + masculine etc.). Memory systems would then use these cues like any others in order to determine which memory chunk should be retrieved during reflexive resolution. Notice that in this case the parser's failures to be uniquely guided by linguistic constraints are due to either the translational procedure used at the interface, or the properties of the general purpose memory system. Correspondingly, we would expect to observe reasonably similar effects across multiple dependencies, which arguably corresponds to actual empirical observations. But even if we question the degree of this alignment, such a unified system is theoretically useful: it allows to easily generate precise predictions for a range of linguistic phenomena and suggests what types of empirical observations we should be looking for. This could make it easier to discover new things.

In addition, a model with a unified treatment of different dependencies my be preferred because it assumes that linguistic constraints *straightforwardly map onto general purpose mechanisms*. As soon as we start making additional assumptions connected to the linguistic nature of the representations - e.g. that different levels of representations (syntax, pragmatics etc.) may have different access procedures or that different dependencies rely on qualitatively different mechanisms, we introduce an additional burden of explaining why such a division exists[18].

is actually more involved than the simplification we used, and as we discuss later, encoding c-command information in terms of cues is notoriously difficult.

[18] Notice that PP2017 of lure-match effects in reflexives (discussed in Section 1.3.1), while largely keeping to the unificational spirit of memory models, already violates the "pure mapping" hypoth-

Ruling out grammatical accounts Grammatical accounts are somewhat harder to rule out. We can think of two arguments against them. First, the argument for preferring memory models we made above is even more relevant for grammatical models: they explicitly state that linguistic constraints affect processing in a non-trivial way (perhaps invoking qualitatively different resolution mechanisms and/or memory access routines) - as we argued, such approach may be undesirable as a working hypothesis. Second, empirical coverage of grammatical models is narrower than that of memory models. As far as we are aware, there are no grammatical accounts of agreement attraction - it is quite unequivocally considered a processing error, and not a manifestation of some not-so-well-described grammatical strategy. Now, if we have accepted the first argument, the fact that at least one phenomenon is a result of a processing error would force us to assume that other lure-match effects stem from the processing side of things as well. These arguments are certainly not rock-solid, but for the moment we will rely on them.

Ruling out non-retrieval memory accounts Ruling out the possibility that lure-match effects stem from the encoding stage is hard to do once and for all, unless we have very fine-grained methods of investigating the contents of memory. However, Jäger et al. (2015) show that at least in some cases the "encoding degradation" account is not supported. The study looked at German reflexive pronoun "sich" in configurations like (11). Since "sich" is unmarked for gender, a reasonable

esis by assuming different weighting schemas for structural cues in subject-verb agreement vs. reflexive resolution assumption would be that this feature is not included in the set of retrieval cues during the search for the antecedent. If inhibitory interference results from encoding processes, this will not matter and we will observe inhibition in (11) when the two nouns match each other in gender. If, on the other hand, inhibitory interference arises due to retrieval dynamics, we should not observe any effect of gender match between the two potential antecedents. This is indeed what Jäger et al. report. This lack of the effect was observed in both self-paced and eye-tracking reading times. The study used unusually large sample sizes (around 150 participants in each of the experiments), which gave them roughly 90% power to detect an effect of 20ms with a standard deviation of 75ms. Thus, the lack of evidence for inhibitory interference is quite reliable.

(11) a. Der Dieb$_i$/Die Diebin$_i$, dem/der der Hehler$_j$/die
the thief-MASC/the thief-FEM whom the dealer-MASC/the
Hehlerin$_j$ befohlen hat zu stehlen, hat überraschenderweise sich$_{i/*j}$
dealer-FEM obliged has to steal has surprisingly self
und die Kollegen angezeigt, berichtete das Hochglanzmagazin.
and the colleagues denounced reported the magazine.
The thief whom the dealer obliged to steal surprisingly denounced himself/herself and the colleagues, reported the magazine.

Jäger et al. (2015) results suggest that interference found in the reading times from feature mismatch paradigm is likely due to retrieval processes. Still, the study only shows that *in some cases* encoding processes are an unlikely culprit, and completely ruling out this possibility or determining the range of phenomena for which it holds may be quite difficult.

Whether repair explanations should or should not be considered is not clear at this point. Choosing between Lewis and Vasishth (2005) model, which does not crucially involve repair, and McElree (2006) model, which does, is complicated. On the one hand, constant speed of memory access, which is only captured by McElree (2006) framework has quite extensive empirical support (McElree et al., 2003; McElree, 2006; Martin and McElree, 2009, 2011). On the other hand, Lewis and Vasishth (2005) model is straightforwardly predicting facilitatory interference RT profile (as observed in agreement attraction), and McElree (2006) model has problems with it. Nicenboim and Vasishth (2018, p.21) suggest that McElree model would not be able to predict it at all, since it is unclear how exactly repair processes would operate in ungrammatical sentences [19].

For the predictions in which the models can be compared directly (inhibitory interference in target match configurations), they appear to be tied. Nicenboim and Vasishth (2018) directly compare the fit of (particular implementations of) Lewis and Vasishth and McElree models to a dataset coming from a study on agreement attraction in target match configurations (Nicenboim et al., 2017). They found that

[19] We think that the following suggestion might work, although it does rely on certain assumptions about repair, and unless we are able to independently validate those assumptions, this suggestion remains purely speculative. Suppose that in target mismatch configurations the system initiates repair quite often, since neither the target or the lure match all of the retrieval cues. Suppose also that fewer repairs will be attempted in lure match conditions, since the error-detection mechanism will be less capable of noticing the violations. Then, in the long run, more time will be spent trying to repair the representation in lure mismatch conditions, giving rise to facilitatory time patterns. For this schema to work, we need to assume that the ease or difficulty of detecting a mis-retrieval is contingent on the number of matching features - i.e. repair detection works in the same way as search result detection. This assumption raises a question: why is feature mismatch not detected during the original search, but is at a later stage? One possible answer could be that the retrieval system only “cares” about finding an item without further knowledge of what has to be done to it, so it may not incorporate on-the-fly sensibility checks. On the other hand, the repair detection mechanism may specifically depend on how smoothly the retrieved item can be integrated in the representation, so it may be more sensitive to such mismatches.

McElree model fit the data somewhat better than a simple variant of Lewis and Vasishth model. However, only a slight complication to the Lewis and Vasishth model allows it to fit the data on par with the McElree model. Thus, at least for target match configurations, neither model is clearly preferred.

Finally, a principled way of distinguishing repair from "first-pass" processes regardless of a model is to look at the time course of lure-match effects. One might argue that an effect appearing early is unlikely to stem from repair processes, unless one assumes they can be really quick. From this point of view, eye-tracking might be the only widely used method allowing for such distinctions: self-paced reading times only gives data about total reading time per region; EEG data, while having perfect time resolution, may be hard to interpret as stemming from a particular time point in the processing. If one accepts this argument, only the studies which find lure-match effect in early eye-tracking measures (such as first fixation or first-pass time) should be taken as evidence on memory search processes.

To sum up my discussion of memory models, lure-match effects can be explained in at least three ways: as stemming from encoding, search or repair mechanisms. Choosing between these possibilities is not always straightforward. We have some evidence against encoding account Jäger et al. (2015), but it is rather limited. Distinguishing between accounts with and without repair as their necessary component has so far been complicated, with no clear winner available. In what follows, we will simplistically assume that any RTs difference we observe stem from memory access processes, but one should keep the issues we have just discussed in mind in order not to become overconfident in the conclusions we draw.

1.2.3 Evidence interpretation: summary

In the last several sections we have been discussing whether we have reasons to choose "structure-defeasible" accounts over competing solutions. We presented the following argument for doing so. Memory accounts arguably provide the best explanation for subject-verb agreement data[20]. There are no grammatical accounts of agreement attraction, and representational accounts are ruled out by the absence of ungrammaticality illusions in subject-verb agreement studies. If we settle on memory accounts, "structure-defeasible" flavors have to be chosen, because otherwise memory accounts would not be able to explain agreement data either. Now, due to the universality of these accounts, a priori it would be tempting to extend them to other linguistics phenomena such as anaphora resolution and NPI licensing. To keep the models' universal coverage, one would also be tempted to preserve "structure-defeasible" assumption we made for agreement. This is essentially the line of explanation pursued in Vasishth et al. (2008) and Parker and Phillips (2017).

This line of argumentation is undercut in two places by several recent studies. First, it has been argued that the generalization about the absence of ungrammaticality illusions is not empirically accurate. Jäger et al. (2017) provide evidence for the existence of ungrammaticality illusions based on the results of a quantitative meta-analysis. Hammerly et al. (draft.april.2018) provide corroborating findings and suggest that the existing generalization is just an artifact of the experimental method. If true, this claim would take away our reasons to prefer memory models to representational ones for the agreement data (and relevant to the main topic of

[20] At least in comprehension

my thesis, “structure-defeasible” to “structure-strict” models). This will have repercussions for our accounts of other dependencies as well: if we do not have to pick “structure-defeasible” models for agreement, we have less ground to a priori assume that other dependencies should be accounted for by “structure-defeasible” models, regardless of the particular architecture we use. Indeed, Sloggett (2017) explains lure-match effects in reflexives with a memory model of “structure-strict” type. It is also the case that these models lose the universality of cue-based models and are very specific. It is not clear whether Hammerly et al. (draft.april.2018) model can be extended to other dependencies apart agreement, and it is clear that Sloggett (2017) model cannot - by design, it is reflexive-specific. Assessing (some of) these recent counter-suggestions is the main goal of this thesis. To start addressing this problem, in the following section I will discuss a prominent model of cue-based parsing (Lewis and Vasishth, 2005) and discuss how it can be used to accommodate (a big part of) the empirical evidence. Then I turn to the recent problematic findings and show in more detail how they might undermine the argument for preferring memory models.

1.3 Lewis & Vasishth (2005) memory model

The central assumption behind cue-based memory models used in psycholinguistic research is that human working memory access is content-addressable and parallel. That is, the items in memory are located by their contents (and not, say, their location) and can be investigated in parallel without the need to go through them one by one. In such an architecture, it is hard or impossible to avoid *interference* - when multiple items are similar enough in terms of their content, it is harder to select the correct one. Therefore, if long-distance dependencies resolution does rely on such system, we can expect processing mistakes similar to those I have reviewed above.

As I have already mentioned, cue-based parsing is theoretically attractive. First, there is independent evidence that domain-general working memory is content-addressable (Jonides et al., 2008, for a review of evidence and arguments). Thus, a theory of long-distance dependencies relying on content-based memory would not postulate any language-specific mechanisms. Second, in cue-based models the dynamics of memory access remain the same regardless of the specific cues one is using. Thus, it might be possible to analyze the processing of different linguistic dependencies in terms of the same underlying mechanisms. In what follows, I will introduce a popular model of cue-based parsing by Lewis and Vasishth (2005) (henceforth, LV05), which I will rely upon in this thesis. I will start with discussing the model itself; then, I will turn to showing how it can account for the data presented above; finally, I will briefly discuss its shortcomings and a few caveats.

In the parsimonious spirit of the discussion above, LV05 model is couched within ACT-R, a general purpose computational model of cognition (Anderson, 2005). According to the model, memory items are represented as bundles of features (for linguistic purposes, it could be "+ NP", "+ subj", "+ animate" etc). Items in memory are not directly available for operations on them - they have first to be "retrieved", i.e. put in a special prominent state. I will call such a state "focus of attention". The capacity of focus of attention is assumed to be extremely limited[21],

[21] The exact capacity of focus of attention is debated. Some proposals suggest that it may only

thus, in order to be able to carry out any remotely complicated procedure, e.g. parsing, memory items have to be constantly shunted in and out of the focus of attention.

In order to put the item in a focus of attention, the system needs to find the item which is most prominent for the current needs. In the model, items' prominence is represented as a quantity called "activation" which is different for each item. This quantity is determined by the history of an item's use, the current search needs and stochastic noise, and is formalized in Eq.1.1. During memory access, the system attempts to boost the activation of the relevant item above a certain threshold. In order to do this, it forms a set of *retrieval cues*[22], which is compared to the features of each memory item in parallel. Each item gets an activation boost proportional to the number of features overlapping with the set of retrieval cues. Formally, S_{ji} is the strength of association item i and the cue j, calculated as shown in Eq.1.2. Each cue has only a limited amount of activation it can "share", denoted by S. Thus, if multiple items match a given cue, the activation boost for each item will be reduced. This corresponds to the intuition that the more items have a certain feature, the less useful this feature is to distinguish between them ("fan effect", Anderson and Reder (1999)). The activation conferred by each matching cue is additionally weighted by w_j. Usually, all retrieval cues are assigned equal weights, but it does not have to be

be able to hold four chunks (e.g. Cowan, 2001), while others claim that the capacity may be even smaller, one or two chunks at most (McElree et al., 2003). ACT-R assumes that the system possesses a number of specialized "buffers", each with capacity of one memory chunk. Of interest to us are two buffers: the goal buffer, in which the set of retrieval cues is placed (see the discussion), and the retrieval buffer which holds the memory access result. Only the item in the retrieval buffer is assumed to be accessible for further operations.

[22] Which is put in the goal buffer.

the case.

$$A_i = B_i + \sum_j w_j S_{ji} \tag{1.1}$$

$$S_{ij} = S - ln(fan_{ji}) fan_{ji} = 1 + items_j \tag{1.2}$$

The boost from retrieval cues determines the item's activation only partially. Items are assumed to have fluctuating baseline activation (B_i in Eq.1.1), which is influenced both by how frequently the item was retrieved in the past and by how recent the last retrieval was. Moreover, the process is assumed to be stochastic, so that activation values are not perfectly predictive of whether the item will be retrieved. The probability that the item will be retrieved on a given access attempt is determined by the assumed distribution of the noise and is described by Eq.1.3, where A_i is the activation value for item i, τ - retrieval threshold, and s - a parameter controlling noise. If the activation for the item does cross the threshold, it is retrieved (i.e. made available for further operations on it) with latency described by Eq.1.4. F is an arbitrary constant, which varies from model to model.

$$P_i = \frac{1}{1 + e^{-(A_i - \tau)/s}} \tag{1.3}$$

$$T_i = Fe^{A_i} \tag{1.4}$$

1.3.1 Explaining RT patterns

LV05 model can successfully explain grammaticality illusions across multiple dependencies. Essentially, a variant of the following (simplified) explanation can be used: when a perfectly matching target is not present in the sentence, an imperfectly matching lure can occasionally be retrieved, leading to faster reading times. Notice that this explanation is "structure-defeasible", since the lures by design located in a structurally inappropriate position; so the fact that they are retrieved at all would mean that the structure fails to constrain the memory access properly.

I will use subject-verb agreement to explain the mechanism in detail and then will show how it could be applied to other dependencies. Consider the ungrammatical sentences from (1), repeated below for convenience. In (1a) the target is a better match to the retrieval cues: it matches the "+subj" (or "+ Nominative") cue, while the lure matches none. On the other hand, in (1b) the target and the lure are each an equally good match[23]. In this situation, response times will be on average faster in (1b) due to statistical facilitation (Raab, 1962). Roughly speaking, if we assume that there are two reaction times distributions, one for the target and the other one - for the lure, and we generate the observed reaction time by drawing a value from each distribution and choosing the smallest one[24], the mean of the distribution of observed RT values will be smaller then the mean of either of the

[23] Assuming that the cues are treated by the system as equally important, i.e. assigned equal weights.

[24] I.e. in this context - the fastest RT.

sampling distributions[25].

1 a. The key to the cabinet are on the table.

b. The key to the cabinets are on the table.

A big selling point for the LV05 model is that it is assumed to be able to explain the absence of ungrammaticality illusions. Wagers et al. (2009) suggested that it could be captured in the following way: since in grammatical sentences the target always matches the retrieval cues better than the lure, it will always be reliably retrieved.

NPI cases work similarly, with the difference that in the lure-match sentences there is no licit target at all. For example, the following explanation is advanced by Vasishth et al. (2008). When people encounter an NPI, they initiate retrieval with "c-commander"[26] and "negative" as cues. In (2c) "no congressman" is indeed a negative element, thus it provides a partial match to the retrieval cues and may be retrieved on a proportion of times.

Finally, lure-match effects in reflexive resolution can be explained in very much the same way as they are for subject-verb agreement (Parker et al., 2017). The only difference is that for reflexives structural cues like "+ local" or "+ c-commander" are

[25] Unless the two sampling distributions do not overlap at all. In the context of the model, we assume that the overlap is bigger in (1b) than in (1a), because the activation values for the lure and the target are closer to each other in the first case, and closer activation values map to more similar retrieval times in ACT-R.

[26] As we discuss in section 3.1, implementing "c-commander" cue may not be straightforward. Vasishth et al. choose to assume that such a cue can be implemented and abstract from the question of how exactly.

weighted higher than morphological cues like “+ sg”. During memory access, the mismatch in a single morphological feature is not enough to make the activation for the target sufficiently low to approach the activation for the lure. But the mismatch in two morphological features suffices to do this, and the system finds itself in the situation I described earlier for subject-verb agreement, with the target and the lure having comparable activation values.

Such a mechanism would explain why the majority of the previous studies did not find evidence for lure-match effects: all of them used designs in which the target mismatched the reflexive at most in one morphological feature. Prima facie, lure-match effects reported in some of the previous studies (Badecker and Straub, 2002; Patil et al., 2016; Cunnings and Felser, 2013; King et al., 2012a) would be problematic for PP2017 account, since they were not observed in 2-feature mismatch configurations. However, most of these findings can be explained away.

Inhibitory interference findings (i.e. lure match leading to slow-down) reported by Badecker and Straub (2002) (Exp.4), Patil et al. (2016) and Cunnings and Felser (2013) (Exp.1) could be discounted on multiple grounds. First, one could argue that these findings reflect fan effect (cf. the discussion above for agreement). Second, they could be dismissed as irrelevant: as I discussed in section 1.2, inhibitory interference can be attributed to other processes apart from memory retrieval, e.g. memory encoding[27]. Finally, this evidence could be dismissed altogether as a fluke: as Jäger et al. (2017, p.326) notice, the meta-review suggests that interference is more likely

[27] This argument is rather weak, especially since Jäger et al. (2017) provides evidence that at least in some cases inhibitory interference *is* informative about retrieval.

to be observed in ungrammatical sentences, and even high-powered studies like Jäger et al. (2015) fail to find reliable evidence for it.

Facilitatory interference observed by Cunnings and Felser (2013) in their Experiment 2 (low-working-memory participants only) could be accounted for if we assume that lures were more prominent than targets in the experimental materials[28]. This assumption would not be too far fetched: the lures were linearly closer to the reflexives and were prominent in the discourse (pronouns were used as lures, while full NPs were used as targets). Within the context of PP2017 model one would have to make an additional assumption that the activation boost from being prominent was functionally equivalent to a single morphological feature mismatch: i.e. combined with an actual single feature mismatch, it was big enough to counteract the higher influence from syntactic cues.

The findings by King et al. (2012a) could be dismissed on different grounds. PP2017 argue that the lure-match effects reported by King et al. (2012a) are of different nature, since they were observed in sentences where the reflexive pronoun was situated in a non-argument position. It is known from the linguistic descriptions that such pronouns may obey a different set of constraints (e.g. Reinhart and Reuland, 1993), therefore they may rely on different processing routines compared to reflexives in argument positions (e.g. it could be the case that parser weighs syntactic cues lower when resolving reflexives in a non-argument position; this would be compatible with the observation that the resolution of such reflexives is often

[28] See discussion in Section 1.5.1.1 about modifications to LV05 model allowing to take linguistic prominence into account

influenced by discourse factors).

1.4 Recent arguments against structure-defeasible models

While LV05 model can successfully explain empirical evidence across multiple dependencies, a number of empirical results are problematic. In this section I focus on two objections to the wholesale adoption of "structure-defeasible" memory models, both of which come from very recent studies. The first objection says that the main argument for choosing memory models over representational ones is invalid. The second objection suggests that "structure-defeasible" memory models should not be used across the board, because different dependencies rely on qualitatively different mechanisms. I discuss them in turn.

1.4.1 LV05 model fit to agreement data is worse than usually thought

The big reason for choosing memory models over representational ones was their ability to explain the existence of lure-match effects in ungrammatical sentences and the absence of lure-match effects in the grammatical sentences. However, this argument is problematic for two reasons. First, as Jäger et al. (2017) discuss, the model's predictions for the ungrammatical sentences are different. In fact, (1d) is not predicted to be as fast as (1c), it is predicted to be slower, since in (1d) the target and the lure share the number feature. As I have discussed in section 1.3, the amount of activation each cue can contribute is limited. If more than one item matches a given cue, this limited amount of activation will be spread evenly between the matching items. In my example, this means that the target in (1d) will on average have lower activation than the target in (1c). And since in LV05 model activation directly determines retrieval times, people should be slower in (1d).

1 a. *The key to the cabinet are on the table.

b. *The key to the cabinets are on the table.

c. The key to the cabinet is on the table.

d. The key to the cabinets is on the table.

Second, empirical evidence appears to be incompatible with either Wagers et al. (2009) account or with the predictions of a basic LV05 model. It appears to be the case that the "absence" of ungrammaticality illusions may be better characterized as "rarer observations" of ungrammaticality illusions. A Bayesian meta-analysis of the available evidence conducted by Jäger et al. (2017) reveals a pattern indicative of ungrammaticality illusions: people read grammatical sentences *faster*, not slower, when the lure matches the target in features. Interestingly, Hammerly et al. (draft.april.2018) report similar observations regarding reaction times; moreover, they suggest a way of capturing the data from *both* grammatical and ungrammatical sentences using a *representational* model. This turns the argument for memory models and against representational models upside down: now it may be the case that it is *representational* accounts which can capture the patterns in both grammatical and ungrammatical sentences, and memory models can do so only in the latter case. We discuss the Hammerly et al. (draft.april.2018) data below and argue that the amount of evidence is not yet sufficient to completely overturn the preference for memory models, but the direction of the argument and the challenges it presents to cue-based accounts are worth noting.

Hammerly et al. (2018) The main claim of Hammerly et al. (draft.april.2018) (henceforth HSD2018) is that the reported absence of ungrammaticality illusions is largely an artifact of people's bias to treat their linguistic input as grammatical. In a series of three grammaticality-judgment experiments and computational simulation, they show that if this grammaticality bias is canceled, lure-match effects appear equally frequently in grammatical and ungrammatical sentences. This result undermines the most important empirical argument for cue-based models and has the potential of bringing the representational models back into limelight. This will mean, of course, returning to the view that structural information categorically guides real-time dependency formation.

HSD2018 argue that to say that ungrammaticality illusions are absent would be too strong, and a more accurate statement would be that both grammaticality and ungrammaticality illusions are observed, although the latter are observed rarer. They suggest that this asymmetry is an artifact of experimental procedure, and not a reflection of real cognitive differences. Namely, the fact that most of experimental sentences are usually grammatical biases participants. In a judgment task that would result in answering "grammatical" more often than "ungrammatical", and this in turn will result in a pattern of responses where ungrammatical sentences are deemed grammatical more often than grammatical sentences are deemed ungrammatical. This explanation suggests that if we manage to eliminate this response bias, the number of grammaticality and ungrammaticality illusions should be roughly the same. This is, indeed, what HSD2018 show.

They discuss the results of three acceptability judgment experiments, where they manipulated participants bias. The first experiment was a replication of previous results, showing the absence of ungrammaticality illusions in a judgment task (Wagers et al., 2009). The sentences followed a standard agreement attraction design, with the grammaticality of the sentence ((12a)-(12b) vs. (12d)-(12c)) and the match between the lure and the verb ((12a)-(12d) vs. (12b)-(12c)) were manipulated. The results were consistent with the previous findings: an interaction of grammaticality and lure match was observed. Grammatical sentences were accurately judged as grammatical regardless of lure match. Ungrammatical sentences were much more likely to be judged as grammatical if the lure matched the verb. That is, the results showed the absence of ungrammaticality illusions.

(12) a. The friend of the nurse frequently visits...

b. The friend of the nurses frequently visits...

c. The friend of the nurse frequently visit...

d. The friend of the nurses frequently visit...

The next two experiments attempted to manipulate people's response bias by changing the proportion of ungrammatical items[29] and drawing participants' attention to

[29] Two thirds of the items were ungrammatical vs. a half in Experiment 1.

this fact in the instructions. In Experiment 2, participants were specifically told that two thirds of the items were ungrammatical. This manipulation led to slight reduction of the asymmetry observed in Experiment 1: now people were a little bit more likely to respond "ungrammatical" to a grammatical sentence with the lure which mismatched the verb. This effect was statistically significant; however, the critical interaction was still present in the data. Thus, even is the asymmetry between the effect of the lure in grammatical and ungrammatical sentences was weakened, it did not disappear completely.

In Experiment 3 the instructions were changed so that instead of receiving the exact proportion of ungrammatical stimuli, the participants were simply told that *the majority* of the items were ungrammatical. This manipulation was more efficient: people became less accurate in grammatical sentences and more accurate in ungrammatical sentences when the lure mismatched the verb. This made the critical interaction disappear: only the effect of lure match was statistically significant. This pattern suggests that both illusions of grammaticality and ungrammaticality were present.

HSD2018 suggest that a representational account by Eberhard et al. (2005) ("Marking and Morphing") could be used to explain their results. This approach assumes that (nominal) number is represented as a continuous value between negative and positive infinity, where negative values correspond to "more singular" and positive values correspond to "more plural". The number of an entire noun phrase, $S(r)$, is determined as in Eq.1.5. $S(n)$ is the notional component of the number valuation, coming from the conceptual message. The sum term represents the syntactic contribution to the number valuation: it is a linear combination of the number information on the head and the embedded phrases, weighted by the distance from the root node. $S(r)$ can be conceptualized as the probability of the plural verb, if we put it through a logistic transformation, thus bounding its values between 0 and 1[30].

$$S(r) = S(n) + \sum_j (w_j \times S(m)_j) \tag{1.5}$$

Consider how this model would predict agreement attraction. For a subject phrase like “The key to the cabinet” both nouns inside it are singular, thus, the number representation of the whole phrase will correspond to “singular” by default. On the other hand, in a phrase like “The key to the cabinets” the embedded noun is plural. This plurality will increase the $S(r)$, making the plural verb more probable.

Results of Experiments 2 and 3 would be captured by this account rather easily. Experiment 1 would be more problematic: as we have already discussed, the changes to the number marking of the subject NP will happen regardless of what the number on the verb actually is, so both illusions of ungrammaticality and grammaticality are predicted. It is also not immediately clear how to incorporate the bias manipulation in the Marking and Morphing model. HSD2018 propose that both goals will be achieved if we map Making and Morphing parameters to the parameters of a drift diffusion model (Ratcliff, 1978). Drift diffusion models model

[30] As far as we understand, $S(r)$ is always greater than 0. Thus, the actual logistic transformation includes a negative bias term in order to bias the representation towards singular in the absence of other sources of number information.

a binary decision process as a diffusion process (a continuous version of random walk) starting at a random point between two decision boundaries. The amount of evidence available to the decision taker determines how quickly the model walks toward the correct decision boundary; however, due to the stochastic nature of the process it is not guaranteed that it will reach it first: it may happen that enough steps occur in the "wrong" direction so that the incorrect decision boundary is hit. Now, we can map the $S(r)$ value from the Marking and Morphing account onto the strength of evidence parameter of the diffusion model: in this way, a more extreme number value on the subject way would lead to (on average) faster and more accurate grammaticality decisions[31]. The crucial assumption is that participants response bias can be mapped to the starting point of the diffusion process. That is, if they are biased to treat sentences as grammatical, the diffusion process will start closer to the "grammatical" decision boundary.

Here is how this model would capture the results of the experiments. In Exp.1, where by hypothesis participants have the strongest grammaticality bias, the diffusion process would start close to the grammatical boundary. This would lead to a situation where for grammatical sentences people would almost always hit the grammatical boundary, even if the number marking on the target NP is less unambiguously singular due to the influence of the lure: the model will have

[31] Notice that $S(r)$ corresponds to the number information, however, the decision boundaries are labeled "grammatical" and "ungrammatical". It seems that the parser would have to decide ahead of time, e.g. based on the information from the verb, which number value should correspond to "grammatical" decision boundary.

little time to accumulate enough drift in the wrong direction. In the ungrammatical sentences, the ambiguous number marking on the head noun as in (12c) would make the drift towards the "ungrammatical" decision boundary slower than in (12c), giving the process enough time to accumulate evidence in the wrong direction and to occasionally hit the "gramamtical" boundary. Overall, we would expect to see the asymmetry between grammatical and ungrammatical sentences: only for the latter the number information on the lure would lead to an increased proportion of incorrect responses.

Exp. 2 and 3, by hypothesis, neutralized the grammaticality bias, shifting the starting point of the diffusion model to a position where it would be roughly equidistant from both boundaries. In this situation, we would expect that in both grammatical and ungrammatical sentences, more ambiguous number marking on the target (as in 12b and 12d) would lead to identical slowdown of the diffusion process in both grammatical and ungrammatical sentences. This would make it roughly equally likely for the wrong decision boundary to be hit, leading to a symmetrical distribution of incorrect responses. People would produce as many "grammatical" responses in (12d) as they would produce ungrammatical responses in (12b), that is, we would observe both illusions of grammaticality and ungrammaticality.

If the reasoning above is correct, it undercuts the argument for choosing "structure-weak" memory models at several points. To begin with, it *eliminates* a crucial argument against representational accounts: if they can capture the absence of ungrammaticality illusions in certain conditions (which, arguably, correspond to the conditions in which no ungrammaticality illusions have been observed), we lose our strongest reason to prefer memory models. HSD2018 findings also provide evidence against memory models. First, the fact that manipulating participants' bias affects the rate of ungrammaticality illusions is prima facie unexpected in a cue-based model. Indeed, if we think that a memory item that perfectly matches the retrieval cues will always outcompete other memory chunks, in grammatical sentences the correct agreement controller should always be retrieved[32].

Second, the observed pattern of reaction times is hard for a cue-based model

[32] In principle, bias can be incorporated in cue-based models, but it is less clear to what degree we would be able to capture all the patterns reported by HSD2018. E.g. one could hypothesize that the weights for retrieval cues are adjusted dynamically depending on the current linguistic (and potentially extra-linguistic) context. The size of the adjustment would correspond to grammatical bias. In a context where most of the sentences are grammatical, the parser can be reasonably sure about its parses, and thus may be willing to rely on structural cues. On the other hand, if most of the sentences are ungrammatical, the parser may lessen its reliance on structural cues. This will result in a situation in which the target has less of an advantage over the lure, and due to stochastic noise, lures will have more chances to be retrieved. In sentences like (12a) this should not affect participants' responses: both the target and the lure match the verb, so whatever noun is retrieved, the result can be judged as grammatical. On the other hand, in sentences like (12b) retrieving the lure would lead to the "ungrammatical" response. Overall, we would observe fewer correct judgments in (12b), than in (12a) - i.e. we would observe ungrammaticality illusions. Unfortunately, this explanation would fail to capture the pattern of results in ungrammatical sentences observed in HSD2018 Exp.3: when the response bias decreased, the proportion of correct responses in the ungrammatical sentences increased. This is the opposite of what my ad hoc account would predict: if the parser does reduce its reliance on structural cues, lures should become more likely to be retrieved regardless of sentence grammaticality. Thus, we would expect that there will be more incorrect judgments in both grammatical and ungrammatical sentences.

to fit. The critical observation is that if the lure matches the target in features, the judgments are given faster, while LV05 model would predict a slow-down in such configurations. This evidence is particularly interesting, since it aligns with the results of the quantitative metaanalysis by Jäger et al. (2017), which we have discussed earlier. Again, it is not the case that cue-based models cannot account for these RT patterns: if we adopt the Engelmann et al. (draft4)'s suggestion that linguistic prominence can affect memory chunks' activation (see section 3.1) AND if we assume that the lure is more prominent than the target, we would be able to capture the HSD2018 RT patterns. But we do not see any arguments for why the lure should be considered more prominent than the target, so practically speaking these RT pattern still remain problematic for the LV05 model.

Thus, HSD2018 account and evidence pose interesting challenges for memory accounts. We think, however, that in their current state they are still too weak to completely rule memory models out. First, HSD2018 explanation relies on and covers only acceptability judgment task. However, the absence of ungrammaticality illusions has been reported in studies relying on other experimental techniques, and it may not be straightforward to extend HSD2018 account to cover these (as the authors themselves discuss). HSD2018 argue that self-paced reading might be accommodated relatively easily: instead of deciding on whether the sentence is grammatical or ungrammatical, people could decide on whether to press or not to press the button to uncover the next word. Eye-tracking might be harder to accommodate, since it is less clear what are the decisions that people are making during reading. Potentially it could be something along the lines of "stay on the current word / make a progressive saccade / make a regressive saccade" etc. But it is clear that the decisions are unlikely to be binary and the model would have to be further complicated to allow for this. Finally, EEG results (e.g Xiang et al., 2008; Tanner et al.), might be the trickiest to cover, since it is not immediately clear how to talk about observed ERP patterns in terms of underlying decision making processes.

More generally, HSD2018 model only accounts for agreement data. One could claim that it should do so, since lure-match effects in subject-verb agreement and reflexive resolution are separate phenomena; this is the route Sloggett (2017) takes. From the perspective we take in this thesis this is rather a disadvantage. It is not clear how the model would apply to reflexives resolution. First of all, as Jäger et al. (2017) suggest, there is very little evidence for illusions of ungrammaticality in reflexive resolution (in comparison, the same analysis does provide evidence for illusions of ungrammaticality in agreement attraction). Second, it is not clear whether Marking & Morphing account could explain lure-match effects in configurations used to study reflexives resolution. Often the lure is structurally higher than the target (e.g. it is located in the matrix clause and the target - in the embedded complement clause): thus, in order for the features of the lure to influence the target, we would need to say that the features information from the lure can percolate downwards. As Wagers et al. (2009) point out, Eberhard et al. (2005) do indeed make this assumption. But even in this case, the lure is more structurally distant from the target than in the agreement cases, where the lure is typically embedded in a PP inside the subject NP. Given that Marking & Morphing account assumes that the influence of the percolating information reduces with structural distance, in the reflexives cases the lure may be too far from the target to noticeably affect its representation. Finally, it is not clear how Marking & Morphing would explain the fact that lure-match effects are not observed in 1-feature target mismatch configurations, but are observed when the target mismatches the reflexive in two features. It would be easier for the information from the lure to bring the mismatching target closer to the match with the reflexive in 1-feature mismatch conditions, and thus, the opposite pattern of the results would be predicted.

To sum up, HSD2018 proposal shifts the focus from "structure-defeasible" cue-based models back to "structure-strict" representational ones, and posits interesting problems for memory accounts. On the other hand, it also gives less universal coverage than the framework offered by cue-based models. I do not feel that the proposal is developed enough to strongly sway the odds against cue-based models, but it may be an important first step towards alternative accounts. As we will see now, the same tendencies are evident in Sloggett (2017) account of the reflexive data.

1.4.2 Reflexives processing is qualitatively different from agreement

The second argument against uniform application of "structure-defeasible" memory models suggests that different dependencies (and in particular - subject-verb agreement and reflexive resolution) rely on qualitatively different mechanisms. This argument has been present in the literature for some time. Originally, it was based on the fact that lure-match effects are rarely observed in reflexives resolution, as well as on some data indicating that the search for the antecedent may not be parallel, as cue-based models assume, but rather serial (Dillon, 2011; Dillon et al., 2013; Dillon, 2014). We will not discuss the previous arguments in detail, focusing instead on two recent studies.

Jäger et al.'s meta-analysis provides evidence for qualitative differences between dependencies. It suggests that for reflexives there is no evidence in the available data for lure-match effects in grammatical sentences[33]. Second, in ungrammatical sentences, the evidence suggests that people are *slower* when the lure matches the reflexive. Both empirical patterns contradict the predictions of the LV05 model. Moreover, both patterns differ from the results of the analysis on the subject-verb agreement data. This discrepancy might be indicative of the underlying differences between dependencies.

We consider the evidence provided by Jäger et al. (2017) as only tentative due to several caveats of their analysis. First, as the authors point out themselves, it is essentially observational, therefore any conclusions should be confirmed using experimental studies specifically designed with this goal in mind. Second, the quantitative meta-analysis relied only on first-pass reading times for the eye-tracking studies, because "it is the most commonly reported eye-tracking measure in the psycholinguistic literature and arguably reflects early cognitive stages of dependency formation" (Jäger et al., 2017, p.323). However, lure-match effects in reflexive reso-

[33] In a way, reflexives are being more catholic than the Pope (subject-verb agreement) - it appears that for reflexives, illusions of ungrammaticality are truly absent

lution appear to be quite variable in timing. E.g. Parker et al. (2017) only observed effects in first-pass reading times in two out of three experiments. Similarly, in Sloggett (2017) the time course of lure-match effects varied across experiments: in two of them, the effects were observed relatively early, (first-pass or regression path times at the critical region), in the other two, the effects became evident later (total times at the critical region or regression path at the spillover region). It is not quite clear whether such differences are due to random variability across participants or experimental items, or whether they reflect meaningful differences in the reflexive resolution process. Whatever the case, an analysis based only on first-pass RTs will miss some of the effects reported in the literature. Finally, Jäger et al. (2017) review does not include Sloggett (2017) data, presumably because it was not available at the time of writing. Sloggett (2017) showed that lure-match effects observed by PP2017 can indeed be reliably observed (modulo the discussion about the exact configurations in which they arise). It is possible that with these data the quantitative analysis would have given different results.

Sloggett (2017) The evidence for qualitative differences between subject-verb agreement and reflexives presented by Sloggett (2017) is more extensive. He builds on and expands empirical observations made by Parker and Phillips (2017). In contrast to the retrieval error explanation for agreement attraction, Sloggett argues that the observed lure-match effects in reflexives processing are a manifestation of a completely grammatical anaphora resolution strategy, namely, logophoric interpretation of reflexive pronouns. Logophors are pronouns which refer to the person "whose speech, thoughts, feelings, or general state of consciousness are reported" (Clements, 1975)[34].

Sloggett's hypothesis is motivated by three crucial observations:

1. Lure-match effects are not observed when the lures are subjects of perception verbs, even if the target mismatches the anaphor in two morphological features. However, the identity of the matrix verb does not matter if the lure refers to the only center of consciousness in the sentence.

2. Lure-match effects are eliminated or at least weakened if the target is an indexical pronoun like "I" or "you".

3. As often as in 30% of the cases, when asked a comprehension question probing the reflexive resolution, people choose the answer compatible with the lure being the intended antecedent, even when the target fully matches the reflexive in morphological features.

[34] The term "logophor" was initially introduced by Hagège (1974) to describe morphologically distinct pronouns in a number of languages spoken in Africa (e.g. Ewe, Yoruba, Igbo; see Culy (1994) for an extensive list of relevant languages). However, the use later was extended to denote non-clause-bound reflexive pronouns in a number of languages (e.g. Japanese *zibun*, Chinese *ziji*, Icelandic *sig*). In this second sense, the term has also been applied to English, to describe the pronouns in configurations like: "Max said that the queen invited both Lucie and himself for tea."(e.g Reinhart and Reuland, 1993). Notice that a priori it is not guaranteed that "logophors" in African languages and non-clause-bound reflexives rely on the same set of constraints (or on the same processing routines), thus one must be careful not to conflate the facts pertaining to the two phenomena.

Here is how logophoric hypothesis can explain these patterns. First, lure match sensitivity to the type of the embedding verb appears to conform to a pattern reported for other languages. Culy (1994) suggested the following implicational hierarchy for a number of languages spoken in West Africa: *Speech* < *Thought* < *Knowledge* < *Direct perception*. That is, if a language allows subjects of verbs on the right end of the scale to be taken as antecedents for logophors, it will allow the same for the subjects of verbs lower on the hierarchy. The first pattern is in line with the Culy hierarchy; and the fact that the type of the verb does not matter if the lure is the only animate entity in the sentence is also naturally accommodated: if there is a single conscious antecedent in the discourse, the logophor "has no choice". The second pattern closely resembles "person blocking", a syntactic phenomenon observed in Chinese long-distance binding of the reflexive "ziji". As Huang and Liu (2001) discuss, "ziji" allows antecedent outside of its local domain, as in (13a). However, an intervening first or second person pronoun may block this resolution option (13b)[35]. Huang and Liu argue that these effects arise when "ziji" is used as a logophor, and not as a syntactic reflexive. Finally, the third piece of evidence is explained almost for free: if resolving the reflexive to the lure is a representation of a *grammatical* strategy, we should expect lure-match effect to appear even in grammatical sentences.

(13) a. Wo/ni$_i$ danxin Lisi$_j$ hui piping ziji$_{i/j}$.
I/you$_i$ worry Lisi$_j$ will criticize SELF$_{i/j}$.

[35] This is not the only environment where blocking effects appear, but it is the most relevant one for the current discussion.

I/you worry that Lisi might criticize me/you/herself.

b. Zhangsan$_i$ danxin wo/ni$_j$ hui piping ziji$_{*i/j}$.
Zhangsani$_i$ worry I/you$_j$ will criticize self$_{*i/j}$.
Zhangsan worries that I/you might criticize my/yourself.

To formalize this hypothesis, S2017 uses a modified version of PP2017 cue-based retrieval account. On the memory retrieval side, the two accounts are virtually identical, with one crucial difference: S2017 makes his model "structure-strict", by assuming that structural features act as gating features and only grammatical antecedents are ever retrieved. As far as I can tell, he does not discuss how this difference is implemented in the model (higher weights on structural cues? different cue combinatorics scheme?)

However, the main explanatory power of S2017 comes from a suggested modification to the grammar of English. Sloggett suggests that a phonologically null operator, OP_{log}, is located in the left periphery of complement clauses (see Charnavel and Sportiche, 2016, for a related proposal). By hypothesis, it tracks perspective centers of the utterance. Each appropriate[36] antecedent is mapped onto one of the three perspective "roles" from Sells (1987): SOURCE, SELF and PIVOT. The first role corresponds to the participant acting as the source of information, the second - to the participant whose mental state is reported and the third one - to the participant whose (physical) location is assumed as the reference point. The roles form a hierarchy, such that a participant takes on all the roles below the role that is assigned to him/her. E.g. if somebody acts as SELF, then he also takes the role of PIVOT, but

[36] Minimally, an animate entity.

not that of SOURCE. OP_{log} obligatorily refers to the highest specified role on this hierarchy. I.e. OP_{log} will refer to whatever participant bears the role of SOURCE; if there are no such participants, it will look for SELF and, failing that, for PIVOT. Finally, S2017 suggests that these roles are assigned probabilistically. E.g. speakers are most likely to be mapped onto SOURCE, but may sometimes get mapped onto SELF; the reverse could hold for indexical pronouns.

Notice that binding by OP_{log} is syntactically local[37] with respect to the reflexive. Thus, the non-local binding is just an appearance, created by the special referential properties of the operator. But if OP_{log} is a licit binder, why are non-local antecedents so inaccessible to native speakers? S2017 explains this in terms of the feature composition of the operator: it is assumed to be underspecified for ϕ-features. This ensures that it is a worse match to the retrieval cues compared to the overt target antecedent. As such, it is less likely to be retrieved; importantly, it does not mean that its retrieval is impossible, just that it will happen only rarely, perhaps, when stochastic noise will happen to bias memory activations in the right direction.

Here is how this formalization would derive the three crucial patterns I mentioned in the beginning of the section. The effect of the embedding verb stems from the way participants are mapped onto Sells's roles. Speech verb subjects are likely to be SOURCEs, so OP_{log} will most often take them as antecedents. On the other

[37] Or at least "more local" than for the lure: OP_{log} is situated in the same clause with the pronoun. However, it does not occupy a typical position for an antecedent, since it is not located in an argument position.

hand, perception verb subjects are more likely to be PIVOTs. Under the assumption that only one participant can be mapped on any given role, by the time the reflexive is encountered, the control of the PIVOT will be taken by the embedded subject. Since there will be no higher roles specified, this is what OP_{log} will refer to, and non-local interpretation will not be available. Person blocking effects are explained in a similar way. S2017 suggests that although speech verb subjects tend to be mapped onto SOURCE's and indexical pronouns - onto SELF, sometimes the mapping gets reversed due to its stochastic nature. It is exactly in these cases that we will observe person blocking effects. Finally, since the resolution process is assumed to follow grammatical constraints, it should come as no surprise that people are able to access non-local interpretations even when the (overt) local target completely matches the features of the reflexive.

1.5 Parker & Phillips (2017) vs. Sloggett (2017)

1.5.1 Empirical observations

We have seen that the recent work has suggested reasons to shift in the direction of "structure-strict" models applicable to isolated linguistic phenomena. They rely on two types of arguments: a) agreement attraction data does not constitutes evidence for (and potentially is evidence against) memory models; b) reflexive resolution relies on different mechanisms and thus there is no reason to capture it and subject-verb agreement with the same kind of "structure-defeasible" model.

While this shift may turn out to be in the right direction, I consider the universality of “structure-defeasible” memory models too attractive to easily give it up. Thus, my goal in this thesis is to look for further evidence supporting or contradicting the recent claims. In order to do so, I will focus on the two accounts of lure-match effects in reflexive resolution, suggested by Parker and Phillips (2017) and Sloggett (2017), which I have already discussed. In this last section of the introduction, I discuss whether any of the empirical facts present a fundamental problem for the accounts and whether we have strong reasons to prefer one over the other.

Altogether, I take the following 7 empirical observations from the previous research to be important (I mark in parentheses the studies they are coming from):

E1 In configurations like THE BOY SAID [THAT THE GIRLS LIKE HIMSELF] people read sentences with feature matching lures faster if the target mismatches the reflexive in two features. This does not happen when the target fully matches the reflexive. (PP2017, S2017)

E2 The speed up also does not occur when the target mismatches the reflexive in a single feature. (PP2017)

E3 Lure-match effects also appear in configurations like THE BOYS [THAT THE GIRL LIKED] PRAISED HERSELF in the cases when the target mismatches the reflexive in two features. It is not known whether the degree of target match modulates lure-match effect in these cases. (PP2017)

E4 Lure-match effects are not observed when the lures are subjects of perception verbs, even if the target mismatches the anaphor in two morphological features. (S2017)

E5 However, the identity of the matrix verb does not matter if the lure refers to the only center of consciousness in the sentence. (S2017)

E6 Lure-match effects are eliminated or at least weakened if the target is an indexical pronoun like “I” or “you”. (S2017)

E7 In an interpretation task, people answer interpretation questions in a way which suggests that they treat the lure as the antecedent as often as in 30% of the cases, even when the target fully matches the reflexive in morphological features.(S2017)

PP2017 interpret findings **E1**, **E2** and **E3** as evidence for faulty memory access in a cue-based memory. S2017 agrees that **E2** should be explained in terms of the properties of the memory access. However, he assumes that retrieval operations only ever return structurally licit antecedents, i.e., his account is “structure-strict”. S2017 suggests that the majority of the facts (except for **E3**) should be seen as evidence for a grammatical strategy of reflexives resolution. In his explanation, a phonologically null operator in the left periphery of complement clauses tracks discourse prominent entities and can create appearances of non-local reflexive binding.

1.5.1.1 PP2017

PP2017 account naturally accommodates observations **E1**, **E2** and **E3**. Prima facie, it would appear that it is unable to explain **E4**, **E5** and **E6**, since they suggest that lure-match effects may depend on things other than the representation of the lure, the target and the reflexive. However, I will argue that in fact PP2017 account could explain these observations. I will start with discussing whether the symbolic component[38] of the ACT-R model can capture S2017 effects.

To capture **E4**, we would have to ensure that the information about the verb enters the cue-matching process in some way. Presumably, it does not do it directly: the parser is likely not looking for verbs when it tries to resolve a reflexive. But we could encode the information about the verb on its subject, e.g. by using a feature like "+ subject of a speech verb". The reflexive would include this feature in the set of retrieval cues, biasing retrieval against subjects of other types of verbs. While possible, I consider this approach too clumsy to be our first choice for a few reasons. First, it requires some additional processing: during the encoding stage, the parser has to retrieve the subject at the verb and update the subject's representation with the verb type information[39]. Second, the information about whether the verb is a speech verb or not may only be useful for logophoric pronouns. But at the point where it has to be encoded, the parser has no idea of whether a reflexive pronoun will come up at all, even less about whether it will receive logophoric interpretation. Thus, often the information will be encoded, but not used, and if we think that memory space available to the parser is limited, using such redundant encodings

[38] I.e. operations defined in terms of features, procedural rules etc.

[39] Although the overhead may be minimal, if we assume that the subject is retrieved at the verb for integration purposes anyway.

would seem like an inoptimal strategy to adopt. Third, the feature I suggest is essentially privative. It is fine if we just want to distinguish between speech and perception verbs. But the Culy (1994) hierarchy that S2017 experiments are inspired by includes other verb types, e.g. verbs of thought and knowledge. Experiments would be required to see whether these verbs support or block lure-match effects, but if at least one more class of verbs does support them, a privative feature of the sort I suggested would not work anymore. Instead, a more general feature like "+ a subject of logophoric verb"[40] would be needed. This would not be too far fetched in languages with grammaticized logophoric pronouns, but would more questionable in English. Alternatively, we could use multiple features, one per verb type (e.g. "+ subject of perception verb", "+ subject of knowledge verb" etc.). But in this case the set of retrieval cues would have to include several verb-type related features, leading to a situation where no antecedent would be a perfect match to the set of retrieval cues. One may also wonder whether it is only the types of verbs on the Culy (1994) hierarchy that even put a feature on their subject. If it is the case, the parser should essentially have access to cross-linguistic generalizations ("these verbs are logophoric in other languages, so I need to pay attention to them in English"). If *all* verbs put some information about their semantic class on the subject, this again raises the question of whether the encodings are optimal from the point of view of memory space usage.

A second approach would be to avoid using features on the verb's subject,

[40] Somewhat sloppily, I use "logophoric verbs" to mean "verbs allowing their subjects to serve as antecedents for logophoric pronouns"

instead accessing the verb's information directly. The following antecedent retrieval algorithm could work: first, retrieve a suitable NP (e.g. matching features of the reflexive). Then, using the information on this NP, retrieve the verb it is the subject of, check whether the verb is the verb of speech or perception. Finally, depending on the result, either return or do not return the NP as the reflexive antecedent. It can not be the algorithm which is attempted first, otherwise it would exclude grammatical local and c-commanding antecedents which happen not to be subjects of speech verbs. Therefore, it is more plausible as a repair algorithm, for the cases when the parser fails to find a licit local antecedent. Interestingly enough, as we discuss in section 5.1.3, some patterns in the data may indeed indicate that lure-match effects are due to repair processes.

Overall, the fact **E4** could be captured even at the symbolic level of ACT-R. It may be harder to capture the fact **E5** (the type of embedding verb is irrelevant if the target antecedent is inanimate) while staying on the same level. I discuss some possibilities below but conclude that they may be too convoluted. Basically, we would have to manipulate the representation of the lure (in my first approach) or the processing procedures (in my second approach) based on whether the lure is the only consciousness center in the sentence. This could be done, but would require introducing yet another feature, something like "+ conscious". During anaphora resolution, the parser would retrieve all antecedents with such feature, and then, if there is more than one, either change the representation of the lure [41] or to change

[41] Which would require retrieving it accurately, which may not be trivial in a "structure-defeasible" model

the processing strategies by removing the check for verb type. While this is all possible, capturing **E5** in symbolic terms appears too ad hoc. As I discuss slightly later, though, these effects could be captured rather naturally by the sub-symbolic level of the model (i.e. in terms of the continuous activation values).

Capturing the fact **E6** at the symbolic level is also problematic. We could assume that indexical pronouns are underspecified for gender and mismatch the reflexive in only one feature, person. However, as S2017 discusses, this may not go through. Sloggett argues that this contradicts the observation that in his experiments sentences with indexical pronouns were judged as slightly less acceptable and were read slightly slower than sentences with *it* as the target (which presumably mismatches *himself/herself* in two features, animacy and gender). Alternatively, we could assume that the parser is aware of the presence of an indexical pronoun in a structurally intervening position with respect to the lure. But that would either mean that the parser tracks indexicals AND their position relative to other elements of the sentences, which is a priori implausible, or that it first accurately retrieves the target (again, not unproblematic given that it mismatches the set of retrieval cues), determines that it is an indexical and for some reason stops looking further, even though the retrieved target mismatches the reflexive.

Finally, capturing **E7** may be most problematic. In grammatical sentences the target always matches the morphological features on the reflexive at least as well as the lure does, and in addition only the target matches the structural cues, even if they are down-weighed. We do not expect to see any considerable amount of lure retrievals in such configurations even when the lure fully matches the morphological features on the reflexive, even less so when it does not. However, S2017 reports roughly 30% of the answers compatible with non-local reflexive resolution in the first case, and roughly 25% in the second. This pattern is hard to explain solely by the dynamics of memory system, unless one assumes that the amount of noise in the system is so high that it ignores a perfectly matching item 30% of the time. It seems odd to assume that memory access is so inefficient: if it performs that poorly in perfect conditions, what would it performance be in real-life situation with potentially less certainty about what the correct outcome is?

Summing up, explaining the empirical effects observed by S2017 purely in terms of memory processes is not impossible, but is rather challenging. But notice that these approaches stayed on the symbolic level of the ACT-R model. Now I would like to argue that it is potentially easier to explain all of the empirical facts if we take a look at the subsymbolic level, i.e. activation dynamics. In a basic ACT-R model, the only way to select an item is to make it more active than all the other competitors. Usual ways of doing so include raising baseline activation due to frequent and/or recent retrievals and getting the boost from matching retreival cues during memory access. As I have discussed, none of these mechanisms would be able to differentially change the lure's activation based on context (e.g. the type of the embedding verb). However, it is possible that activation can be directly affected by other factors as well, e.g. linguistic prominence (Engelmann et al., draft4). One could hypothesize that subjects of speech verbs are more prominent in the discourse being sources of information. If the boost in activation from this source is sufficient, only the subjects of speech verbs may act as efficient lures. One could also argue that animate entities are more prominent than inanimate - that would explain why the type of the embedding verb does not matter if its subject (the lure) is the only animate entity in the sentence. Finally, it is quite plausible that first and second person indexical pronouns are more prominent than third-person entities, being more central to the communicative situation - this could explain person blocking effects. Thus, discourse (or other related kind of) prominence might in principle explain effects for which S2017 postulates OP_{log}. I have to admit that these speculations remain only that - speculations - unless we are a) able to quantify the hypothesized changes in prominence and map them to specific values within an ACT-R model and b) show by simulations that the tentative predictions I make above indeed hold.

To conclude, empirical facts that S2017 uses as a strong empirical evidence for his model could potentially be explained even without recurring to OP_{log} and the claim that lure-match effects in reflexive resolution have grammatical nature. Explicit simulations and more theoretical work are required to test whether my way of accounting for these effects within a "structure-defeasible" cue-based model is viable.

1.5.1.2 S2017

S2017 account can accommodate all of the empirical facts listed above, except for **E3**. S2017 crucially relies on a null operator tracking prominent discourse entities in the left periphery of *complement* clauses. However, in Parker and Phillips (2017) Exp.2 the lure is embedded in a *relative* clause. Without OP_{log} present, only the local target should be returned during retrieval, and no lure-match effects should be observed. Sloggett has to fall back on an ad hoc explanation. He suggests that these effects represent a case of sub-command binding, a phenomenon occurring in Chinese. In sub-command configurations, an animate NP embedded inside an inanimate subject may still bind a reflexive. If this explanation turns out to be false, we will have to conclude that even if some lure-match effects reflect logophoric interpretations of the reflexives, some others still arise as a result of processing errors. That alone would be enough to argue for "structure-defeasible" models (following PP2017 and contra S2017).

Could the OP_{log} account be extended to cover the lure-match effects from lures inside relative clauses? In principle, nobody prevents us from saying that OP_{log} is located in the left periphery of *all* clauses, not just complement ones. This possibility would be supported by Charnavel and Huang (2018) who argue that sub-command binding is actually an instance of logophoric binding. If that is the case, and if one assumes that all instances of logophoric interpretations are mediated by something like OP_{log}, S2017 account would be able to explain PP2017 findings in a way analogous to the explanation of the fact **E5**: since in their stimuli the embedded lure was the only animate entity, it would also be the only possible logophoric antecedent. But on the other hand, this account would fail to explain the contrast in (14 Huang and Liu (2001): since, by hypothesis, OP_{log} is able to refer to Zhangsan in the first case, it is not clear why it would not be able to do so in the second. It would be hard to argue that OP_{log} is only present in the first case: by the time the parser reaches the embedding NP, the semantics of which licenses the logophoric interpretation[42], OP_{log} would have already been introduced to the parse. Thus, while postulating OP_{log} in all clauses could improve empirical coverage of S2017 account, it would also introduce new problems. In addition, as S2017 himself notices, it would go against the tendency for logophoric antecedents to be subjects of attitude predicates. I will discuss these issues in more detail in Chapter 2.

(14) a. Zhangsan$_i$ de baogao biaoshi tamen dui ziji$_i$ mei xinsin.
Zhangsan DE report indicate they to self no confidence
Zhangsans report indicates that they had no confidence in self.

b. *Zhangsan$_i$ de shibai biaoshi tamen dui ziji$_i$ mei xinxin.
Zhangsan DE failure indicate they to self no confidence
Zhangsans failure indicates that they have no confidence in him.

With respect to the fact **E7**, although S2017 uses it to support his account, this empirical observation may be even more problematic for him than for PP2017. According to Sloggett's hypothesis, the silent logophoric operator does not carry any ϕ-features. Thus, while in the PP2017 scenario the target had a single feature advantage over the lure (only the target being the reflexive's clausemate), in S2017 scenario the target is a better match than OP_{log} in at least two features - gender and number (and potentially person and/or animacy). Thus, one would have to assume an even noisier retrieval system than in PP2017 scenario to account for these patterns.

[42] As Huang and Liu (2001) put it: "This is because [the first sentence] implies that Zhangsan himself indicates that they had no confidence in him. (If his report indicates P then he indicates P.) No similar implication holds of the unacceptable [second sentence]."

1.5.2 Empirical coverage: summary

To sum up, neither account covers all of the available data. Nor is any a clear winner. The original PP2017 account fails to explain the evidence for factors other than the number of matching features affecting lure-match effects. However, I have suggested a way in which this account could be extended to cover most of the problematic evidence. The S2017 account covers most of the data to begin with, but extending it to cover the data from the non-c-commanding lures is not straightforward. It either requires introducing an additional mechanism, not tightly related to the main proposal, or else complicating the interpretation of other empirical facts. In addition, I have argued that the interpretation data reported by S2017 may be problematic for both accounts.

1.5.3 My experiments

The main body of the thesis is an attempt to tease the two accounts apart using additional evidence. I report the following experimental work.

First I investigate one of the weak points of the S2017 account. Chapter 2 reports the results of an experiment investigating the configurations with the lures inside relative clauses, which Sloggett's account does not cover directly. S2017 argues that in this instance another grammatical option is used (sub-command binding, which I discuss in more detail in the corresponding experimental chapter). This grammatical option crucially requires the head of the relative clause to be inanimate, and indeed, this is the configuration PP2017 had. Therefore, I am going to test same configurations but with animate RC heads. To preview, the results will be taken to support PP2017 account.

Second, given this support, I investigate further predictions of PP2017 account to validate the conclusions from Chapter 2. The experiments by PP2017 do not provide reliable evidence regarding the use of c-command information. Their results are compatible either with a model which faithfully encodes c-commands but does not manage to use this information to rule out illicit antecedents, and a model which encodes c-command information only approximately (thus not covering relevant cases) but follows this approximation faithfully. I attempt to distinguish between these two possibilities by considering the case of quantificational (QP) lures. As I discuss in detail in Chapter 3, experimental evidence suggests that QPs in the wrong position (non-scoping/non c-commanding) cannot bind pronouns like "him" (Kush, 2013). From the syntactic point of view, binding works in the same way for pronouns and reflexives. If the c-command information fails to uniquely direct the parser's decisions, as PP2017 suggest, we should observe lure-match effects even from non-c-commanding QP lures. If, on the other hand, some approximation, such as Kush et al. (2015) ACCESSIBLE is used, non-c-commanding QP lures will not elicit lure-match effects. Third, and finally, I perform a series of two experiments addressing an issue I ran into: I failed to find strong evidence for lure-match effects even from NP lures. Thus, I perform a direct replication of PP2017 Experiment 3 to determine the robustness of the effects they report. These replications also help to verify whether extra-syntactic factors, such as (non-)nativeness of experimenter, might have affected the results. If such factors do affect lure-match effects, it would be more readily compatible with PP2017 account. Two experiments addressing replication issues are reported in Chapter 4.

Chapter 2: Interference effects from non-c-commanding lures

I am going to start the thesis with investigating the only strong evidence against S2017 account: lure-match effects for the lures embedded inside a relative clause (we will call them "RC-lures" for short), e.g. "the students" in "The tea that the students drank calmed themselves" (Parker and Phillips, 2017, Exp.2). This evidence is problematic for S2017: his account crucially relies on a null operator in the left periphery of *complement* clauses to account for lure-match effects and says nothing about *relative* clauses. To save the situation, S2017 suggests an ad hoc explanation: perhaps, the lure-match effects observed with RC-lures reflect a grammatical strategy of sub-command binding (see below). While different in substance, this explanation is similar in spirit to the main S2017's account: it attributes lure-match effects to a grammatical mechanism, and not to a faulty processing routines. It is important to empirically test this suggestion. If it turns out that S2017 is incorrect and we have to invoke processing errors to cover that evidence, the support for his account weakens. At the very least, it would have to be modified to explain why in some cases the structure does accurately constrain the search and in others, very similar cases, it does not. In what follows, I briefly explain the phenomenon of sub-command binding and discuss my experiment testing S2017's hypothesis.

Sub-command binding is a grammatical phenomenon observed in Chinese. Several empirical facts are relevant for this discussion. First, the anaphor *ziji* requires animate antecedents (Tang, 1989)[1], as demonstrated by (1).

(1) a. Wo taoyan ziji.
I dislike ANA
I dislike myself.

b. Xiaomao zai tian ziji de lian.
little_cat is lick ANA DE face
The kitten is licking its own face.

c. *Men guanshang le ziji.
door close PER ANA
The door closed itself.

Second, *ziji* can have a non-c-commanding NP (say, NP$_1$) as its antecedent, as long as this NP$_1$ is embedded within another NP$_2$ which does c-command *ziji* - this phenomenon is referred to as "sub-command binding" (2). Third, sub-command binding is only possible if the embedding NP$_2$ is inanimate (3) Tang (1989).

(2) a. [[Zhangsan$_i$ de] jiaoao]$_j$ hai le ziji$_{i/*j}$
Zhangsan$_i$ DE pride hurt PER ANA$_i$
Zhangsan$_i$'s arrogance harmed him$_i$

b. [[Zhangsan$_i$ zuoshi xiaoxin de] taidu]$_j$ jiu le ziji$_{i/*j}$ yiming
Zhangsan do_thing careful DE attitude save PER ANA one life
Zhangsan's cautious attitude saved him.

(3) a. [[Zhangsan$_i$ de] baba]$_j$ dui ziji$_{*i/j}$ mei xinxin.
Zhangsan DE father to ANA no confidence
Zhangsan's$_i$ father$_j$ has no confidence in himself$_{*i/j}$

[1] At least, this is the accepted generalization. But see the discussion of Charnavel and Huang (2018) later in this section.

b. [Zhangsan$_i$ pengdao de] nage ren]$_j$ dui ziji$_{*i/j}$ mei xinxin
Zhangsan meet DE that-CL man to ANA no confidence
The man$_i$ that Zhangsan$_j$ met had no confidence in himself$_{*i/j}$

To capture these patterns, Tang (1989) argues that in Chinese the structural requirements on the anaphor's binders should be slightly relaxed. She suggests the following binding principles for the Chinese:

A. β SUB-COMMANDS α iff

1. β c-commands α or
2. β is an NP contained in an NP that c-commands α or that sub-commands α, and any argument containing β is in subject position.

B. A POTENTIAL BINDER for α is any NP which satisfies all conditions of being a binder of α except that it is not yet coindexed with α.

C. A reflexive α can be BOUND BY β iff

1. β is coindexed with α, and
2. β sub-commands α, and
3. β is not contained in a potential binder of α.

The sub-command stipulation obviously captures the structural requirement. The condition C3 together with the definition B2 helps to explain why sub-command binding is only possible in (2), but not in (3): only in the first case the embedded antecedent is NOT contained in a potential binder of *ziji*.

Parker and Phillips (2017) only used inanimate matrix subjects when investigating RC-lures. Given this, Sloggett suggests that the sub-command binding can underlie lure-match effects from RC-lures observed by Parker and Phillips (2017). As S2017 himself points out, if this hypothesis is correct, the lure-match effects should not be observed (or at least be observed to a smaller degree) with the animate matrix subjects. Investigating such configurations is the primary goal of the current chapter. However, before turning to the presentation of the experiment, I discuss a potential *theoretical* counter-argument to S2017 suggestion.

Charnavel and Huang (2018) argue that the cases of sub-command binding in fact represent instances of logophoric interpretation of the pronoun. They observe (following Huang and Liu (2001)) the following contrast:

(4) a. Zhangsan$_i$ de baogao biaoshi tamen dui ziji$_i$ mei xinsin.
Zhangsan DE report indicate they to ziji no confidence
Zhangsans report indicates that they had no confidence in self.

b. *Zhangsan$_i$ de shibai biaoshi tamen dui ziji$_i$ mei xinxin.
Zhangsan DE failure indicate they to self no confidence
Zhangsans failure indicates that they have no confidence in him.

They argue that the contrast arises because in (4a) the matrix subject creates conditions for a logophoric interpretation: namely, it allows for an implication that Zhangsan's perspective is conveyed by the report, while no such implication is possible in (4b). So, only in the first case Zhangsan is the perspective holder in the sentence, which makes it a possible antecedent for a logophoric pronoun.

Charnavel and Huang (2018) support their claim with the results of an acceptability judgment study. They first show that *ziji* is acceptable with inanimate antecedents, in contrast to generally accepted claims in Tang (1989): sentences like (5a) received high ratings (4.69 on average, with the judgments being made on a scale between 1 (worst) to 6 (best)). Then they show that sentences with inanimate sub-commanding antecedents (5b) receive (significantly) lower ratings as compared to (5a) (3.35 on average). From this they conclude that apparent sub-command binding results from the logophoric interpretation of the pronoun: logophoric pronouns require their antecedents to be animate, while *ziji* does not. Thus we only expect the inanimate sub-commanding antecedent to cause judgments to degrade if *ziji* is being interpreted logophorically.

(5) a. [Zhe ke shu de shuguan]$_i$ tai chen, ya wan le ziji$_i$.
this CL tree DE tree_crown too heavy, burden bent ASP REFL
[The crown of this tree]$_i$ is too heavy. It$_i$ bent itself$_i$.

b. *[Zhe ke shu]$_i$ de guoshi ya wan le ziji$_i$.
this CL tree DE fruit press bent ASP REFL
The fruits of [this tree]$_i$ bent it$_i$.

If Charnavel and Huang (2018) conclusions are correct, PP2017 results have to reflect an extra-grammatical strategy: RC-lures should only be accessible when the matrix subject creates appropriate logophoric context. However, Parker and Phillips (2017) used stimuli like “The broken zipper that the skilled tailors tried to fix pinched themselves” or “The safety net that the brave policeman used saved themselves”. Arguably, in such sentences the embedding NPs do not make the embedded NPs prominent perspective holders (in any case, no more than “failure” in (4b)). Thus, the lure-match effect Parker and Phillips (2017) can not be reduced to the properties of syntactic configuration.

While Charnavel and Huang (2018) arguments are interesting, we think that they are insufficient to reduce all cases of sub-command binding to instances of logophoric interpretations. First, Tang (1989) gives the following example:

(6) a. [[Zhangsan$_i$ de] baba$_j$ de] qian]$_k$ bei Ziji$_{*i/j/*k}$ de pengyou touzou
Zhangsan DE father DE money BEI ANA DE friend steal
le.
PER
Zhangsan's father's money was stolen by his friend.

In this example it is again not clear how "money" would encourage to treat "Zhangsan's father" as a prominent perspective center, in the way that "report" in "Zhangsan's report" does. Second, as Huang and Liu (2001) discuss, logophoric interpetation of *ziji* is subject to person-blocking effects (i.e. if a first or second person pronoun intervenes between *ziji* and another antecedent, *ziji* obligatorily refers to the first/second person pronoun) (7). However, such effects are not present in cases of sub-commanding antecedents (8).

(7) a. Zhangsan$_i$ dui wo shuo ziji$_i$ piping-le Lisi.
Zhangsan to me say self criticize-Perf Lisi
Zhangsan$_i$ said to me that he$_i$ criticized Lisi.

b. *Zhangsan$_i$ shuo wo piping-le ziji$_i$
Zhangsan say I criticize-Perf self
Zhangsan$_i$ said that I criticized him$_i$

(8) Zhangsan$_i$ de biaoqing gaosu wo$_j$ [ziji$_{i/*j}$ shi
Zhangsan DE expression tell me self is innocent

wugude].

Zhangsan$_i$s [facial] expression tells me that he$_i$ is innocent.

These counter-examples to Charnavel and Huang (2018) conclusions still make it possible for S2017 hypothesis to be viable. Therefore, I now turn to its experimental evaluation.

2.1 Experiment 1

2.1.1 Participants

53 members of University of Maryland participated in the experiment for a class credit or a payment of $12 (11 M, 42 F; age range: 18-27; mean age: 20.1, SD: 1.76). The experimental session took around 45 minutes, including setup and calibration. I excluded data from 3 participants: two failed to finish the experiment due to problems with calibration and for one the recorded reading patterns were too erratic to be used. The remaining 50 participants entered the analysis[2].

2.1.2 Materials

The study used 2x2 factorial design with TARGET ANIMACY (whether the matrix subject was animate or inanimate) and LURE MATCH (whether the lure matched or mismatched the reflexive in features) as factors. Overall, I constructed 24 sets

[2] I aimed for 60, to have double the number of participants in Parker and Phillips (2017) Exp.2 which I model this study after; however, I couldn't reach that goal due to time limitations

of 4 items, exemplified in Table 2.1. The target antecedent always mismatched the reflexive in features (animacy and number for sentences with inanimate matrix subjects, and gender and number for sentences with animate matrix subjects); thus, all critical sentences were ungrammatical. The items with inanimate matrix subjects were borrowed directly from Parker and Phillips (2017) Exp.2 materials; I made sure to select items which did not contain pronouns or gaps in the spillover regions, in order not to force additional memory retrievals. The items with animate matrix subjects were constructed from scratch. Thus, the two subsets of items were not lexically aligned; moreover, the critical pronouns was "themselves" in the first group and "himself/herself" in the second[3]. While this design choice potentially introduces more noise to the comparisons, I felt it was justified: it gives us both the opportunity to test the effect of interest (lure-match effects from lures in relative clauses with animate RC heads) and to replicate previous findings (which is important to establish the reliability of PP2017 findings, given that no other study has investigated similar configurations).

In addition to the sentences with reflexives, which were of the primary interest, I used a control set of 24 agreement items borrowed from Wagers et al. (2009). An example is given in (9). A standard 2x2 design was used, with the factors being

[3] Changing the pronouns from "themselves" to "himself/herself" is inevitable for the sentences with animate matrix subjects. Since the target is animate, we need a reflexive which can differ from the target in gender and number to be able to create a 2-feature mismatch configuration. But "themselves" does not carry gender, so if we used "themselves", we would only be able to induce a 1-feature mismatch (in number) with the target.

LURE NUMBER and VERB NUMBER. Thus, I had 6 items per conditions. Head nouns were always singular. The lure and the verb were always separated by an adverb to avoid potential confounds[4].

(9) The key to the cell(s) unsurprisingly was/were rusty from many years of disuse.

Finally, I had 66 fillers, falling in the following categories. 12 fillers had the same structure as experimental items with animate matrix subjects, so that people would not only see this construction in ungrammatical sentences. In addition, 6 fillers with the same structure used "themselves" as the reflexive, so that "themselves" was not uniquely associated with ungrammatical items. 6 fillers were of the structure I used in the previous experiments: "NP said that [NP reflexive]"; these sentences were introduced to make sure that the target was not always the linearly farthest NP. 12 fillers started with an inanimate noun heading a relative clause; these were introduced to make sure that inanimate subjects were not uniquely associated with ungrammatical sentences. 12 fillers were sentences with no constraints on the structure containing a reflexive pronoun. Finally, 12 fillers were just sentences with no special properties. All of the fillers were grammatical.

Overall, I had 24 critical sentences with reflexives (all ungrammatical), 24

[4] Briefly, people may experience slowdown on "cells" just because it's plural, and thus arguably more complex morphologically and semantically. If the critical verb immediately follows the noun, this slow-down can affect the reading times on the verb and potentially mask intrusion effects. See Wagers et al. (2009) for more details.

Animate targets

Lure match

The skilled businesswomen that the inexperienced ***manager*** criticized at the meeting blamed ***himself*** for the inaccurate report.

Lure mismatch

The skilled businesswomen that the inexperienced ***secretary*** criticized at the meeting blamed ***himself*** for the inaccurate report.

Inanimate targets

Lure match

The thank you letter that the helpful ***secretaries*** received praised ***themselves*** for a great job.

Lure mismatch

The thank you letter that the helpful ***secretary*** received praised ***themselves*** for a great job.

Table 2.1: Experiment 1 materials example
Targets are underlined, lures are bolded.

control agreement attraction sentences (12 grammatical) and 66 fillers (all grammatical), for a total of 108 sentences. Thus the grammatical-to-ungrammatical ratio was 2:1. All sentences were accompanied by a forced choice Yes/No question.

2.1.3 Procedure

For the purposes of randomization, reflexive and agreement sentences were treated as same. The 48 stimuli were distributed into 4 lists in a Latin Square design, for 6 agreement and 6 reflexive sentences per condition. Fillers were added to these lists, which were then pseudo-randomized with the constraint that no more than 2 experimental items occur in a row. Each experimental list started with 7 practice items.

The items were presented on an LCD screen with the resolution of 1280x720 in a 12-point fixed width font (Courier). The maximum length of a line fitting on the screen was 139 symbols, so all items fit on a single line.

Eye movements were recorded using EyeLink 1000 tower-mounted eye-tracker. The distance between the tower and the screen was 37 inches. The sampling rate was 1000 Hz. Participants had binocular vision during the experiment, but only the samples from one eye[5] were recorded. Participants' heads were immobilized using a chin rest and a forehead restraint. In the beginning of the session, the eye-tracker was calibrated using 9-dot calibration display. Calibration was repeated throughout the experiment as necessary.

Each trial started with a fixation mark on the left of the screen. Participants had to fixate it in order for the stimuli to appear on the screen. After having read the sentence, participants pressed a button in order to display a question related to the sentence. Participants answered the question by pressing one of two associated buttons. After that, the next trial started.

2.1.4 Analysis

I analyzed two regions of interest in each set of conditions: critical and spillover. For agreement sentences, the critical region included the verb and the following word (two following words, if the second word was a determiner). For reflexive sentences, the critical region included the reflexive and the last three letters of the preceding

[5] Normally, the right one, unless tracking the left eye was the only option giving a stable calibration.

word. This was done to increase the chances of having at least one fixation in the critical region, following Kush et al. (2015) and S2017. The spillover region was defined for both types of stimuli as two words following the critical region (three, if the second word was a determiner). Spaces were included in the region to the right.

In each of the regions, I looked at the following eye-tracking measures: first pass, the sum of all fixations before the region is exited to the right or to the left; regression path, the sum of all fixations on the region from the moment it is first entered from the left and to the moment it is exited to the right, including fixations in the previous regions; total time, the sum of all fixations on the region. I chose these measures following S2017. PP2017 analyze first-pass, regression path, right bound (the sum of all fixations on the region from the moment it is first entered from the left and to the moment it is exited to the right, NOT including fixations in the previous regions) and re-read (the sum of all fixations on the regions after it has been exited once). However, I decided not to include right bound and re-read in the analysis to reduce the number of comparisons I was making.

Prior to statistical analysis, the data were manually preprocessed using Eye-Doctor[6] to remove blinks and ensure that fixations align with the text. Fixations shorter than 80 ms or longer than 1000ms were automatically rejected before calculating eye-tracking measures using custom scripts[7]. Missing values were discarded form the analyses. See section 4.1.6 for a detailed discussion of how analyses choices might affect the final conclusions. In particular, rejecting missing values (instead of

[6] `https://blogs.umass.edu/eyelab/software/`

[7] `https://github.com/UMDLinguistics/EyePy`

replacing them with zeros, as PP2017) might lead to more precise model estimates.

I chose to log-transform reading times to bring the models' residuals closer to normality. The transformed values were analyzed with linear mixed effects models in R (R Core Team, 2014) using *lme4* package (Bates et al., 2015b). Separate models were fit to the data from reflexive and agreement conditions. The following fixed effects were specified (numerical values in parentheses indicate contrast coding coefficients; sum coding was used for all fixed effects). Agreement conditions: TARGET MATCH (match = -0.5 vs. mismatch = 0.5), LURE MATCH (match = -0.5 vs. mismatch = 0.5) and their interaction. Reflexive conditions: TARGET ANIMACY (animate = -0.5 vs. inanimate = 0.5), LURE MATCH (match = -0.5 vs mismatch = 0.5 in features with the reflexive) and their interaction. A fixed effect was taken to be significant if the absolute value of the associated t statistic was ≥ 2 (Gelman and Hill, 2007). Models random effects structure was fully specified, including random intercepts and slopes for each fixed effect per subjects and items, following recommendations by Barr et al. (2013). If the model with the maximal random effect structure did not converge[8], I simplifed it by removing random slopes. To be internally consistent, I always dropped the random effects in the order, shown in Table 2.2.

I start with removing the interactions as the most complex term (see, e.g. Bates et al. (2015a), p.6, for a claim that this can be considered standard ap-

[8] As determined by diagnostics reported by `lmer`. One has to access the internal structure of the model fit and look at the information contained in `m@optinfo$conv$lme4`, where `m` is the model object.

Agreement	Reflexives
$verb.num : lure.num \mid item$	$target.animacy : lure.match \mid item$
$verb.num : lure.num \mid participant$	$target.animacy : lure.match \mid participant$
$verb.num \mid item$	$lure.match \mid item$
$lure.num \mid item$	$lure.match \mid item$
$verb.num \mid participant$	$target.animacy \mid participant$
$lure.num \mid participant$	$target.animacy \mid participant$

Table 2.2: Order of random effect structure simplification in Exp.1.

lme4 syntax is used: ":" indicates interactions and "—" separates grouping factors. Effects were removed from the model, starting from the top.

proach[9]). Then, I remove random effects by item, making an assumption that the variability among items is smaller than among participants, and random slopes by items would make a smaller contribution to the model. As a final step, all models were automatically assessed to check whether any random effects had correlations of 1 or -1, suggesting that a given random effects structure may be too complicated to estimate given the data (Bates et al., 2015a). Such random effects were dropped from model specification, starting with the interactions [10].

2.1.5 Results

Mean question answering accuracy was 91%. Fig.2.1 and 2.2 show the observed reading times means for agreement and reflexive sentences respectively. The full set

[9] The following StackExchange discussion may also be of interest: `https://stats.stackexchange.com/questions/323273/what-to-do-with-random-effects-correlation-that-equals-1-or-1`

[10] Admittedly, this is a somewhat simplistic approach to deal with overly complex random effect structures. See Bates et al. (2015a) for a more principled way to identify redundant random effects in `lmer` models.

of model coefficients is reported in Appendix C.

I will start the discussion with the results for control agreement sentences. I found evidence for the effect of grammaticality (ungrammatical sentences being read slower): the main effect of TARGET MATCH was significant in all three reading measures at the critical region and in regression path at spillover. The effect of LURE MATCH was significant in regression path and total times in the critical region: when the lure matched the verb in number, the sentences were read faster. Interestingly, these effects were not moderated by a significant interaction: matching lures were read faster regardless of whether the sentence was grammatical or not. This means that I did not find enough evidence for grammatical assymetry and the data is consistent with the presence of both illusions of grammaticality and ungrammaticality. The TARGET MATCH X LURE MATCH interaction did reach significance but only in regression path at the spillover region. Pairwise comparisons confirm that the interaction is driven by the simple main effect of LURE MATCH within ungrammatical sentences: sentences with the lure matching the verb were read faster. This pattern of results is interesting: the interaction in the spillover region is indicative of the classical attraction pattern (only illusions of ungrammaticality arise). On the other hand, the main effect of LURE MATCH and no interaction at the critical region is in line with the conclusions of Hammerly et al. (draft.april.2018): both illusions of grammaticality and ungrammaticality can be observed for subject-verb agreement. I return to this pattern in the discussion section.

Now turning to reflexive conditions, three effects reached statistical significance. First, two main effects of TARGET ANIMACY in total times at the critical

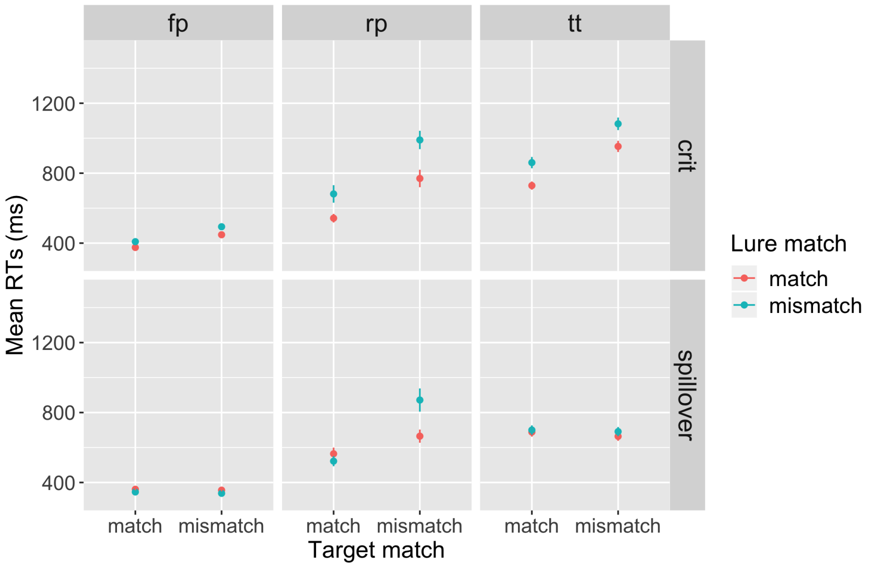

Figure 2.1: Experiment 1. RT means for agreement sentences.

Columns correspond to eye-tracking measures: fp - first-pass, rp - regression path, rp - total times. Rows correspond to ROIs. Errors bars represent standard error of the mean, adjusted for participant variability (Cousineau, 2005; Morey, 2008)

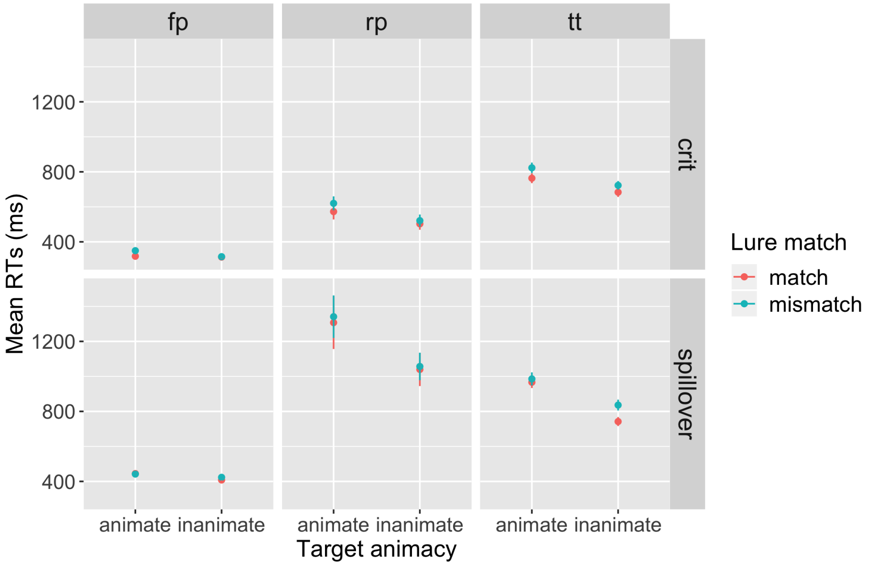

Figure 2.2: Experiment 1. RT means for reflexive sentences.

Columns correspond to eye-tracking measures: fp - first-pass, rp - regression path, rp - total times. Rows correspond to ROIs. Errors bars represent standard error of the mean, adjusted for participant variability (Cousineau, 2005; Morey, 2008)

and spillover regions: sentences with animate targets were read slower. Second, the main effect of LURE MATCH in total times at spillover: sentences with matching lures were read faster. This last effect provides some evidence that a) lures within RCs can provoke interference and b) this effect does not depend on the animacy of the matrix subjects, not supporting Sloggett's suggestion. I will argue that these conclusions are tentatively correct, although I will provide more fine-grained discussion below.

Multiple comparison corrections Von der Malsburg and Angele (2017) suggest that corrections for multiple comparisons should be applied in eye-tracking studies when several models are fit to different regions of interest and eye-tracking measures. Their simulations suggest that Bonferroni correction is an appropriate one and does not result in losing overly much power. In the current experiment I fit 2 (reflexive and agreement) x 2 (ROIs) x 3 (measures) = 12 models. Bonferroni correction would result in the following corrected α: $0.05/12 = 0.004$. The corresponding t-value for a two-sided test is ± 2.87[11]. Almost all significant effects survive the correction with two exceptions: in agreement sentences, the effect of lure match in regression path at the critical region; in reflexive sentences, main effect of LURE MATCH in total times at the spillover region.

[11] Approximated from normal distribution following Jäger et al. (2015). The corresponding R command is `qnorm(0.004/2)`.

2.1.6 Exploratory analysis

In my main analysis I followed the procedure from S2017 since I consider it better along several dimensions (see Chapter 4 for discussion). However, I want to check whether the choice of analysis procedure matters. Thus, in addition to the main analysis, I also perform a set of exploratory analyses to better understand the alignment between my results and those by Parker and Phillips. First, I calculated average reaction times for right-bound and re-read times: these measures were not included in the main analyses to reduce the number of statistical comparisons, but they would allow for a closer comparison of average RT patterns with PP2017. Second, I analyzed the extended set of ROIs with the same procedure that PP2017did: the critical region was defined as including the reflexive alone, without any material to the left; missing values were replaced with zeros; no trimming of exceedingly long reading times was performed.

I have also performed additional exploratory statistical analyses. In addition to the main model, I look at simple main effects of LURE MATCH by levels of TARGET ANIMACY: in Parker and Phillips (2017) design the lure-match effect was a simple main effect, while in my case I am looking at main effects and interactions. The coefficients for statistical models are presented in Fig. 2.3. Each dot in the plot corresponds to a $\hat{\beta}$ value from some model. We consider models fit to 4 different datasets, resulting from the variation of pre-processing choices. These 4 datasets are plotted along the Y axis of each plot. Each line in the plot corresponds to a variant of analysis, with the variable treatments delimited by underscores. The first component corresponds to the study; the second component corresponds to the critical region markup ("nonext" - critical region includes only the reflexive itself; "ext" - critical region includes the reflexive and three characters to the left; the second component corresponds to missing values treatment ("narm" - remove, "nazero" - replace with zero). Each column corresponds to a fixed effect in a given region: TA - target animacy; LM - lure match; TA x LM - their interaction; prws_A/IA - pairwise comparisons, corresponding to simple main effects of lure match within target animacy conditions. The first five columns display estimates obtained for the critical region, the last five panels - estimates for the spillover region.

Notice that the coefficients are not easily interpretable on their own, since they are on log scale. Thus, I will only discuss what inferences one would make, if one was simply looking at the coefficients and checking whether they are significantly different from zero. I take coefficients with $|t| > 2$ to be significantly different from zero, and mark them in red.

I discuss first what conclusions Parker and Phillips (2017) would have made based on my data. We need to look at the "nonext_na.zero" version of analysis (i.e., critical region comprising only the reflexive itself with missing values replaceed with zeros), the "Prws_IA" column (since PP2017 only had that single condition) and all eye-tracking measures except for total times. In their analyses, PP2017 found significant lure-match effect in first-pass, right-bound and re-read times at the critical region and re-read times at spillover. As can be seen from Fig. 2.3, in my data only the last effect is replicated.

Turning to other possible analysis variants, we can see that the TARGET AN-

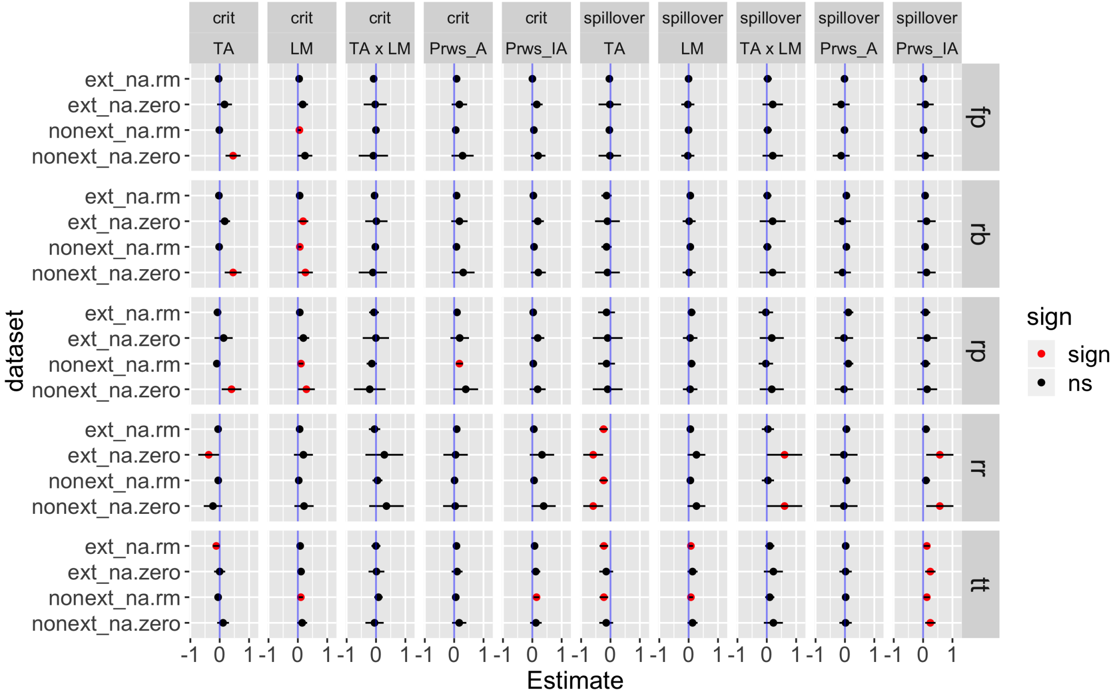

Figure 2.3: Sensitivity analysis: Model coefficients in Exp.1

Errors bars represent standard deviation of the estimates. See text for further description of columns and row labels.

imacy x Lure match interaction virtually never reaches significance, except in regression path at spillover for analysis variants where missing values are replaced with zeros. Main effect of lure match is mostly detectable when one chooses a critical region comprised of reflexive alone (all eye-tracking measures except total times if one removes the missing values, and right-bound and regression path if one replaces them with zeros). If one chooses an extended critical region, the only significant effect of lure match is observed in right-bound times. At spillover, the effect is observable at total times for the variants of analysis removing missing values - this is what I reported in the main analysis.

Simple main effect of lure-match for sentences with animate matrix subjects is only significant at regression path if one chooses to look at reflexive only and to remove missing values. Simple main effects of LURE-MATCH for sentences with inanimate matrix subjects are a little bit easier to show up as significant: at the critical region, they appear in total times if one looks at the reflexive only and removes the missing values. At the spillover, they appear in regression path for the variants of analysis replacing missing values with zeros and in total times for all variants of analyses.

2.1.7 Discussion

I start the discussion with the control agreement attraction sentences. I have found reliable evidence for agreement attraction, which suggests that the experiment worked as intended and interference effects can be detected in the data I collected. Interestingly, I did not always observe the classical pattern: lure-match effects only within ungrammatical sentences. I did observe it in regression path at the spillover; however, at the critical region there was evidence for lure-match effects within both grammatical and ungrammatical conditions. That is, I observed both illusions of grammaticality and ungrammaticality, supporting the claims by Hammerly et al. (draft.april.2018) that grammatical asymmetry is observed only in a proportion of the experiments. Also in line with their conclusions is the fact that the current experiment had the lowest grammatical-to-ungrammatical ratio among all my experiments which included agreement conditions (2:1 vs. 3:1 in Exp.3 and 5:1 in Exp.4). Hammerly et al. (draft.april.2018) observed that as the proportion of ungrammatical items increases, the illusions of ungrammaticality are more likely to be observed. Notice though that in their study, illusions of ungrammaticality were not observed even when the grammatical-to-ungrammatical ratio was 1:1 (even lower than in the current experiment), and only appeared when the ratio lowered to 1:2. Thus, while the direction of the influence between my experiment and Hammerly et al. (draft.april.2018) appears to be similar, the cut-off points for the emergence of ungrammaticality illusions do not agree. Therefore, I prefer not to make theoretical claims about the cause of the symmetrical lure-match effect I observe, only noting that the effect is, indeed, symmetrical, contra what the accepted generalization is.

This symmetry bears on the main question I address in my thesis. As I discussed in Chapter 1, the cue-based model by Lewis and Vasishth (2005) is not able to accommodate such patterns. While the addition of the prominence correction suggested by Engelmann et al. (draft4) would allow the model to capture those patterns, it is not at all clear whether there are linguistic reasons for applying the prominence correction (i.e., why would we think that "cabinets" is more prominent than "key" in "The key to the cabinets are..."?) Thus, symmetrical agreement attraction patterns are a point against Lewis and Vasishth (2005) model. Notice that this model underlies both PP2017 and S2017 explanations for the reflexive data, so if we end up doubting it, we will have to doubt these accounts as well. It is especially so for PP2017, who explain the lure-match effects in terms of faulty memory search. It is less critical for S2017: while he uses the cue-based architecture to capture the fact that lure-match effects are only observed in sentences with targets mismatching the reflexive in two features, the lure-match effect itself does not rely on the properties of cue-based memory search. In principle, any search method which only searches the local domain of the reflexive and mostly returns the overt target, but occasionally returns the logophoric operator, could work for S2017 model. (Although it would still have to explain somehow why this other search method is more prone to return the logophoric operator when the target mismatches in features). Thus, symmetrical agreement attraction may be slightly more problematic for PP2017.

Now I turn to reflexive conditions, which are of main interest. The central goal of the experiment was to further investigate the lure-match effects from lures inside a relative clause ("RC-lures"). This configuration is interesting, since this is the only instance of lure-match effects not covered by the S2017 account. As such, it may be the only evidence for lure-match effects emerging from ungrammatical dependency resolution and for the sentence processing not respecting structural constraints.

I were asking two questions. First, how strong is the evidence for the existence of lure-match effects from RC-lures? Unlike with the effects with c-commanding lures, that have been replicated in several experiments, the RC-lures configuration has only been tested once by Parker and Phillips (2017). If it turns out that the effects were artifactual, they cannot be used as an argument against the S2017 account. Second, does the animacy of the matrix subject affects the lure-match effect? If the lure-match effect truly reflects ungrammatical resolution of the dependency, it should not. On the other hand, if, as S2017 suggested, the effect reflects a grammatical phenomenon of sub-command binding, we should only observe it when the matrix subject is inanimate. I discuss these questions in turn.

The first question I ask is - have I managed to replicate PP2017 and do we have reasons to believe that lure-match effects from RC-lures in “inanimate” conditions are real? I argue that the experiment does provide the evidence for this, albeit limited. On the one hand, statistical analysis does confirm the presence of lure-match effects in total times at spillover. An exploratory analysis also shows that if I had followed Parker and Phillips (2017) procedure, I would have found a simple main effect of LURE MATCH within sentences with inanimate matrix subjects in the regression path at the spillover - this is in line with PP2017 results, who also find such an effect. It also appears that regardless of the pre-processing decisions, at least *some* effect indicative of interference would come out significant.

I have to note, though, that these conclusions should be taken with a grain of salt. First, the main effect of lure match in total times I observed in my primary analysis would not survive the Bonferroni correction for multiple comparisons. If I were being conservative, I would not claim that I found evidence for the interference. Furthermore, it may be worrying that the prima facie inconsequential choices one makes during pre-processing (i.e. whether the critical character does or does not include the three last characters of the word preceding the reflexive) affect when and what exact effects indicative of the interference are becoming significant. To us, this indicates that the amount of evidence available in the data is not overwhelming, and more highly powered experiments are needed to make better conclusions. One final worry is that the patterns I observe do not completely align with what Parker and Phillips (2017) report. They observed significant lure-match effects in four eye-tracking measures in the critical region, while I did not. Also, as Fig.2.4 shows, numerical magnitudes of the lure-match effects are about twice as small than what

PP2017 report. I will defer the discussion of these numerical differences until later chapters, where such differences emerge again and are investigated in more detail.

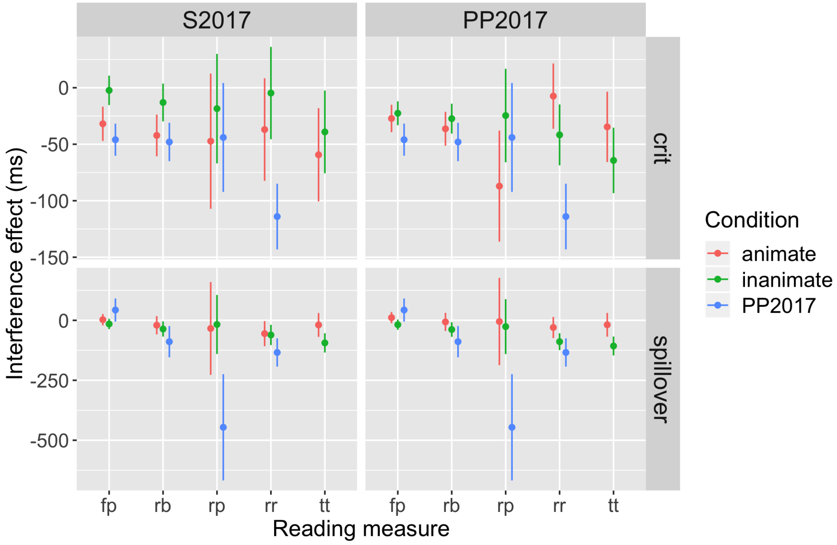

Figure 2.4: Experiment 1. Comparisons of the interference effect magnitudes in reflexive conditions with Parker and Phillips (2017).

Interference effect is calculated as the difference between lure match and lure mismatch conditions. Errors bars represent standard error of the difference of the means (calculated under the assumption that RTs in lure match and lure mismatch conditions are not correlated. This is likely false, since they come from the same subjects, thus, the SEs are overestimates).

Now that I have tentatively concluded that my results are roughly in line with PP2017 conclusions, the next question is: do I have any evidence for the difference for the lure-match effect from RC-lures within "animate" conditions? This is the critical question to test S2017 subcommand hypothesis: if it is correct, I should not find any lure-match effects in "animate" conditions. The evidence appears to speak against S2017 hypothesis, although again, in a rather limited way.

The only statistically significant effect pointing out in this direction is the main effect of LURE MATCH in the total times at spillover, not accompanied by a significant TARGET ANIMACY X LURE MATCH interaction. This results suggests that there is not enough evidence to conclude that LURE MATCH effect behaves differently depending on the animacy of the matrix subject[12]. Numerical values suggest similar conclusions. Fig.2.4 shows the magnitude of lure-match effect in "animate" and "inanimate" conditions along with the effects from the original PP2017 study. We can see that regardless of the pre-processing procedure, lure-match effects are present in the "animate" conditions. The lure-match effect in conditions with inanimate matrix subjects is small, as I have already discussed. On the contrary, in the sentences with animate matrix subjects the effect is around 50ms in magnitude, which, as we will see, brings it within the range of interference effects I observe in the rest of my experiments. In the early eye-tracking measures (first-pass, right-bound, spillover) it aligns rather closely with the original lure-match effect from PP2017. In the late measures (re-read, total times) it rather patterns together with the lure-match effect in the sentences with inanimate heads. I take these patterns to suggest that lures within an RC do provoke lure-match effects even when the matrix subject is animate. This conclusion holds if I adopt the PP2017 pre-processing routine, although in this case the magnitude of the effect in animate conditions goes down.

Just the presence of lure-match effect in "animate" conditions does not necessarily contradict S2017's hypothesis: the effect could be merely reduced, not elim-

[12] If we look at the patterns of the interference effects in Fig.2.4, we will notice that this effect is likely driven uniquely by the lure-match effect within "inanimate" conditions. Exploratory pairwise comparisons confirm this impression. This would support the S2017 subcommand hypothesis. However, these impressions come from exploratory analysis, and as I discuss below, other patterns in these analysis still suggest that lure-match effects do arise in "animate" conditions.

inated, when the matrix subject is animate. However, the numerical patterns do not seem to support this possibility. Almost in all measures in both pre-processing variants (except for late measures with PP2017 preprocessing) the lure-match effect in "animate" conditions is equal or bigger to the effect in "inanimate" conditions. Overall, I conclude that the experiment provides some evidence against S2017's sub-command explanation, although the evidence is merely suggestive and should be verified in a highly powered confirmatory experiment.

Chapter 3: Interference effects with quantificational lures

In the previous chapter we have shown that S2017 sub-command explanation for relative clause lures is likely wrong, and at least some lure-match effects might be better explained as instances of fallible memory access. This serves as a support for PP2017 account, which claims that structural cues do not have special status in the system: they do constrain memory access to some degree, but can be outweighed by other sources of information. In this chapter I am going to further test PP2017 account by looking at quantificational (QP) lures. The main question I address here is: are all structural cues treated equally by the parser? In particular, I will argue that available evidence tells us little about how c-command information is used, and I will report three experiments designed to fill this gap.

In this chapter, I am going to adopt Chomskyan Binding theory view, in which possible antecedents should c-command the anaphor and be local to it in some relevant sense. For the purposes of my experiments, locality can be reduced to clause-mateness. This simplification has an additional advantage: it makes it easy to represent locality as a cue in content-addressable architecture; but theoretical descriptions of locality conditions are more involved and can be trickier for cue-based models to accommodate (see, e.g., Cunnings et al. (2015) for an exploration of less canonical configurations). Having clarified these assumptions, what evidence does PP2017 provide for the nature of these structural constraints?

Their evidence, undoubtedly, suggests that locality constraints are being used but do not uniquely guide memory access: without making this assumption, we would have hard time explaining the contrast between the absence of lure-match effects in 1-feature target mismatch conditions and their presence in 2-feature target mismatch sentences. The situation with c-command is more complicated. Prima facie, Exp.2 (where lures were located inside a subject relative clause) suggests that c-command acts fails to uniquely guide retrieval as well. However, this pattern of results does not tell us anything about whether or how c-command information was used.

First, the same pattern of results could hold if the parser never even included c-command information in the set of retrieval cues and instead only relied on locality. In Exp.2 PP2017 only looked at 2-feature target mismatch configurations. There, locality would fail to rule out a non-local antecedent, and if no c-command information were used to further restrain memory access, both c-commanding and non-c-commanding lures could be potentially retrieved. While theoretically possible, we consider this possibility unlikely, based on empirical evidence from other dependencies. Antecedents c-commanded by the pronoun appear to be excluded from consideration, in accordance with Binding Theory Principle C (Kazanina et al., 2007; Aoshima et al., 2009; Kazanina and Phillips, 2010). Similarly, Kush et al. (2015) show that non-c-commanding quantified NPs do not appear to induce lure-match effects in pronouns resolution. This evidence suggests that at least some way of representing c-command constraints should be available to the parser. Since reflexives resolution is also constrained by c-command, a priori it might be weird to assume that the parser can represent such information but prefers not to use specifically with reflexives.

However, the fact that c-command is used by the parser in some manner does not necessarily imply that it is represented in such a way that it faithfully captures all and only licit c-command relations in the grammar. As we will discuss, faithfully representing c-command relations in cue-based memory models is not straightforward. Accordingly, various approximate encodings have been suggested (Alcocer and Phillips, 2012; Kush et al., 2015). Some of these approximations (e.g. Kush et al. (2015) ACCESSIBLE feature) are formulated in such a way that in the configurations used by Parker and Phillips (2017) they would not render the lures illicit. If this is the case, it would be hard to talk about the feature failing to categorically constrain processing: it would do so, only the constraints it would follow would not perfectly map to c-command constraints defined in the grammar. This conclusion would not align with the PP2017 account of lure-match effects which assumes that any structural cue can fail to categorically restrict retrieval. Thus, it is important to empirically evaluate the sensitivity of lure-match effects to the c-command status of the antecedent. In what follows, I will first discuss the approaches one might use to represent c-command information, and then show how we could experimentally differentiate them.

3.1 Representing c-command in cue-based models

C-command is defined as follows (Reinhart, 1976): a node A c-commands a node B if neither A or B dominate each other, and the first branching node which dominates A dominates B. From the point of view of Chomskyan Binding Theory, reflexives are bound by their antecedents, and for that, a c-command relation between the binder and the bindee is necessary. Thus, a priori we might expect that the parser will be able to encode and use c-command information.

On the other hand, c-command is inherently relational: one has to know the position of *two* nodes in a tree in order to figure out whether one of them c-commands the other. And as Alcocer and Phillips (2012) discuss, representing relational information in an efficient way in cue-based architectures may not be straightforward: one cannot just use a feature like "+ c-commanded" or "+ c-commander", it would have to be much more specific, something like "+ c-commanded by NP_{17}". Given the number of potential c-commanders a given node can have, if one attempts to encode these relations explicitly, the size of the representation and the processing effort required to update the information on the older nodes as the new nodes come in may quickly get out of hand. Additionally, as Alcocer and Phillips note, exhaustive encoding might be inefficient in restricting the range of constituents being considered during memory retrieval.

To demonstrate this, they discuss a scenario in which each node in the tree carries a complete list of its c-commanders. When a c-commander needs to be retrieved, this list (or its individual components) is included in the set of retrieval cues. However, this could lead to several problems. In ACT-R model, the activation for the items which do not match some of the retrieval cues is *decreased*, in proportion to the number of the mismatching cues. With this encoding approach, any single item would only match one of these cues, receiving a sizeable mismatch penalty for all the others. This activation drop could lead to longer retrieval times or even retrieval time-out. In addition, retrieving an *incorrect* item would be easier: a constituent might have many c-commanders utterly irrelevant to the dependency at hand, such as complementizers or functional projections. Yet they will receive an activation bump and might end up being misretrieved.

A related approach (not discussed by Alcocer and Phillips (2012)) would be to annotate each chunk with a list of phrases it is c-commanding. When the parser reaches the reflexive, it would use its ID to create a single retrieval cue, e.g. "+ c-commands NP_{17}". This approach would escape some of the problems from the previous one, but would also bring new issues. The mismatch penalty problem of the previous approach would be circumvented, since only a single c-command-related retrieval cue would be used. However, the efficiency of this cue will be very low, practically making it all but useless[1]: due to the fan effect among all of the items matching the c-command cue, activation boost to any single one will be rather small.

[1] That is, if we accept the standard assumption that cues are combined additively. If instead we assume that they are combined multiplicatively, so that a mismatch on any single cue drops the activation significantly, we will avoid the fan effect problem. But we will also make the structural cues rather powerful, potentially removing ways for ever receiving a structurally inappropriate constituent. This may or may not be desired: e.g. this change would fit rather well in the S2017 model, but not the PP2017 one.

The problems of contacting completely irrelevant items and of spreading redundant information across the representation are shared with the previous approach.

The above challenges to exhaustive encoding of c-command information would suggest that if one wants to efficiently represent relational information in a cue-based system, one would have to rely on approximations. Alcocer and Phillips (2012) suggest two such approximations. The first one relies on a binary "command-path" feature, which every node carries. It is switched to ON just in case a node c-commands the node which is currently being introduced to the parse, otherwise it is set to OFF. This encoding schema has two disadvantages. First, it does not allow the parser to use c-command information to make attachment decisions. Second, more importantly, it only allows the system to know which nodes c-command the node which is just being introduced. It does not let the parser know which nodes the current node c-commands (which may be necessary in head-final languages), not does it allow to figure out c-command relations for nodes which are already part of the parse (which might be needed in special configuration, e.g. those involving phonologically null pronouns, whose existence is only recognized when the following node is being processed).

A second heuristic by Alcocer and Phillips (2012) relies on "dominance spines". A dominance spine is a chain of nodes dominating one another along a right branch in the tree. Every time a left branch is created, it is assigned a new dominance spine index. In order to figure out whether the two nodes are standing in a c-command relation, one just needs to know whether their parents share the same dominance spine. In our case, when a parser reaches the dependency tail, it would look up the dominance spine index of the tail's parent and look for other nodes whose parents also have this index[2]. The limitations of this approach are as follows. It does not capture certain kinds of c-command relations; specifically, when a node is embedded too deeply in a left branch, the dominance spine heuristic will fail to capture its c-command relations with the nodes which c-command the top of the branch. In addition, this approach still has (arguably minor) problems with potentially leading to retrievals of irrelevant c-commanding nodes.

All of the approaches above, while differing in details, have one thing in common: they are trying to capture c-command relations proper, and the encoding schemas they suggest could be used by any kind of mechanism requiring c-command information. This generality is an advantage of these approaches and it brings them in closer alignment with grammatical theories. It also leads to the situation where c-command information fails to be a strong filter, making only minor contributions to constraining memory access due to fan effect (at least as long as one assumes that retrieval cues are combined additively). This may be seen as another advantage, if we take empirical evidence from subject-verb agreement, reflexives and NPIs to indicate that, indeed, c-command information may fail to strictly constrain the range of retrieved constituents. It would be more of a disadvantage to hypotheses which postulate that structural information helps to categorically filter out illicit dependency elements (e.g Sloggett, 2017).

A different kind of approach is advanced by Kush et al. (2015). They ac-

[2] In order for this schema to work, each node would also have to encode some information about its parent, including the parent's dominance spine

knowledge the complications with exhaustive encoding of relational information in cue-based models and also use an approximation. However, the feature they suggest, ACCESSIBLE, is designed to capture a very specific use case of c-command: binding by quantifiers. In a series of experiments, Kush et al. (2015) show that the parser is sensitive to the binding constraints on quantificational antecedents. First, they compare sentences like (1a) where the QP c-commands the pronoun to sentences like (1b), where it does not. They show that people experience more processing difficulties in (1b) as compared to (1a), as indexed by a slow-down in reading times in first-pass and right-bound times in the spillover region. This is interpreted as evidence for people not considering the grammatically inaccessible QP as the antecedent and either trying to bind the pronoun to a gender-mismatching "Kathi" or trying to coerce a sentence-external referent. These results show that the parser can distinguish between grammatically appropriate and grammatically inappropriate QP antecedents.

(1) a. Kathi didnt think any janitor liked performing his custodial duties, but he had to clean up messes left after prom anyway.

b. Kathi didnt think any janitor liked performing his custodial duties when he had to clean up messes left after prom anyway.

Second, Kush et al. also show that the parser appears to rule out inappropriate QPs categorically: people appeared to experience similar processing difficulties in both (2a) and (2b), despite the fact that in the second case the QP antecedent fully matches the pronoun in features. That is, QP lures do not appear to induce lure-match effects. Similar findings have been reported by Cunnings et al. (2015).

(2) a. The troop leaders that no boy scout had no respect for had scolded her after the incident at scout camp.

b. The troop leaders that no girl scout had no respect for had scolded her after the incident at scout camp.

To capture these patterns, Kush et al. suggest that a single feature ("±ACCESSIBLE") is used to mark the status of potential referents. NPs are always + ACCESSIBLE for retrieval. QPs, however, start out as + ACCESSIBLE, but are switched to - ACCESSIBLE as soon as the parser detects the edge of their c-command (or scope) domain. This implementation avoids several problems from Alcocer and Phillips (2012) (fan effect would be reduced due to the fact that only NPs and QPs carry the ACCESSIBLE feature; feature update requires minimal computations (unlike in some cases of "command-path")) at the cost of generality (this approach would not be able to rule out binding by non-c-commanding NPs).

Before we turn to our experiments, we briefly address one more issue. As we have discussed, Kush et al. (2015) findings do not allow to decide whether ACCESSIBLE tracks c-command or scope, because these two things are perfectly confounded in their stimuli. However, recent experiments by Moulton and Han (toappear) may disambiguate between these two possibilities. They suggest that at least in *some* cases (including configurations examined by Kush et al. (2015)) c-command is the relation tracked by the parser. They use feature mismatch paradigm to see whether scoping but not c-commanding QPs would be accessed during pronoun resolution. They show that in configurations like (3a), where the QP "each boy" does c-command the pronoun, people are faster to read the pronoun if it matches the QP in gender. However, in (3b), where the QP scopes over but not c-commands the pronoun, no such effects were observed. This can mean that only c-commanding QPs are considered by the parser.

(3) a. It seems each boy brought fresh water from the kitchen quickly right before he/she went on an early break.

b. After each boy brought fresh water from the kitchen quickly it seems that he/she went on an early break.

c. After the boy brought fresh water from the kitchen quickly it seems that he/she went on an early break.

A follow up experiment ruled out the possibility that the lack of the gender mismatch effect from scoping but not c-commanding QPs is not related to the reduced prominence of QPs within adjunct clauses: referential NPs in the same position (as in (3c)) did elicit gender mismatch effects. Moulton and Han (toappear) interpret these results as meaning that scope alone cannot explain constraints on QPs antecedents, and c-command is an important component of these constraints[3].

[3] They also provide results which are hard to explain if we think that c-command alone is used. They show in an interpretation study that in both (3a) and (3b) people choose co-varying interpretation equally frequently, about 60-65% of the times. Whether the QP does or does not

3.2 Outline of the experiments

The distinction between the general approaches by Alcocer and Phillips (2012) and the specific one by Kush et al. (2015) brings us back to the interpretation of the data from Parker and Phillips. To remind, they observed lure-match effects from non-c-commanding lures in configurations like (4), which prima facie could be used as evidence for the c-command acting as a violable constraint. However, this interpretation will depend on the encoding schema we choose. If a general strategy of the sort discussed by Alcocer and Phillips (2012) is used, Parker and Phillips (2017) conclusions will be supported[4]. If, on the other hand, the parser relies on a task-specific approximation a-la Kush et al.'s ACCESSIBLE, Parker et al. lure match effects from Exp.2 would rather indicate that the parser faithfully follows whatever encoding is available to it. Since NPs are always + ACCESSIBLE, the parser would have legitimate "right" to choose them as antecedents, given that locality constraints have already been overridden. That is, the choice is between a parser which follows a faithful representation less than perfectly, or a parser which faithfully follows a less than perfect representation.

(4) The soothing tea [that the nervous students drank] calmed themselves down after the test.

I try to tease these two possibilities apart in three experiments. I do this

c-command the pronoun does not seem to affect the judgments.

[4] To the degree that we can also rule out Sloggett (2017) explanation.

by looking at QP lures in c-commanding and non-c-commanding positions. If the system is relying on domain-general c-command information which sometimes fails to perfectly guide retrieval, we will expect to observe lure-match effects from the QPs, regardless of where they are located. If, on the other hand, the system is faithfully following a more domain-specific feature like ACCESSIBLE, the c-command status of the lures will determine the outcome: we would only expect to observe lure-match effect from c-commanding QP lures.

The presence or absence of lure-match effects from QP lures may also be informative for Sloggett (2017)'s model. Empirically, QPs appear to be acceptable as logophoric antecedents across languages (we have been able to find examples from Icelandic (Sells, 1987, p.467), Chinese (Huang and Liu, 2001, p.165), Yoruba (Adesola, 2006, p.2091)). As as (5) shows[5], QPs antecedents are also acceptable in the sub-command configuration in Chinese. Given these facts, S2017 account would predict lure-match effects from c-commanding lures and from sub-commanding lures embedded under an inanimate noun. For sub-commanding lures embedded under animate nouns (as in our experiments), no lure-match effects should be observed. (Notice that our Exp.1 has already suggested that sub-command explanation is not on the right track. However, since the evidence there was mostly based on numerical patterns, we discuss the corresponding predictions for S2017 account as well, to be able to determine whether further evidence is (in)compatible with it).

(5)	a.	mei	yi-ming	jizhe$_i$	xie	de	baodao	dou	hai-le	ziji$_i$
		every	one-CL	reporter	write	MOD	report	ADVQuant	harm-PFV	self

[5] I would like to thank Nick Huang, Chia-Hsuan Liao and Yu'an Yang for these judgments.

The report that every reporter$_i$ wrote harmed him$_i$

b. mei yi-ming yuangong$_i$ huode de huahong zuizhong dou
every one-CL employee receive MOD bonus ultimately ADVQuant
hai-le ziji$_i$
hurt-PFV self
The bonus that every employee$_i$ received ultimately hurt him$_i$

A potential counter-example to the cross-linguistic facts above comes from Postal (2006), who briefly discusses examples like (6)[6]. These examples suggest that in English, long-distance antecedents of reflexives cannot be quantificational. If we think that English-specific evidence should receive priority over cross-linguistic evidence, S2017 account would predict no lure-match effects from QP lures regardless of their position.

(6) a. *Every/No woman$_1$ claimed that HERSELF$_1$, most people could never understand.

b. *Every/No woman$_1$ claimed that they had praised no one but herself$_1$.

c. *Every/No woman$_1$ claimed that they had defaced carvings of herself$_1$.

d. *Every/No recent president$_1$ claimed that the Queen was inferior to himself$_1$.

e. *Every/No woman$_1$ claimed that it was HERSELF$_1$ that people should vote for.

f. *It was HERSELF$_1$ that every/no woman claimed that people should vote for.

[6] Caps indicate strong stress.

g. *Every/No woman$_1$ claimed that there was still HERSELF$_1$ for people to vote for.

h. *Every/No woman$_1$ claimed that Bob wanted to interview Carl and herself$_1$.

i. *Every/No woman$_1$ claimed that as for herself$_1$ she would prefer to drink beer.

j. *Every/No waitress$_1$ claimed that workers like herself$_1$ deserved a raise.

The rest of the chapter reports three experiments. In the first one I consider configurations like (7a), to determine whether QP lures embedded within relative clauses - i.e. in a non-c-commanding position - provoke lure-match effects. To preview the results, I did not find any interference from QP lures, supporting the possibility that a process-specific feature like ACCESSIBLE is being used. The interpretation of these results is complicated by the fact that NP lures did not provoke any lure match effects either, contra results reported in Parker and Phillips (2017). To determine possible reasons for this, in the next two experiments I look at c-commanding lures, as in (7b). If ACCESSIBLE is, indeed, in use, we should observe interference effects. Indeed, I did observe numerical patterns suggestive of lure-match effects in both QP and NP stimuli, although these patterns were not supported by statistical analysis. I suggest several possible reasons for this, which lead to the follow-up experiments discussed in Chapter 4.

(7) a. The stuntmen [that **no/the actress** had worked with] introduced **herself** to the rest of the cast.

b. No/the understanding **doctor** would complain that expectant mothers distress himself during stressful medical examinations.

3.3 Experiment 2

The goal of my first experiment is to check whether the parser faithfully obeys structural relational constraints as specified in the grammar; in particular, whether it obeys the c-command/scope restrictions on QP antecedents. I contrast two hypotheses. The first one is that c-command/scope constraints are implemented only approximately (in particular, I choose Kush et al. (2015) ACCESSIBLE feature hypothesis), but the parser strictly follows this approximation. If this hypothesis is true, I will not observe lure-match effects from non-c-commanding/non-scoping QP antecedents. The second hypothesis is that the parser does not follow the c-command/scope constraints, regardless of whether they are implemented approximately or in strict accordance with grammatical generalizations[7]. If this hypothesis is correct, I will observe lure-match effects from non-c-commanding/non-scoping QP lures[8]. Only the second scenario would constitute strong support for Parker and Phillips (2017) claim that structural features are of limited use to the parser when it has to select suitable antecedents.

[7] A third possible hypothesis – that the c-command/scope constraints are implemented in full accordance with the grammar and the parser DOES faithfully obey them appears to be ruled out by Parker and Phillips (2017) Exp.2.

[8] Notice that we would have made the same prediction if we assumed that c-command/scope information is never used in the reflexive resolution at all. However, a priori this possibility seems unlikely, and I do not pursue it further.

3.3.1 Participants

32 members of University of Maryland (13 M, 19 F; age range: 18-28; mean age: 20.2, SD: 2.15) participated in the experiment for a class credit or a payment of $10. The experimental session took about one hour, including setup and calibration. Data from 7 participants were excluded: one because of a software failure; two because they did not manage to complete the experiment in an hour; two because the preprocessing suggested that the data is extremely noisy; one because of an experimenter's error; one due to low question answering accuracy (64% correct).

3.3.2 Materials

Critical sentences were modeled after Parker and Phillips (2017, Exp.2) and Kush et al. (2015, Exp.2)[9]. An example of a critical sentence is given in (8). An example of a full set of stimuli along with accompanying contexts (see details below) is given in Table 3.1.

(8) The stuntmen that **the actress** had worked with introduced **herself** to the rest of the cast.

In all critical sentences the target was the subject of the matrix clause; the lure was the subject of a relative clause modifying the target; the reflexive was the object of the matrix verb. Thus, the lure is neither local relative to the reflexive nor

[9] I would like to thank Julia Buffinton for her great help with constructing the materials and providing native speaker judgments on them.

c-commanding it.

The study used 2x2 factorial design with lure type (NP or QP) and lure match (the lure matching or mismatching the anaphor in gender) as factors. The target always mismatched the anaphor in gender and number, thus, all the critical sentences in the experiment were ungrammatical. I used nouns with both definitional ("woman") and stereotypical ("secretary") genders. The gendered nouns came partly from previous studies (Parker and Phillips, 2017; Osterhout et al., 1997), and partly from a native speaker judgments. Target NPs were always plural, lures were always singular. Half of the reflexives were masculine, and half were feminine. QP lures were always created with the negative quantifier "No". This quantifier was chosen to make sure that binding is indeed the only linking option (see Kush (2013); Kush et al. (2015) for a discussion of how other quantifiers, like "every", sometimes allow for alternative readings which resemble binding, but do not behave exactly like it). Additionally, I made sure that the interpretation of the reflexive is not biased towards one of the potential antecedents: matrix verbs denote an action which can plausibly be performed by the target on the target itself (in case of grammatical binding of the reflexive; e.g. stuntmen can introduce themselves) and by the target on the lure (in case of the ungrammatical binding of the reflexive; e.g. stuntmen can introduce the actor/actress to others).

The presence of a subject relative clause makes the sentences sound less natural outside of a context. It suggests that the matrix subject is selected from a larger class, in such a way that only the selected members have the property denoted by the relative clause. But this larger class is never mentioned explicitly, which may tax the processing, if people try to come up with it on the fly. This may distort the reaction times for the critical sentences and mask potential intrusion effects. To avoid this, I add two context sentences which explicitly introduce the larger class, to make the relative clause modification of the critical sentence subject more felicitous. The contexts are structured in the following way.

For the NP lures, we first indicate the existence of a group of people which the target NP will be related to (stuntmen in Table 3.1, and a salient individual (actor/actress)). We then describe an action/attitude of that individual which subdivides the target group in two. In the example above, some stuntmen have worked with one of the actors/actress, and some have not. This allows us to single out stuntmen based on the property of (not) having worked with the actor in question. For the QP lures, we indicate existence of two groups of people (e.g., stuntmen and actors/actresses). Then, we describe a situation in which some (but not all!) members of the second group have a relation/attitude to some members of the first group. Importantly, there must be some members of the first group that no member of the second group has relation to. E.g. in QP LURE conditions in Table 3.1 some stuntmen have worked with some of the actors, but there are stuntmen who haven't worked with any of the actors. The existence of this latter subgroup of stuntmen allows us to felicitously use a relative clause with a negative quantifier to single this subgroup out.

As I mentioned above, all of the critical sentences are ungrammatical. Thus, if I don't find any interference effect, I am not able to check whether the experiment worked as intended. In order to remedy this, I included a control set of 12 agreement attraction items, mostly adapted from Wagers et al. (2009) (see the description for Experiment 1 for more details). Thus, I had 3 items per conditions[10]. Agreement attraction sentences were embedded into short contexts as well.

NP lures

The action movie had a very large cast, only some of whom knew each other from previous films, so they had a round of introductions. Most of the stuntmen had already worked with one of the actors/actresses, so they decided to talk first.

Lure match
The stuntmen that ***the actress*** had worked with introduced ***herself*** to the rest of the cast.

Lure mismatch
The stuntmen that ***the actor*** had worked with introduced ***herself*** to the rest of the cast.

QP lures

The action movie had a very large cast, only some of whom knew each other from previous films, so they had a round of introductions. Most of the stuntmen had already worked with some of the actors/actresses, but some of the stuntmen were novices, so they decided to talk first.

Lure match
The stuntmen that ***no actress*** had worked with introduced ***herself*** to the rest of the cast.

Lure mismatch
The stuntmen that ***no actor*** had worked with introduced ***herself*** to the rest of the cast.

Table 3.1: Experiment 2 materials example
Targets are underlined, lures are bolded.

Finally, I used a variety of fillers type to make the manipulation less noticeable and to prevent people from adopting an unusual reading strategy, which is possible if they notice that all sentences with reflexives are ungrammatical. Similar to other

[10] This is fewer than usual, but the size of the experiment did not allow to include more.

sentences in this experiment, filler sentences were embedded into short contexts. Some of the fillers contained grammatical violations which could occur in any of the three sentences of the text. Appendix E provides detailed information on the fillers used.

Overall, I had 24 experimental stimuli with reflexives, 12 control agreement stimuli and 66 fillers for a total of 102 items. Each item consisted of two context sentences and one critical sentence, for a total of 306 sentences. The grammatical-to-ungrammatical ratio was roughly 3.5:1 if counting individual sentences and roughly 1:2 if counting items (I count any item containing at least one ungrammatical sentence as ungrammatical). All of the items in the experiment were accompanied by a forced-choice Yes/No question to make sure that the participants are paying attention while reading.

3.3.3 Procedure

The procedure was mainly the same as in Exp.1 with the following minor differences. Since the stimuli consisted of three sentence contexts, they did not fit on a single line. The maximum length of a line fitting on the screen was 139 symbols, so all items were broken down to several lines of texts (in most of the cases three, but sometimes more). I ensured that critical and spillover region did not occur immediately before or after the line break. Each experimental list started with 5 practice items.

.

3.3.4 Analysis

Analysis procedure was identical to that in Exp.1 with the following differences due to the design of the experiment. Separate models were fit to the data from reflexive and agreement conditions. Agreement data analysis was identical to that in Exp.3. The following fixed effects were specified for the model fit to the reflexive data (numerical values in parentheses indicate contrast coding coefficients; sum coding was used for all fixed effects): LURE TYPE (NP = -0.5 vs. QP = 0.5), LURE MATCH (match = -0.5 vs mismatch = 0.5 in features with the reflexive) and their interaction. Models random effects structure was fully specified, including random intercepts and slopes for each fixed effect per subjects and items, following recommendations by Barr et al. (2013). If the model with the maximal random effect structure did not converge, the random effects were dropped in the order specified in Table 2.2 (replacing TARGET ANIMACY withLURE TYPE to account for differences in the experimental design).

3.3.5 Results

Mean question answering accuracy was 92%. Figures 3.1 and 3.2 show the observed reading times means for agreement and reflexive stimuli respectively. Statistical analyses indicate that there is very little evidence for any effects: the only effect which reaches significance is the main effect of LURE MATCH in the agreement stimuli at the critical region in regression path. The positive coefficient means that on average, logRTs are smaller (i.e. people are faster) in lure match conditions

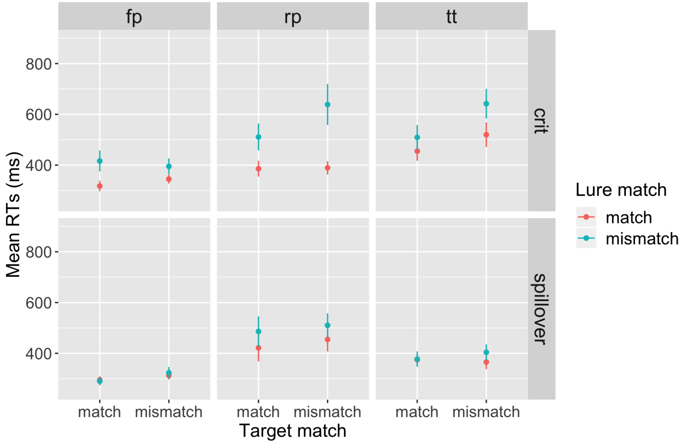

Figure 3.1: Experiment 2. RT means for agreement sentences.

Columns correspond to eye-tracking measures: fp - first-pass, rp - regression path, rp - total times. Rows correspond to ROIs. Errors bars represent standard error of the mean, adjusted for participant variability (Cousineau, 2005; Morey, 2008)

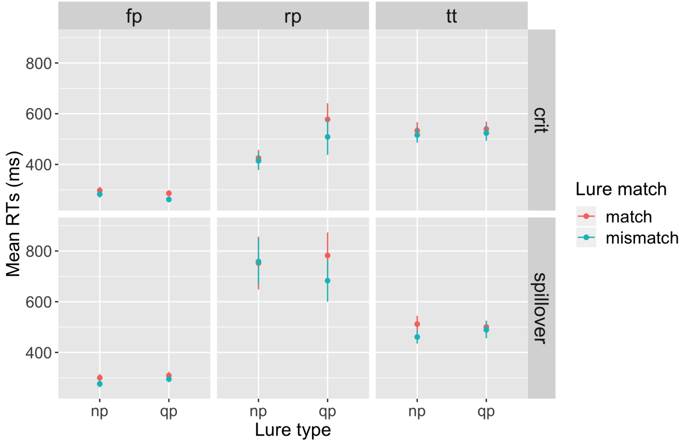

Figure 3.2: Experiment 2. RT means for reflexive sentences.

Columns correspond to eye-tracking measures: fp - first-pass, rp - regression path, rp - total times. Rows correspond to ROIs. Errors bars represent standard error of the mean, adjusted for participant variability (Cousineau, 2005; Morey, 2008)

regardless of whether the target matches or mismatches the verb.

Multiple comparison corrections As in the previous experiments, I fit 12 models, thus the Bonferroni corrected α-value remains the same: 0.004. The corresponding t-value for a two-sided test is ±2.87. If I apply this correction, no statistical comparison in the experiment reaches significance. Thus, after Bonferroni correction the intrusion effects I observed in agreement sentences do not receive statistical support.

3.3.6 Discussion

Agreement attraction stimuli I used agreement attraction sentences to confirm that the experiment worked as intended. I did find evidence for the intrusion effect at the critical region: lure match led to faster RTs in regression path at the critical region, which would indicate that the manipulation was successful. As in Experiment 1, the evidence suggested that this facilitation happened in both grammatical and ungrammatical sentences. This further strengthens Hammerly et al. (draft.april.2018) conclusions that the absence of ungrammaticality illusions may be artifactual. That said, the grammatical-to-ungrammatical sentence ratio was higher than in Experiment 1 and much higher than in Hammerly et al. (draft.april.2018) experiments, thus it is not clear whether their explanation in terms of grammaticality bias can be used to explain these data.

It may be worrying that statistical support for the lure-match effect in agreement attraction is not very strong: I only observe it in one region and reading time measure, and even there it disappears if I apply multiple comparisons correction. However, it is not inconsistent with previous eye-tracking studies. Both Dillon et al. (2013) and Parker and Phillips (2017) report facilitation in only one measure (total times and re-read times, correspondingly). It is also the case that I had fewer data points than the other studies: only 3 per condition per participant. This might have made it more complicated for the effects to reach statistical significance.

Overall, I conclude that the data from agreement sentences did provide evidence for intrusion effect, and thus indicates that the experiment worked as expected, and it is possible to observe lure-match effects in my data.

Reflexive stimuli Let us now turn to the main question of interest — intrusion effects in the reflexive sentences. I did not find any statistical evidence for intrusion effect in reflexive resolution with QP lures. This fact alone could indicate that the parser faithfully follows constraints on binding by quantifiers, contra PP2017 and in line with Kush et al. (2015) ACCESSIBLE account. However, this interpretation is complicated by the absence of reliable interference effects in the sentences with NP lures. This may be worrying, since the stimuli were structurally identical to those used in Parker and Phillips (2017), and thus a priori I might expect to observe a similar effect. Thus, before making my final conclusions about QP stimuli, I investigate the lure-match effects in NP stimulli in more detail.

To better understand the alignment between my study and PP2017, I perform a set of exploratory analyses similar to those in Experiment 1. Fig.3.3 shows the size of the interference effect in these exploratory analyses. Since there is not as much information in this figure, I change the plot layout: eye-tracking measures are plotted along the x-axis, rows correspond to ROIs, and columns — to pre-processing variants. Averages for PP2017 data stay the same in both rows.

In the critical region my main pre-processing procedure yields interference effects which are consistently smaller than in PP2017 data and which concentrate around 0. When I adopt their pre-processing procedure, the effect in my data shifts towards negative values (indicating facilitation) and the difference with PP2017 reduces but does not disappear completely. The difference remains biggest in re-read times — the only measure in which PP2017 did find statistically significant facilitation. In the spillover region, my effects consistently concentrate at or above zero regardless of the analysis procedure. If anything, this would be indicative of inhibition, rather than facilitation observed in PP2017 data. Overall, Fig.3.3 suggests that in my data the effect of lure match in reflexive sentences is consistently smaller than in PP2017.

There are several possibilities of why this could be the case. First, it could be that my experiment did not work at some basic level. However, I do not think this is the case. Basic reading effects are visible in the data: people's reading times increase and the proportion of first-pass skips decreases as the length of the words increases. Question answering accuracy is reasonably high: except for one person which I excluded from the analysis, everybody has at least 83% of correct answers, and most people are above 90% accuracy. Finally, I did observe the expected patterns in the agreement attraction sentences, although statistical evidence for them was limited. Second, it could be that the presence of sentence contexts has affected

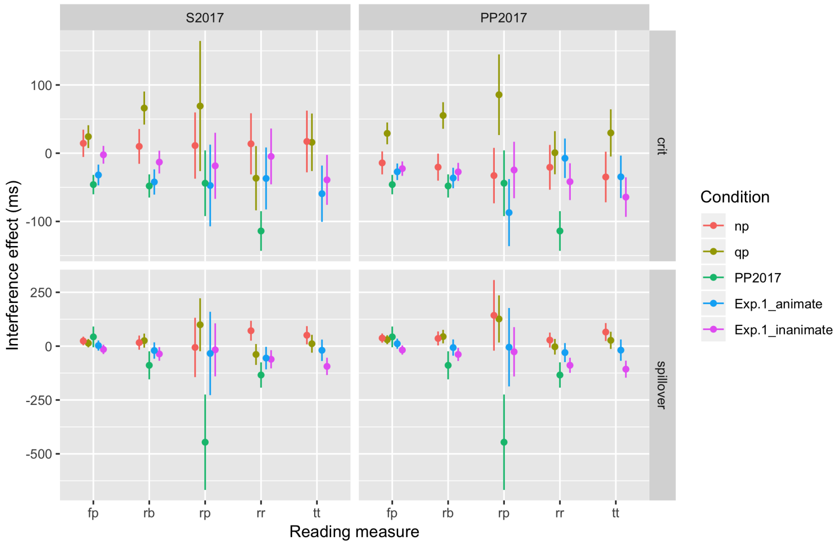

Figure 3.3: Experiment 2. Comparisons of the interference effect magnitudes in reflexive conditions with Parker and Phillips (2017).

Interference effect is calculated as the difference between lure match and lure mismatch conditions. Errors bars represent standard error of the difference of the means (calculated under the assumption that RTs in lure match and lure mismatch conditions are not correlated. This is likely false, since they come from the same subjects, thus, the SEs are overestimates).

people's reading strategies. E.g. it might have made people process the sentences more deeply and tobe more discriminating in the choice of antecedents. Or it might have just led to higher fatigue - people had to read three times more sentences compared to PP2017 experiment. Third, it might be that the intrusion effects from non c-commanding lures are more fragile, e.g. because it's harder to ignore c-command information.

However, most of these concerns are assuaged by the data from Experiment 1. It relied on the exact same stimuli as PP2017 did, with no additional contexts, and still observed very small numeric effects of interference. Fig.3.4 shows the magnitude of the interference effects in NP and QP lure conditions from the current experiment and from Experiment 1. As in the other places in the thesis, I look at two pre-processing variants in order to check whether they have any noticeable effect on the conclusions I make.

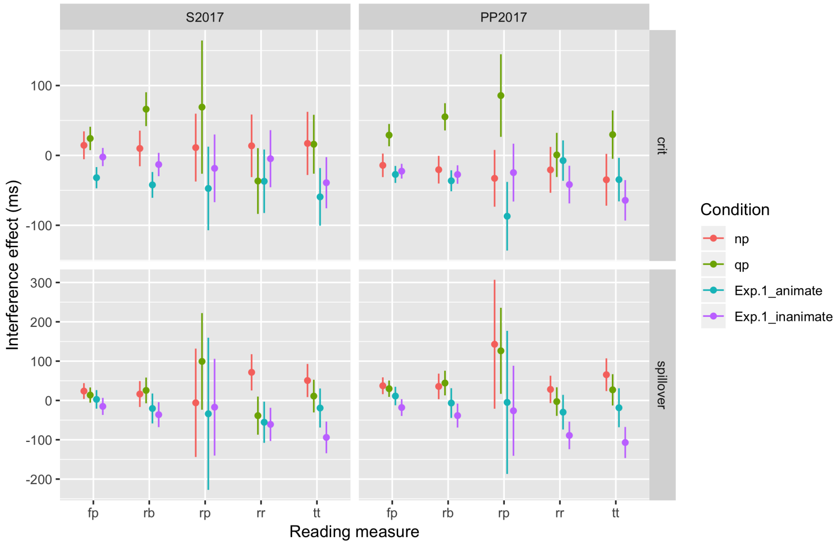

Figure 3.4: Experiment 2. Comparisons of the interference effect magnitudes in reflexive conditions with Parker and Phillips (2017).

Interference effect is calculated as the difference between lure match and lure mismatch conditions. Errors bars represent standard error of the difference of the means (calculated under the assumption that RTs in lure match and lure mismatch conditions are not correlated. This is likely false, since they come from the same subjects, thus, the SEs are overestimates).

I will compare the stimuli most similar between the two experiments - those with animate matrix subjects. Unfortunately, the results do depend on the preprocesing procedure. If one chooses the procedure from S2017 (extended critical region, removed missing values), the interference effects from the current effects are smaller in magnitude than in Experiment 1 by roughly 50 ms, which puts them in the vicinity of zero. If, on the other hand, one chooses the PP2017 procedure (critical region comprises only the reflexive, missing values are replaced with zeros), the effects from the two studies align rather closely. Perhaps the only conclusion I am comfortable making here is the following. To the degree that we believe that Experiment 1 provided evidence for interference in conditions with animate matrix subject, we should also believe that smaller interference effects in the current experiment are not stemming simply from the fact that I used animate matrix subjects. I.e. it is unlikely that the sub-command hypothesis by S2017 could be used to explain the reduced magnitude of the effects.

Returning to the QP stimuli, their lure-match effects quite consistently go in the opposite direction as compared to the NP stimuli in both the current experiment and Experiment 1. In most cases, lure-match effects from QP lures are around 0 or positive (indicating inhibitory interference), while lure-match effects from NP stimuli are generally negative[11]. I interpret this as suggesting that NP and QP lures differentially affect processing. E.g. it could be the case that people can access non-c-commanding NP, but not QP lures. If these conclusions are on the right track, it would mean that PP2017 claim about structural cues failing to categorically restrict dependencies resolution is overly general and that some structural information does play a gating role. It may be that a feature like Kush et al. (2015) is used to approximate c-command relations. In combination with my previous conclusions about the presence of lure-match effects in "animate" conditions, it may mean that some kinds of structural information, such as (approximated) c-command can accurately guide retrieval, while some other kinds (like locality) cannot.

In order to validate these conclusions, in the next two experiments I turn to configurations with c-commanding NP and QP lures. If my conclusions about the

[11] And in cases when they do float around zero, the QP lure-match effects are still different, going in the positive direction

representation of c-command are correct, I expect to observe lure-match effects from both NPs and QPs.

3.4 Experiment 3

3.4.1 Participants

38 members of University of Maryland participated in the experiment for a class credit or a payment of $10 (13 M, 25 F; mean age: 20.2, SD: 1.2). The experimental session took around 45 minutes on average, including setup and calibration. Data from 3 participants were excluded due to poor quality (high amount of trials with trackloss / missing data points in the critical regions).

3.4.2 Materials

24 experimental sentences, modeled after Parker and Phillips (2017, Exp.3)[12] were included in the experiment. An example of a full set of stimuli is given in Table 3.2.

Several modifications were made to the design in order to address the concerns from Experiment 2. The main idea was to create conditions which would maximally favor the resolution of the dependency to the QP lures. First of all, I put the lures in a position which makes them good binders from the grammatical point of view: subject position of the matrix clause, with both the target and the reflexive being

[12] Many thanks to Hanna Muller for helping with the stimuli construction and providing native speaker judgments.

inside a complement clause. This configuration ensures that the QP c-commands the reflexive, and the only thing stopping the QP from being linked to the reflexive are locality restrictions on the reflexive. Second, I got rid of accompanying contexts, shortening the experiment and reducing the risk of participants being overstrained. As in the previous experiment, all experimental stimuli with reflexives are ungrammatical.

The properties of the sentences were similar to those in Experiment 2. All lures and reflexives were always singular, targets were always plural. Matrix verbs were verbs of speech or belief. In the QP conditions, matrix predicates had the form "would V" ("No X would V that…"), while in the NP conditions, the predicates were of the form "did not V" or "would never V". This was done to make the sentences in the two sets more parallel in meaning. There was equal number of feminine and masculine reflexives. As in the previous experiment, I tried to make sure that the target and the lure are equally plausible antecedents for the reflexive, given the verb of the embedded clause. E.g. in Table 3.2 expectant mothers can plausibly distress themselves (target antecedent) or the doctor/nurse (lure antecedent). Spillover regions were aligned in structure. The first word after the reflexive was always a preposition, optionally followed by a determiner, then by an adjective or nominal modifier. I made sure that spillover regions did not contain null elements or gaps like in "…humiliated himself _ trying to get a spot", since the presence of a gap could initiate additional memory access and retrieval operations.

I also tried to make the sentences as natural out of context as possible. In particular, I tried to construct situations in the following way: they describe something that a certain group would never do or think. The group has to be selected in such a way that this generalization naturally follows from the membership in the group alone. E.g. we can expect that *understanding* doctors will be patient; this does not necessarily hold for all doctors. We found that such subgroups are often well defined by adjectives which describe a function (flower girls, cleaning ladies), social status / power (powerful, influential), common sense (reasonable, discrete).

As in the previous experiment, I include an additional set of agreement attraction sentences. The same conditions as in Experiment 2 were used, but the overall reduction of the experiment size allowed us to include 24 control items (6 per conditions). Additionally, I used 96 filler sentences of different types. They included sentences with grammatical reflexives bound by NP and QP antecedent and sentences with QP subjects. All of the fillers were grammatical. Overall, the experiment had 144 sentences and grammatical-to-ungrammatical ratio was 3:1. All sentences were accompanied by forced-choice Yes/No questions.

3.4.3 Procedure

The procedure was identical to Experiment 1.

3.4.4 Analysis

Same analysis procedure as in Experiment 2 was used.

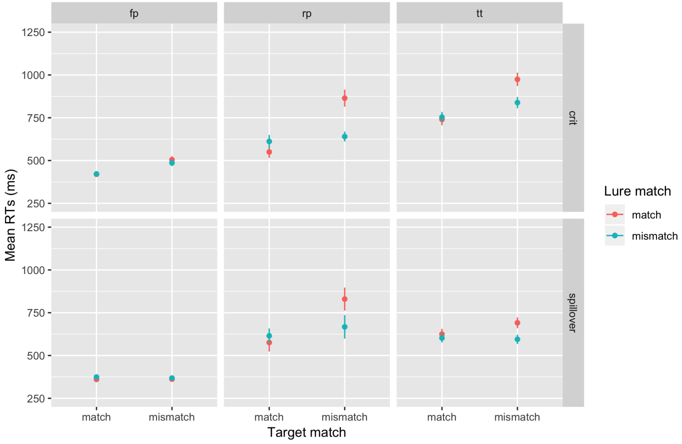

Figure 3.5: Experiment 3. RT means for agreement sentences.

Columns correspond to eye-tracking measures: fp - first-pass, rp - regression path, rp - total times. Rows correspond to ROIs. Errors bars represent standard error of the mean, adjusted for participant variability (Cousineau, 2005; Morey, 2008)

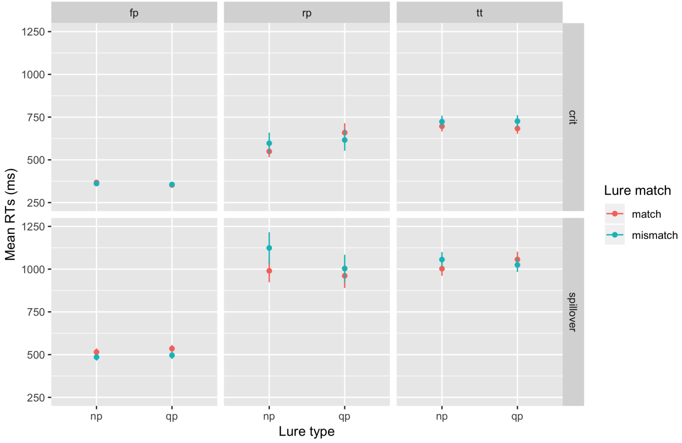

Figure 3.6: Experiment 3. RT means for reflexive sentences.

Columns correspond to eye-tracking measures: fp - first-pass, rp - regression path, rp - total times. Rows correspond to ROIs. Errors bars represent standard error of the mean, adjusted for participant variability (Cousineau, 2005; Morey, 2008)

NP lures
Grammatical, Lure match The understanding ***doctor*** would not complain that expectant mothers distress ***himself*** during stressful medical examinations.
Grammatical, Lure mismatch The understanding ***nurse*** would not complain that expectant mothers distress ***himself*** during stressful medical examinations.
QP lures
Grammatical, Lure match No understanding ***doctor*** would complain that expectant mothers distress ***himself*** during stressful medical examinations.
Grammatical, Lure mismatch No understanding ***nurse*** would complain that expectant mothers distress ***himself*** during stressful medical examinations.

Table 3.2: Experiment 3 materials example
Targets are underlined, lures are bolded.

3.4.5 Results

Mean question answering accuracy was 94%. Figures 3.5 and 3.6 show the observed reading times means for agreement and reflexive stimuli respectively.

Starting with the control agreement sentences, I did find statistical support for multiple effects. In the critical region, the main effect of TARGET MATCH reached significance in all three eye-tracking measures; the positive coefficient indicates that on average target match conditions were read faster than target mismatch conditions. This is the grammaticality effect. This main effect was qualified by a significant TARGET MATCH x LURE MATCH interaction in regression path. Pairwise comparisons confirmed that this effect was driven by the classical agreement attraction pattern: lure match conditions were read faster only within target mismatch conditions. I also observed a trend for this interaction in total times, but it did not reach significance there. In the spillover, I observed a significant main effect of TARGET MATCH and a significant TARGET MATCH x LURE MATCH interaction in regression path. These effects receive the same interpretation as in the critical region.

Turning to reflexive conditions, I found no reliable evidence for any effects, as in Experiment 2.

Multiple comparison corrections I follow the same logic as in Experiment 2, correcting for 12 comparisons, which results in critical t-value of ±2.87. All the main effects reported above survive this correction, and neither of the interactions does.

3.4.6 Discussion

Agreement attraction stimuli As in Experiment 2, I used agreement attraction sentences to confirm that the experiment worked as intended. This time I received a stronger support for this being the case. First of all, I observed reliable grammaticality effects at the critical region in all eye-tracking measures. More importantly, I also observed the TARGET MATCH x LURE MATCH interaction in regression path at the critical and spillover regions, which is indicative of the presence of agreement attraction. Together with the trend in total times, I take this as evidence that people did behave as expected at least with subject-verb agreement. I note that these interactions do not survive Bonferroni correction for multiple comparisons, but I return to further discussion of this issue after presenting the results for reflexive sentences.

Reflexive stimuli The main question I addressed in this experiment was: do c-commanding QP lures give rise to lure-match effects? There is very little indication that they do. Even if we ignore the lack of statistically reliable effects and look at numerical magnitudes of the interference effects, most of them are very close to zero, with the exception of total times, where I observe lure-match effect of roughly 50 ms. The magnitudes of the interference effects are rather similar between the stimuli with NP and QP lures, suggesting that they were treated similarly by the parser.

The interpretation of these results is complicated by the fact that I observe almost no evidence for lure-match effects from NP lures, contrary to what PP2017 and S2017 report. In this respect, this experiment is similar to Experiments 1 and 2; but now this lack of the effects may be more worrying: both PP2017 and S2017 consistently observed lure-match effects from c-commanding lures in multiple experiments. This again raises the possibility that my current experiment had some confounds which affected the manifestation of the lure-match effects. If this is the case, the data may be uninterpretable.

As before, I conduct more detailed comparisons with the original study (Parker and Phillips (2017) Experiment 3). This exploratory analysis followed the outline and modifications to the pre-processing procedure described in Experiment 2. I also included data from (S2017) Experiment 1 (conditions with speech verbs) in comparison, since their stimuli are structurally identical to ours. The comparisons are presented visually in Fig. 3.7. We can observe that regardless of the preprocessing procedure, interference effects for NP lures at the critical region are consistently smaller than in the original studies by PP2017 and S2017.

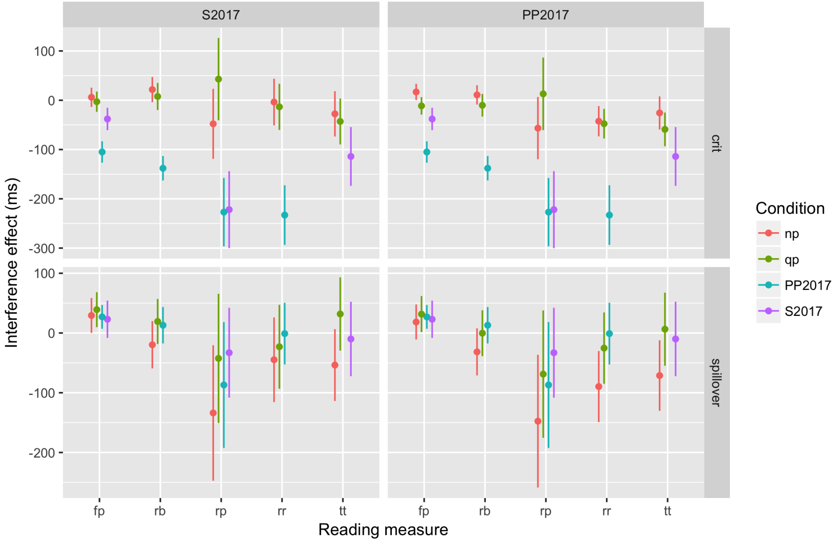

Figure 3.7: Experiment 3. Comparisons of the interference effect magnitudes in reflexive conditions with Parker and Phillips (2017).

Interference effect is calculated as the difference between lure match and lure mismatch conditions. Errors bars represent standard error of the difference of the means (calculated under the assumption that RTs in lure match and lure mismatch conditions are not correlated. This is likely false, since they come from the same subjects, thus, the SEs are overestimates).

As in Experiment 2, I can readily rule out the possibility that the experiment did not work at a basic level: I did observe he expected effects of length on RTs and first-pass skipping probability. I also did find evidence for agreement attraction, suggesting that the experimental procedure was fine and allowed me to detect lure-match effects.

However, my stimuli did include an important confound. I did not control the type of embedding verb, and only half of the verbs were speech verbs. This may be critical: as S2017 demonstrates, in configurations similar to mine lure-match effects are only observed with embedding verbs of speech, and not perception. If in my stimuli lure-match effects were effectively elicited only by half of experimental sentences, it could lead to the reduction ot the average magnitude of the effect. I rule this confound out in the next experiment and delay further discussion of patterns in QP sentences until I have done that.

3.5 Experiment 4

3.5.1 Participants

40 members of University of Maryland participated in the experiment for a class credit or a payment of $10 (19 M, 21 F; range: 18-26; mean age: 20.3, SD: 1.44). The experimental session took around 45 minutes on average, including setup and calibration. I excluded data from 2 participants: one because of software issues, another one due to extremely low proportion of correct responses (11%).

3.5.2 Materials

I had two sets of 24 experimental items, following the structure from Experiment 2. The examples of the items are given in Table 3.3. One of the sets had QP lures, and was of main interest; the other one had NP lures and was used as control.

The QP set was based on the stimuli from Experiment 3 with two differences. First, all matrix verbs were verbs of communication. Most of the verbs (18 out of 24) were taken from S2017. The change of the verbs required changing some of

NP lures

Grammatical, Lure match
The worried ***doctor*** reported that the delirious hiker talked to ***himself*** in the operating room before the procedure.

Grammatical, Lure mismatch
The worried ***midwife*** reported that the delirious hiker talked to ***himself*** in the operating room before the procedure.

Ungrammatical, Lure match
The worried ***doctor*** reported that the delirious mothers talked to ***himself*** in the operating room before the procedure.

Ungrammatical, Lure mismatch
The worried ***midwife*** reported that the delirious mothers talked to ***himself*** in the operating room before the procedure.

QP lures

Grammatical, Lure match
No shy ***girl*** would mention that the prom queen embarrassed ***herself*** in the school cafeteria after class.

Grammatical, Lure mismatch
No shy ***boy*** would mention that the prom queen embarrassed ***herself*** in the school cafeteria after class.

Ungrammatical, Lure match
No shy ***girl*** would mention that the schoolyard bullies embarrassed ***herself*** in the school cafeteria after class.

Ungrammatical, Lure mismatch
No shy ***boy*** would mention that the schoolyard bullies embarrassed ***herself*** in the school cafeteria after class.

Table 3.3: Experiment 4 materials example
Targets are underlined, lures are bolded.

the stimuli to maintain the naturalness of the scenarios. Second, in addition to ungrammatical sentences with the target mismatching the anaphor in two features, the experiment included grammatical sentences, with the target matching the anaphor. Thus, two factors were manipulated: target match (the anaphor either matches the target fully or mismatches it in gender and number) and LURE MATCH (rhe anaphor either matches the lure fully, or mismatches it in gender). The NP set consisted of a subset of stimuli from S2017, adapted virtually verbatim[13]. Similarly to the first set, all verbs were speech verbs, and the same two factors were manipulated.

Overall, I had 24 sentences with QP lures (12 grammatical), 24 sentences with NP lures (12 grammatical) and 96 fillers (adapted from Experiment 2; all grammatical), for a total of 144 sentences. Thus the grammatical-to-ungrammatical ratio was 5:1. As in the previous experiments, all sentences were accompanied by a forced choice Yes/No question.

3.5.3 Procedure

The procedure was identical to Experiment 1.

3.5.4 Analysis

I used the same preprocessing procedure as in Experiment 1, and analyzed the same regions (critical, spillover) and measures (first pass, regression path, total time) of interest. The definitions of the regions were the same as in Experiment 1

[13] The only change was the following. In two sentences, I exchanged the genders of the targets in grammatical and ungrammatical conditions. E.g. if the original combinations of target-anaphor were "actor - himself" and "actresses - himself", I replaced it with "actress - herself" and "actors - herself". This was done to counter-balance the number of feminine and masculine reflexives in the experiment. I do not think that this change affects the results of the experiment, since presumably what matters is the gender-match between the noun and the anaphor, not the lexical identity of the noun.

for the QP stimuli; for the NP stimuli, I followed the region mark-up reported in Sloggett and Dillon (in prep). NP and QP stimuli were analyzed separately (since the materials were not lexically matched) with the same modeling procedure as in Experiment 3.

3.5.5 Results

Mean question answering accuracy was 90%. Fig.3.8 shows the observed reading times means.

Statistical analyses provide robust evidence for the grammaticality effect: main effect of target match reached significance in the regression path and total times across virtually all regions of interest for both NP and QP conditions (except regression path in the critical region of NP sentences). The positive sign of the coefficient indicates that on average, 2-feature target mismatch conditions were read slower than target match conditions. However virtually no evidence for the interference effect was found. Two interactions have reached significance: NP stimuli, first-pass times at the critical region and QP stimuli, total times at the spillover region. Pairwise comparisons indicate that the first difference is likely driven by *slower* reading times in lure match conditions within target match conditions only. In the second case neither of the pairwise comparisons was significant, and Figure 3.8 suggests that the interaction effect is driven by the cross-over pattern of the means: lure match conditions are slower in target match sentences, but faster in target mismatch ones.

Multiple comparison corrections The number of the models I am building in this experiment is the same as in the previous one, thus, the critical t-value for Bonferroni correction remains the same: ±2.87. Grammaticality effects mainly survive the correction (only three do not: NP stimuli, total time at the critical region; QP stimuli, first pass and regression path at the critical stimuli). On the other hand, neither of the observed interactions (NP stimuli, first pass at the critical region and QP stimuli, total time at the spillover) survives the correction.

3.5.6 Discussion

The primary question that this experiment addressed was: is the type of the embedding verb responsible for the lack of reliable lure match effects from c-commanding lures that I observed in Experiment 3? The answer appears to be negative. Despite ensuring that all of the embedding verbs were verbs of communication, and despite using a subset of stimuli from S2017, which have already been shown to provoke lure-match effects, I failed to find statistical support for interference effects.

The only two statistically significant effects I observed are not in line with PP2017 and S2017. The interaction effect in the first-pass times at the critical region may serve as a weak evidence for intrusion effects in NP sentences. However, the intrusion pattern observed there does not match that reported by PP2017and S2017. If anything, it is rather more similar to the effects reported by Badecker and Straub (2002, Exp.3) and Patil et al. (2016): lure-match conditions are read slower than lure-mismatch in grammatical sentences. Such effects are predicted by some memory models of sentence processing(see, e.g. Jäger et al., 2017), but there is a discussion in the literature whether they should be treated as evidence for memory retrieval problems (see section 1.2.2 for more details). For now, I am hesitant about whether these effects are meaningful and if they are, why the same set of stimuli would give rise to facilitation in some studies and inhibition in others.

Similarly, the interaction I observed in total times at the spillover regions for the QP sentences is likely driven by the crossover pattern of the means alone. Neither of the pairwise comparisons reaches significance, and the t-value for the interaction is right at the critical threshold. Incidentally, numerically the pattern of reaction rimes was similar to the one I observed in NP stimuli in first-pass at the critical region: lure match conditions are slower in target match sentence but faster in target mismatch . But since the pairwise comparisons did not reach significance, right now I prefer not to ascribe meaning to these patterns.

As in the previous experiments, I perform more detailed comparisons of my results with PP2017 and S2017. Fig.3.9 shows the magnitude of the interference effect in 2-feature target mismatch conditions with NP lures. The results look virtually identical to the previous studies: at the critical region, interference effect is much smaller in my case than in either PP2017 or S2017, and the choice of the pre-processing procedure has practically no effect. We can also notice that lure-match effects are somewhat more robust in comparison with Experiment 3: they more consistently reach the magnitude of at least 50 ms, and sometimes they get even bigger. This suggests that while the type of the embedding verb is not responsible for the lack of statistically significant results, it might have still played a role in reducing the average size of the interference effects.

Controlling for the embedding verb type allows for a cleaner comparison of the NP and QP stimuli. As in Experiment 3, they align rather closely: interference effects go in the same direction and have comparable magnitudes, regardless of the choice of the preprocessing procedures. I take this as an indication that both lure types were treated by the parser similarly. That said, the similarity is not absolute: in regression path (both pre-processing procedures) and re-read times (S2017preprocessing only), the interference effect for NP stimuli is twice as big as in QP sentences. This may indicate that for some reason QP lures are less accessible to the parser. However, given that this difference at least partially depends on the analysis procedure and given that I do not have reliable evidence for interference within either pre-processing procedure, I prefer to remain conservative and conclude that the current dataset does not provide strong evidence for differential interference effects with NP and QP lures.

Overall, based on the numerical patterns I tentatively conclude that a) lure-match effects are provoked by NP and QP in roughly the same degree; b) this fact, together with the findings of no lure-match effect from non-c-commanding QP lures in Experiment 2 suggests that a feature similar to Kush et al. (2015) ACCESSIBLE categorically controls the access to reflexives antecedents. These conclusions are somewhat dependent on my interpretation of the reduced magnitude of the interference effect in the stimuli with NP lures. If it turns out that I have missed some important confound, the conclusions above may not hold. I discuss these issues in more detail in the general discussion.

3.6 General discussion

In this chapter I have presented three experiments attempting to clarify Parker and Phillips (2017) conclusions about the nature of structural constraints, focusing on c-command. I contrasted two possibilities: c-command is encoded in a process-agnostic fashion, or c-command is encoded as a process-specific feature, like Kush et al. (2015) ACCESSIBLE. To differentiate between these possibilities, I investigated QP lures in both c-commanding and non-c-commanding positions.

The evidence for lure-match effects in sentences with QP lures was extremely weak across all three experiments. The only statistically significant effect which could be interpreted as evidence for interference was observed in Exp.4 in total times at the spillover. Even there the pairwise comparisons were not significant, not allowing to conclude that the overall interaction indeed reflected lure-match effects. Additionally, this interaction effect stopped being significant after Bonferroni correction for multiple comparisons. Thus, the best evidence for intrusion I have comes from the numeric pattern of the means.

In Exp.2 I observed that the numeric patterns of the RTs were different for QP and NP lures. While NP lures mostly provoked facilitation, QP lures provoked inhibition (i.e. sentences with matching lures were read *slower*). I interpreted these patterns as indicating that the parser treats non-c-commanding NP and QP lures, and this may be evidence for the use of ACCESSIBLE. If this conclusion were correct, we should be able to observe interference from both NP and QP lures when they c-command the reflexive.

The numerical patterns observed in Experiments 3 and 4 were consistent with this prediction: both NP and QP c-commanding lures appeared to elicit interference effects. The evidence was clearest in Experiment 4: lure-match effects in NP and QP stimuli had similar direction and magnitude, with the effects reaching 50 ms in most eye-tracking measures regardless of the pre-processing procedure I used. These patterns were somewhat weaker in Experiment 3: effect sizes were smaller and some of the effects were grouped around zero. However, the alignment between NP and QP stimuli was largely the same. I took this to mean that the parser treated the two types of lures similarly. Weaker evidence in Experiment 3 could be explained by the properties of the stimuli I used: in Experiment 3 only have of the predicates I used were verbs of communication, while in Experiment 4 all of them were. Since S2017 showed that lure-match effects most consistently appear when the lures are subject of communication verbs, the reduced proportion of such verbs in Experiment 3 could lead to the reduction of the average lure-match effect sizet. If these conclusions are correct, these data would support the hypothesis that ACCESSIBLE is used as a proxy to c-command in reflexives resolution.

A second possible interpretation of my QP data is that the effects are not real, and the numerical patterns I observe are due to random noise. This interpretation would basically mean that for some reasons QPs are not efficient lures. If this is the case, my data cannot be used to make *any* conclusions about how c-command information is encoded in the system. Exactly *why* QPs would be bad lures is not clear, but I discuss several speculative possibilities. The first has been already mentioned: QPs may be bad lures not inherently, but because they lead people to weigh structural information more highly during reflexive resolution. This variant assumes that PP2017 story is correct, since in S2017 account structural features are already perfectly constraining retrieval to select the appropriate antecedents. It would be quite easy to test: e.g. one could look at the sentences with quantificational *targets*, e.g. "The actress said that the/most directors like herself". If the presence of a QP does indeed change processing strategies, we would only observe interference from "the actress" with referential targets ("the directors"). A second possibility for why QPs might be bad lures assumes that S2017 account is correct. In this case it may be the case that QPs make bad logophoric antecedents. E.g. OP_{log} might not be able to efficiently track the kind of entities that QPs introduce in the discourse. If this is the case, QPs would not lead to interference in *any* configurations considered by S2017. This would be the conclusion compatible with the data from Postal (2006) I mentioned in the introduction to this chapter, although it would go against cross-linguistic evidence.

Deciding whether the QP effects are real (and thus - should be taken into account at all) depends on the interpretation I assign to the evidence coming from the sentences with NP lures. The evidence for intrusion in the sentences with NP lures was also very weak, unlike in previous studies by PP2017 and S2017, but similarly to my Experiment 1. What do we make of this fact? There are several questions we need to address. First, do we have any reason to believe that the previous findings are not reliable? I do not think so. The only explanation I can think of how a nonexistent effect can receive statistical support is a Type I error. While it is not inconceivable that all of the previous findings are due to Type I errors, it is rather implausible, since the intrusion effects look very similar in the previous experiments, at least in terms of direction and the magnitude (the timing of the effects appears somewhat more subject to variability for unclear reasons).

If reliability of the previous finding is not a concern, the next question is: why did I not observe statistically supported evidence for intrusion effects in my study? To begin with, it could be that I have some basic problems with experimental procedure and my data cannot be trusted at all. I do not think this to be the case for several reasons. On a basic level, people did exhibit basic reading effects (longer words provoking increase of reading times and decrease of first-pass skips). Question answering accuracy was generally high[14]. Next, I did find support for interference effects in subject-verb agreement sentences. Finally, it is not the case that no statistically significant effects were observed at all. Reliable grammaticality effects were observed in Exp. 3 and 2, and this suggests that people were sensitive to the grammatical properties of the input.

If the general experimental procedure is fine, the next question is: are there any systematic confounds which could have weakened intrusion effects in reflexive stimuli, or made them disappear altogether? I have already ruled out two such confounds. First, numerically small effects in Exp.2 from the lures embedded inside a relative clause could be due to the fact that the embedding nouns were animate. It would be consistent with the hypothesis advanced by Sloggett (2017): lure-match

[14] Mean response accuracies: Exp.1 — 92%, Exp.2 — 94%, Exp.3 — 90%.

effects from lures within a subject relative clause represent a case of sub-command binding, similarly to *ziji* in Chinese (Huang and Liu, 2001). However, my results from Chapter 2 do not support this possibility: there, I observed lure-match effects from lures inside a relative clause regardless of the embedding NP animacy. Also notice that the sub-command explanation would not account for reduced lure-match effects in Exp. 3 and 4: in these experiments the lure in fact c-commanded the reflexive, being the subject of the matrix clause which embedded the complement clause containing the reflexive. Second, I have considered the effect of the embedding verb type (speech vs. perception) on the lure-match effect from c-commanding lures. This factor appeared to be of some importance: Experiment 4, which controlled for it, showed more consistent and numerically bigger lure-match effects. However, even there the effects were smaller than in the previous studies. This is even more telling since I have used a subset of the stimuli from S2017, which have been shown to produce lure-match effects. Thus, the verb type alone cannot account for the reduction of the magnitude of lure-match effects in comparison with the previous results.

I identify two further possibilities, which I broadly classify as extra-linguistic context effects. The first possibility is that merely the presence of QP stimuli in the experimental materials might have affected lure-match effects in the whole experiment. E.g. the presence of QPs might have forced the parser to weigh structural features more highly, since they may be more important for dependencies involving QPs. The second possibility has to do with the properties of the experimental procedure, namely, the language status of the experimenter. In my case, the participants were instructed in non-native English, and it is possible that they might have adjusted their processing strategies. There exists evidence that such adjustments can in fact happen. E.g. in an EEG study on Dutch Hanulíková et al. (2012) gender violations elicited a P600[15], if the stimuli were read by a native Dutch speaker, but not when they were read by an L2 Turkish speaker. It is conceivable that something similar happened in my experiments, although additional qualifications are necessary to explain why grammaticality effects were observed, if we think that people corrected for non-nativeness of the experimenter. One could say that adjustments were selective, such that violations were ignored in only one of the two morphological features I manipulated (number and gender). In this case, the sentences would essentially behave as one-mismatch sentences for which Parker and Phillips (2017) only found grammaticality effects. Another possibility compatible with Parker and Phillips feature weighting account would be that the adjustment effectively downweighted morphological features; in this case, the speakers would have to solely rely on syntactic features, which would predict only grammaticality effects without interactions. Yet another possibility is that repair processes are adjusted. It has been argued that agreement attraction arises as a result of repair and not initial misretrieval (e.g Lago et al., 2015). The evidence is based on the fact that attraction seems to only arise in a small proportion of trials with longest reaction times, while grammaticality effects are seen in a larger proportion of reaction times, including shorter ones. If people are able to selectively suppress repair (e.g. if they realize that with non-native speech repair will have to happen too often), we could expect

[15] ERP component, often associated with grammatical violations.

to only see grammaticality effects. It is unclear whether this last explanation would work for reflexives.

Notice that as discussed the confounds only suggest that something might have gone wrong. They do not tell us whether the effect is real, but weakened, or whether the confounds affected the experiment to such degree that the effects were gone altogether, and the numerical patterns represent just noise. To figure this out, I could compare the patterns of my results to the previous findings. Presumably, if the effects in my study are real, they would be similar to the effects observed in the previous studies. If my effects are not real and the patterns of the means just reflect random noise, we might expect the patterns to be incongruent, either between themselves, or in comparison with the previous literature, or both. In the latter case the random nature of noise does in principle allow the situation where the effects are not real, but by chance alone the means pattern consistently with the previous studies. I did perform such comparisons, but will defer the discussion until the end of the next chapter, when more relevant data, including those from direct replications of previous studies, will have been discussed.

To summarize, I tentatively concluded that the data from QP lures indicate that c-command information is represented in an approximate way. However, these conclusions are uncertain due to the lack of statistical evidence for interference effects from NP lures, which may indicate that some unknown factor has confounded my results. If this is the case, it may not be safe to make any conclusions about QP stimuli at all. The lack of lure-match in NP stimuli is also worrying, since it may indicate that the interference effects in reflexive resolution may be spurious or at least subject to higher degree of variability than PP2017 and S2017 results could lead to believe. In order to better understand the reason for my non-replication, I attempt two direct replications of PP2017 findings, which I discuss in the following chapter.

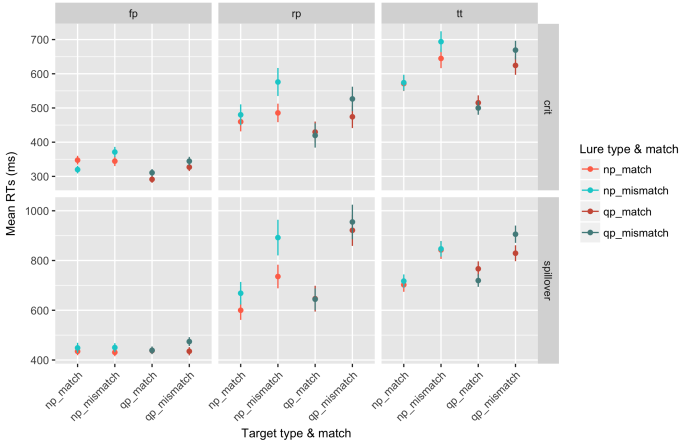

Figure 3.8: Experiment 4. RT means for agreement sentences.

Columns correspond to eye-tracking measures: fp - first-pass, rp - regression path, rp - total times. Rows correspond to ROIs. Errors bars represent standard error of the mean, adjusted for participant variability (Cousineau, 2005; Morey, 2008)

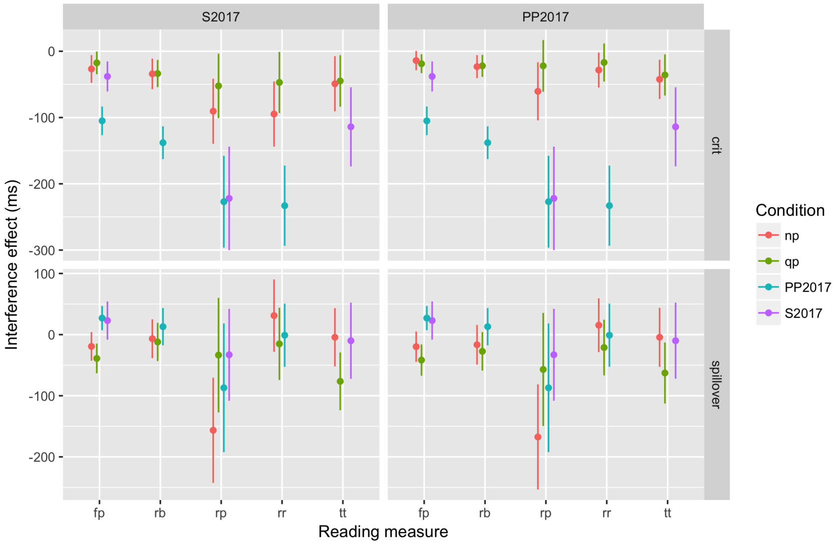

Figure 3.9: Experiment 4. Comparisons of the interference effect magnitudes in 2-feature target mismatch conditions with PP2017and S2017.

Interference effect is calculated as the difference between lure match and lure mismatch conditions. Errors bars represent standard error of the difference of the means (calculated under the assumption that RTs in lure match and lure mismatch conditions are not correlated. This is likely false, since they come from the same subjects, thus, the SEs are overestimates).

Chapter 4: Replicability of lure-match effects in reflexive resolution

In the previous chapter I reported experiments aimed at understanding the role of c-command information in the reflexive resolution. The focus was on the behavior of quantificational lures - I predicted that only if PP2017 is correct, we would see lure-match effects. I did not find any reliable evidence for such effects, while the numerical patterns indicated that PP2017 approach may not be on the right track. However, these conclusions were weakened by the fact that I found no detectable interference from *referential* lures, using stimuli similar or identical to those used in the previous studies. It is not clear whether lure-match effects did not receive statistical support because of some confounds, which could have affected the effect sizes. This means that the QP results from the previous chapter cannot be used to make strong conclusions about my main question until I have a better understanding of what could have caused the smaller lure-match effect magnitudes.

As I have discussed in the previous chapter, it is possible that the lack of lure match effects is due to people adjusting their processing strategies in response to some property of the experiment, e.g. the presence of QPs in the materials or the non-nativeness of the experimenter. Ruling out these confounds is the primary goal of this chapter: Experiment 5 addresses the first concern, and Experiment 6 - the second one.

In addition to helping with interpretation of the earlier findings, these experiment may also be used to distinguish between PP2017 and S2017 accounts. Adjustment of processing strategies in response to experimental context factors is relatively easy to explain if we adopt PP2017 account: one could say that people re-weigh their retrieval cues, giving more priority to structural information, and with this new set of weights even a mismatch in two morphological features is not enough to outweigh structural features. On the other hand, S2017 account would have hard time accommodating this finding. To remind, the account postulates that lure-match effects arise because people retrieve a structurally accessible null operator in a proportion of cases. In order to account for (hypothesized) effects of the experimental context, we would have to come up with a mechanism which would reduce the preference for choosing logophoric interpretation depending on the broader (extra-linguistic) context properties.

The only straightforward way of doing this would be to somehow ensure that OP_{log} is less accessible for retrieval. One way of doing it would be to lower the degree of match of OP_{log} with the retrieval cues. There are two ways of doing this: modifying featural representation of OP_{log} and modifying the set of retrieval cues. The first way is rather implausible. OP_{log} is hypothesized to by ϕ-deficient and to only carry a structural feature, encoding both c-command and locality information. It is hard to imagine that OP_{log} will or will not include information about its clausehood depending on the context. The second way could in principle work; however, in this case, it would imply **not** using structural cues to guide retrieval. If this were to happen, lures could be accessed directly, and we would still expect to see lure-match effects even in context hypothesized to produce processing adjustments.

Alternatively, one could try to raise the activation of the overt target to the degree where it would always outcompete OP_{log}. Again, I do not think it is plausible that the featural composition of the target will be affected by the context, and it is not clear what additional cues specific only to the overt target could be added to the retrieval cues set. Thus, this possibility seems to fail as well.

Finally, one could hypothesize that the activation of OP_{log} and/or overt target is changed because the change in experimental context affects their linguistic prominence (see the discussion in section 1.5.1.1). I *might* see how the presence of QP sentence could do it: perhaps, in a sentence with a QP lure and an NP target the latter becomes more prominent because it refers to a specific entity; then, the comprehenders would have to make this shift even for the sentences with the NP lures. It is not clear to me how the (non-)nativeness of the experimenter would change linguistic prominence of OP_{log} or overt targets.

4.1 Experiment 5

In this experiment I attempt a direct replication[1] of PP2017 Exp.3 in order to assess whether the presence of QP sentences in the stimuli set of my Exp.2-4 could have affected the results. I use their exact materials with minor corrections (typos etc., as specified below). I also collect a sample which is twice as big as theirs (48

[1] I would like to thank Dan Parker for generously sharing the eye-tracking scripts, materials and data from his studies.

vs. 24 people) which gives better statistical power and reduces the likelihood of over-estimating the magnitude of the effects.

4.1.1 Participants

52 members of University of Maryland community (38F, 14M; age range: 18-25, average age: 20.1) participated in the experiment for a class credit or a payment of $10. The experimental session took about one hour, including setup and calibration. Data from four participants were removed: two participants did not manage to complete the experiment in an hour, and two participants gave incorrect answers to more than 30% of comprehension questions.

4.1.2 Materials

Stimuli and fillers were directly adopted practically verbatim from PP2017 Exp.3 (I literally used the same eye-tracking script). The examples of the stimuli are given in Table 4.1. The only modifications I made consisted in correcting typos and inconsistencies in the stimuli sets (the full list of modifications can be found in Appendix D). Overall, the stimuli set had 36 experimental sentences and 72 fillers. 24 out of 36 experimental sentences were ungrammatical, all fillers were grammatical, giving a ungrammatical-to-grammatical ratio of roughly 1:4.

4.1.3 Procedure

The procedure was identical to Experiment 1.

Target match, Lure match

The talented ***actor*** mentioned that the attractive *spokesman* praised ***himself*** for a great job.

Target match, Lure mismatch

The talented ***actress*** mentioned that the attractive *spokesman* praised ***himself*** for a great job.

Target mismatch (1 feature), Lure match

The talented ***actor*** mentioned that the attractive *spokeswoman* praised ***himself*** for a great job.

Target mismatch (1 feature), Lure mismatch

The talented ***actress*** mentioned that the attractive *spokeswoman* praised ***himself*** for a great job.

Target mismatch (2 features), Lure match

The talented ***actor*** mentioned that the attractive *spokeswomen* praised ***himself*** for a great job.

Target mismatch (2 features), Lure mismatch

The talented ***actress*** mentioned that the attractive *spokeswomen* praised ***himself*** for a great job.

Table 4.1: Experiment 5materials example
Targets are italicized, lures are bolded.

4.1.4 Analysis

In my initial analyses, I followed PP2017 in choice of regions and measures of interest, and statistical comparisons. Although I do have concerns about their procedure, which I will discuss later, I decided to first follow their procedure as closely as possible to create a baseline comparison between results reported in PP2017 and my replication attempt. Three regions of interest (ROIs) were defined: precritical, which included the embedded subject and predicate (i.e. four words before the reflexive pronoun); critical, including only the reflexive pronoun itself; and spillover, including two words after the pronoun. Four eye-tracking measures of interest (MOIs) were analyzed: first-pass, right-bound, regression path and total times. Definitions were given earlier.

Prior to statistical analysis, the data were manually pre-processed using Eye-Doctor[2] to remove blinks and ensure that fixations align with the text. Fixations shorter than 80 ms or longer than 1000 ms were automatically rejected before calculating eye-tracking measures using custom scripts[3]. Then, reading times were log-transformed and missing observations were replaced with zeros.

A linear mixed effect model with TARGET.MATCH, LURE.MATCH and their interaction were then fit to the data. Treatment coding was used with baseline level being TARGET: MATCH, LURE: MISMATCH to every ROI and MOI. I will refer to this model as "full model". Additionally, a series of smaller models was fit to subsets of conditions. One model, including the same predictors, was fit to ungrammatical conditions only (baseline level was recoded to TARGET: ONE MISMATCH, LURE: MISMATCH). The interaction term from this model was used to assess the difference in interference effect between one-mismatch and two-mismatch sentences. I will refer to this model as "ungrammatical-only" model. Two more models corresponding to pairwise comparisons were fit separately to the ungrammatical TARGET: ONE MISMATCH and TARGET: TWO MISMATCH conditions, to assess simple main effects of LURE MATCH. No corrections for multiple comparisons were performed.

[2] `https://blogs.umass.edu/eyelab/software/`

[3] Avaliable at: `https://github.com/UMDLinguistics/EyePy`. Notice that PP2017 relied on the previous generation of these scripts, so this pre-processing step was not *exactly* identical across the two studies.

$target.match : lure.match \mid item$
$target.match : lure.match \mid participant$
$target.match \mid item$
$lure.match \mid item$
$target.match \mid participant$
$lure.match \mid participant$

Table 4.2: Order of random effect structure simplification, Exp.5. Effects were removed from the model, starting from the top.

If the model with the maximal random effect structure did not converge[4], I simplify it by removing random slopes. PP2017 did not specify their procedure in much detail, stating only: “random slopes for items or participants were removed”. To be internally consistent, I always dropped the random effects in the order shown in Table 4.2. As a final step, all models were automatically assessed to check whether any random effects had correlations of 1 or -1. The procedure was described in more detail in section 2.1.4.

4.1.5 Results

Mean question accuracy was 91%. Observed patterns of mean reaction times are shown in Fig.4.1. Numerical values for the means are given in Appendix A.

Fig.4.1 is laid out as follows. Columns correspond to reading times measures (see figure description for details). They have to be arranged in two separate subpanels because the time scales are quite different between early and late measures (especially in the critical and spillover regions), and plotting them all in the same

[4] As determined by diagnostics reported by `lmer`. One has to access the internal structure of the model fit and look at the information contained in `m@optinfo$conv$lme4`, where `m` is the model object.

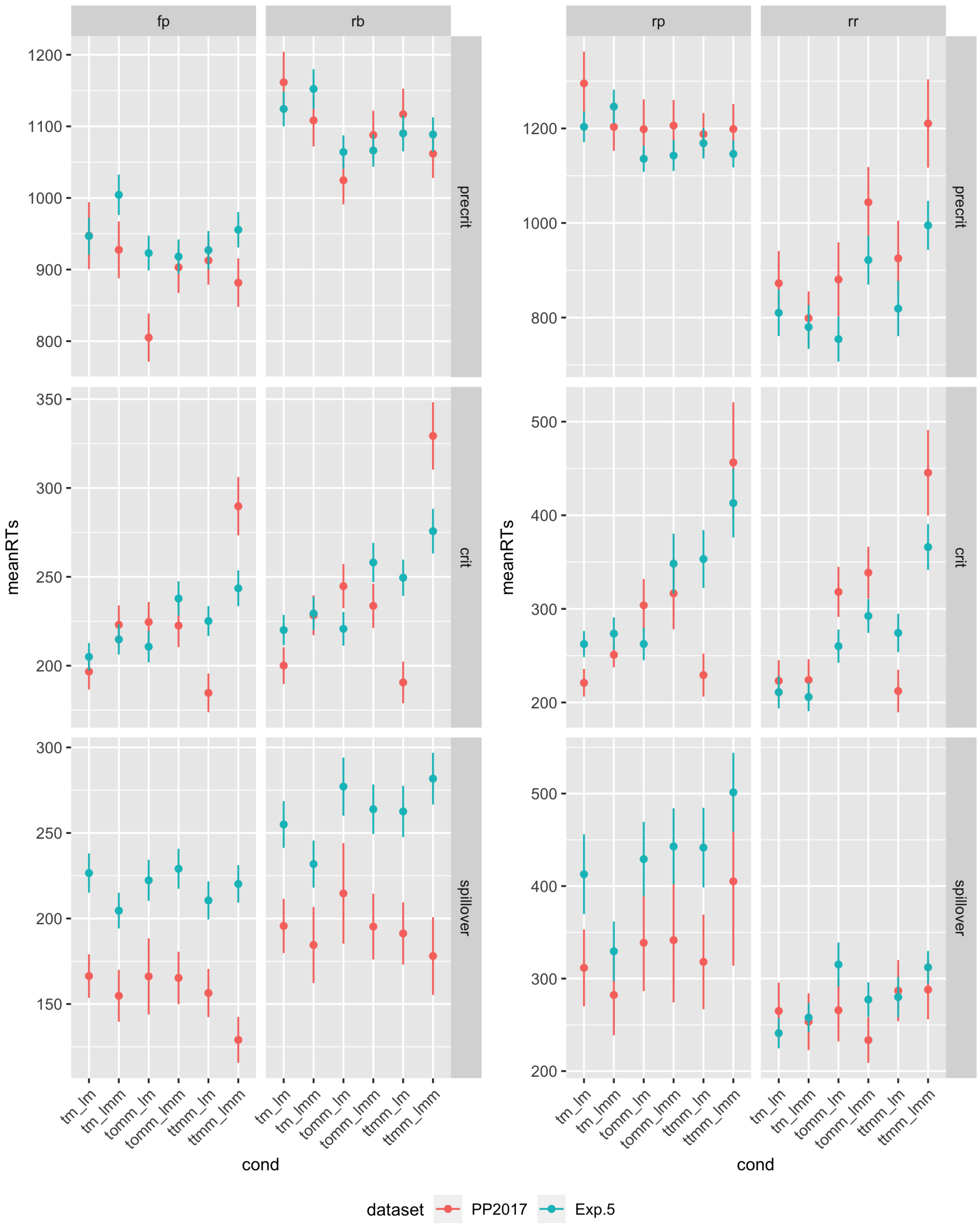

Figure 4.1: Mean RTs in PP2017 (red) and my replication (blue).

Error bars represent standard error of the mean. Conditions: tm/tomm/ttmm - target match/one-mismatch/two-mismatch; lm/lmm - lure match/mismatch. Columns correspond to eye-tracking measures: fp - first-pass, rb - right-bound times, rp - regression path, rr - re-read times. Rows correspond to ROIs: precritical, critical, spillover.

plot would make it harder to see the patterns; the sub-panelling does not have any other purpose except making the exposition clearer. The rows correspond to ROIs; we will be mostly interested in the central row, corresponding to the critical region (i.e., the reflexive). X axis corresponds to different conditions (see figure description); each condition is associated with two RTs: red dots represent means from PP2017, blue dots - from the current replication. We will be mostly interested in comparing two rightmost conditions in each facet - they correspond to 2-target mismatch conditions, where we expect to see interference effects. If PP2017 findings are replicated, we expect to see that the RTs estimate is much higher in the rightmost condition, indicating that lure mismatch sentences are read slower than lure match sentences within 2-target mismatch conditions.

I draw attention to two things in these plots. One is that there is less uncertainty in the replication, as suggested by the smaller error bars - this is to be expected, since I had twice as many participants as PP2017 did. Second, in both studies the estimates in the critical region follow similar patterns, but the replication estimates seem to be less extreme (e.g. they appear to be bigger than small PP2017 estimates and smaller than large PP2017 estimates). As a result, the numerical magnitude of the interference effect goes down in the replication, although it is still appears to be present in the pattern of the means.

Fig.4.2 shows the size of the interference effect in TARGET: TWO MISMATCH conditions, calculated as the difference between LURE: MATCH and LURE: MISMATCH conditions. Negative values indicate that LURE: MATCH conditions are read faster, an indication of facilitatory interference. Since there is not as much infor-

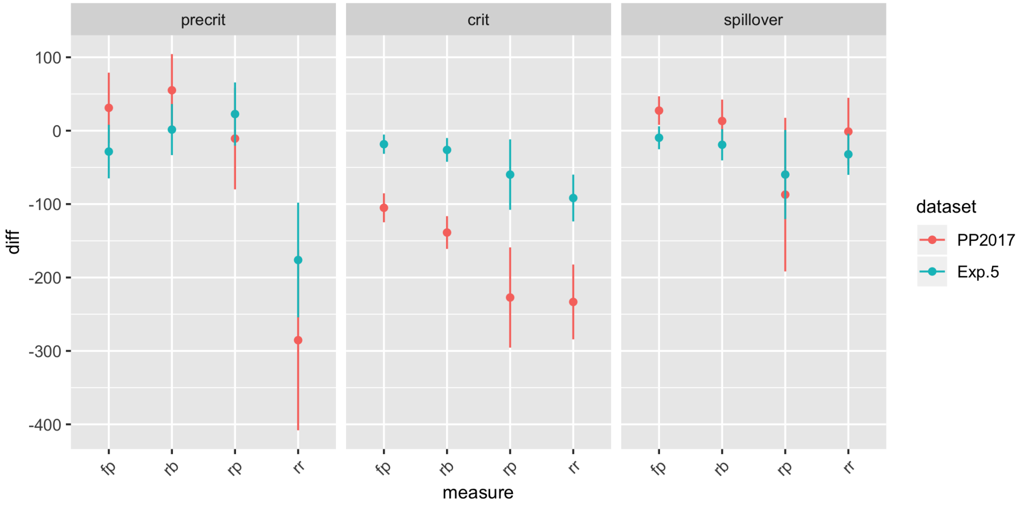

Figure 4.2: Interference effect in TARGET: TWO MISMATCH conditions (lure match - lure mismatch) in PP2017(red) and my replication (blue).

Error bars represent standard error of the difference of the means. Columns correspond to ROIs. Rows correspond to eye-tracking measures: fp - first-pass, rb - right-bound times, rp - regression path, rr - re-read times.

mation in this figure, I change the plot layout: eye-tracking measures are plotted along the x-axis, and plot facets correspond to ROIs. Critical region - the central facet - is of most interest, since this is where PP2017 observed the interference effect most consistently. One can see that interference effects in the replication are much smaller than in the original PP2017 study. Even the biggest one, observed in re-read times, is roughly twice as small as the corresponding estimate from PP2017.

Let us now turn to the results of the statistical analysis. For ease of comparison, I present model estimates from two experiment side-by-side on Fig.4.3. The dots correspond to $\hat{\beta}$ values ("Estimate" values from `lmer` output). In this figure I am mostly interested in showing which effects reach statistical significance (coefficients for which $|t| > 2$ are represented with square markers) and whether these effects survive multiple comparisons correction (triangles indicate coefficients which stop being significant after applying the correction; I discuss this issue further down the text). Intercept estimates are not represented, because they are numerically much bigger and including them in the overall plot forces the x-axis to stretch, so that minor differences between other estimates become less noticeable. As in Fig.4.1 columns correspond to eye-tracking measures, and rows - to ROIs. The size of the estimate is displayed on the X axis (since the models were fit to log-transformed RTs, the estimates are numerically small). Different effects from the model are plotted along the Y axis (see figure description for more details).

I will focus my discussion on the critical region (central row), since this is were the lure-match effects were the clearest in PP2017, and this is what PP2017 base their conclusions on. For the full models, I can expect three estimates to reflect the interference effect:

- Main effect of LURE MATCH, which would indicate that reading times on the reflexive differ for matching and mismatching lures, if I compare TARGET: MATCH and TARGET: TWO MISMATCH conditions;

- TARGET MATCH X LURE MATCH interaction for TARGET: ONE MISMATCH conditions. It would indicate that reading times are differentially affected by lure mis(match) in TARGET: MATCH and TARGET: ONE MISMATCH conditions.

- Similarly for TARGET MATCH X LURE MATCH interaction for TARGET: TWO MISMATCH conditions.

In the replication data, only two effects of interest reach significance. The

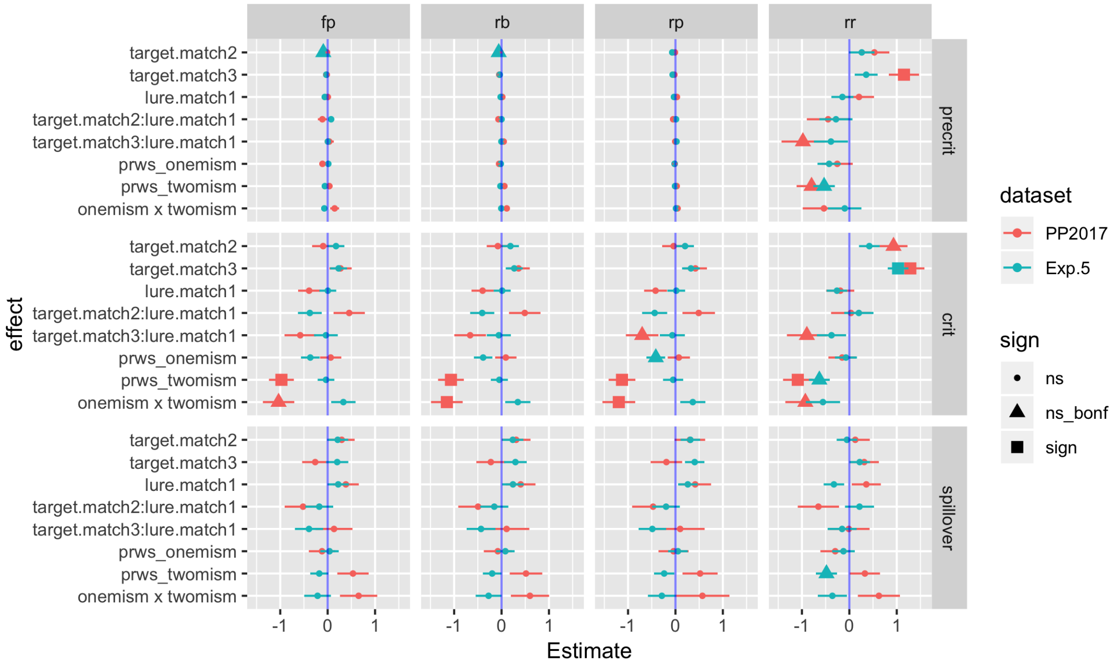

Figure 4.3: Model estimates in PP2017 (red) and my replication (blue).

First five effects come from the full model (in order: main effect of target match in 1-mismatch conditions; in 2-mismatch conditions; main effect of lure match; two interactions of LURE MATCH and TARGET MATCH for 1-mismatch and 2-mismatch conditions correspondingly); next two effects are the simple main effects of lure match; finally, the last effect is the interaction from the ungrammatical-only model. The order of the coefficients is the same as in PP2017, to make comparisons across papers easier. Error bars represent standard error of the estimate. Coefficients for which $|t| \geq 2$ are represented with squares if they remain significant after a Bonferroni correction (with the corrected $|t| = 3.34$), and with triangles, if they don't.

first one is the pairwise comparison within TARGET: ONE MISMATCH conditions in regression path, indicating that the pronoun is on average read faster in LURE: MATCH conditions. The second effect is the pairwise comparison within TARGET: TWO MISMATCH conditions in total times, which receives a similar interpretation. This is suggestive of the lure-match effects. However, for neither of those cases, the interaction term for the full model or the interaction term for the ungrammatical-only model reach significance. Together with the absence of main effects of LURE

MATCH this suggests that the evidence for lure-match effects is at best weak.

Compare these results with PP2017 data. They do find significant TARGET MATCH X LURE MATCH interactions for TARGET: TWO MISMATCH conditions in regression path and re-read times. It indicates that in TARGET: MATCH and TARGET: TWO MISMATCH conditions reading times are differentially affected by lure match. Pairwise comparison within TARGET: TWO MISMATCH conditions reach significance across all ROIs. This indicates that within TARGET: TWO MISMATCH the pronoun is read faster if it matches the features of the lure. Finally, the interaction term from ungrammatical-only model also reaches significance across all ROIs. This indicates that the speed-up associated with matching lures has different magnitude in TARGET: ONE MISMATCH and TARGET: TWO MISMATCH conditions. PP2017 interpret these statistics as providing evidence for two claims: first, that interference effects do exist, and two, that they only appear in ungrammatical sentences with target noun mismatching the reflexive in two features. Generally, I agree with this interpretation. But before I proceed, I address two issues with PP2017 analysis which might affect the conclusions I reach.

Lack of multiple comparisons corrections Despite carrying on quite a lot of comparisons, PP2017 do not report any multiple comparisons correction. This is somewhat traditional in eye-tracking studies, but, as von der Malsburg and Angele (2017) argue, this is not optimal, and if one does run analyses for multiple ROIs and MOIs, these multiple analyses should be corrected for. von der Malsburg and Angele (2017) simulations indicate that Bonferroni correction is not overly conservative and can be applied in such situations. I use it and assess whether PP2017 and my conclusions hold after a correction for multiple comparisons. For a given combination of ROI and MOI PP2017 fit 4 models. One could argue that the full model should be counted as two comparisons for the purposes of multiple comparison corrections: we are interested in three effects, coming from this model, but we will be making only two inferences based on them: whether we have evidence for interference (main effect of LURE MATCH), and whether we have evidence for the interference differing between grammatical and ungrammatical conditions (two TARGET MATCH X LURE MATCH INTERACTIONS). Ungrammatical-only model and the simple main effects models correspond to three more comparisons. Thus, we are making at least five inferences per a combination of ROI and MOI. Overall, PP2017 analyze 3 ROIs and 4 MOIs, which leaves us with 5x3x4 = 60 comparisons. Given the original $\alpha = 0.05$, a Bonferroni corrected $\alpha = 0.05/60 = 0.00083$. The corresponding t-value[5] for a two-sided test is ±3.34.

For my data, neither of the significant effects remain significant after applying the correction. In contrast, for PP2017 data many of the significant comparisons reported for the critical region remain significant after this correction, except the following four (indicated with triangles in Fig. 4.3): two interaction effects for the full model in regression path and re-read time, and two interaction effect ungrammatical-

[5] Approximated from normal distribution following Jäger et al. (2015). The corresponding R command is `qnorm(0.00083/2)`

only model in first-pass and re-read times[6].

Choice of preprocessing procedures The most notable pre-processing choice by PP2017 I do not agree with is the decision to replace missing values with zeros. Although PP2017 are certainly not the first to make this decision (Sturt, 2003; Cunnings and Felser, 2013; Cunnings et al., 2014; Cunnings and Sturt, 2014, see, e.g.), there are reasons for which it may be problematic, and I discuss some of them below.

First of all, at a risk of stating the obvious, missing values are not the same as zero reading times. In the eye-tracking experiments, there are at least two ways in which missing values can arise: people did not fixate on a given region, or the fixation was not recorded due to technical difficulties. I may be willing to ignore this source of noise and assume that all missing values represent the lack of fixation. But even if we make this assumption, literally assuming that the dataset includes RTs of 0 ms does not really make sense. Fixations below 50ms are likely non-existent in actual reading (Rayner, 1998, plot at p.376), so any and even if they did exist, the preprocessing procedure PP2017used[7], removed any fixations shorter than 80 ms, so we do not expect to find any smaller values in the data.

Now, one may argue that "the region was not read" is essentially the same as

[6] One of the main effects of TARGET MATCH in re-read times also stops being significant. However, we are not interested in this effect, so I do not discuss it further

[7] PP2017 do not directly report doing that, but Parker (2014), which is the original source of these experiments, does. Given that the average RTs reported in both sources are the same, it is reasonable to assume that the pre-processing procedures were the same, or very similar, as well.

"the region was read for 0 ms", and amounts to the difference in wording, but it does not. If we consider the way reading times may be generated in an eye-tracking experiment, this difference will be clearer. For a given word, people either make a fixation on it or not, with a certain probability. If people did fixate the word, they look at the word for a certain time. Of course, how these specific decisions are made is probably a very complicated process depending on multiple factors. But we believe that this general two-step description of the generative process is not unreasonable. So we are dealing with at least two distributions: one determining the probability of fixation, and another one - determining the duration of a fixation, given that it has occurred[8]. So the statement "the region was not read" corresponds to a situation where one drew 0 from the distribution specifying the probability that the region is re-read[9], and the statement "the region was read for 0 ms" corresponds to a situation where one drew 1 from the probability of re-read distribution, and then drew 0 from the distribution of re-read times. Notice that we assume that the distribution of re-read times even includes 0, which it may not, given the data I mentioned above on the fixations duration.

A better model for the data of such structure would reflect the structure of the generative process and could be constructed in two steps: first, use logistic regression to predict whether a certain word will or will not be skipped; then, for

[8] It is also the case that not all missing values result from people skipping a region. E.g. sometimes they may result from a failure of the equipment to record a fixation. So one could also assume distributions for the probability of such events, e.g., specifying the probability of eye-tracker losing track for a particular fixation

[9] Assuming 0 corresponds to "no re-read".

the words which were not skipped, use linear regression or any other appropriate model to predict reading times for these words ((see e.g. Gelman and Hill, 2007, p.126) for a similar approach).

It may be argued in response to the above that the goal of statistical analyses (in these particular experiments) was not to come up with a good model of the process that generated the data, but rather to capture some generalization about them, regardless of whether it's ecologically valid. E.g. by collapsing missing and non-missing values in the same analysis, one might be asking the question: on average, how much time do people spend re-reading a given region? Which is different from the question: *given that people re-read the word*, how much time on average did they spent doing so? However, replacing missing values with zeros may have consequences to the inferences as well.

Replacing missing observations with zeros will likely drag the estimates down and increase their uncertainty. Now, we do not know whether the values are missing at random. It may be, for example, that people are less likely to re-read the reflexive region in grammatical conditions. If this is the case, fewer zeros will be introduced to the RTs coming from grammatical conditions; and similarly for other possible differences in re-read rates between conditions. Ultimately, the value and the certainty of the model estimates may be differentially affected in different conditions, and it is not immediately clear whether and in what way this may influence the inferences. (To be fair, the same logic applies to rejecting the missing values: since we do not know whether they are missing at random, rejecting them can bias the conclusions we have. But since replacing missing values with zeros introduce other potentially undesirable effects, I advocate this method of dealing with them. Of course, this is not the only method: e.g. one could try to use imputation procedures of various degrees of sophistication. I do not know whether and in what way that would affect the conclusions I would make.)

An additional complication with 0 values is due to the fact that PP2017 conduct their analysis on log-transformed RTs. Due to non-linearity of log function, small changes on log scale can translate to quite big changes on linear scale. Fig.4.4 demonstrates this: it shows the difference between the observed RTs (red markers) and the RTs predicted by the model fir to the data with missing values replaced with zeros (blue markers)[10]. The figure layout is mostly analogously to Fig.4.1: experi-

[10] In order to generate model predictions, I do the following. First, for each fixed effect defined in the model I obtain a sample from $Normal(\hat{\mu}, \hat{\sigma})$, where $\hat{\mu}$ and $\hat{\sigma}$ are the estimate and standard error of the estimate, provided by the model. In this case, this will give us 7 samples: intercept and six other effects. I combine these values according to the contrast coding schema, to get the predicted reading time (on log-scale) for each condition. I.e. the estimate for the intercept would correspond to the reading time in the baseline, TARGET: FULL MATCH, LURE: MISMATCH condition. If we want to get an estimate for TARGET: TWO-MISMATCH, LURE: MATCH, we have to sum the four corresponding effects (intercept, two main effects and an interaction), etc. In the end, we obtain six predicted reading times on log-scale, one for each condition, which we exponentiate to convert them back to the linear scale. I repeat the above procedure 3000 times, obtaining 3000 predicted RTs for each condition. I average these predictions and compute their standard deviation. These averages are my estimates of the mean reaction times on the linear scale as predicted by the model. Notice that I do not incorporate uncertainty associated with participant and items into these predictions; essentially, I am predicting reading times for an average participant reading an average item. I am also ignoring variability in individual RTs: I am not trying to see what range

mental conditions are plotted along the X axis (see figure description for details), RT estimates are plotted along the Y axis. Columns correspond to eye-tracking measures. The only major difference from Fig.4.1 is in the rows: all the data are coming from the critical region, so now the rows represent datasets (the upper row corresponds to the data from the current replication; the lower one - to PP2017 data.). The main observation I make about this figure is that the RTs predicted by the model are clearly much smaller than the observed RTs. In some cases they are nonsensically small. E.g if we look at the the original study, Target Match, Lure Mismatch condition in the re-read times ("tm-lmm" condition in the figure), we will see that the model predicts average reading times of 20 ms.

4.1.6 Sensitivity analysis

In order to assess the influence of analysis decisions on the inferences I can derive from PP2017 data, I conduct a small scale sensitivity analysis. In addition to the two factors discussed above - corrections for multiple comparisons and treatment of missing observations - I consider two more: treatment of extreme values and the definition of the critical region (PP2017 and S2017 differ in these last two things). Here is the summary of the analysis decisions I manipulate:

1. **Definition of critical region.** I contrast two possible definitions: critical region includes only the pronoun or the pronoun plus three additional characters to the left[11]. PP2017 uses the first variant, S2017 - the second.

of RTs my model predicts, rather what range of mean RTs it predicts.

[11] Since reflexive pronouns are close-class words, they may be skipped relatively frequently during

2. **Dealing with missing values.** I contrast replacing missing values with zeros and rejecting them. PP2017 adopted the first approach; S2017do not report how they treated the missing values.

3. **Dealing with extreme values.** I contrast no trimming with fixed threshold trimming, following PP2017 and S2017 correspondingly. For fixed threshold, any values above 2000ms for first-pass and above 4000ms for total times are rejected.

4. **Multiple comparisons corrections**. I contrast analyses with no correction at all and analyses with correction for 60 comparisons, as discussed earlier.

4.1.6.1 Qualitative analysis

I start with discussing numerical patterns in the data. For the purposes of the discussion I focus on the change in the size of interference effects in TARGET: TWO MISMATCH conditions, shown in Fig. 4.5. To remind, interference effects is calculated as the RT in the target mismatch lure match condition minus the RT in the target mismatch lure mismatch conditions. Thus, negative values are indicative of the facilitatory interference. The columns correspond to variation in extreme data trimming decisions[12]; the rows - to missing data removal decisions, the colors - to the combination of studies and critical region mark-up decisions (see figure description

the reading. Extending the critical region to the left may reduce the amount of missing data resulting from skips.

[12] Notice that cut-offs for extreme data points were only established for first-pass and total times data, thus for other eye-tracking measures there is no change across the columns.

for details). Different eye-tracking measures are represented along the x axis.

When looking at this figure, we are mostly interested in whether the preprocessing decisions affect the size of the interference effects. The answer is: preprocessing choices naturally do have influence on the estimates, but it is not huge, and the patterns of the results remain relatively stable across different pre-processing variants. Choosing a shorter region naturally leads to shorter reaction times, but the intrusion effect remains almost unaffected. Choosing to reject missing values leads to the reduction of the effect size in some cases, not exceeding roughly 30ms. Finally, the trimming of extreme values does not affect first pass, and leads to a reduction in total times (roughly 40 ms for PP2017data and roughly 20ms for the replication). As we will see later, we have reasons to believe that the size of lure-match effect in reflexives lies in the range of roughly 50-100ms, so a reduction of 30 or 40 ms is quite sizeable. Finally, I would like to point out that for all preprocessing variants I observe numerically big intrusion effects in PP2017 and much smaller effects in my replication. This suggests that the discrepancies we observe cannot be explained by di

4.1.6.2 Quantitative analysis

Now I turn to the discussion of statistical estimates obtained in different analysis procedures. The primary question I am interested in this section is: do analysis choices influence the inferences we make from the models? I use the same analysis procedure as described earlier, although I automate the simplification of the random effects structure for the non-converging models (it was done manually in the main analysis).

Fig. 4.6 displays the estimated coefficients for the models. Each dot in the plot corresponds to a $\hat{\beta}$ value from some model, in a way similar to Fig.4.3. We consider models fit to 12 different datasets, resulting from the variation of pre-processing choices and the study the data was coming from. These 12 datasets are plotted along the Y axis of each plot. The names of analyses variants are compositional: each sub-component corresponds to one dimension we manipulated. The first component corresponds to the study ("repl" - current replication, "pporig" - original study by PP2017); the second component corresponds to the critical region markup ("nonext" - critical region includes only the reflexive itself; "ext" - critical region includes the reflexive and three characters to the left. Notice that only "nonext" variant is available for PP2017); the third component corresponds to the treatment of extreme values ("notrim" - include all values in the analysis, "trim" - trim values exceeding 2000ms in first-pass and 4000 in total times); the last component corresponds to missing values treatment ("narm" - remove, "nazero" - replace with zero). Sub-panels of the plot display estimates from models fit to different eye-tracking measures. Each column displays model estimates for a given fixed effect: TM1/2 - target match , one/two feature mismatch; LM - lure match; ung_int - interaction term from ungrammatical only model; prws1/2 - pairwise comparisons, corresponding to simple main effects of lure match within the corresponding target match conditions.

Notice that the coefficients are not easily interpretable on their own, since they are on log scale. Thus, I will only discuss what inferences one would make, if one was simply looking at the coefficients and checking whether they are significantly different from zero. I take coefficients with $|t| > 2$ to be significantly different from zero, and mark them in red on Fig. 4.6. In addition, I consider how the use of a Bonferroni correction affects the conclusions. Following the logic I described earlier for the main analysis, I correct for 60 comparisons, resulting in a Bonferroni corrected $\alpha = 0.05/60 = 0.00083$. The estimates which stop being significant are highlighted in yellow in Fig. 4.6.

Let me specify a decision algorithm I will use to make conclusions from the models. As far as I can say, this algorithm does not correspond exactly to how decisions were made in PP2017, and may lead to somewhat more conservative conclusions then reached in the original paper. I will discuss more liberal variants of it as well. I will keep using the model specification from PP2017, although some of the questions we are going to ask might be better addressed with slightly different models. I will use the following decision procedure:

1. First, I will look at three effects in the full models, to answer the question: **Do lures affect the processing of reflexive pronouns?** Main effect of LURE MATCH would suggest that LURE MATCH and LURE MISMATCH conditions differ within TARGET MATCH sentences. Two TARGET MATCH X LURE MATCH interactions, one for each non-baseline level of TARGET MATCH would indicate that the effect of LURE MATCH differs between grammatical sentences and - depending on the interaction - one-mismatch or two-mismatch ungrammatical

sentences. If I do find at least one significant interaction, I will declare that I have evidence that interference effects exist, and that they behave differently in grammatical and ungrammatical sentences[13]; in this case, I will proceed to the following step of the algorithm. If I do not find any interactions, I will only conclude that lure-match effects exist and stop.

2. If I find at least one interaction in the previous step, I will ask the next question: **Does interference behave differently depending on the degree of ungrammaticality?** To answer this, I will look at the interaction coefficient from ungrammatical-only model. If it is significant, I will claim that not only interference differs between grammatical and ungrammatical sentences, it also behaves differently in different types of ungrammatical sentences. If the interaction term is not significant, I will stop and declare that I don't have any evidence to believe that the degree of ungrammaticality affects interference effects.

3. If I do find the interaction in the previous step, I will want to resolve it. To do this, I will look at the simple main effect of LURE MATCH within TARGET: ONE MISMATCH and TARGET: TWO MISMATCH sentences.

[13] Notice that in this model specification, we reach this conclusion somewhat indirectly. We look at differences between grammatical sentences and one-mismatch ungrammatical sentences; similarly for two-mismatch ungrammatical sentences; then, we conclude that grammatical and *both types* of ungrammatical sentences do or do not differ. A better way to address this question might be to use Helmert contrasts to first compare grammatical conditions to the average or the two ungrammatical conditions, and then compare the ungrammatical conditions between themselves.

A more liberal version of the algorithm would allow to proceed to consider the ungrammatical only model even if no interactions in the full model are significant. This would roughly correspond to assuming that we are not interested in whether LURE MATCH differentially affects grammatical and ungrammatical sentences, only in whether LURE MATCH behaves differentially in ungrammatical sentences depending on the degree of ungrammaticality. Another way to make more liberal decision is not to take Bonferroni correction into account. I will apply the decision algorithm to all eye-tracking measures, and make the corresponding conclusions if for any single measure my algorithm allows us to do so[14]. I start with analyzing PP2017 data.

Main effect of LURE MATCH never reaches significance regardless of pre-processing procedure. TARGET MATCH X LURE MATCH interactions do reach significance for TARGET: TWO MISMATCH conditions in all regions and measures, only if we choose to remove missing values from the data[15]. These effects survive Bonferroni correc-

[14] This is yet another degree of freedom in the analysis. As von der Malsburg and Angele (2017) notice, it is somewhat common to make claims about some effect if it is present at any given measure and/or region. In principle, it would be better if we could make predictions about in which exact measure the effect of interest will appear.

[15] Notice that in the analyses reported in PP2017, this interaction reached significance in regression path and re-read times, even though PP2017 chose to replace missing values with zeros. This discrepancy potentially results from the differences in random effect structures. I discuss it on the example of regression path model. As far as I could tell, PP2017 only specified random intercepts for both subjects and items, while in my case, I additionally had random slopes, $1 + target.match + lure.match \mid subject$. This was the most complex model that converged during the simplification process described earlier. This difference in random effects specification was apparently enough to push the t-value for the interaction estimate from -2.04 (which would

tion in right-bound and re-read times, but not in first-pass and regression path. For the models based on data with rejected missing values (and, if I am being liberal, for the models with missing values replaces with zeros), I can continue to the next step of the decision algorithm and look at the interaction term for the ungrammatical only model. It turns out to be significant in almost all measures, except for re-read time. However, in neither of the reading measures does it survive Bonferroni correction. Thus, if I am being conservative, I would stop here and only be able to claim that intrusion effects do differ between grammatical and ungrammatical sentences, but that I do not have enough evidence for differential effects of LURE MATCH depending on the degree of ungrammaticality. If I am being more liberal and ignore Bonferroni correction, I can look at the simple main effects of LURE MATCH within ungrammatical conditions, which always turn out to be significant. Thus, in a liberal version of the decision algorithm, I would be able to claim that interference effects are only observed in two-mismatch conditions. Summing up, if I ignore multiple comparisons correction, I would be able to reach the same conclusions as PP2017 do; if I am being conservative, I only have strong enough evidence for the difference in the effect of LURE MATCH between 2-feature target mismatch and target match conditions, but not the difference between the two ungrammatical conditions. A possible interpretation for this weaker result would be: the experiment provides evidence for interference in reflexive resolution contra previous findings (while the original PP2017 conclusions are able to accommodate previous findings due to the

correspond to a significant effect) to -1.58 (non-significant). This suggests that random effects specification has to be carefully considered and preferably reported for all models in a paper.

claim of differential interference effect). Overall, it looks like I can reach the original conclusions of PP2017 regardless of the pre-processing, although one needs to be more liberal in the decision-making when using the models with missing observations replaced with zeros.

Let us now turn to the replication data. Similarly to PP2017, the main effect of LURE MATCH in the full model never reaches significance. TARGET MATCH X LURE MATCH interactions do not reach significance in any analysis variant and eye-tracking measure. Thus, if I am being conservative, I would have to stop here and claim that I do not have strong enough evidence for interference in reflexive resolution. Being somewhat more liberal does not help: even if I look directly at ungrammatical only model, I would not be able to claim that I have found evidence for interference effect, since under no pre-processing scenario does the interaction from the reduced model reach significance. Only if I relax the standards even further and decide the simple main effect of LURE MATCH are good enough evidence for interference, I would be able to claim the I found the effect - but only without the correction for multiple comparison. Summing up, for my data I would have to be very liberal to claim that I have enough evidence for interference.

Overall, while the sensitivity analysis suggests that the evidence presented by PP2017 may be somewhat weaker than claimed in the paper, the replication study provides even weaker evidence for interference effect (in the conservative case - provides no reliable evidence at all).

4.1.7 Discussion

The main goal of this experiment was to check whether it was a mere presence of QPs in the set of experimental materials which prevented us from observing reliable interference effects in the previous chapter. The answer is negative. While the overall patterns of (average) reading times appeared to be similar, the magnitudes of the reading times were less extreme in the replication. The magnitude of the interference effect was at least twice as small as compared to the original study, and was comparable to that observed in my Exp.3-4. The statistical analyses also did not provide a strong support for the existence of interference effects. While the choice of analysis procedure did have some influence on the conclusions (e.g. only allowing for weaker claims in case with missing values replaced with zeros), the overall impression still held: the data from the replication provides less statistical support for the existence of the interference effect than the data from the original study.

I also intended to use this experiment to tease apart PP2017 and S2017 accounts. If I did observe reliable interference effects here, it could be taken to support PP2017: their account can accommodate the influence of context on the resolution process much more readily than S2017. However, since it does not appear that the presence of QP affected the size of the lure-match effects, I cannot use these data to help distinguish between the two accounts. In the next experiment I address a different possible context effect: the influence of the experimenter language characteristics. I will use the same exact materials as in the current experiment, only this time all the data will be collected by a native experimenter.

4.2 Experiment 6

In the previous experiment I failed to find strong support for interference effects from NP lures, despite using the same exact materials as Parker et al. (2017). I did find numerical trends in the predicted direction, but they were not supported by statistics. The only systematic explanation for the lack of replication I can think of is the influence of the experimenter: perhaps, instructions in non-native English affected participants' processing of ungrammatical sentences. The current experiment aims at testing this possibility. It exactly replicates my previous experiment, with a single difference: all the data were collected by native English speakers[16].

4.2.1 Participants

35 members of University of Maryland community (23F, 12M; age range: 18-34, average age: 21.4) participated in the experiment for a class credit or a payment of $12. The experimental session took about one hour, including setup and calibration. Data from two participants were removed: one participant did not pay attention during calibration and was distracted during the experiment; for the other, experimental software could not adequately process the data. Thus the analysis was

[16] I would like to thank Lalitha Balachandran and Cassidy Wyatt for their help with data collection.

based on the data from 33 participants[17].

4.2.2 Materials

The materials were identical to Experiment 5.

4.2.3 Procedure

The procedure was identical to Experiment 5.

4.2.4 Analysis

The analysis procedure was identical to Experiment 5. Again, I first follow the analysis procedure as outlined in Parker and Phillips (2017), to be as close as possible to the original study. Then I perform an additional set of analyses with a different pre-processing procedure, which I think to be better for the reasons discussed earlier in this chapter (its two main differences are a wider critical region and removing missing values instead of replacing them with zeros).

4.2.5 Results

I start with observed patterns of mean reaction times, which are shown in Fig. 4.7. Numerical values for the mean are given in Appendix A. I compare the results from the current experiment (green dots), my previous replication in Exp.5 (blue dots) and the means from the original study by PP2017 (red dots).The layout

[17] I aimed at 48 in order to have the same power as in the previous replication, but could not reach the goal due to time limitations

of the figures is identical to Figs.4.1 and 4.2 from the previous chapter and was described in detail there. Here, I will focus the discussion on the critical region, since this is where PP2017 observed the interference effect most consistently. The overall patterns of means seem to be roughly similar in all three studies.

The first thing I note about the numerical values is that for some reason the average reading times are smaller than in both PP2017 and Exp.5. All three experiments look most similar within 2-feature target mismatch conditions in terms of the pattern of the RTs: lure match conditions are read faster than lure mismatch conditions. However, in terms of the magnitude of the interference effect[18] in 2-feature target mismatch conditions, the current experiment is closer to my first replication than to the original study (Fig.4.8[19]), and again it is smaller in magnitude than the interference effects from Parker and Phillips (2017).

Let us now turn to the statistical analyses. As in Exp.5, I present model estimates from the current experiment and from the original PP2017 experiment side-by-side on the upper panel of Fig.4.9. The layout is essentially the same as in 4.3 from the previous experiment, except that we are presenting two sets of of comparisons in the upper and the lower panels, and the estimates are only coming from the models fit to the data from the critical region. Coefficients for which $|t| \geq 2$ are represented with square markers; triangles indicate coefficients which

[18] As a reminder, calculated as the difference between LURE: MATCH and LURE: MISMATCH conditions. Negative values indicate that LURE: MATCH conditions are read faster.

[19] Notice the change in the plot layout: eye-tracking measures are plotted along the x-axis, and columns correspond to ROIs.

stop being significant after a multiple comparisons correction. Intercept estimates are not included in the plot.

I will focus my discussion on the critical region. To remind, for the full models, I can expect three estimates to reflect interference effect:

- Main effect of LURE MATCH, which would indicate that reading times on the reflexive differ for matching and mismatching lures, if we compare TARGET: MATCH and TARGET: TWO MISMATCH conditions;
- TARGET MATCH X LURE MATCH interaction for TARGET: ONE MISMATCH conditions. It would indicate that reading times are differentially affected by lure mis(match) in TARGET: MATCH and TARGET: ONE MISMATCH conditions.
- Similarly for TARGET MATCH X LURE MATCH interaction for TARGET: TWO MISMATCH conditions.

In the current experiment, no effects of interest reach significance. Two other effects, unrelated to LURE MATCH were significant: the two main effects of TARGET MATCH, indicating that ungrammatical sentences with mismatching lures are being read slower than grammatical sentences with mismatching lures. Both these effects survive Bonferroni correction based on the same calculations as in the previous experiment[20]

Compare these results to the original study and Exp.5. The pattern of results in re-read times is perhaps the most similar between the three. Two main effects

[20] To remind, I am correcting for 3 ROIs x 4 MOIs x 5 comparisons in each = 60 comparisons, which results in critical α of 0.00083 and the corresponding t-value of ± 3.34

of TARGET MATCH are most consistent. For one-mismatch conditions, they reach significance in two studies, and remain significant after Bonferronni correction in one of them; for two-mismatch conditions, they reach significance in all three studies and remain significant in two of them after the multiple comparisons correction. The magnitude of the estimates is also very similar. These results suggest that the ungrammaticality of the sentences strongly affects participants. In other eye-tracking measures, it appears that the two replications are closer to each other than to the original study: the magnitude of the estimates in the replications is smaller; very few effects reach significance and none survive the multiple comparisons correction.

Exploratory analysis The sensitivity analysis I performed for the previous experiment suggested that out of three preprocessing parameters I have manipulated - definition of the critical region, trimming of the extreme values and treatment of missing values - the last one had the biggest impact on statistical estimates. Thus, I briefly discuss what would happen had I chosen to remove the missing values instead of replacing them with zeroes in the main analysis. Fig. 4.10 shows the mean RTs, Fig. 4.11 shows the size of the interference effect and Fig.4.12 shows the coefficients from statistical models.

As we can see, the pattern of the RTs and interference effects are not considerably affected by the preprocessing routine, except that the average RTs get somewhat more similar between the three studies in regression path and re-read. Statistical patterns also remain roughly the same: the only effects that reach significance are the main effects of TARGET MATCH (i.e. grammaticality effects). They are now significant in all eye-tracking measures, but most of them do not survive the Bonferroni correction. Still, it indicates that there is evidence for grammaticality effects in the data.

It is interesting to note that the estimates for the effects which did not reach significant are rather similar across all three studies, even more so than in the main analysis. On the other hand, the estimates for the effects indicative of interference are bigger in the original study as compared to the two replications. This again makes us think that the effects reported by PP2017 may be overestimated.

4.2.6 Discussion

The results of my second replication appear to pattern with my first replication and not with the original study. The lure-match effect in 2-feature target mismatch conditions was numerically going in the same direction as in the original PP2017 study, but the magnitude of the effect was at least two times smaller. Statistical analyses do not provide support for lure-match effects. This is like my first replication, and unlike the original, where statistics supported the presence of the effects in all analyzed eye-tracking measures. Overall, the study does not indicate that the language of the experimenter was what caused the absence of lure-match effects in my Exp.2-4. These conclusions are supported by the results of the exploratory data analysis: regardless of the pre-processing procedure I choose, my conclusions hold.

This experiment, as the previous one, fails to provide evidence for the influence of experimental context on the processing strategies. Thus, these results do not provide support for the PP2017 account, as I hypothesized they might. However, they provide valuable evidence on the size of lure-match effects we may be expecting in reflexives resolution, and this information may be used to guide power analyses for the future studies. Next section discusses these issues in more detail.

4.3 Interference magnitude across studies

In this section I compare the magnitude of the intrusion effect in the few studies which have looked at reflexive resolution in 2-feature target mismatch configuration, to see whether the results I have obtained in the studies I have run are unexpectedly small. I start by looking at five experiments, which were maximally close to each other in terms of materials and manipulation: PP2017 Exp.3, its two replications (my Exp. 5 and 6), S2017 Exp.1c (speech verbs), Exp.4 from this thesis. PP2017 Exp.3 and the current experiment used the same materials (experimental sentences and fillers). My Exp.4 overlapped with S2017 Exp.1c in a subset of experimental sentences (both experiments included additional conditions which were not similar; the fillers were also different). In all studies, the stimuli shared the following characteristics:

- Two clause sentences with the reflexive inside a complement clause;
- The matrix predicate is a predicate of report (e.g. "say", "mention"; etc.)[21];

[21] This parameter was controlled in S2017; it was not in PP2017, but in their materials about 70% percent of the verbs were report verbs, so I consider the materials to be comparable along

- The reflexive and lure are always singular;
- The lure is the subject of the matrix clause;
- The target is the subject of the embedded clause;
- The lure can match or mismatch the reflexive in gender;
- The target mismatches the reflexive in gender and number [22];
- The target and the lure are referential NPs.

Fig.4.13 shows the interference effect in 2-feature target mismatch conditions, calculated as the difference between lure match and lure mismatch conditions. Different studies report different reading time measures, that is why there are gaps in the plots. Two observations are of interest. First, intrusion effect is generally much bigger in PP2017 than in the other studies (except in regression path, where the magnitude of the intrusion is similar in PP2017 and S2017). Second, the magnitude of the intrusion appears to be reduced in my studies as compared to the corresponding original studies. Importantly, the magnitudes of facilitation from the two replications are very close to each other. This suggests that PP2017 estimates might be inflated, and that my estimates might be closer to the underlying effect.

this dimension.

[22] PP2017 and its replication had an additional condition with the target mismatching the target only in number, but S2017 and Exp.4 from this thesis didn't have it. Since intrusion effect appears only in the two-mismatch condition, I will not focus on the additional one-mismatch condition in PP2017 and its replication

Let us now consider interference effects in all studies investigating reflexive resolution in 2-feature target mismatch conditions with c-commanding lures. In addition to the experiments above, this list includes Exp.3 from this thesis, two more experiments by PP2017 and four more experiments by S2017. The magnitude of interference effect in these studies is displayed in Fig.4.14. The picture is perhaps most consistent in total times: the effect reported by PP2017 completely overshadows the other effects, but otherwise the size of the effect is surprisingly similar in experiments by S2017, and the magnitude of the effect from two replications I report is in line with them as well. The effects from Exp.3 and 4 are somewhat smaller, especially the first one. Finally, lure-match effect in S2017 Exp.1 for sentences with perception embedding verbs is almost non-existent, but this difference is accounted by S2017 account (and, as I have argued in Chapter 1, by PP2017 as well). Turning to other reading times measures, the picture is very similar in first-pass time, right-bound and re-read times: the magnitude of the interference effect in PP2017 Exp.3 is bigger than in all other studies, which are relatively similar (with the exception of Exp.3 from this thesis). Regression path shows the most heterogeneous picture, although even here PP2017 effect in Exp.3 is numerically the biggest, contested only by the effect from S2017 Exp.1 for sentences with speech verbs.

The comparisons discussed above do suggest that the intrusion effect in PP2017 Exp.3 is inordinately big. I do not think it is due to the particular manipulation used: S2017 Exp.1b and 2b rely on the same manipulation, and yet generally yield numerically smaller effects. The native language of the experimenter, at least in this set of studies, does not uniquely determine the size of the effect either: in both my Exp.6 and S2017 the data were collected by native English speakers, and the intrusion effects are smaller than in PP2017, It thus seems more likely that the difference in the effect size is due to some unmeasured factors; the fact that PP2017 only collected data from 24 participants might have contributed to the surprising size of the effect. As Gelman and Carlin (2014, (a.o.)) discuss, when the power is low, significant effects tend to be overestimates; the magnitude of the overestimation tends to infinity as power tends to zero. While we would need to conduct a formal power analysis to make specific claims, it is true that PP2017 Exp.3 relies on the smallest sample size among all the studies discussed in this section, so presumably has the lowest power and would have the biggest overestimation ratio among them all.

Another suggestive piece of information comes from a simulation study by Engelmann et al. (draft4). They investigate the influence of the LV05 model parameters on the size of the lure-match effect predicted in the model (p.13). The magnitude for the majority of the predicted facilitation effects (i.e. the type of effect I am interested in here) lies below 100ms, and the effects of the size reported by PP2017 are rather rare.

4.4 Pooled data analysis

The previous exploratory analyses indicated that the numerical lure-match effects I observed might be real but fail to receive statistical support due to the lack of power. In this section I report an exploratory analysis on the pooled data from both replications - this will more than triple the original sample size used by Parker and Phillips (2017).

First, I wanted to make sure that the data from the two replications is homogeneous enough to be pooled. To this end, I run a bootstrapping simulation, sampling 48 subjects with replacement from each of the three experiments - the original one by Parker and Phillips (2017) and the two replications I report here. I draw a thousand samples from each dataset, and calculate the magnitude of the lure-match effect in 2-feature target mismatch conditions for different eye-tracking measures in the critical region. Fig. 4.15 represents the distributions of bootstrapped lure-match effects magnitudes. As in most plots in this chapter, columns correspond to different eye-tracking measures. The magnitudes of the interference effects observed in the simulation are plotted along the Y axis; the height of the distribution at any given point roughly[23] corresponds to the number of observed effects of a given magnitude. Colors of the distributions correspond to the dataset which the simulations were based on (see the figure's legend).

We can see that the distributions of possible effect sizes are very similar between the two replications. They look almost identical in regression path and total times; they are not as similar in other eye-tracking measures but are still located and shaped quite similarly to each other. On the other hand, possible effect sizes one can obtain from PP2017 data are bigger and practically do not overlap with the replications distributions. I conclude that the data from the two replications is homogeneous enough to be pooled for an exploratory statistical analysis.

[23] Although not precisely, since it's not a histogram, but rather a smoothed version of it.

Fig.4.16 shows the results of statistical modeling with the pooled dataset. The layout is similar to Fig.4.3. The data only come from the critical region. As in Exp.6 for the sake of exploration I compare whether the treatment of the missing values would affect my conclusions. We can see that grammaticality effects (main effects of TARGET MATCH) reach significance in all measures at the critical region; this is not surprising given that even in the individual experiment the evidence for the grammaticality effects was quite robust. On the other hand, evidence for lure-match effects still remains rather elusive.

In the analysis with missing values removed, the TARGET MATCH X LURE MATCH interaction for TARGET: 2-MISMATCH conditions reaches significance in re-read and total times. Neither of these is significant if I choose to replace missing values with zeros. This is probably due to a larger standard error of the estimate in this case, since the magnitude of the estimates is rather similar between the two analysis procedures. Overall, the analysis of the pooled data suggests that a moderate increase in sample size may boost the statistical power enough to make the lure-match effect more readily detectable.

4.5 General discussion

The goal of the two experiments reported in this chapter was two-fold. First, I wanted to determine how reliable my results from Exp.2-4 were. To remind, my question there was whether QP lures elicit lure-match effects in the same way as NP lures do. I failed to find statistical support for interference from either NP or QP lures. Numerically the RTs did go in the right direction (lure match sentences were read faster within target-mismatch conditions only) when the QP lure c-commanded the pronoun, but the magnitude of the effect was reduced as compared to PP2017. This made me worry that the experimental manipulation did not work in some subtle way, and the results were completely uninterpretable. The experiments and analyses reported in the current chapter allowed to partially assuage this worry.

In both replication experiments I found a lure-match effect which is numerically smaller than that reported in PP2017. As in Exp.2-4, this effect does not reach significance in the statistical analyses. However, the comparison of the effect sizes across multiple related studies suggests that it is PP2017 and not my results which are outliers: effects of about 70-100ms (as in my studies) are more common than those of 200-250ms (as reported by PP2017).

Importantly, the magnitude of the lure-match effects in Exp.4 lies in the same range. It suggests that the experiment did, indeed, work as intended. Thus, I conclude that the interpretation of the data I tentatively suggested in the end of the previous chapter holds: c-commanding QP and NP lures elicit similar lure-match effects, suggesting that both types of lures influence the reflexive resolution process. This supports the conclusion that the differential patterns of RTs between NP and QP lures I observed in Experiment 2 are not due to QP being inefficient lures, and may truly indicate the inaccessibility of non-c-commanding QP to the parser.

Influence of the context A second question I addressed in this chapter was whether experimental context (in terms of the stimuli set or experimental environment) has an effect on the magnitude of the lure-match effect. I argued that if contextual factors did indeed affect the magnitude of the interference, it would provide support for PP2017 account, which could explain such influence much more easily than S2017. However, the results suggest that the context (at least the two factors I explored) does not influence the size of the lure-match effect. The size of the lure-match effect I observed was consistent regardless of whether the experiment was run with a stimuli set containing QPs, with a set without them, by a native or non-native experimenter. Therefore, there is no direct evidence for preferring PP2017 over S2017 coming from this set of experiments.

Lack of statistical significance While the comparison of effect sizes is informative, the six experiments I conducted so far provided extremely limited statistical evidence for the existence of lure-match effects. Given that lure-match effects do appear to be relatively consistent in direction and magnitude across my and some of the previous studies, I think that the lack of statistical support is in big degree due to statistical reasons. As I have already mentioned, under-powered studies tend to produce over-estimates of significant effects (Gelman and Carlin, 2014). A recent report of a large reproducibility project by Collaboration (2015) provides empirical evidence for this claim: out of 100 replication attempts, 97 original effects were significant at 0.05 level, while only around 35% replication effects were so. Similarly, in 82 out of 99 original studies for which effect size could be calculated showed a bigger effect size than the replication.

While this explanation is tempting, and, I think, has substance to it, it cannot be exactly right: S2017 used very similar sample sizes, but consistently found statistically robust effects. I return to this issue in the closing chapter of the thesis.

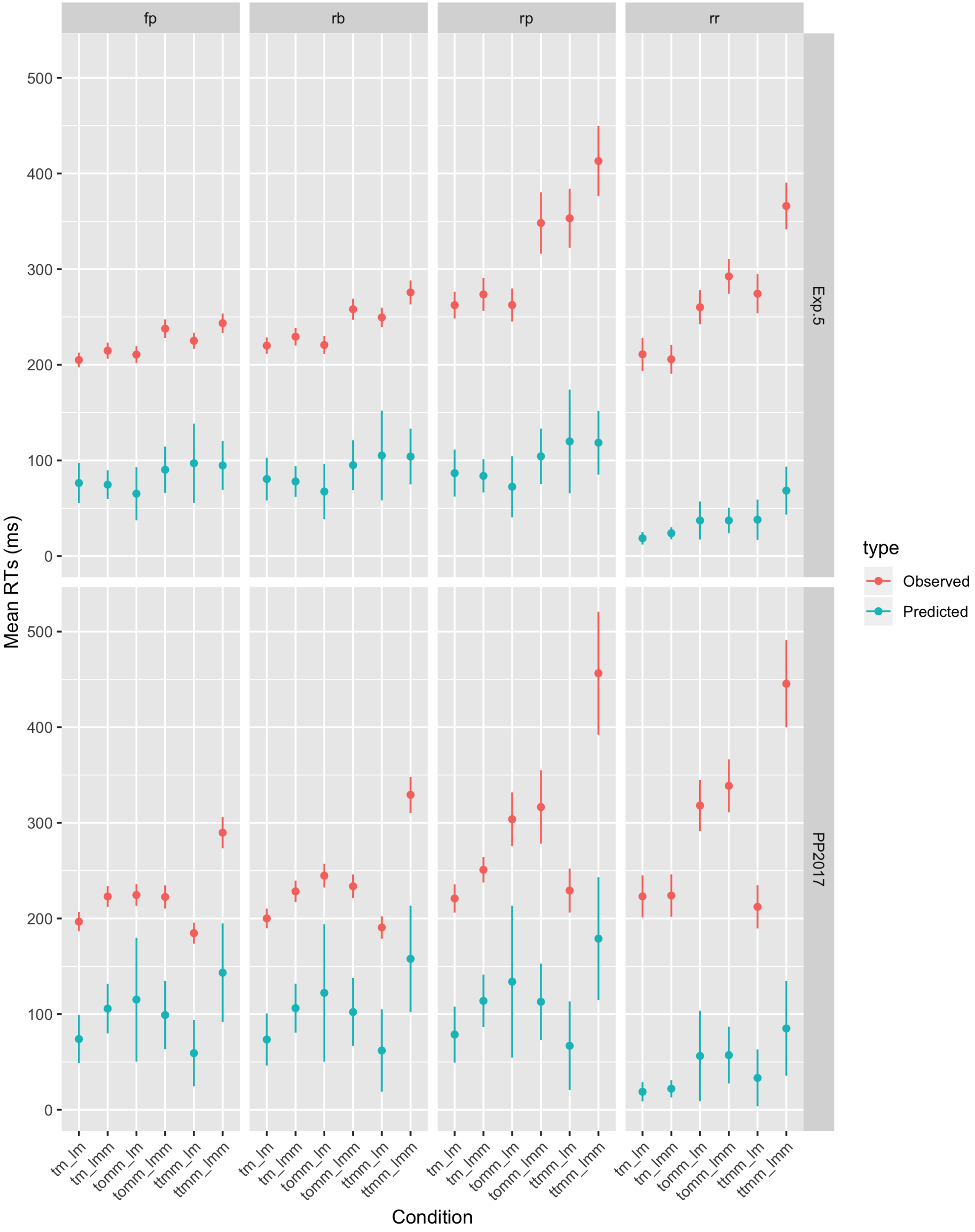

Figure 4.4: Predicted and observed RT values for PP2017 and my replication.

Columns correspond to eye-tracking measures: fp - first-pass, rb - right-bound times, rp - regression path, rr - re-read times (all RTs come from the critical region). Rows correspond to datasets (the upper one - current experiment; the lower one - original PP2017 data). Error bars for the observed data represent standard error of the estimate, adjusted for participant variability (Cousineau, 2005; Morey, 2008). Error bars for the predicted data represent standard deviation for the simulated means. Conditions: tm/tomm/ttmm - target match/one-mismatch/two-mismatch; lm/lmm - lure match/mismatch.

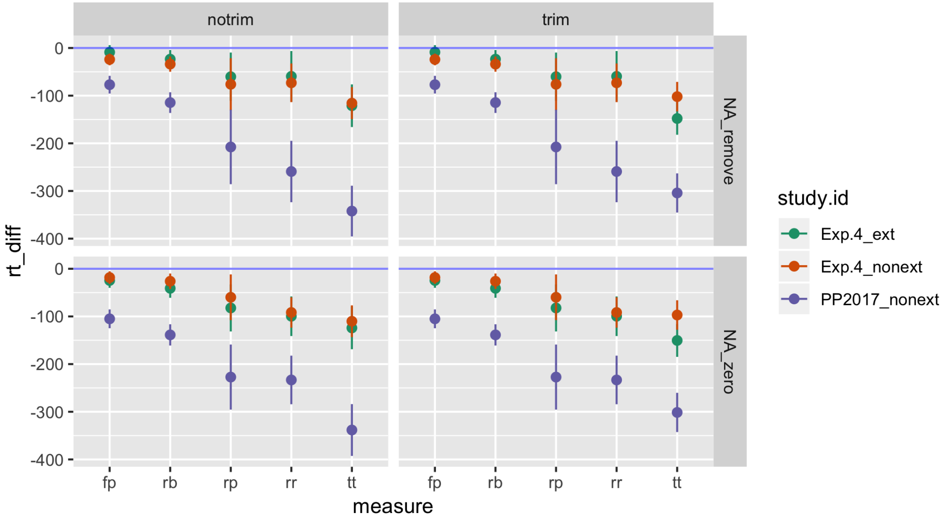

Figure 4.5: Sensitivity analysis: Intrusion effect (lure mismatch minus lure match) in 2-feature target mismatch conditions.

Columns correspond to variation in extreme values treatment (“notrim” - include all values in the analysis, “trim” - trim values exceeding 2000ms in first-pass and 4000 in total times); rows correspond to missing values treatment (remove or replace with zero); colors correspond to the combination of study and crtical region markup (“nonext” - critical region includes only the reflexive itself; “ext” - critical region includes the reflexive and three characters to the left. Only “nonext” variant is available for PP2017). Error bars represent standard error of the difference of the means.

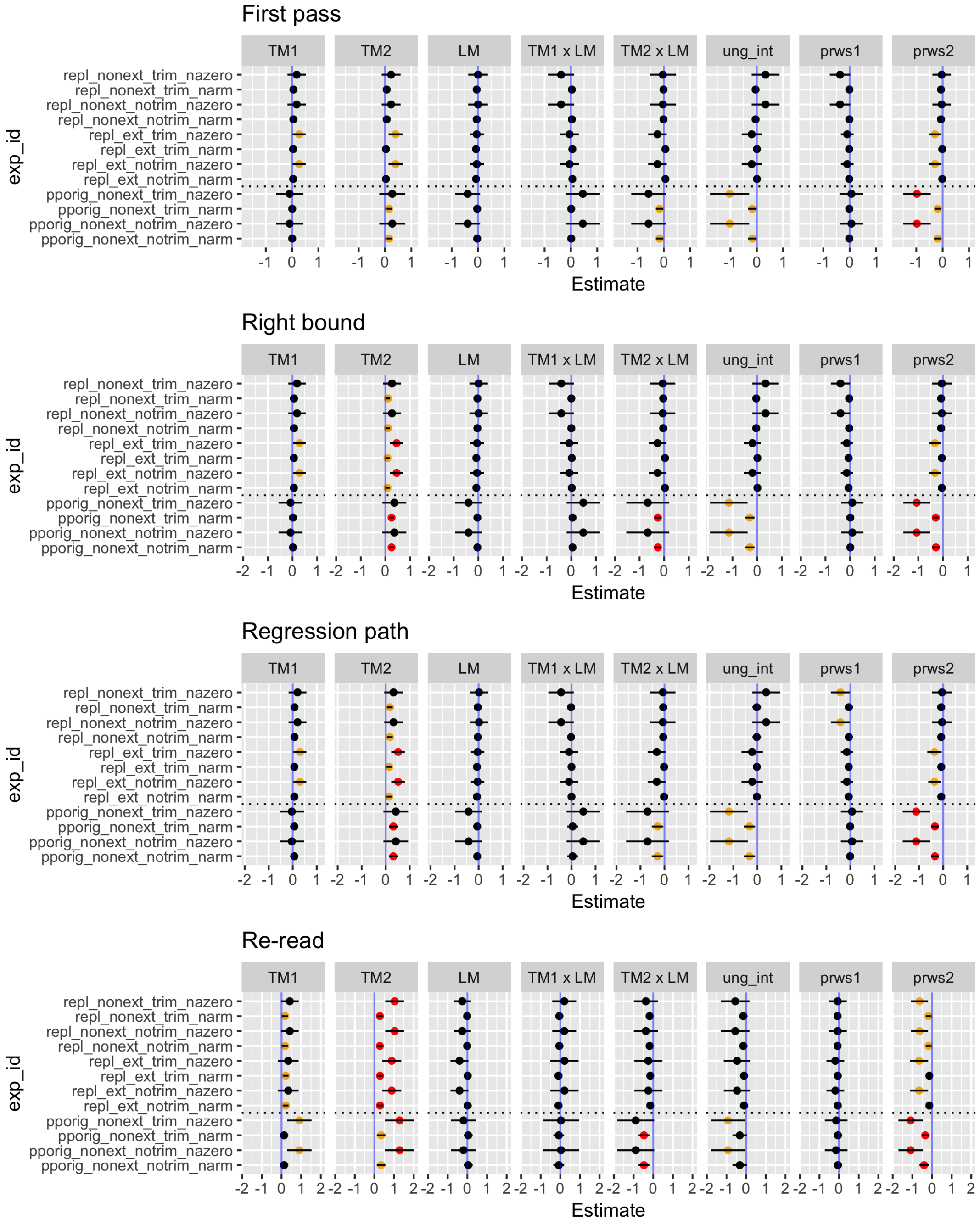

Figure 4.6: Sensitivity analysis: Model coefficients in PP2017 Exp.3 and replication.

Errors bars represent standard deviation of the estimates. See text for further description of columns and row labels. Coefficients for which $|t| \geq 2$ are represented with red markers if they remain significant after a Bonferroni correction (with the corrected $|t| = 3.34$), and with yellow markers, if they don't.

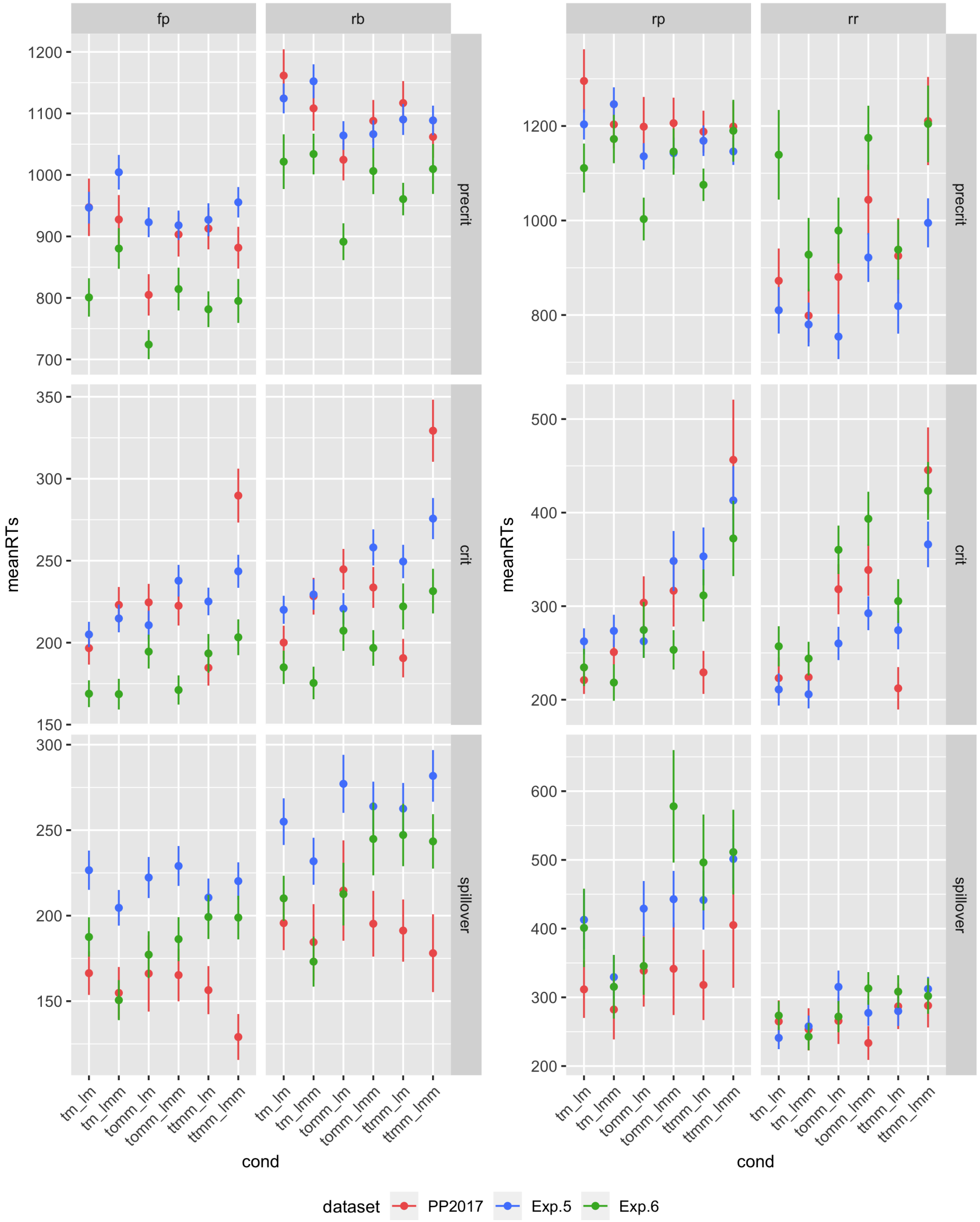

Figure 4.7: Mean RTs in PP2017 (red), my previous replication (blue) and my current replication (green).

Error bars represent standard error of the mean. Conditions: tm/tomm/ttmm - target match/one-mismatch/two-mismatch; lm/lmm - lure match/mismatch.

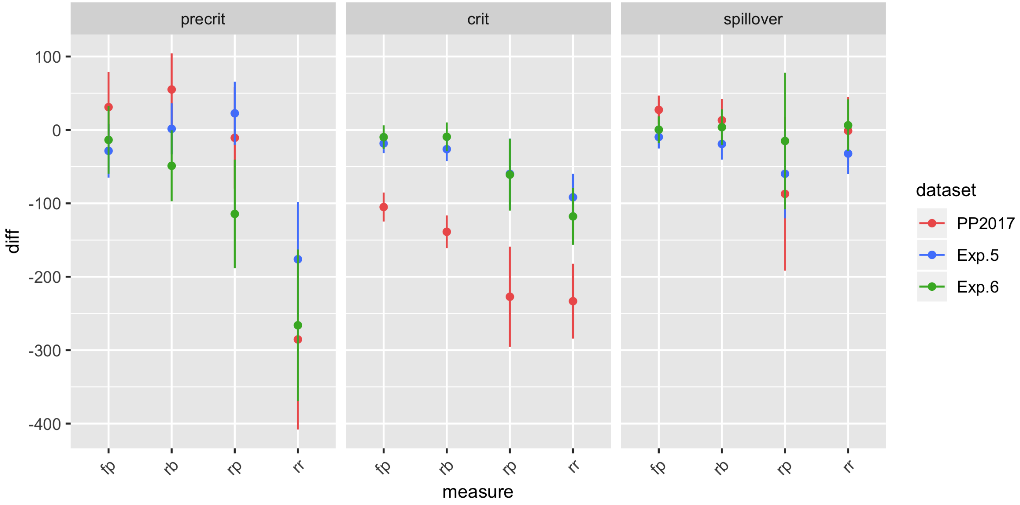

Figure 4.8: Interference effect in TARGET: TWO MISMATCH conditions (lure match - lure mismatch) in PP2017 (red), my previous replication (blue) and my current replication (green).

Error bars represent standard error of the difference of the means.

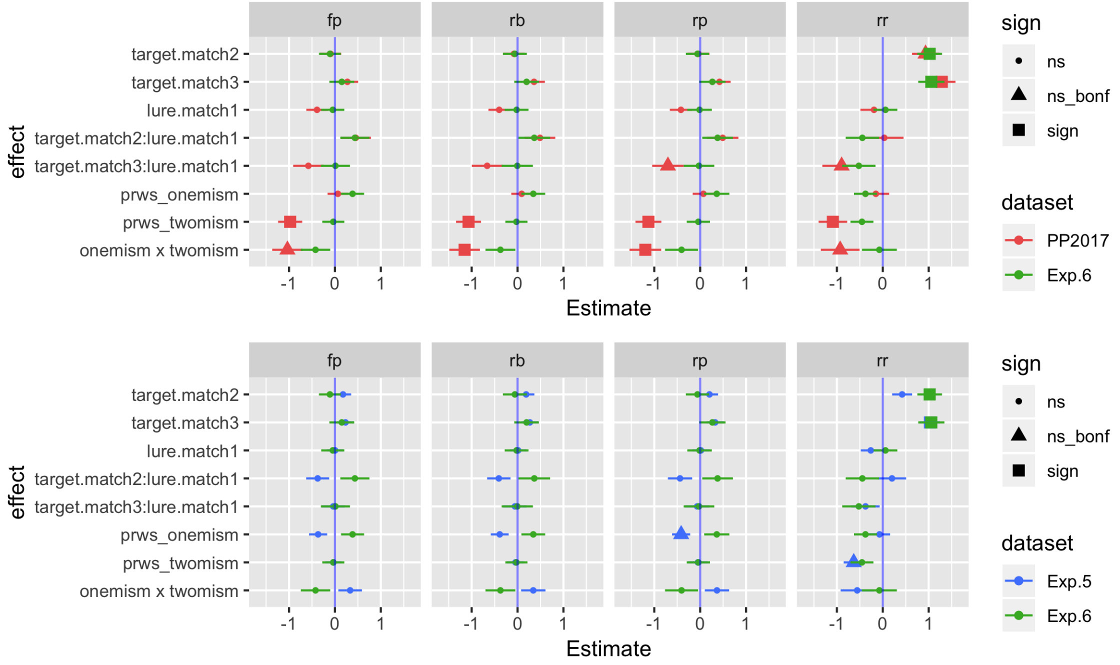

Figure 4.9: Model estimates in PP2017 (red) and my replication (blue).

First five effects come from the full model; next three represent the simple main effects of lure match and the interaction from ungrammatical-only model. The order of the coefficients is the same as in PP2017, to make comparisons across papers easier. Error bars represent standard error of the estimate. Coefficients for which $|t| \geq 2$ are represented with squares if they remain significant after a Bonferroni correction (with the corrected $|t| = 3.34$), and with triangles, if they don't.

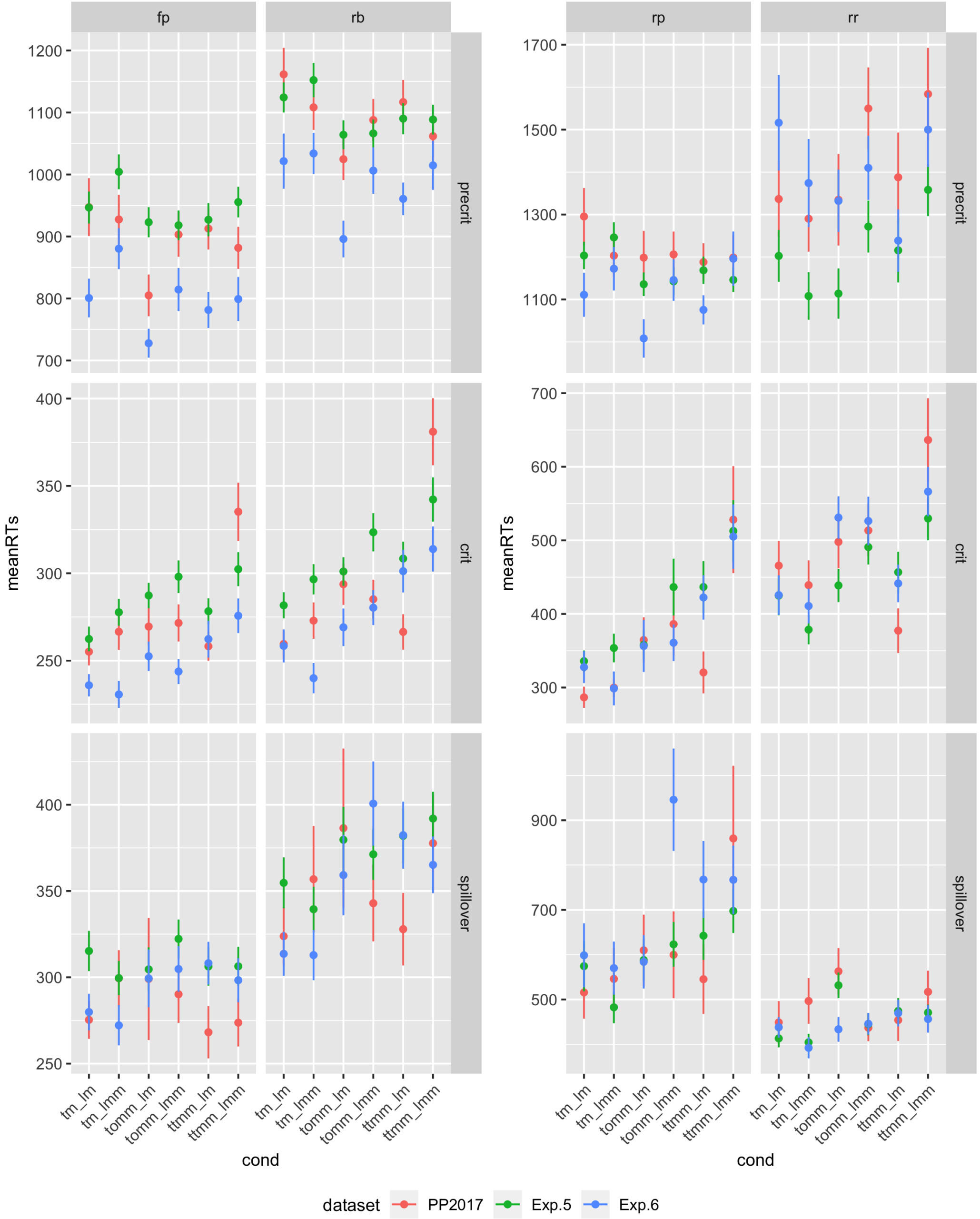

Figure 4.10: Mean RTs in PP2017(red), my previous replication (blue) and my current replication (green).

Exploratory analysis: removing missing values instead of replacing them with zeros. Error bars represent standard error of the mean. Conditions: tm/tomm/ttmm - target match/one-mismatch/two-mismatch; lm/lmm - lure match/mismatch.

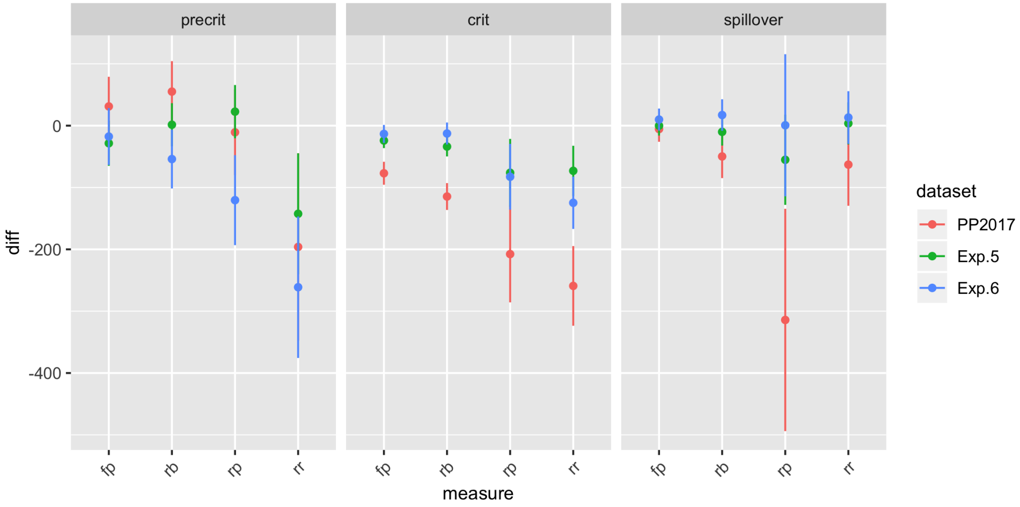

Figure 4.11: Interference effect in TARGET: TWO MISMATCH conditions (lure match - lure mismatch) in PP2017 (red), my previous replication (blue) and my current replication (green).

Exploratory analysis: removing missing values instead of replacing them with zeros. Error bars represent standard error of the difference of the means.

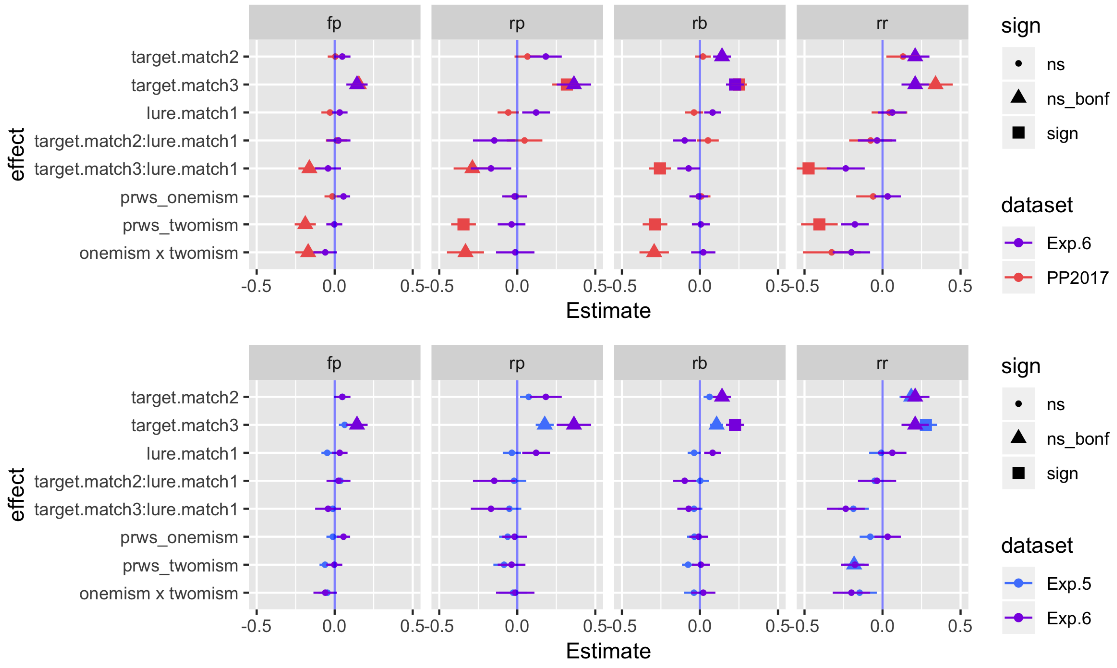

Figure 4.12: Model estimates in PP2017 (red) and my replication (blue). Exploratory analysis: removing missing values instead of replacing them with zeros.

First five effects come from the full model; next three represent the simple main effects of lure match and the interaction from ungrammatical-only model. The order of the coefficients is the same as in PP2017, to make comparisons across papers easier. Error bars represent standard error of the estimate. Coefficients for which $|t| \geq 2$ are represented with squares if they remain significant after a Bonferroni correction (with the corrected $|t| = 3.34$), and with triangles, if they don't.

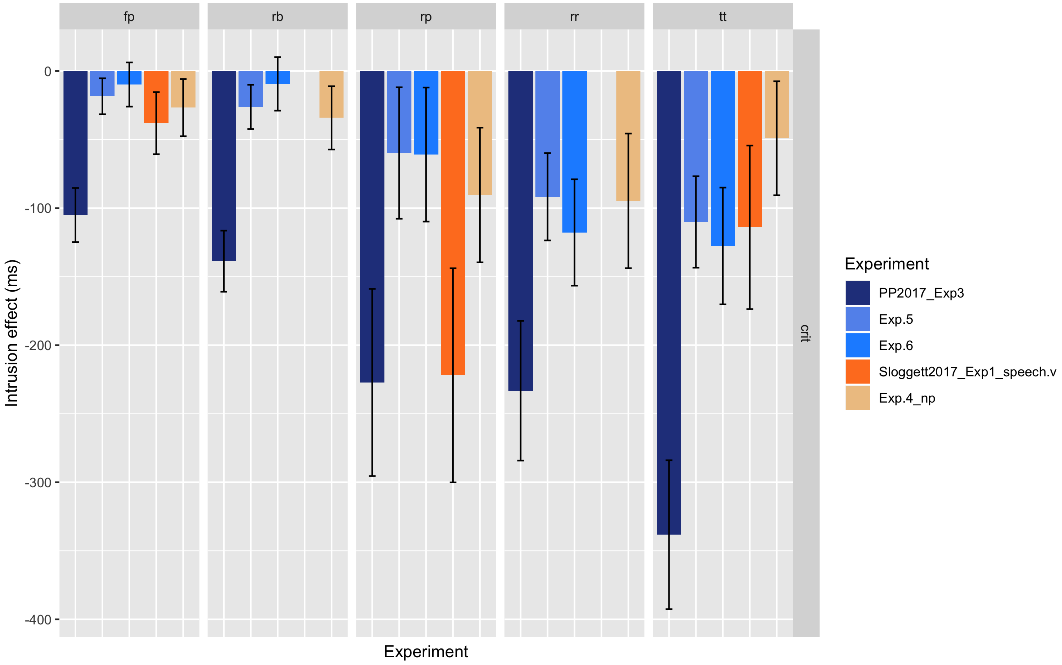

Figure 4.13: Intrusion effect in five studies with most similar design (PP2017 Exp.3, its two replications (this thesis Exp. 5 and 6), S2017 Exp.1c, this thesis Exp.4).

Errors bars represent standard error of the difference of the means. Reaction times are shown for the reflexive region. Columns correspond to reading times measures: fp - first-pass, rb - right bound time, rp - regression path, rr - re-read time, tt - total time.

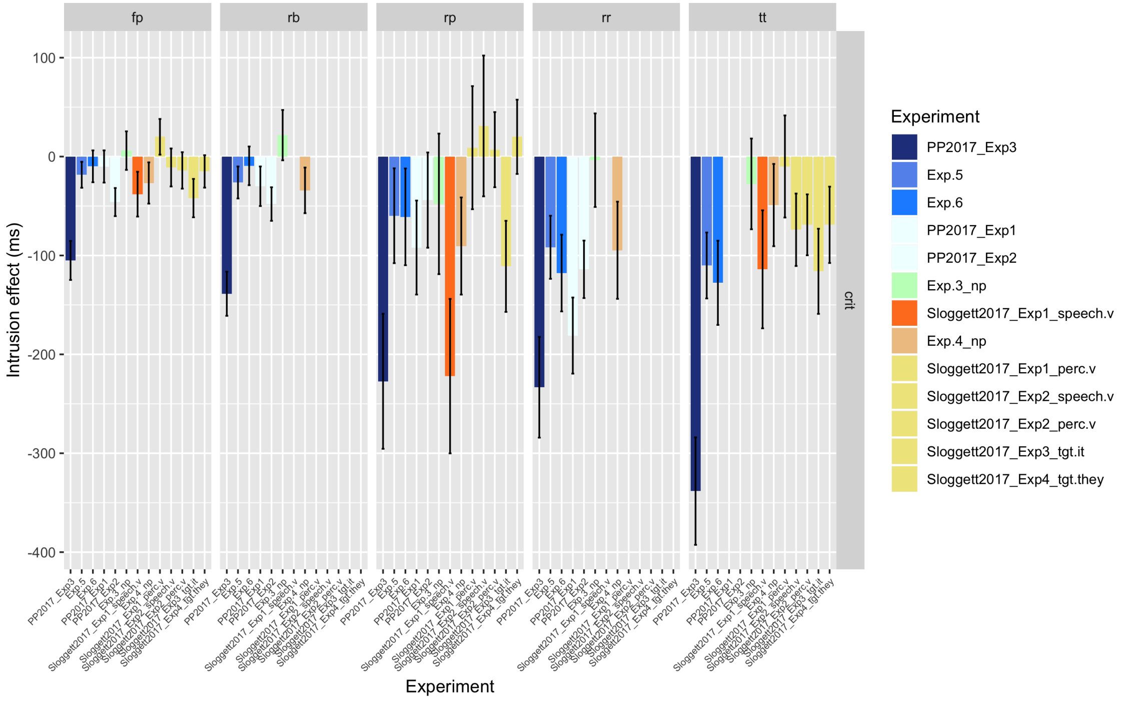

Figure 4.14: Intrusion effect in all studies investigating 2-feature target mismatch configurations (PP2017, S2017, the current study).

Errors bars represent standard error of the difference of the means. Reaction times are shown for the reflexive region. Columns correspond to reading times measures: fp - first-pass, rb - right bound time, rp - regression path, rr - re-read time, tt - total time.

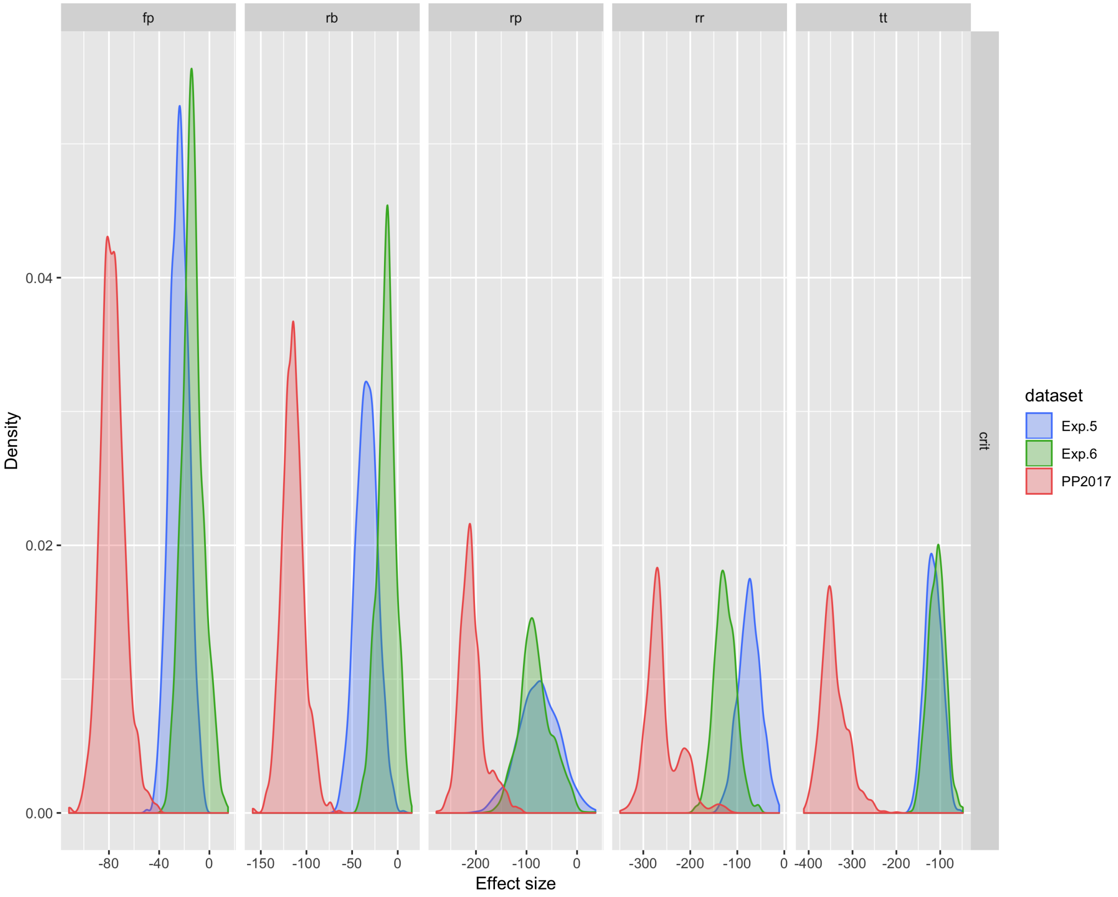

Figure 4.15: Distribution of lure-match effect sizes in 2-feature target mismatch conditions obtained in a bootstrapping simulation.

Sampling 48 participants with replacement. Columns correspond to reading times measures: fp - first-pass, rb - right bound time, rp - regression path, rr - re-read time, tt - total time.

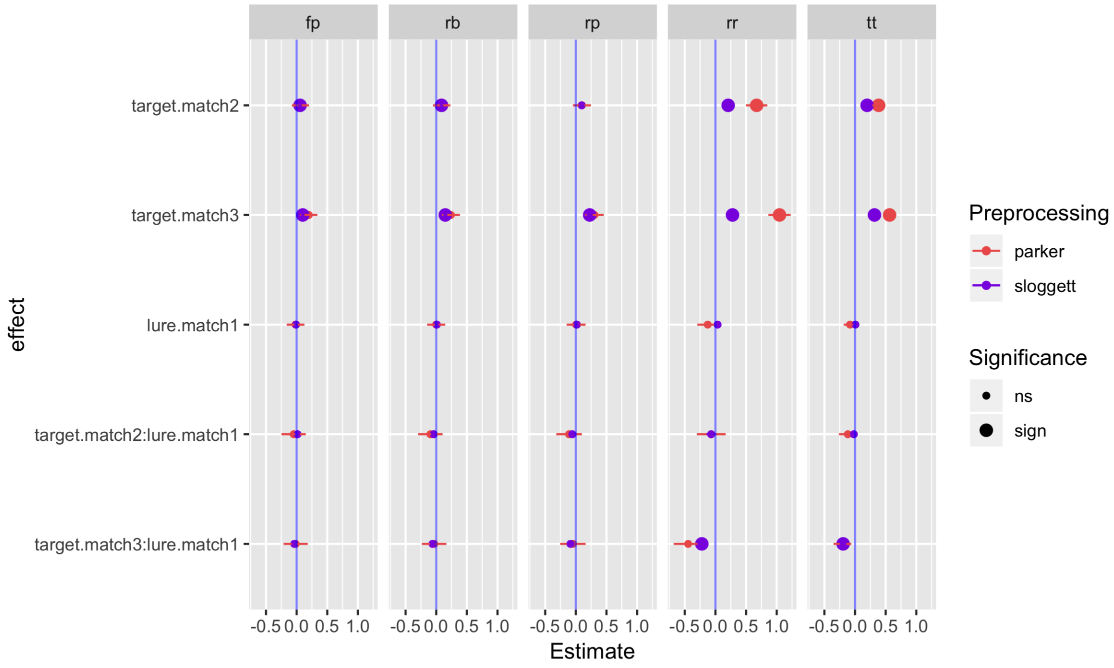

Figure 4.16: Model estimates for the models fit to the pooled data from Exp.5 and 6.

Error bars represent standard error of the estimate. Dot size indicates whether the estimate was statistically significant ($|t| \geq 2$) (big dots) or not (small dots). Colors correspond to the pre-processing variant used ("parker" - critical region comprises the reflexive alone, missing values are replaced with zeros; "sloggett" - critical region includes the reflexives plus three last characters of the preceding word, missing values are removed). Models are fit to the data for the critical region only.

Chapter 5: Conclusions

In this thesis I attempted to answer the question: does structural information constrain real-time sentence processing in a way consistent with grammatical generalizations? Prima facia, the question has already been answered: a wealth of experimental evidence suggests that sometimes structural information appears not to rule out structurally inappropriate antecedents. We know that measures of processing difficulty, such as reading times and acceptability judgments are affected by structurally irrelevant material. Agreement attraction is the primary example, and similar behavior has been reported for NPI licensing and reflexive resolution. However, as I discussed, this evidence can be explained in a multitude of ways, only some of which - in particular, cue-based parsing - do indeed assume that structural information fails to uniquely direct dependency resolution. I have argued that cue-based models should be preferred both on theoretical and empirical grounds: they provide wide empirical coverage and serve as a solid working hypothesis, allowing to easily formulate and test new predictions. In addition, they have been argued to be the only kind of models which explained the absence of ungrammaticality illusions in agreement attraction. If this preference is correct, we have quite good reasons to believe that structurally inappropriate phrases are considered by the human parser.

However, several recent studies have raised doubts about whether cue-based models do indeed constitute our best choice. Two papers attack what had been the empirical cornerstone of the argument for cue-based models: their ability to explain the evidence from grammatical agreement attraction sentences. A meta-analysis and computational modeling by Jäger et al. (2017) suggest that cue-based models do not in fact accommodate the reading times patterns in grammatical sentences with agreement attraction. Hammerly et al. (draft.april.2018) go in a similar direction: they suggest that the absence of ungrammaticality illusions in the data is artifactual, and that the cause of agreement attraction lies in a badly constructed sentence representation which is accessed in full agreement with the grammar. Another paper attacks the cue-based explanation for the reflexive data (Sloggett, 2017): instead of processing error account they suggest a model which explains empirical evidence from reflexive resolution in terms of a fully grammatical resolution strategy - interpreting reflexives as logophors.

In this thesis I was looking for evidence which could help to support or weaken these recent claims. To do so, I focused on recent findings by Parker and Phillips (2017) and Sloggett (2017). Both of them show that reflexive resolution can be affected by structurally inappropriate antecedents. But the explanations they give are diametrically opposite. Parker and Phillips (2017) suggest that the effects they observed are a result of a processing error - faulty memory access which fails to be uniquely constrained by the structural information. In contrast, Sloggett (2017) argues that these effects constitute evidence for logophoric interpretation of the pronoun, which he hypothesizes to be fully grammatical, although underutilized, strategy of anaphora resolution in English. This approach assumes that structural information always uniquely leads the parser to only consider grammatically appropriate antecedents.

Neither of the accounts fully covers available evidence. Thus, the primary goal of this thesis was to investigate the predictions of these two accounts in order to choose between them (and as a consequence - between "structure-defeasible" and "structure-strict" models). I have reported six experiments: four original and two direct replications of the previous studies. One experiment investigated interference from lures embedded inside relative clauses under animate matrix subject - configuration where PP2017 and S2017 accounts make differential predictions. Three experiments investigated the behavior of QP lures, with the aim of better understanding the role of c-command in on-line reflexive resolution. Two experiments were direct replications of Parker and Phillips (2017) Experiment 3, investigating possible roles of extra-linguistic context in lure-match effects, as well as reliability of the previously reported findings. The experiments I report are quite intimately intertwined: most of them both provide a piece of information which is valuable on its own, and help to narrow down the interpretation of the previous results. In the following section, I summarize my main findings and discuss how they could help us choose between the two types of accounts.

5.1 Choosing between PP2017 and S2017

5.1.1 On sub-command binding

The first question I asked in my thesis was: do we have reasons to believe that lure-match effects from non-c-commanding lures reported by PP2017 are evidence for the parser accessing structurally illicit antecedents? If they are, it would constitute a problem for S2017 logophoric account. To capture them without resorting to a processing problem explanation, S2017 suggested that they represent an instance of sub-command binding. This phenomenon occurs in Chinese; an anaphor can be bound by a non-c-commanding animate antecedent if it is contained within a c-commanding inanimate NP. If S2017 explanation were true, the lure-match effects would only appear when the embedding NP was inanimate. The results of Experiment 1 speak against this explanation. I observed lure-match effects regardless of the animacy of the matrix subject. The only significant effect of lure-match was likely driven by the effect within "inanimate" conditions only. However, at the critical region the magnitude of the effect in the "animate" conditions was mostly as big or bigger than in the "inanimate" conditions, and approached the size of the effect from c-commanding NP lures in my other experiments.

Therefore, I tentatively conclude that lure-match effects from non-c-commanding lures are real and are not instances of sub-command binding. Taken in isolation, this finding would support a processing error explanation over the grammatical one, i.e. "structure-defeasible" models over "structure-strict" ones. Such a conclusion would weaken the S2017 logophoric account: if we have evidence that at least in some cases the parser can violate structural constraints (or maybe: cannot help but violate them), what forces it not to do so when resolving the very same dependency in similar configurations? However, as I discuss in the next section, these conclusions would be too strong. Instead, as I will argue, they rather support a model in which the parser faithfully and accurately follows constraints that imperfectly align with the grammar.

5.1.2 On the role of c-command

In my Experiments 2-4 I focused on QP lures, asking the question: are PP2017 correct in claiming that any kind of structural information can be outweighed by competing factors (such as morphology), or can some syntactic features categorically rule out inappropriate antecedents?

Experiment 2 investigated non-c-commanding QP lures. I hypothesized that only if morphology can completely out-weigh structural information would we observe lure-match effects from non-c-commanding QP lures. The results do not appear to support this hypothesis. First of all, I did not find robust statistical support for lure-match effects from QP lures. This is not very telling, since in general my experiments failed to produce statistically significant results in most of the cases. However, QP lures appeared to behave differently than NP lures in numeric patterns. QP lures which matched the reflexive either had no effect or caused a slow-down. In contrast, matching NP lures caused facilitation in both Experiment 1 and 2.

I have established that the lack of facilitation is not due to several possible confounds. It cannot be due to the fact that QPs are not good lures for some reason: c-commanding QP lures which matched the reflexive did cause a speed-up relative to non-matching ones. Second, Experiment 5 and 6 suggest that context factors such as composition of the stimuli set and experimenter's language do not affect the magnitude of the lure-match effect. Third, small effect sizes cannot be the result of the lures being embedded under animate subjects as S2017's sub-command hypothesis would suggest, as shown by Experiment 1 results. Finally, the fact that QP lures provoked inhibitory, rather than facilitatory interference is reminiscent of the numerical trend observed in Kush et al. (2015) Exp.2. There, people were slower to read pronouns like "he" in sentences like "The troop leaders [that no boy/girl scout respected] scolded him" when the embedded QP matched the gender of the pronoun, suggesting that out observations are not spurious.

Experiments 3 and 4 showed that in contrast to non-c-commanding QP lures, the c-commanding ones do cause lure-match effects. The magnitude and the direction of the effects was mostly consistent for NP and QP lures, suggesting that the parser treats the two lure types similarly when they c-command the reflexive. While the magnitude of the effects was smaller than in the original PP2017 study, it was consistent with the two replications I ran in Exp.5 and 6.

This contrast between c-commanding and non-c-commanding lures makes me conclude that c-command information is able to successfully restrict the range of possible antecedents. The fact that the parser appears to treat non-c-commanding NP and QP lures differently also suggests that c-command information is represented in an approximate way, perhaps using Kush et al. (2015) ACCESSIBLE.

5.1.3 On the reliability of lure-match effect

5.1.3.1 Magnitude of the effect

Throughout the thesis I have observed lure-match effects which were going in the right direction numerically (faster RTs in conditions with matching lures) but were consistently smaller than those reported in the original study by Parker and Phillips (2017). Almost in no case did these effects in my experiments received statistical support. I have ruled out several potential explanations for the reduced magnitude. First, I made sure that the experiments worked on a basic level - as expected, people spent more time reading longer words and skipped them less frequently. Grammatical characteristics of the stimuli did affect people's RTs - quite consistently, I observed grammaticality effects when ungrammatical sentences were read slower than grammatical ones. Moreover, lure-match effect were observed for subject-verb agreement, suggesting that I did not accidentally selected participants immune to such effects. Second, I ruled out potential systematic confounds. Controlling for the composition of stimuli set (in particular, the presence of QPs in it) and characteristics of experimenter ((non-)nativeness) did not have an effect on the magnitude of lure-match effects.

It is tempting to conclude that the observed discrepancies between my and previous results are likely due to purely statistical reasons. As I have mentioned in Chapter 4, it is well known that in low-powered studies statistically significant effects are more likely to be overestimates (Gelman and Carlin, 2014; Collaboration, 2015). The original studies by PP2017 that I tried to replicate had sample sizes of 24 and 30 people, which is rather low even by psycholinguistics standards. Even my studies which had doubled sample sizes still failed to provide robust statistical evidence. Only in the exploratory analysis on the pooled data, comprising the data from 81 participants, did some lure-match effects reach significance. On the other hand, S2017's experiment may speak against such interpretation: he used sample sizes comparable to mine and consistently found statistical support for lure-match effects. While in a low-powered setting this could be attributed to chance alone, the consistency with which lure-match effects are supported by statistics in S2017 on the one hand and in my experiments on the other makes me think that additional factors may be at play. I could think of several directions which could be investigated, although I do not have strong hypotheses about how the differences in these parameters would have affected the inferences.

One potentially important factor is that both PP2017 and S2017[1] sentences contained gaps or pronominal elements in the spillover region: e.g. "The flustered nurse complained that the elderly veterans considered herself _ to be very attractive and made suggestive comments" or "The broken zipper that the skilled tailors tried to fix pinched themselves repeatedly on **their** fingers.". It is possible that people could preview the spillover regions with their parafoveal vision, or that the interpretation of the reflexive was still on-going while their gaze had already shifted on the

[1] To the degree that I can see: I only had access to the materials they used in their Experiments 1 and 3

spillover. In these cases the presence of an element depending on the reflexive's interpretation could affect the processing. For example, null or overt pronouns might have provoked additional retrievals of the lures and artificially increased their accessibility in S2017's experiments. Alternatively, given that the interpetation of the reflexive would influence the interpretation of some other element, people might engage in additional checks of whether the antecedent they chose is permissible by the grammar. In this case, the intrusion effects might be smaller or absent altogether.

The first alternative might receive some support from the data. In my Experiment 2 12 out of 24 items had an empty element in the spillover, while in the in Parker and Phillips (2017, Exp.2), using similar stimuli, only 4 out 36 items did. In my Exp.4 we borrowed a subset of NP items from Sloggett (2017, Exp.1). In that experiment, 14 out of 48 items (i.e. roughly a quarter) had an empty element in the spillover. It turns out that I borrowed 9 of such items into my experiment. Since I had fewer NP stimuli (24), those items accounted for almost half of them. It could be that the reduction of the intrusion effect in my case was strong enough to make the substantial evidence for it disappear, while Sloggett (2017, Exp.1) was affected less due to a higher proportion of stimuli without an empty element in the spillover. It would be trivial to check whether this is the case if I had the original data from Sloggett (2017): in that case, I could look at whether there are differences in the effect size between the stimuli I did and did not select. This account would fail to explain the absence of evidence for intrusion in Exp.3: in constructing the stimuli for it, I made sure that the spillover did not contain any empty elements. However, these results may be independently explained by the fact that I used a mix of communication and other types of verbs in the materials for Exp.3.

Another possibility could be that the gender stereotypes on the nouns I used were weaker than on the nouns in PP2017 and S2017 experiments. It could lead to them being worse lures and thus producing on average smaller lure-match effects. While possible in principle, I think this explanation is a non-starter: I failed to observe statistically robust effects in two direct replications of PP2017, thus, the strength of the gender stereotype alone can not be the issue. In other experiments I either directly borrowed a subset of PP2017 and S2017 materials, or was heavily influenced by their materials in creating mine. Therefore, I think that the strength of gender stereotypes is not an issue here.

It is possible that subtle differences in the experimental procedure might have changed the results. As Hammerly et al. (draft.april.2018) show, the change in wording from “two thirds of the items are ungrammatical” to “the majority of the items are ungrammatical” apparently was enough to change the observed agreement attraction effects. Such biases do not have to be restricted to the instructions and may involve other factors: how often do participants take breaks, how often is recalibration performed, how does the experimenter interacts with the participants before the experiment starts etc., and may be hard to catch.

Overall, I conclude that while low statistical power may be an issue, unrecognized biases have likely affected my results. It is hard to make stronger conclusions in the absence of formal or simulation-based power analysis, but I suggest that since lure-match effects may apparently be hard to detect in smaller samples, sample sizes bigger than 40-50 participants are desirable in order to find reliable evidence for interference when looking at reflexives resolution.

5.1.3.2 Timing of the effect

The time course of the lure-match effects appears to be variable across the studies. Some experiments (Parker and Phillips (2017) Exp.1,3; Sloggett (2017) Exp.1b and 4b) show detectable lure-match effects as early as in first-pass and regression path at the critical region. Some others do not show reliable effects until later - re-read and total times at the critical region (Parker and Phillips (2017) Exp.2, Sloggett (2017) Exp.2b), or even at spillover in regression path (Sloggett, 2017, Exp.3b). My experiments rather fall in the second group: numerically, the lure-match effects got biggest in re-read or total times at the critical region, and the only instance of a significant main effect of LURE MATCH (Exp.1) was detected in the total times at spillover.

What is the reason behind such variability? The first possibility is that it is due to some systematic differences in the stimuli set. However, the evidence does not support this hypothesis: in multiple cases the exact same or very similar stimuli set produced variable results. The best example is PP2017 Exp.3, where lure-match effects were observed as early as in first pass at the critical region, and my two replications (Exp.5 and 6) where the magnitude of the effects was biggest in late eye-tracking measures. These studies used the exact same set of stimuli. PP2017 Exp.2 and my replication in Exp.1 had similar variability, with Exp.1 using a subset of the original stimuli. Finally, S2017 observed such variability in his Exp.3b and

4b, despite using same critical stimuli.

It is possible that individual characteristics of the participants lead to the variability in the timing of the lure-match effects. Cunnings and Felser (2013) investigated the effect of working memory on lure-match effects in reflexive resolution. They only considered 1-target mismatch configurations, and had a rather small sample sizes (32 participants by experiment, in each case split into two groups depending on their working memory capacity), so the results may not be very reliable. For what it's worth, they report differential effects depending on the working memory span: low-span group was faster to detect the mismatch with the target. The effect first became significant in first fixation duration at the reflexive for the low-span group and in re-read times for the high-span group[2].

Finally, one could think that the lure-match effects observed in early measures are artifactual. As I have just discussed, for all experiments where lure-match effects were observed in early measures there is a related experiment where they were not. On the other hand, in the late measures lure-match effects are observed more consistently. Such a pattern may indicate that lure-match effect are, in fact, a reflection of a repair strategy. If it is the case, that would constitute evidence against PP2017 account: if a repair is initiated, the original resolution attempt must have failed; it would only fail, if the parser correctly accessed the target; therefore we have to assume that structural information is able to strictly rule out

[2] The authors suggest that this may reflect the application of least effort strategy from low-span readers, in which they simply select the closest noun as the antecedent instead of fully processing the dependency.

ungrammatical antecedents at least during initial resolution attempts (cf. with "the defeasible filter" hypothesis by Sturt (2003)).

Several arguments can be made against such possibility. As Vasishth et al. (2012) notice, effects which become apparent in late measures are not necessarily stemming from late processes; instead, they might stem from a process which starts early but unfolds slowly enough to only become apparent later. Thus, simply observing some effect emerging in late measures does not allow to uniquely attribute it to repair processes. Further, S2017 provides three arguments against repair account. First, he notes that some experiments do show early lure-match effects. Second, S2017 discusses the results of his interpretation study, in which people appeared to choose non-local interpretation as frequently as in 30% of the cases even when the sentences were grammatical (i.e. the target fully matched the reflexive in features). And third, he notices that reflexives with clearly non-local interpretations are sometimes observed in natural production[3].

Overall, the tendency of the lure-match effects to be more prominent in late eye-tracking measures is interesting, but it lacks sufficient evidence for firmer conclusions. While it could potentially be problematic for PP2017 account, further studies are necessary to decide whether the tendency does, indeed, exist, and if so,

[3] While these arguments would help the claim I am making by ruling out the repair explanation of lure-match effects, I tend to be somewhat skeptical about them. As I have just noted, lure-match effects in early measures appear to be less reliable than late effects and might be artifactual. I discuss my concerns about the interpretation of these results in section 5.2. Finally, the third argument is potentially the strongest one, but it assumes that lure-match effects in comprehension and logophoric use of reflexives in production rely on the same mechanisms, which is not given.

what are its causes.

5.1.4 Choosing between the two accounts

Based on the evidence I have discussed above, I suggest that neither PP2017 nor S2017 account is completely supported. On the one hand, my findings from Experiment 1 suggest that sub-command binding may not be a good explanation for lure-match effects from non-c-commanding lures. On the other hand, the results from Experiments 2-4 suggest that the parser does faithfully follow an approximate version of c-command to categorically rule out inaccessible QP antecedents.

This interpretation goes against grammatical models, such as S2017: the parser does appear to perform operations which would not be licensed by the grammar. However, it also goes against "structure-defeasible" memory models such as PP2017: structural information appears to accurately guide the retrieval at least in some cases. I suggest that a memory type of account is still a better explanation for the available data.

Two main empirical facts supporting this conclusion are a) the presence of lure-match effects from non-c-commanding lures embedded in an animate subject; b) differential lure-match effects from non-c-commanding QP and NP lures in the same configurations. S2017 sub-command hypothesis would not be able to explain them, since it only predict interference from the lures embedded inside an RC with an inanimate head. The same goes for the logophoric explanation of the sub-command facts by Charnavel and Huang (2018).

But let us imagine that one of those explanation could account for lure-match effects from RC-lures, perhaps because in English sub-command binding is possible regardless of the animacy of the (head of the) embedding NP. Even in this case we would not be able to explain why non-c-commanding QP lures do not cause lure-match effects: as we showed in Chapter 2, in Chinese sub-commanding QPs can serve as antecedents for “ziji”. We could build another layer of speculation: suppose, in English QPs are not good as non-local antecedents for reflexives, as Postal (2006) suggests. But now we would not be able to explain why c-commanding QPs *do* provoke lure-match effects. Let us build yet another counterfactual. Suppose, we suggest that the null operator mediating logophoric interpretation, OP_{log}, is located in the left periphery of all clauses, and not only complement ones, as S2017 suggests. We think, it would still fail to capture interference from the lures embedded under an animate noun: arguably, the embedding noun is more prominent in the discourse and is more likely to control OP_{log} reference. Thus, even if OP_{log} is occasionally misretrieved, as S2017 suggests, the resolution would still pick the matrix subject, and not the embedded lure, as the antecedent. The OP_{log} account would also not be able to capture the contrast between c-commanding and non-c-commanding QP lures. Although S2017 is not very specific about the mechanism which is used to determine the reference of OP_{log}, he speculates that the mechanism should be similar to the one used by regular pronouns looking for their antecedents in the discourse Sloggett (2017, p.140). This remark suggests that OP_{log} resolution relies on co-reference. Since QPs may only bind, they would not be able to serve as antecedents for OP_{log}. This conclusion would be in line with Postal (2006) observations.

Thus, it appears that it is hard to explain my findings in terms of some grammatical strategy, and a memory models might be preferred. But we would have to choose a model which is only partially "structure-defeasible": instead of assuming that all structural information is weighted uniformly and can be out-competed, we would have to assume that the parser differentiates between different kinds of structural information, with some being able to act as gating features.

How could this suggestion be implemented in a memory model? Under the usual assumptions cue-based models do not make it easy for a cue to categorically block access to some memory item: even if a feature on an item mismatches the retrieval cues, other features of the item may match them well enough to still make the item retrievable. Kush et al. (2015) discuss two possibilities of how a feature could act as a gating feature, categorically restricting access only to matching items. One could weigh the gating feature extremely highly, so that the items that match it receive such a high activation boost that no other item can compete with that. Alternatively, one can change the way cues are combined to determine an item's activation. Most often it is assumed that the activation is determined by a weighted sum of activation values provided by the matching cues, with weights corresponding to their importance, and activation values - the strength of association between the item and the cue in question. However, one could assume that a weighted product is used instead. In this case, if a single feature mismatches the set of retrieval cues, it would make the activation of the item very small or kill it completely, thus preventing its retrieval in the majority of the cases[4]. Kush et al. (2015) mention

[4] Presumably, such an item could still be retrieved occasionally, if the random noise happens to

that it is impossible to tell these possibilities apart from their data.

I think that the data in the thesis may favor the first hypothesis: super-weighting of the gating cues. In order to explain the fact that mismatch on locality information may be out-weighed by other factors, one would have to conclude that locality cue(s) contribute to item's activation in an additive fashion. Now, ACCESSIBLE cue has to be able to categorically rule some items out from consideration. In principle it is possible that it combines multiplicatively with other cues, in a manner described by Eq. 5.1, where GCA_i[5] is the binary indicator of whether all gating cues[6] allow the $i-th$ item to be retrieved (i.e. whether the $i-th$ item matches these cues). However, this would create a situation where different cues rely on different combinatorics rules, and without strong evidence for such mixed schemas, it seems preferable to keep cue combinatorics rules uniform within a model. Since we have to assume an additive schema to capture locality facts, ACCESSIBLE would

push its activation high enough to outcompete other items in memory.

[5] Standing for "gating cues activation".

[6] It seems possible that if the system relies on gating cues at all, there is more than one such cue. In this case, we need to know how these cues combine. The schema I suggest assumes that if even a single gating cue fails to match, the item is ruled out. This could be achieved if zero activation is spread from a mismatching gating cue, and the values of all gating cues are multiplied to produce the final GCA_i value. More involved combinatorics schemas could in principled be produced, so that, say, a item has two mismatch two gating cues together in order to be ruled out. Also notice that when several items *do* match a gating cue, the activation it spreads to these items will be less than 1 due to fan effect. If the system uses multiple gating cues at once, and it happens so that each gating cue is matched by multiple items, the items' activation values can be driven very low, since the product of several values smaller than 1 may indeed be small.

have to combine additively as well. In this case, super-weighting is the only way of making it act as the gating feature. This hypothesis could in principle be tested empirically: if super-weighting is, indeed, the underlying mechanism, we might find ways to boost the activations from other features highly enough to out-weigh even the super-weighted cues. E.g. we might try to make the match with the true target even stronger, and/or put the lure in a linguistically prominent position.

$$A_i = B_i + GCA_i \times (\sum_j w_j S_{ji}) \tag{5.1}$$

Concluding, I have to note that several empirical points may be taken as speaking against my suggested account, although very speculatively. First, I observed the tendency for lure-match effects to be more readily detectable in late eye-tracking measure. So far it is simply a suggestive observation, but if confirmed experimentally, it could indicate that lure-match effects stem from repair processes. As discussed, this would imply that at least at some point structural information does categorically constrain parser's actions. Second, it appears that the magnitude of lure-match effects from NP lures is bigger when the lure c-commands the reflexive. This is true both in PP2017 and in my replications: despite that the effect sizes we observe are overall reduced compared to PP2017, the difference between c-commanding and non-c-commanding lures holds. This may suggest that non-c-commanding lures are less accessible in general, with non-c-commanding QP lures being completely inaccessible and c-commanding NP lures being relatively inaccessible. If this intuition is correct, it will weaken my conclusions about the use of ACCESSIBLE: accounts relying on this feature do not predict any access difficulty for + ACCESSIBLE phrases regardless of their c-command relations. Third, in both Experiment 1 and Experiment 2 I did observe an illusion of ungrammaticality in agreement attraction stimuli. This is problematic for cue-based models, of which PP2017 account is an instance. This is also in line with Hammerly et al. (draft.april.2018), who suggest a "structure-strong" account for agreement attraction. As I have discussed in Chapter 1, so far their account is not fully developed and, for example, is not easily extendable to eye-tracking data. But such data points support their model and if it is better developed, it could constitute a strong competitor to cue-based modes.

Finally, I remind that my conclusions necessarily remain tentative, being based almost entirely on numerical patterns of RTs.

5.2 Future directions

Given that most of my conclusions are based on numerical patterns or RTs, the most important next step is to conduct adequately powered confirmatory experiments. The numerical estimates of the lure-match effect magnitude I obtained in this study[7] could be used for power calculations. The most important thing is to further investigate the interference from non-c-commanding lures - this would help to improve the evidence quality of evidence for the only pattern not covered by S2017 account.

[7] As well as other estimates, such as random effects and variance-covariance matrices for participants and items

Another possible line of investigation is to look for lure-match effects from lures in a possessor position, e.g. “The librarian’s brothers praised herself during the meeting”. S2017 account does not predict any interference: the subject NP is not embedded in a complement clause, so there is no null operator which could mediate the dependency between “librarian” and “herself”. Even if such operator were present, we do not think “librarian” is prominent enough to control its reference. Sub-command explanation would not work either: the configuration is of the right type, but the embedding NP is animate and thus should block sub-command binding. PP2017 account, on the other hand, would predict lure-match effects from “librarian” unless c-command information is accurately used to guide the retrieval. Thus, in the suggested experiment the presence of an effect would speak for PP2017 model, but the absence of the effect would be hard to interpret since it could be compatible with both S2017 and PP2017 accounts.

A third possible line of investigation would be to consider whether the mechanism behind lure-match effects in reflexive resolution is purely formal (e.g. just checking the morphological features) or whether it affects interpretation. PP2017 account does not make strong predictions either way. On the other hand, S2017 account necessarily predicts that interpretation is affected, since S2017 argues for a grammatical option of reflexive resolution. S2017 supports this point of view with a results of interpretation study: he shows that when presented with sentences like ”The librarian said that the schoolgirl misinterpreted herself...” and asked a question like ”Who was misrepresented at the meeting? – The schoolgirl / The librarian”, people choose the answer compatible with non-local binding of the reflexive (”librarian" in this example) as frequently as 30% of the time. S2017 uses this finding to support his argument about the logophoric nature of interference effects in reflexive resolution: if people can choose the non-local option even in grammatical sentences, it this choice has to be a grammatical possibility.

We have doubts about this claim for several reasons. First, question answers do not necessarily stem from the same processes which underlie reflexive resolution during sentences processing. Instead, they may come from some extra-grammatical strategy (e.g. if upon seeing the question, people choose the answer not based upon how the reference was resolved but based on some other factors - familiarity, the answer being a more sensible choice etc.). Second, according to the Sloggett's hypothesis, non-local reference is mediated by a silent logophoric operator. It is hypothesized not to carry any ϕ-features. Thus, if the overt local noun matches the reflexive in features, it will be a better match on at least two features - gender and number (and potentially person and/or animacy). If S2017 results do indeed indicate that people retrieve the silent operator, it would mean that even a target which is the best match along multiple dimensions can be ignored during memory access as frequently as 30% of the time. It seems odd to assume that memory access is so inefficient: if it performs that poorly in perfect conditions, what would it performance be in real-life situation with potentially less certainty about what the correct outcome is? Third, in all of his acceptability judgment experiments people give rather low ratings[8] for sentences with target mismatch - even when the lure

[8] To give an idea, target mismatch conditions received ratings in the range of 3-4, while target match conditions were all above 5 on a Likert scale from 1 to 7.

matches the reflexive the ratings are lower than for grammatical sentences. This fact is hard to explain if, as S2017 argues, lure-match effects arise from a completely grammatical anaphora resolution strategy[9].

I suggest several ways of checking whether S2017's interpretation results do indeed reflect the outcome of reflexive resolution during sentence reading or whether they are due to some extragrammatical strategy. Perhaps, the best option would be to use an eye-tracking study with interpretation questions. After collecting the data, one could separately analyze the conditions depending on the antecedent choice people made. If S2017 is correct, and RTs patterns and interpretations stem from the same mechanism, we will only observe facilitatory interference in conditions where people chose the non-local antecedent. Second, one could turn to a visual world paradigm: if lures are indeed chosen as antecedents in real-time, we should observe increased amount of looks for the picture representing the lure. Finally, one could make people's task as easy as possible[10]. In S2017 experiment, people had to read the sentence in a self-paced fashion before answering the question. I.e. they never saw the sentence in its entirety, and had limited time to process it. If S2017 is correct, and non-local interpretations do stem from a grammatical interpretation option, these factors should not matter a lot and people will produce roughly the same interpretation patterns in easier conditions, e.g. when having unlimited time to look at the sentence. Otherwise, we would see a reduction in the number of non-local interpretations.

[9] I would like to thank Colin Phillips for this observation.

[10] I would like to thank Colin Phillips for this idea.

5.3 Conclusion

In this thesis, I attempted to distinguish between two types of real-time long-distance dependency resolution models - those which assume that structural information categorically rules out structurally inappropriate elements and those which do not. Specifically, I contrasted two recent instantiations of these models explaining lure-match effects in reflexives resolution. Parker and Phillips (2017) attributed those effects to a processing error, while Sloggett (2017) attributed them to a grammatical strategy. The evidence I obtained in six experiments suggested that neither account is entirely correct. Contrary to what S2017 claims there seem to be cases (namely, lure match effects from non-c-commanding lures) which are best explained in terms of a processing error. Contrary to what PP2017 say, structural information appears to have flavors - some kinds of it (e.g. locality), may not always succeed in categorically ruling out inappropriate antecedent, while others (e.g. some approximation of c-command), may do so. I suggested that the best explanation for this evidence is a mixed memory model, in which some structural information (e.g. locality) is weighted highly but can be outcompeted relatively easily, and other structural cues (e.g. c-command) is super-weighted so that it is hard or impossible for other factors to out-weigh it.

These studies have also shown that lure-match effects may not be as reliable and big as previous studies suggest. In fact, most of my conclusions are based on numerical patterns in the data. I failed to find statistical support for lure-match effects, despite having sample sizes comparable to or bigger than in the previous studies. I do not think this indicates that the effects are bogus - they have been replicated multiple times, the numerical patterns I observe in my studies are going in the expected direction, and an exploratory analysis of pooled data from about 80 participants suggests that such sample size is enough for the statistics to support even the reduced effects. Rather, these results may indicate that the lack of statistically significant effects is due to low statistical power or unidentified subtle biases in experimental materials and procedures. I stress the necessity to conduct a priori power analyses; the data that I have collected can be used to set expectations for such analyses. I would certainly welcome higher-powered replications of my own studies which could verify the conclusions I make.

Appendices

Appendix A: Mean reaction times in PP2017and replication (PP2017analysis)

Notice that standard errors reported here for PP2017data are slightly smaller than those reported in the original paper. The difference comes from the fact that we calculated them taking into account subject variability, by using the method suggested in Cousineau (2005) and Morey (2008)[1]. In contrast, as far as we can say, PP2017used the standard formula for calculating SEs: $\frac{sd}{\sqrt{n}}$.

[1] The code implementing this procedure was taken from: `http://www.cookbook-r.com/Graphs/Plotting_means_and_error_bars_(ggplot2)/`.

Mean RTs in PP2017

			Region	
	Cond	precrit	crit	spillover
fp	tm, lm	947.4 (46.68)	196.7 (10.05)	166.4 (12.74)
	tm, lmm	927.5 (39.73)	223 (10.93)	154.8 (15.16)
	tomm, lm	805 (33.65)	224.6 (11.29)	166.2 (22.25)
	tomm, lmm	903.1 (35.85)	222.5 (12.08)	165.2 (15.33)
	ttmm, lm	912.8 (33.82)	184.7 (10.91)	156.4 (14.06)
	ttmm, lmm	881.7 (33.92)	289.7 (16.44)	129.1 (13.42)
rb	tm, lm	1162 (42.82)	200.1 (10.3)	195.7 (15.84)
	tm, lmm	1108 (36.39)	228.3 (11.22)	184.5 (22.23)
	tomm, lm	1025 (33.64)	244.8 (12.42)	214.7 (29.35)
	tomm, lmm	1088 (33.93)	233.7 (12.46)	195.3 (19.23)
	ttmm, lm	1117 (35.72)	190.6 (11.76)	191.3 (18.19)
	ttmm, lmm	1062 (33.93)	329.3 (18.94)	178.1 (22.75)
rp	tm, lm	1295 (67.08)	221 (14.8)	311.6 (41.56)
	tm, lmm	1203 (50.55)	250.9 (13.29)	282.3 (43.47)
	tomm, lm	1199 (62.91)	303.8 (28.14)	338.7 (52.22)
	tomm, lmm	1206 (53.94)	316.5 (38.28)	341.5 (67.3)
	ttmm, lm	1188 (44.37)	229.2 (22.99)	318.1 (51.11)
	ttmm, lmm	1199 (53.05)	456.5 (64.29)	405.2 (91.27)
rr	tm, lm	872.5 (68.21)	223.1 (22.04)	265 (30.71)

tm, lmm	798.7 (56.6)	224 (22.17)	253.4 (30.72)
tomm, lm	880.6 (78.36)	318.1 (26.77)	265.8 (33.63)
tomm, lmm	1044 (74.39)	338.7 (27.66)	233.6 (24.64)
ttmm, lm	925.1 (79.73)	212.2 (22.67)	286.9 (33.07)
ttmm, lmm	1210 (93.32)	445.5 (45.64)	288 (31.86)

Table A.1: Mean RTs in PP2017

Reading measures: fp - first pass, rb - right-bound, rp - regression path, rr - re-read. Conditions: tm/tomm/ttmm - target match/one-mismatch/two-mismatch; lm/lmm - lure match/mismatch.Values in the parenthesis represent standard error of the mean, corrected for within-subjects variability(Cousineau, 2005; Morey, 2008)

Mean RTs in the replication of PP2017

		Region		
	Cond	precrit	crit	spillover
fp	tm, lm	946.7 (25.82)	205 (7.73)	226.6 (11.48)
	tm, lmm	1004 (28.14)	214.8 (8.47)	204.6 (10.43)
	tomm, lm	923.1 (24.3)	210.7 (8.8)	222.3 (12.01)
	tomm, lmm	918.1 (23.91)	237.8 (9.67)	229.1 (11.66)
	ttmm, lm	927.1 (26.78)	225.1 (8.43)	210.6 (11.13)
	ttmm, lmm	955.5 (24.76)	243.5 (10.07)	220.2 (10.98)
rb	tm, lm	1124 (24.23)	220.1 (8.56)	255 (13.65)
	tm, lmm	1152 (27.64)	229.4 (9.3)	231.8 (13.74)
	tomm, lm	1064 (23.33)	220.7 (9.5)	277.1 (16.97)
	tomm, lmm	1066 (22.51)	258.1 (11.03)	263.9 (14.48)
	ttmm, lm	1090 (25.18)	249.5 (10.2)	262.6 (15.02)
	ttmm, lmm	1089 (24.01)	275.7 (12.56)	281.7 (15.1)
rp	tm, lm	1203 (32.16)	262.4 (13.95)	412.8 (43.15)
	tm, lmm	1246 (35.75)	273.6 (17.24)	329.5 (32.07)
	tomm, lm	1136 (27.78)	262.5 (17.34)	429.1 (40.17)
	tomm, lmm	1143 (32.61)	348.3 (32.09)	442.8 (41.18)
	ttmm, lm	1169 (32.19)	353.2 (30.91)	441.6 (43.06)
	ttmm, lmm	1146 (28.6)	413 (36.68)	501.3 (42.68)
rr	tm, lm	810.1 (49.34)	211 (17.24)	241 (16.41)

tm, lmm	779.9 (46.35)	205.8 (15.09)	257.9 (15.62)
tomm, lm	754.3 (47.66)	260.1 (17.78)	315.2 (23.77)
tomm, lmm	921.7 (51.72)	292.4 (18.11)	277.3 (18.65)
ttmm, lm	818.9 (58.33)	274.4 (20.47)	280 (21.7)
ttmm, lmm	995 (51.92)	366.1 (24.45)	312.2 (17.63)

Table A.2: Mean RTs in the replication of PP2017

Reading measures: fp - first pass, rb - right-bound, rp - regression path, rr - re-read. Conditions: tm/tomm/ttmm - target match/one-mismatch/two-mismatch; lm/lmm - lure match/mismatch.Values in the parenthesis represent standard error of the mean, corrected for within-subjects variability(Cousineau, 2005; Morey, 2008)

Appendix B: Mean RTs

			Region	
Measure	Target match	Lure match	Critical	Spillover
fp	match	match	318 (20)	295 (14)
		mismatch	416 (41)	290 (16)
	mismatch	match	345 (18)	313 (16)
		mismatch	395 (31)	324 (23)
rp	match	match	386 (31)	422 (53)
		mismatch	511 (53)	486 (60)
	mismatch	match	389 (26)	455 (47)
		mismatch	639 (81)	510 (47)
tt	match	match	455 (38)	375 (23)
		mismatch	509 (48)	377 (30)
	mismatch	match	520 (48)	366 (28)
		mismatch	642 (58)	404 (31)

Table B.1: Experiment 2 means for agreement stimuli.
Numbers in parentheses are standard errors of the mean, corrected for within-subjects variability (Cousineau, 2005; Morey, 2008).

			Region	
Measure	Lure type	Lure match	Critical	Spillover
fp	NP	match	297 (14)	300 (15)
		mismatch	283 (14)	276 (13)
	QP	match	286 (12)	309 (15)
		mismatch	262 (11)	295 (12)
rp	NP	match	425 (32)	752 (104)
		mismatch	414 (36)	758 (91)
	QP	match	578 (64)	782 (91)
		mismatch	509 (71)	683 (83)
tt	NP	match	533 (33)	512 (33)
		mismatch	516 (30)	461 (26)
	QP	match	540 (29)	500 (25)
		mismatch	524 (30)	489 (33)

Table B.2: Experiment 2 means for reflexives stimuli.
Numbers in parentheses are standard errors of the mean, corrected for within-subjects variability (Cousineau, 2005; Morey, 2008).

			Region	
Measure	Target match	Lure match	Critical	Spillover
fp	match	match	420 (12)	361 (15)
		mismatch	422 (14)	374 (16)
	mismatch	match	505 (20)	362 (14)
		mismatch	486 (18)	368 (13)
rp	match	match	550 (35)	576 (52)
		mismatch	612 (38)	615 (43)
	mismatch	match	864 (50)	830 (67)
		mismatch	640 (29)	668 (69)
tt	match	match	741 (34)	625 (31)
		mismatch	754 (31)	602 (26)
	mismatch	match	974 (39)	691 (32)
		mismatch	839 (33)	595 (27)

Table B.3: Experiment 3 means for agreement stimuli.
Numbers in parentheses are standard errors of the mean, corrected for within-subjects variability (Cousineau, 2005; Morey, 2008).

			Region	
Measure	Lure type	Lure match	Critical	Spillover
fp	NP	match	368 (13)	515 (21)
		mismatch	361 (15)	485 (20)
	QP	match	353 (15)	535 (21)
		mismatch	356 (15)	496 (21)
rp	NP	match	549 (34)	990 (66)
		mismatch	597 (63)	1124 (92)
	QP	match	659 (55)	961 (72)
		mismatch	616 (63)	1004 (81)
tt	NP	match	696 (30)	1002 (42)
		mismatch	724 (35)	1056 (43)
	QP	match	683 (30)	1056 (46)
		mismatch	726 (35)	1025 (41)

Table B.4: Experiment 3 means for reflexives stimuli.
Numbers in parentheses are standard errors of the mean, corrected for within-subjects variability (Cousineau, 2005; Morey, 2008).

			Region	
Measure	Target match	Lure match	Critical	Spillover
NP lures				
fp	match	match	347 (13)	435 (16)
		mismatch	320 (11)	449 (20)
	mismatch	match	345 (14)	431 (15)
		mismatch	371 (15)	450 (18)
rp	match	match	460 (28)	600 (39)
		mismatch	480 (30)	668 (46)
	mismatch	match	486 (27)	736 (48)
		mismatch	576 (41)	892 (72)
tt	match	match	571 (22)	702 (28)
		mismatch	574 (23)	717 (27)
	mismatch	match	645 (28)	843 (36)
		mismatch	694 (30)	847 (31)
QP lures				
fp	match	match	291 (10)	438 (14)
		mismatch	311 (11)	439 (15)
	mismatch	match	327 (11)	435 (17)
		mismatch	344 (13)	474 (18)
rp	match	match	430 (31)	647 (52)
		mismatch	420 (35)	644 (43)
	mismatch	match	474 (33)	921 (63)
		mismatch	526 (36)	955 (70)
tt	match	match	515 (22)	767 (30)
		mismatch	500 (20)	719 (25)
	mismatch	match	624 (27)	829 (32)
		mismatch	669 (27)	906 (35)

Table B.5: Experiment 4 mean reaction times.

Reading measures: fp - first pass, rp - regression path, tt - total times. Numbers in parentheses are standard errors of the mean, corrected for within-subjects variability (Cousineau, 2005; Morey, 2008).

Appendix C: Model coefficients

Tables start on the next page.

		Critical		Spillover	
Measure	Effect	Estimate	t value	Estimate	t value
Agreement					
fp	(Intercept)	5.91 (0.05)	129.23	5.71 (0.04)	141.49
	Target match	0.18 (0.04)	4.42	-0.06 (0.03)	-1.8
	Lure match	0.03 (0.04)	0.72	-0.03 (0.04)	-0.9
	Target match x Lure match	0.02 (0.08)	0.28	-0.04 (0.06)	-0.57
rp	(Intercept)	6.30 (0.06)	104.16	6.13 (0.05)	112.58
	Target match	0.29 (0.05)	5.58	0.20 (0.06)	3.46
	Lure match	0.13 (0.05)	2.7	0.11 (0.06)	1.86
	Target match x Lure match	0.10 (0.09)	1.13	0.32 (0.10)	3.22
tt	(Intercept)	6.60 (0.07)	95.08	6.31 (0.07)	86.88
	Target match	0.28 (0.05)	6.06	-0.06 (0.04)	-1.47
	Lure match	0.13 (0.03)	3.93	0.05 (0.04)	1.21
	Target match x Lure match	0.01 (0.08)	0.18	0.04 (0.08)	0.47
Reflexives					
fp	(Intercept)	5.64 (0.03)	196.79	5.85 (0.05)	116.57
	Target animacy	-0.04 (0.04)	-0.82	-0.05 (0.06)	-0.8
	Lure match	0.05 (0.04)	1.36	-0.01 (0.04)	-0.14
	Target animacy x Lure match	-0.09 (0.07)	-1.42	-0.00 (0.08)	-0.05
rp	(Intercept)	6.00 (0.05)	122.01	6.44 (0.09)	71.99
	Target animacy	-0.06 (0.07)	-0.86	-0.14 (0.15)	-0.95
	Lure match	0.05 (0.05)	0.93	0.07 (0.07)	1.01
	Target animacy x Lure match	-0.10 (0.10)	-0.91	-0.06 (0.15)	-0.39
tt	(Intercept)	6.42 (0.05)	119.22	6.56 (0.07)	94.72
	Target animacy	-0.11 (0.06)	-1.76	-0.22 (0.08)	-2.81
	Lure match	0.08 (0.04)	1.77	0.07 (0.04)	1.61
	Target animacy x Lure match	-0.00 (0.08)	-0.01	0.05 (0.07)	0.65

Table C.1: Model coefficients for Experiment 1.

Reading measures: fp - first pass, rp - regression path, tt - total times. Contrasts coding: “animate” = -0.5, “inanimate” = 0.5; “match” = -0.5, “mismatch” = 0.5. Statistically significant comparisons are highlighted: yellow - significant only before a Bonferroni correction, red - significant before and after Bonferroni correction (see text). Values in the parentheses represent standard error of the estimate.

		Critical		Spillover	
Measure	Effect	Estimate	t value	Estimate	t value
Agreement					
fp	(Intercept)	5.73 (0.09)	64.27	5.62 (0.05)	108.3
	Target match	0.08 (0.09)	0.93	0.06 (0.06)	1.09
	Lure match	0.09 (0.07)	1.23	0.01 (0.07)	0.17
	Target match x Lure match	-0.04 (0.13)	-0.27	0.08 (0.12)	0.65
rp	(Intercept)	5.89 (0.10)	61.64	5.89 (0.06)	94.91
	Target match	0.12 (0.11)	1.13	0.06 (0.08)	0.73
	Lure match	0.21 (0.08)	2.77	0.09 (0.11)	0.87
	Target match x Lure match	0.11 (0.19)	0.6	0.03 (0.16)	0.2
tt	(Intercept)	6.01 (0.11)	56.35	5.77 (0.07)	77.87
	Target match	0.17 (0.09)	1.92	0.03 (0.07)	0.44
	Lure match	0.11 (0.07)	1.63	0.04 (0.06)	0.66
	Target match x Lure match	0.10 (0.12)	0.79	0.10 (0.13)	0.8
Reflexives					
fp	(Intercept)	5.53 (0.04)	142.43	5.55 (0.05)	118.83
	Lure match	-0.06 (0.04)	− 1.64	-0.06 (0.05)	− 1.22
	Lure type	-0.04 (0.05)	− 0.92	0.04 (0.04)	1.01
	Lure match x Lure type	-0.04 (0.08)	− 0.47	0.03 (0.08)	0.32
rp	(Intercept)	5.85 (0.05)	111.20	6.11 (0.07)	85.43
	Lure match	-0.08 (0.06)	− 1.30	-0.07 (0.09)	− 0.74
	Lure type	0.10 (0.07)	1.47	-0.02 (0.08)	− 0.21
	Lure match x Lure type	-0.08 (0.12)	− 0.63	-0.17 (0.16)	− 1.07
tt	(Intercept)	6.07 (0.05)	115.34	5.97 (0.07)	91.59
	Lure match	-0.05 (0.06)	− 0.85	-0.09 (0.06)	− 1.52
	Lure type	0.02 (0.06)	0.41	0.01 (0.05)	0.12
	Lure match x Lure type	-0.03 (0.10)	− 0.29	-0.04 (0.11)	− 0.36

Table C.2: Model coefficients for Experiment 2.

Reading measures: fp - first pass, rp - regression path, tt - total times. Contrasts coding: "match" = -0.5, "mismatch" = 0.5. Statistically significant comparisons are highlighted: yellow - significant only before a Bonferroni correction, red - significant before and after Bonferroni correction (see text). Values in the parentheses represent standard error of the estimate.

		Critical		Spillover	
Measure	Effect	Estimate	t value	Estimate	t value
fp	(Intercept)	5.91 (0.05)	129.23	5.71 (0.04)	141.49
	Target match	0.18 (0.04)	4.42	-0.06 (0.03)	-1.8
	Lure match	0.03 (0.04)	0.72	-0.03 (0.04)	-0.9
	Target match x Lure match	0.02 (0.08)	0.28	-0.04 (0.06)	-0.57
rp	(Intercept)	6.30 (0.06)	104.16	6.13 (0.05)	112.58
	Target match	0.29 (0.05)	5.58	0.20 (0.06)	3.46
	Lure match	0.13 (0.05)	2.7	0.11 (0.06)	1.86
	Target match x Lure match	0.10 (0.09)	1.13	0.32 (0.10)	3.22
tt	(Intercept)	6.60 (0.07)	95.08	6.31 (0.07)	86.88
	Target match	0.28 (0.05)	6.06	-0.06 (0.04)	-1.47
	Lure match	0.13 (0.03)	3.93	0.05 (0.04)	1.21
	Target match x Lure match	0.01 (0.08)	0.18	0.04 (0.08)	0.47
fp	(Intercept)	5.64 (0.03)	196.79	5.85 (0.05)	116.57
	Target animacy	-0.04 (0.04)	-0.82	-0.05 (0.06)	-0.8
	Lure match	0.05 (0.04)	1.36	-0.01 (0.04)	-0.14
	Target animacy x Lure match	-0.09 (0.07)	-1.42	-0.00 (0.08)	-0.05
rp	(Intercept)	6.00 (0.05)	122.01	6.44 (0.09)	71.99
	Target animacy	-0.06 (0.07)	-0.86	-0.14 (0.15)	-0.95
	Lure match	0.05 (0.05)	0.93	0.07 (0.07)	1.01
	Target animacy x Lure match	-0.10 (0.10)	-0.91	-0.06 (0.15)	-0.39
tt	(Intercept)	6.42 (0.05)	119.22	6.56 (0.07)	94.72
	Target animacy	-0.11 (0.06)	-1.76	-0.22 (0.08)	-2.81
	Lure match	0.08 (0.04)	1.77	0.07 (0.04)	1.61
	Target animacy x Lure match	-0.00 (0.08)	-0.01	0.05 (0.07)	0.65

Table C.3: Model coefficients for Experiment 3.

Reading measures: fp - first pass, rp - regression path, tt - total times. Contrasts coding: “match” = -0.5, “mismatch” = 0.5. Statistically significant comparisons are highlighted: yellow - significant only before a Bonferroni correction, red - significant before and after Bonferroni correction (see text). Values in the parentheses represent standard error of the estimate.

		Critical		Spillover	
Measure	Effect	Estimate	t value	Estimate	t value
NP lures					
fp	(Intercept)	5.72 (0.04)	159.58	5.92 (0.06)	101.46
	Target match	0.04 (0.04)	0.97	0.00 (0.04)	0.12
	Lure match	-0.01 (0.03)	-0.25	0.01 (0.03)	0.38
	Target match x Lure match	0.16 (0.07)	2.38	0.00 (0.08)	0.06
rp	(Intercept)	5.96 (0.05)	122.63	6.27 (0.07)	85.02
	Target match	0.07 (0.05)	1.46	0.18 (0.05)	3.37
	Lure match	0.04 (0.06)	0.68	0.08 (0.06)	1.36
	Target match x Lure match	0.08 (0.09)	0.86	-0.00 (0.12)	-0.02
tt	(Intercept)	6.23 (0.06)	107.23	6.46 (0.07)	87.34
	Target match	0.13 (0.06)	2.32	0.16 (0.04)	4.23
	Lure match	0.02 (0.04)	0.54	0.04 (0.04)	1.13
	Target match x Lure match	0.04 (0.07)	0.59	0.03 (0.07)	0.47
QP lures					
fp	(Intercept)	5.64 (0.03)	162.94	5.94 (0.06)	98.23
	Target match	0.09 (0.04)	2.46	0.01 (0.04)	0.34
	Lure match	0.05 (0.04)	1.55	0.03 (0.04)	0.85
	Target match x Lure match	-0.01 (0.06)	-0.14	0.08 (0.07)	1.2
rp	(Intercept)	5.86 (0.05)	116.63	6.32 (0.08)	78.04
	Target match	0.14 (0.06)	2.29	0.30 (0.05)	6.23
	Lure match	0.05 (0.05)	0.92	0.04 (0.06)	0.58
	Target match x Lure match	0.11 (0.08)	1.26	0.02 (0.09)	0.28
tt	(Intercept)	6.15 (0.06)	108.98	6.49 (0.07)	88.28
	Target match	0.23 (0.05)	4.36	0.14 (0.03)	4.28
	Lure match	0.03 (0.04)	0.79	0.04 (0.05)	0.84
	Target match x Lure match	0.12 (0.08)	1.54	0.15 (0.08)	2

Table C.4: Model coefficients for Experiment 4.

Reading measures: fp - first pass, rp - regression path, tt - total times. Contrasts coding: “match” = -0.5, “mismatch” = 0.5. Statistically significant comparisons are highlighted: yellow - significant only before a Bonferroni correction, red - significant before and after Bonferroni correction (see text). Values in the parentheses represent standard error of the estimate.

Appendix D: List of changes to the Parker and Phillips (2017) stimuli in the replication experiment

"→" indicates the changes made (the left-hand side: original; the right-hand side: replacement). "Q" indicates that the changes were made not to the sentence itself, but to the accompanying comprehension question.

- Item 13, cond 2, Q: "spokeswoman" → "spokesman" (to align with the sentence)
- Item 13, cond 3, Q: "spokesman" → "spokeswoman" (to align with the sentence)
- Item 13, cond 5, Q: "spokesmen" → "spokeswomen" (to align with the sentence)
- Item 15, cond 2,4,6: "rock star" → "pop diva" (fixing factorial manipulation: matrix subject should vary in gender across sentences)
- Item 22, cond 1: "The friendly waitress mentioned that the outspoken hostess recommended herself for the new position" → "The clumsy waitress mentioned that the outspoken hostess criticized herself for the horrible mistake" (to make the sentence as close as possible to the other sentences in the set)

- Item 23, cond 6: “worried” → “said” (to align with the other sentences in the set)

- Item 25, cond 1: “flustered” → “noisy” (to align with the other sentences in the set)

- Item 28, cond 4 and 6. Remove “pretty” (in the original materials, conds 1,2,4,6 had “pretty” modifying the embedded noun, and conds 3,5 didn’t. But the noun in conds 1,2 is “cheerleader”, and in 4-6 - “football players”. I decide to keep the length of the NP to 3 words, thus, removing “pretty” in conds 4,6)

- Item 28. Added complementizers.

- Item 51: “refunded” → “refund” (typo)

- Item 88, Q: “... become popular with the teenagers?” → “... become popular?” (the sentence doesn’t mention teenagers)

Appendix E: Appendix: Experiment 2 fillers types

The following types of fillers were used:

1. **With reflexives and NP antecedents**

    - Sentences either have an embedded complement clause or no embedding at all;
    - If there is a complement clause, the reflexive is inside it;
    - The reflexive matches the grammatically accessible antecedent in some sentences and mismatches it in gender in others (see Table E.2 for counts); it always mismatches any other nouns in the sentences.
    - Half of the items have masculine reflexives, half — feminine.

    Example[1]: *Everybody in the receiving team could see that the woman could hardly stand by herself.*

2. **With reflexives and QP antecedent**

    Similar to the above, but with QP antecedents.

[1] The examples in this section were embedded in larger contexts, so they may sound unnatural by themselves

Example: *Interestingly, every politician seemed to portray himself as a trustworthy and absolutely honest person, according to the journalist.*

3. **Other**

   Three-sentence stories with no special constraints on them. Some of the sentences included a mistake of one of the following types (see Table E.2 for counts):

   - Wrong verb form: *After **having** unexpectedly **losing** her husband in a car crash, Jane were utterly depressed.*
   - Verb number marking: ***Tracy were** babysitting for a new family in the neighborhood who had a very large and old house.*
   - Noun number marking: *By the end of his studies he already spoke **three European language** fluently and was taking lessons in Chinese.*

	Context1	Context2	Critical
Experimental sentences	24	24	24
Agreement attraction	6	6	6
	6	6	6
Fillers, referential	6	6	6
	3	3	3
	3	3	3
	6	6	6
Fillers, quantificational	6	6	6
	3	3	3
	3	3	3
	6	6	6
Fillers, other	18	18	18
	6	6	3
	6	6	6

Table E.2: Experiment 2 stimuli types and counts.

White cells correspond to grammatical sentences, pink - to ungrammatical.

Bibliography

Adesola, O. P. A-bar dependencies in the Yoruba reference-tracking system. *Lingua*, 116(12):2068–2106, December 2006. doi:10.1016/j.lingua.2005.06.001.

Alcocer, P. and Phillips, C. Using relational syntactic constraints in content-addressable memory architectures for sentence parsing. Master's thesis, University of Maryland, 2012.

Anderson, J. R. Human symbol manipulation within an integrated cognitive architecture. *Cognitive science*, 29(3):313–341, 2005. doi:10.1207/s15516709cog0000_22.

Anderson, J. R. and Reder, L. M. The fan effect: New results and new theories. *Journal of Experimental Psychology: General*, 128(2):186, 1999. doi:10.1037//0096-3445.128.2.186.

Aoshima, S., Yoshida, M., and Phillips, C. Incremental processing of coreference and binding in Japanese. *Syntax*, 12(2):93–134, 2009. doi:10.1111/j.1467-9612.2009.00123.x.

Badecker, W. and Straub, K. The processing role of structural constraints on interpretation of pronouns and anaphors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(4):748–769, 2002. doi:10.1037/0278-7393.28.4.748.

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, 68(3):255–278, 2013. doi:10.1016/j.jml.2012.11.001.

Bates, D., Kliegl, R., Vasishth, S., and Baayen, H. Parsimonious mixed models. *arXiv preprint arXiv:1506.04967*, 2015a.

Bates, D., Maechler, M., Bolker, B., and Walker, S. *lme4: Linear mixed-effects models using Eigen and S4*, 2015b. URL `http://CRAN.R-project.org/package=lme4`. R package version 1.1-8.

Charnavel, I. and Huang, Y. Inanimate ziji and Condition A in Mandarin. In *35^{th} West Coast Conference on Formal Linguistics*, pages 132–141. Cascadilla Proceedings Project, 2018.

Charnavel, I. and Sportiche, D. Anaphor binding: What French inanimate anaphors show. *Linguistic Inquiry*, 2016. doi:10.1162/ling_a_00204.

Chow, W.-Y., Lewis, S., and Phillips, C. Immediate sensitivity to structural constraints in pronoun resolution. *Frontiers in Psychology*, 5, June 2014. doi:10.3389/fpsyg.2014.00630.

Clements, G. N. The logophoric pronoun in Ewe: Its role in discourse. *Journal of West African Languages*, 10:141–177, 1975.

Collaboration, O. S. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716–aac4716, August 2015. doi:10.1126/science.aac4716.

Cousineau, D. Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson's method. *Tutorials in quantitative methods for psychology*, 1 (1):42–45, 2005.

Cowan, N. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24(1):87–185, November 2001.

Culy, C. Aspects of logophoric marking. *Linguistics*, 32(6):1055–1094, 1994. doi:10.1515/ling.1994.32.6.1055.

Cunnings, I. and Felser, C. The role of working memory in the processing of reflexives. *Language and Cognitive Processes*, 28(1-2):188–219, January 2013. doi:10.1080/01690965.2010.548391.

Cunnings, I. and Sturt, P. Coargumenthood and the processing of reflexives. *Journal of Memory and Language*, 75:117–139, August 2014. doi:10.1016/j.jml.2014.05.006.

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

Cunnings, I., Patterson, C., and Felser, C. Structural constraints on pronoun binding and coreference: Evidence from eye movements during reading. *Frontiers in Psychology*, 6, June 2015. doi:10.3389/fpsyg.2015.00840.

Daneman, M. and Carpenter, P. A. Individual differences in working memory and reading. *Journal of verbal learning and verbal behavior*, 19(4):450–466, 1980.

Dillon, B. *Structured access in sentence comprehension.* PhD thesis, University of Maryland, College Park, 2011.

Dillon, B. Syntactic memory in the comprehension of reflexive dependencies: an overview. *Language and Linguistics Compass*, 8(5):171–187, May 2014. doi:10.1111/lnc3.12075.

Dillon, B., Mishler, A., Sloggett, S., and Phillips, C. Contrasting intrusion profiles for agreement and anaphora: Experimental and modeling evidence. *Journal of Memory and Language*, 69(2):85–103, 2013. doi:10.1016/j.jml.2013.04.003.

Eberhard, K. M., Cutting, J. C., and Bock, K. Making syntax of sense: Number agreement in sentence production. *Psychological Review*, 112(3):531–559, 2005.

Engelmann, F., Jäger, L. A., and Vasishth, S. The effect of prominence and cue association in retrieval processes: A computational account (former title: The determinants of retrieval interference in dependency resolution: Review and computational modeling) [jan 4 2018 draft]. *Journal of Memory and Language*, draft4.

Franck, J., Vigliocco, G., and Nicol, J. Subject-verb agreement errors in French and English: The role of syntactic hierarchy. *Language and Cognitive Processes*, 17 (4):371–404, 2002.

Gelman, A. and Carlin, J. Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors. *Perspectives on Psychological Science*, 9(6):641–651, November 2014. doi:10.1177/1745691614551642.

Gelman, A. and Hill, J. *Data analysis using regression and multilevel/hierarchical models*, volume 1. Cambridge University Press New York, NY, USA, 2007.

Hagège, C. Les pronoms logophoriques. *Bulletin de la Société de Linguistique de Paris*, 69(1):287–310, 1974.

Hammerly, C., Staub, A., and Dillon, B. The grammaticality asymmetry in agreement attraction reflects respones bias: Experimental and modeling evidence. draft.april.2018.

Hammerly, C., Staub, A., and Dillon, B. Response bias modulates the grammaticality asymmetry: Evidence for a continuous valuation model of agreement araction. Presented at CUNY 2018, 2018.

Hanulíková, A., Van Alphen, P. M., Van Goch, M. M., and Weber, A. When one person's mistake is another's standard usage: The effect of foreign accent on syntactic processing. *Journal of Cognitive Neuroscience*, 24(4):878–887, 2012. doi:10.1162/jocn_a_00103.

Huang, C.-T. J. and Liu, C.-S. L. Logophoricity, attitudes, and ziji at the interface. In *Long-distance reflexives*, volume 33, pages 141–195. 2001.

Jäger, L. A., Benz, L., Roeser, J., Dillon, B. W., and Vasishth, S. Teasing apart retrieval and encoding interference in the processing of anaphors. *Frontiers in Psychology*, 6, June 2015. doi:10.3389/fpsyg.2015.00506.

Jäger, L. A., Engelmann, F., and Vasishth, S. Retrieval interference in reflexive processing: Experimental evidence from Mandarin, and computational modeling. *Frontiers in Psychology*, 6, May 2015. doi:10.3389/fpsyg.2015.00617.

Jäger, L. A., Engelmann, F., and Vasishth, S. Similarity-based interference in sentence comprehension: Literature review and Bayesian meta-analysis. *Journal of Memory and Language*, 94:316–339, June 2017. doi:10.1016/j.jml.2017.01.004.

Jonides, J., Lewis, R. L., Nee, D. E., Lustig, C. A., Berman, M. G., and Moore, K. S. The mind and brain of short-term memory. *Annual Review of Psychology*, 59(1):193–224, January 2008. doi:10.1146/annurev.psych.59.103006.093615.

Kazanina, N. and Phillips, C. Differential effects of constraints in the processing of russian cataphora. *The Quarterly Journal of Experimental Psychology*, 63(2): 371–400, 2010.

Kazanina, N., Lau, E. F., Lieberman, M., Yoshida, M., and Phillips, C. The effect of syntactic constraints on the processing of backwards anaphora. *Journal of Memory and Language*, 56(3):384–409, April 2007. doi:10.1016/j.jml.2006.09.003.

King, J., Andrews, C., and Wagers, M. Do reflexives always find a good antecedent for themselves? Presented at CUNY 2012, 2012a.

King, J., Andrews, C., and Wagers, M. Do reflexives always find a grammatical antecedent for themselves? In *25th annual CUNY conference on human sentence processing*, page 67. The CUNY Graduate Center New York, NY, 2012b.

Kush, D. *Respecting Relations: Memory Access and Antecedent Retrieval in Incremental Sentence Processing.* PhD thesis, University of Maryland, College Park, 2013.

Kush, D., Lidz, J., and Phillips, C. Relation-sensitive retrieval: Evidence from bound variable pronouns. *Journal of Memory and Language*, 82(82):18–40, 2015. doi:10.1016/j.jml.2015.02.003.

Lago, S., Shalom, D. E., Sigman, M., Lau, E. F., and Phillips, C. Agreement attraction in Spanish comprehension. *Journal of Memory and Language*, 82:133–149, July 2015. doi:10.1016/j.jml.2015.02.002.

Lewis, R. L. and Vasishth, S. An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 29(3):375–419, 2005. doi:10.1207/s15516709cog0000_25.

Martin, A. E. and McElree, B. Memory operations that support language comprehension: Evidence from verb-phrase ellipsis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(5):1231–1239, 2009. doi:10.1037/a0016271.

Martin, A. E. and McElree, B. Direct-access retrieval during sentence comprehension: evidence from sluicing. *Journal of Memory and Language*, 64(4):327–343, May 2011. doi:10.1016/j.jml.2010.12.006. URL `https://doi.org/10.1016%2Fj.jml.2010.12.006`.

McElree, B. Accessing recent events. *Psychology of Learning and Motivation*, 46: 155–200, 2006. doi:10.1016/s0079-7421(06)46005-9.

McElree, B., Foraker, S., and Dyer, L. Memory structures that subserve sentence comprehension. *Journal of Memory and Language*, 48(1):67–91, 2003.

Morey, R. D. Confidence intervals from normalized data: A correction to cousineau (2005). *Tutorial in Quantitative Methods for Psychology*, 4(2):61–64, 2008. doi:10.20982/tqmp.04.2.p061.

Moulton, K. and Han, C.-h. C-command vs. scope: An experimental assessment of bound variable pronouns. *Language*, toappear.

Nairne, J. S. A feature model of immediate memory. *Memory & Cognition*, 18(3): 251–269, 1990.

Nicenboim, B., Vasishth, S., Engelmann, F., and Suckow, K. Exploratory and confirmatory analyses in sentence processing: A case study of number interference in German. *Open Science Framework.*, 2017. doi:10.17605/OSF.IO/MMR7S.

Nicenboim, B. and Vasishth, S. Models of retrieval in sentence comprehension: A computational evaluation using Bayesian hierarchical modeling. *Journal of Memory and Language*, 99:1–34, April 2018. doi:10.1016/j.jml.2017.08.004.

Nicol, J. and Swinney, D. The role of structure in coreference assignment during sentence comprehension. *Journal of Psycholinguistic Research*, 18(1):5–19, 1989.

Nicol, J., Foster, K., and Veres, C. Subject–verb agreement processes in comprehension. *Journal of Memory and Language*, 36:569–587, 1997.

Osterhout, L. and Mobley, L. A. Event-related brain potentials elicited by failure to agree. *Journal of Memory and language*, 34(6):739–773, 1995.

Osterhout, L., Bersick, M., and McLaughlin, J. Brain potentials reflect violations of gender stereotypes. *Memory & Cognition*, 25(3):273–285, 1997.

Parker, D. and Phillips, C. Negative polarity illusions and the format of hierarchical encodings in memory. *Cognition*, 157:321–339, 2016. doi:10.1016/j.cognition.2016.08.016.

Parker, D. and Phillips, C. Reflexive attraction in comprehension is selective. *Journal of Memory and Language*, 94:272–290, 2017. doi:10.1016/j.jml.2017.01.002.

Parker, D., Shvartsman, M., and Van Dyke, J. A. The cue-based retrieval theory of sentence comprehension: New findings and new challenges. *Language processing and disorders. Newcastle: Cambridge Scholars Publishing*, 2017.

Parker, D. *The Cognitive Basis for Encoding and Navigating Linguistic Structure.* PhD thesis, University of Maryland: College Park, 2014.

Patil, U., Vasishth, S., and Lewis, R. L. Retrieval interference in syntactic processing: The case of reflexive binding in English. *Frontiers in Psychology*, 7, May 2016. doi:10.3389/fpsyg.2016.00329.

Postal, P. M. Remarks on English long-distance anaphora. *Style*, 40(1-2):7–18, 2006. doi:10.5325/style.40.1-2.7.

R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2014. URL `http://www.R-project.org/`.

Raab, D. H. Division of psychology: Statistical facilitation of simple reaction times. *Transactions of the New York Academy of Sciences*, 24(5 Series II):574–590, 1962.

Ratcliff, R. A theory of memory retrieval. *Psychological Review*, 85(2):59–108, 1978. doi:10.1037/0033-295x.85.2.59.

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

Reinhart, T. and Reuland, E. Reflexivity. *Linguistic inquiry*, 24(4):657–720, 1993.

Reinhart, T. M. *The syntactic domain of anaphora.* PhD thesis, Massachusetts Institute of Technology, 1976.

Sells, P. Aspects of logophoricity. *Linguistic Inquiry*, 18(3):445–479, 1987.

Slioussar, N. and Malko, A. Gender agreement attraction in Russian: production and comprehension evidence. *Frontiers in psychology*, 7, 2016. doi:10.3389/fpsyg.2016.01651.

Sloggett, S. *When errors aren't: how comprehenders selectively violate Binding Theory.* PhD thesis, University of Massachussetts, Amherst, 2017.

Sloggett, S. and Dillon, B. Do comprehenders violate the Binding Theory? Depends on your point of view. in prep.

Sturt, P. The time-course of the application of binding constraints in reference resolution. *Journal of Memory and Language*, 48(3):542–562, April 2003. doi:10.1016/s0749-596x(02)00536-3.

Tang, C.-C. J. Chinese reflexives. *Natural Language & Linguistic Theory*, 7(1): 93–121, 1989.

Tanner, D., Nicol, J., and Brehm, L. The time-course of feature interference in agreement comprehension: Multiple mechanisms and asymmetrical attraction. *Journal of Memory and Language*, 76:195–215. doi:10.1016/j.jml.2014.07.003.

Tucker, M. A., Idrissi, A., and Almeida, D. Attraction effects for verbal gender and number are similar but not identical: Self-paced reading evidence from modern standard Arabic. submitted. URL `https://matthew-tucker.github.io/files/papers/gender-attraction-msa-comprehension.pdf`.

Van Dyke, J. A. and Johns, C. L. Memory interference as a determinant of language comprehension. *Language and Linguistics Compass*, 6(4):193–211, 2012.

Van Dyke, J. A. and McElree, B. Retrieval interference in sentence comprehension. *Journal of Memory and Language*, 55(2):157–166, 2006.

Vasishth, S., Brussow, S., Lewis, R., and Drenhaus, H. Processing polarity: How the ungrammatical intrudes on the grammatical. *Cognitive Science: A Multidisciplinary Journal*, 32(4):685–712, June 2008. doi:10.1080/03640210802066865.

Vasishth, S., von der Malsburg, T., and Engelmann, F. What eye movements can tell us about sentence comprehension. *Wiley Interdisciplinary Reviews: Cognitive Science*, 4(2):125–134, December 2012. doi:10.1002/wcs.1209.

Vasishth, S., Jaeger, L. A., and Nicenboim, B. Feature overwriting as a finite mixture process: Evidence from comprehension data. *arXiv preprint arXiv:1703.04081*, 2017. URL `https://arxiv.org/pdf/1703.04081.pdf`.

Vigliocco, G. and Nicol, J. Separating hierarchical relations and word order in language production: Is proximity concord syntactic or linear? *Cognition*, 68: B13B29, 1998.

von der Malsburg, T. and Angele, B. False positives and other statistical errors in standard analyses of eye movements in reading. *Journal of Memory and Language*, 94:119–133, 2017. doi:10.1016/j.jml.2016.10.003.

Wagers, M. W., Lau, E. F., and Phillips, C. Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language*, 61:206–237, 2009.

Xiang, M., Dillon, B., and Phillips, C. Illusory licensing effects across dependency types: ERP evidence. *Brain & Language*, 108:4055, 2008. doi:10.1016/j.bandl.2008.10.002.

Xiang, M., Grove, J., and Giannakidou, A. Dependency-dependent interference: NPI interference, agreement attraction, and global pragmatic inferences. *Frontiers in Psychology*, 4, 2013. doi:10.3389/fpsyg.2013.00708.