Agreement and reflexives in non-native sentence processing

It is advisable to refer to the publisher's version if you intend to cite from the work. See Guidance on citing.

To link to this article DOI: http://dx.doi.org/10.1017/S136672892400049X

Publisher: Cambridge University Press

www.reading.ac.uk/centaur

CentAUR

Agreement and reflexives in non-native sentence processing

Shatha Alaskar[1] and Ian Cunnings[2]

[1]Department of English Language, College of Education, Majmaah University, Majmaah, Saudi Arabia and [2]School of Psychology and Clinical Language Sciences, University of Reading, Reading, UK

This research article was awarded Open Data and Open Materials badges for transparent practices. See the Data Availability Statement for details.

CAMBRIDGE
UNIVERSITY PRESS

Abstract

How native (L1) and non-native (L2) readers utilise syntactic constraints on linguistic dependency resolution during language comprehension is debated, with previous research yielding mixed findings. To address this discrepancy, we report two large-scale studies, using self-paced reading and grammaticality judgements, investigating subject-verb agreement and reflexives in L1 English speakers and Arabic learners of L2 English. We manipulated sentence grammaticality and the properties of 'distractor' constituents (The key(s) to the cabinet(s) were rusty) in two studies testing number in agreement and gender/number in reflexives. Study 1 showed that L2ers' performance largely patterned with L1ers'. Although grammaticality effects were smaller for agreement in L2ers than in L1ers, proficiency modulated L2 performance. Study 2 revealed no significant between-group differences. Contrasting some L1 studies, significant distractor effects were only detected for reflexives in Study 1. Together, these results imply that L2ers compute syntactic dependencies similarly to L1ers, and potential differences might be driven by L2 proficiency.

Highlights

- We investigated subject–verb agreement and reflexives in L1/L2 sentence processing
- Two large-scale studies used offline judgements and online self-paced reading
- Results suggest both groups use similar parsing mechanisms
- Group differences driven by individual differences in L2 proficiency

1. Introduction

Forming syntactic dependencies between non-adjacent constituents is a prerequisite for successful comprehension. For instance, matching a verb with its grammatical controller, the subject, as in (1), or linking anaphoric expressions, such as reflexives, to their antecedents as in (2), correctly requires the integration of different information sources in real-time.

(1)   The key to the cabinet was rusty.
(2)   The friend of the boy hurt himself.

The processing of such dependencies has informed debate about native (L1) and non-native (L2) processing, but how to explain potential L1/L2 differences is contested. An influential account posits that L2 speakers (L2ers) have difficulty applying syntactic constraints during processing relative to L1 speakers (L1ers) (The Shallow Structure Hypothesis, SSH, Clahsen & Felser, 2006, 2018). Others attribute L1/L2 differences to cognitive demands or individual differences in proficiency or lexical processing ability (Hopp, 2014; Lim & Christianson, 2015; McDonald, 2006). More recently, Cunnings (2017) argued that differences between L1 and L2 processing can be explained in terms of the working memory operations that underpin sentence comprehension.

Whether L2ers violate constraints on linguistic dependencies during processing has been important in assessing these theories, but existing research has mixed findings. Some studies suggest L1ers and L2ers resolve dependencies in a similar manner, while others suggest L1/L2 differences (Dallas & Kaan, 2008; Felser, 2015, 2019; Roberts, 2013). Various factors may influence these findings. L2 performance may vary across dependencies, as in subject-verb (S-V) agreement and reflexives (Felser et al., 2009; Felser & Cunnings, 2012; Jiang, 2004; Tanner et al., 2012). Diverse methods, L1/L2 combinations, and/or individual variation may additionally cause conflicting findings across studies. We contribute to this debate by examining S-V agreement and reflexive-antecedent dependencies. We are unaware of any existing study that has examined these two dependencies in parallel experimental settings in the same L2 group. In two studies, we focus on whether working memory operations, especially memory retrieval, and/or individual differences in proficiency can explain potential L1/L2 differences. We also

consider whether our results support other L2 processing models, such as the SSH. Our findings suggest that L2ers largely mirror L1ers, and potential differences might arise from individual differences in L2 proficiency. We begin by discussing the real-time processing of linguistic dependencies, before discussing existing work in S-V agreement and reflexives in turn.

2. Resolving linguistic dependencies in real-time

Resolving linguistic dependencies involves storing and retrieving information from memory, and it has been argued that processing such dependencies relies on cue-based memory retrieval (Lewis et al., 2006; McElree, 2000; Vasishth et al., 2019). Cue-based parsing predicts that items during incremental processing are stored in memory as chunks or groups of features. These features include lexical content and structural relations that function as symbols or pointers to other related items. For example, in (3) from Pearlmutter et al. (1999), the verb needs to be connected to its target subject 'the key' to complete the S-V agreement dependency.

(3) a. **The key** $\begin{pmatrix} +subject \\ +singular \end{pmatrix}$ to the cabinet $\begin{pmatrix} -subject \\ +singular \end{pmatrix}$ **was** $\begin{Bmatrix} +subject \\ +singular \end{Bmatrix}$ rusty …

b. **The key** $\begin{pmatrix} +subject \\ +singular \end{pmatrix}$ to the cabinets $\begin{pmatrix} -subject \\ -singular \end{pmatrix}$ **was** $\begin{Bmatrix} +subject \\ +singular \end{Bmatrix}$ rusty …

c. *****The key** $\begin{pmatrix} +subject \\ +singular \end{pmatrix}$ to the cabinets $\begin{pmatrix} -subject \\ -singular \end{pmatrix}$ **were** $\begin{Bmatrix} +subject \\ -singular \end{Bmatrix}$ rusty …

d. *****The key** $\begin{pmatrix} +subject \\ +singular \end{pmatrix}$ to the cabinet $\begin{pmatrix} -subject \\ +singular \end{pmatrix}$ **were** $\begin{Bmatrix} +subject \\ -singular \end{Bmatrix}$ rusty …

The reflexive-antecedent dependency in (4) also requires resolution for successful comprehension. Here, the reflexive 'himself/ herself' should be bound by a c-commanding subject within the local domain, 'the schoolboy', as per Binding Principle A (Chomsky, 1981, 1986).

(4) a. **The schoolboy** $\begin{pmatrix} +subject \\ +c-command \\ +masculine \end{pmatrix}$ who asked the man $\begin{pmatrix} -subject \\ -c-command \\ +masculine \end{pmatrix}$ had prepared **himself** $\begin{Bmatrix} +subject \\ +c-command \\ +masculine \end{Bmatrix}$ …

b. **The schoolboy** $\begin{pmatrix} +subject \\ +c-command \\ +masculine \end{pmatrix}$ who asked the woman $\begin{pmatrix} -subject \\ -c-command \\ -masculine \end{pmatrix}$ had prepared **himself** $\begin{Bmatrix} +subject \\ +c-command \\ +masculine \end{Bmatrix}$ …

c.

*****The schoolboy** $\begin{pmatrix} +subject \\ +c-command \\ +masculine \end{pmatrix}$ who asked the woman $\begin{pmatrix} -subject \\ -c-command \\ -masculine \end{pmatrix}$ had prepared **herself** $\begin{Bmatrix} +subject \\ +c-command \\ -masculine \end{Bmatrix}$ …

d. *****The schoolboy** $\begin{pmatrix} +subject \\ +c-command \\ +masculine \end{pmatrix}$ who asked the man $\begin{pmatrix} -subject \\ -c-command \\ +masculine \end{pmatrix}$ had prepared **herself** $\begin{Bmatrix} +subject \\ +c-command \\ -masculine \end{Bmatrix}$ …

According to cue-based parsing, readers access memory representations when reaching the verb in (3) or reflexive in (4) to retrieve the target item ('the key' in (3) or 'the schoolboy' in (4)) for dependency resolution. Retrieval cues derived from the local syntactic context and the item that triggered this process will be used to guide retrieval. For instance, {+*subject*} and {+*singular*} are set out by the verb in (3a/b) to seek out a matching noun that can act as a subject of the verb 'was'. Similarly, the reflexive in (4a/b) may use {+*subject*}, {+*c−command*}[1] and {+*masculine*} to find an antecedent with matching features. However, the presence of similar items or 'distractors' in memory that partially match the retrieval cues may decrease the probability of retrieving the target item and potentially cause retrieval errors. This is known as similarity-based interference (Lewis et al., 2006; Jäger et al., 2017; Vasishth et al., 2019). For example, in (3a/c), the complex noun phrase headed by 'the key' includes a distractor ('the cabinet/s') that may interfere in dependency resolution when it matches the verb's number. The gender overlap between the distractor ('man/woman') and the reflexive in (4a/c) can also increase the possibility of wrongly retrieving the partially matching distractor.

Similarity-based interference manifests as inhibition or facilitation. Inhibitory interference occurs in grammatical sentences when access to the target item is disrupted by distractors that also match a subset of the retrieval cues (Jäger et al., 2017). This predicts longer reading times in (3a/4a) relative to (3b/4b). Facilitatory interference (Jäger et al., 2017) occurs in ungrammatical sentences containing a partially-matching distractor and manifests as a speed-up in reading time (Hammerly et al., 2019; Lewis & Phillips, 2014). Accordingly, the matching distractor should facilitate processing time in (3c/4c) compared to (3d/4d).

L1 S-V agreement has been widely studied (see Hammerly et al., 2019 for review). Across studies, a frequent asymmetrical pattern of interference effects was observed. That is, facilitated processing has usually been found in ungrammatical sentences like (3c) compared to (3d). However, interference in grammatical sentences has rarely been observed (Dillon et al., 2013; Parker & An, 2018; Parker & Phillips, 2017; Wagers et al., 2009).

---

[1]Technically, c-command is a relational concept between constituents, rather than a static feature. For further discussion on implications of this, see Kush (2013).

Given that the cue-based model predicts facilitatory interference in ungrammatical but not grammatical sentences, this grammatical asymmetry has been taken to argue in favour of cue-based parsing (Wagers et al., 2009).[2] Though cue-based retrieval also predicts longer reading times in grammatical sentences like (3a) than (3b) due to inhibitory interference, this has not been well attested. Wagers et al. claimed that in grammatical sentences, as the target provides a complete match to the retrieval cues, very little or no interference occurs. Alternatively, Nicenboim et al. (2018) claimed inhibitory effects may be small, and most studies are underpowered to detect it.

How the cue-based account of interference in agreement generalises to other linguistic dependencies has been debated. Studies investigating reflexives have been influential in this regard (Dillon et al., 2013; Jäger et al., 2020; Parker & Phillips, 2017; Sturt, 2003). Some studies failed to observe interference effects consistent with cue-based memory retrieval for reflexives, suggesting that structural constraints are weighted more strongly than gender/number features during reflexive resolution (Cunnings & Sturt, 2014; Dillon et al., 2013; Parker & Phillips, 2017; Sturt, 2003). For example, Dillon et al. (2013) compared both dependencies and found that S-V agreement was susceptible to facilitatory interference but reflexives were not. They argued that reflexive-antecedent retrieval involves only structural cues, guiding retrieval to a grammatical antecedent only, while S-V agreement also utilises agreement features, leading to interference. However, a large sample replication by Jäger et al. (2020) found similar facilitatory interference profiles in both S-V agreement and reflexives, which they took as indicating a similar retrieval mechanism, utilising both structural cues and agreement features, is employed in both dependencies. They argued that previous claims about insensitivity to interference for reflexives from low-powered studies should be considered with caution.

In sum, there has been debate about the nature of interference effects across dependencies in L1 processing. Dillon et al.'s (2013) results were influential as they suggested it may be premature to draw conclusions about the general architecture of memory retrieval during sentence processing based on a single linguistic dependency. While recent work by Jäger et al. (2020) suggests Dillon et al.'s conclusions that reflexives are insensitive to interference were too strong, these studies demonstrate the need to compare different linguistic dependencies when drawing conclusions about dependency resolution. While this issue has been well studied during L1 processing, whether L2 processing shows the same pattern of sensitivity to linguistic cues across dependencies has received less attention.

3. Resolving linguistic dependencies in L2 processing

Several models have been proposed to explain how L2ers process linguistic dependencies. Cunnings (2017) argued that L2ers may face difficulty in resolving linguistic dependencies if they weight memory retrieval cues differently to L1ers. Specifically, Cunnings claimed that L2ers underweight syntactic retrieval cues compared to L1ers, leading them to be more susceptible to interference during memory retrieval. This would predict larger similarity-based interference effects in L2ers as compared to L1ers.

The SSH (Clahsen & Felser, 2006, 2018) has also been an influential account of L2 processing. This account predicts that L2ers may adopt shallow processing routines that underutilise syntactic information in favour of other non-syntactic information sources. This would predict that L2ers may not adhere to syntactic constraints on linguistic dependencies in the same way as L1ers during processing.

Alternatively, if L2 processing is modulated by individual differences in factors such as proficiency (Hopp, 2006, 2014), L2ers should behave more nativelike as proficiency increases. The current study aims to tease apart these accounts by testing S-V agreement and reflexive-antecedent dependencies in the same L2 group.

3.1. S-V agreement in L2 processing

Research on L2 S-V agreement has shown inconsistent findings. In self-paced reading, Jiang (2004) found Chinese L2ers of English insensitive to number disagreement between non-adjacent subjects and verbs relative to L1ers (see also Chen et al., 2007). Tanner et al. (2012) examined interference in S-V agreement in Spanish L2ers of English using sentences like (3a-d) in an ERP study. Unlike Jiang (2004) and Chen et al. (2007), significant grammaticality effects were found for both groups. Although P600 effects were smaller in L2ers than L1ers, both groups showed reduced grammaticality effects when the verb followed a plural distractor, indicating facilitatory interference.

From this, one might infer that L2ers show native-like S-V agreement processing only if their L1 features agreement, like Spanish, but not when it does not, as in Chinese. However, recent studies suggest L2ers can process S-V agreement similarly to L1ers even if their L1 lacks agreement. For instance, Korean L2ers of English were examined by Lim and Christianson (2015) on S-V agreement, that is absent in their L1. They used sentences like "The teacher(s) who instructed the student(s) were very strict", manipulating distractor number across grammatical and ungrammatical conditions. During reading, L1ers showed clear grammaticality effects at the verb and spillover regions, with similar effects observed in L2ers at the spillover region. L2 grammaticality effects were however affected by proficiency, such that they appeared more clearly as proficiency increased. In response to the distractor, both groups showed reduced ungrammaticality effects with matching plural distractors, indicating facilitatory interference. These results suggest L2ers can detect agreement violations regardless of their L1, but this is modulated by L2 proficiency.

Lee and Philips (2022) reported that Korean L2ers could even outperform L1ers in speeded judgement tasks. They found that acceptance rates for ungrammatical sentences with relative clauses (e.g., 'The artist who made the sculpture/s are very talented') indicated facilitatory interference in both groups, but not in L2ers in sentences with prepositional phrases (PP) (e.g., 'The artist with the tall sculpture/s are very talented'), unlike L1ers. Lee and Phillips argued that L2ers may utilise an additional monitoring mechanism that helps filter out ungrammatical structures compared to L1ers, at least when making explicit judgements.

Finally, in Reifegerste et al. (2020) L1 and L2 German speakers were shown sentence fragments (e.g. "Der Brief/Die Briefe des diplomatischen Anwalts/der diplomatischen Anwälte…" – "The letter/s from the diplomatic lawyer/s…") and had to choose appropriate continuations ("hat/haben" – "has/have"). L1ers more frequently chose the incorrect plural verb for sentences containing singular subjects and plural distractors, compared to incorrectly selecting a singular verb for sentences with plural subjects and

---

[2]Another class of models predict facilitated processing when a distractor matches the verb's number in both grammatical and ungrammatical sentences (see Hammerly et al., 2019). For reasons of space, and because our focus is on retrieval-based accounts, we do not discuss these accounts in further detail.

singular distractors, replicating the often observed 'mismatch asymmetry' with larger interference effects from plural than singular distractors (Eberhard, 1997). L2ers, however, exhibited similar interference effects from both singular and plural distractors. Reifegerste et al. interpreted these results as suggesting L2ers were more likely to assign a shallow structure to the complex subject noun phrase.

In summary, although early studies (e.g. Jiang, 2004) suggested insensitivity to agreement during L2 processing, which might indicate shallow L2 processing, more recent studies indicate the opposite across a variety of L2 groups. Sensitivity to agreement is also influenced by L2 proficiency (Lim & Christianson, 2015). The precise pattern of interference effects observed has however differed across studies, though the results of one study (Lee & Phillips, 2022) suggest L2ers may be less sensitive to interference in some cases (contra Cunnings, 2017). Reifegerste et al. (2020) proposed alternatively that L1/L2 differences in interference effects in certain circumstances may indicate shallow L2 processing (Clahsen & Felser, 2006, 2018).

3.2. Reflexive resolution in L2 processing

Fewer studies have examined reflexives. Felser et al. (2009) examined Japanese L2ers of English sensitivity to syntactic constraints on reflexives compared to L1ers. In an offline antecedent choice task, both groups performed accurately, however, their reading patterns differed during processing. In an eye-tracking experiment with only grammatical sentences, two factors were manipulated: the gender match between distractors and reflexives and the syntactic structure, such that the distractor either c-commanded the reflexive (e.g., John/Jane noticed that Richard had cut himself…) or did not (e.g., It was clear to John/Jane that Richard had cut himself…). L2ers had longer reading times at the reflexive when a c-commanding but non-local distractor matched the reflexive's gender compared to when it did not. This pattern suggests inhibitory interference, as predicted by cue-based parsing. No gender match effects were found for the non-c-commanding antecedent conditions, and L1ers showed no effects of distractor gender. Felser et al. concluded that L2ers' native-like performance in the offline task indicated the use of grammatical knowledge while their reading times revealed that they were initially influenced by the presence of a matching discourse-prominent antecedent. L1 influence, however, could not be precluded since Japanese reflexives allow non-local antecedents.

Felser and Cunnings (2012) investigated whether German L2ers of English, whose L1 is similar to English in that reflexives must be bound by a local antecedent, would be affected by distractors during processing. Responses in an offline antecedent choice task confirmed L2 knowledge of binding Condition A. Analysis of eye-tracking data, however, showed that L2 reading patterns differed from L1ers regardless of whether an illicit antecedent c-commanded the reflexive or not. Unlike L1ers who showed significantly longer reading times when a local (target) antecedent mismatched in gender with a reflexive, L2ers showed initially a main effect of a gender-mismatching distractor whereas effects of the target were only seen at later processing measures. Given that German allows only local binding, L1 transfer was precluded as a cause for L2ers' early preference for the distractor, and Felser and Cunnings instead argued that the L2ers had difficulty applying binding constraints.

Finding that L2ers have difficulty applying binding constraints and instead initially prefer discourse-prominent distractors is compatible with the SSH (Clahsen & Felser, 2006, 2018). Cunnings

(2017) alternatively argued that L2ers may construct fully specified parses, but instead are more prone to interference than L1ers, if they weight non-syntactic retrieval cues, especially those related to discourse prominence, more highly than L1ers. How individual differences in proficiency influence the L2 processing of reflexives has however not previously been examined.

4. The present study

Against this background, we aimed to tease apart the different accounts of dependency resolution during L2 processing and elucidate the contrasting existing L2 findings between agreement and reflexives. Previous studies on S-V agreement have revealed mixed findings, while L1/L2 differences have been observed for reflexives. However, existing studies have examined either S-V agreement or reflexives, making it difficult to compare L2 processing across these two dependencies. Studies have also used different experimental designs, and methodologies, and tested different L2 groups, with inconsistent examination of proficiency across both dependencies. Studies comparing S-V agreement and reflexives have been influential in the L1 processing literature (Dillon et al., 2013; Jäger et al., 2020), and we adopt a similar approach to examine L2 processing, to assess the extent to which different accounts of L2 processing can explain L2 dependency resolution across different linguistic dependencies. Therefore, we tested L1 Arabic speakers' L2 processing of both S-V agreement and reflexive-antecedent dependencies.

Arabic L2ers of English were chosen because S-V agreement and reflexives behave similarly in Arabic and English, thus reducing the possibility of L1 influence. In Arabic and English, retrieval of the local subject is required for both S-V agreement and reflexives. The local subject and dependent element (the verb or reflexive) should also have the same agreement features. For S-V agreement, the number is realised morphologically in Arabic on nominals and verbs[3], as in English. Reflexives are locally bound in both languages by the closest c-commanding subject.

We report two large-scale studies. Study 1 tested S-V number agreement and gender congruency in reflexives, while Study 2 tested reflexives but with a number manipulation. In both studies, participants completed a grammaticality judgement task (GJT) to assess grammatical knowledge and a self-paced reading (SPR) experiment that examined real-time processing.

5. Study 1

The GJT tested sentences as in (5) and (6). In both cases, there were two 'no distractor' conditions that tested the basic understanding that subjects and verbs must agree in number and that reflexives require a gender-matching antecedent. For agreement, in the grammatical condition (5a), the subject matches the verb in number, while in ungrammatical (5b) there is a mismatch. (6a/b) have a similar manipulation using gender for reflexives. The two 'distractor' conditions aimed to test L2 understanding that only certain

---

[3]Modern Standard Arabic (MSA) has two word-orders, SVO and VSO. In SVO, the verb agrees with the subject in person, gender, and number. In VSO, the verb matches the subject in gender and person but not number, remaining singular even with dual/plural nouns. Both word-orders are common, and the verb in many local varieties of Arabic agrees fully with dual/plural subjects regardless of subject position (Benmamoun, 1992; Musabhien, 2008). Therefore, we do not expect the L1 VSO word-order to affect L2 processing, given availability of the SVO word-order, which parallels English.

constituents must match in number/gender with either the verb or reflexive. For agreement, we included a grammatical condition (5c), where the subject matched the verb's number while the distractor was mismatched. We also included an ungrammatical condition (5d) in which the subject mismatched the verb's number, while the distractor matched. (6c/d) have a similar manipulation using gender and reflexives.

(5)   **S-V Agreement**
  a. *No Distractor, Grammatical*
  The boys were hurt yesterday afternoon.
  b. *No Distractor, Ungrammatical*
  *The boy were hurt yesterday afternoon.
  c. *Distractor, Grammatical*
  The boys near the girl were hurt yesterday afternoon.
  d. *Distractor, Ungrammatical*
  *The boy near the girls were hurt yesterday afternoon.

(6)   **Reflexives**
  a. *No Distractor, Grammatical*
  The man cut himself two hours ago.
  b. *No Distractor, Ungrammatical*
  *The man cut herself two hours ago.
  c. *Distractor, Grammatical*
  The man behind the lady cut himself two hours ago.
  d. *Distractor, Ungrammatical*
  *The man behind the lady cut herself two hours ago.

Accuracy rates would help clarify if L2ers' real-time processing reflects their grammatical knowledge or is a result of other processing mechanisms.

To examine real-time processing, the SPR task tested sentences as in (7/8). Grammaticality was manipulated by varying the number feature of the subject (The waitresses/waitress) in S-V agreement, such that it matched the verb in (7a/b) but not in (7c/d), and the gender feature of the subject (The man/lady) for reflexives, as in (8a/b) versus (8c/d). The distractor's properties were also manipulated, such that it matched in number with the verb in (7a/c) but mismatched it in (7b/d). Similarly, the distractor matched the gender of the reflexive in (8a/c) but mismatched in (8b/d).

(7)   **S-V Agreement**
  a. *Grammatical, Distractor Match*
  The waitresses near the schoolgirls unsurprisingly were unhappy about all the noise.
  b. *Grammatical, Distractor Mismatch*
  The waitresses near the schoolgirl unsurprisingly were unhappy about all the noise.
  c. *Ungrammatical, Distractor Match*
  *The waitress near the schoolgirls unsurprisingly were unhappy about all the noise.
  d. *Ungrammatical, Distractor Mismatch*
  *The waitress near the schoolgirl unsurprisingly were unhappy about all the noise.

(8)   **Reflexives**
  a. *Grammatical, Distractor Match*
  The man near the policeman strangely isolated himself from society for many years.
  b. *Grammatical, Distractor Mismatch*

The man near the policewoman strangely isolated himself from society for many years.
  c. *Ungrammatical, Distractor Match*
  *The lady near the policeman strangely isolated himself from society for many years.
  d. *Ungrammatical, Distractor Mismatch*
  *The lady near the policewoman strangely isolated himself from society for many years.

We expected L1ers and L2ers to show grammaticality effects, with longer reading times in (7c/d) and (8c/d) compared to (7a/b) and (8a/b). Interference effects would most obviously be exemplified by shorter reading times in ungrammatical conditions when the distractor matches the properties of the verb or reflexive. This would predict shorter reading times in (7c/8c) than (7d/8d), as evidence of facilitatory interference. Inhibitory interference would be evidenced by longer reading times in (7a/8a) than in (7b/8b). If L2ers are more susceptible to interference than L1ers (Cunnings, 2017), they should show larger interference effects. That is, L2ers should show larger differences between (7c/8c) and (7d/8d), and between (7a/8a) and (7b/8b), than L1ers. If L2ers assign a shallow structure to complex noun phrases (Reifegerste et al., 2020), L2 reading times may be more affected by the distractor and less influenced by the subject, than L1ers. This would be manifested as reduced or absent grammaticality effects in L2 readers, with increased reading times in distractor mismatch than match conditions. This pattern of results would indicate shallow processing (Clahsen & Felser, 2006, 2018). Finally, if individual differences explain between-group differences between L1 and L2 processing, any L1/L2 differences should be ameliorated at higher levels of L2 proficiency.

5.1. Participants

188 Saudi-Arabic L2 English speakers and 189 L1 English speakers were recruited either from the University of Reading student community or via social media. Participants took part voluntarily or received course credit.

Using a background information questionnaire, only Arabic L2ers of English who started learning English at age 5 or after were included. This led to the removal of eight participants. For L1 English speakers, only those who identified English as their only native language and considered themselves not bilingual, meaning that they do not have a native-like command of languages other than English, were included. This led to the removal of ten participants. Two more L2ers were excluded because they did not complete all tasks and two L1ers were also removed; one for incorrect button pressing in the GJT and the other for incomplete participation. Before analysis, we also excluded two L2ers and one L1er with fast median reaction times in the GJT, indicating inattention.

Accordingly, 176 L2ers (44 males, mean age = 30, range = 18–43) and 176 L1ers (37 males, mean age = 28, range = 18–62) were included. The L2ers also completed the Oxford Quick Placement Test (*Quick Placement Test:Version*1, 2004) with a mean score of 42/60 (SD = 9.7, range = 22–60). Most L2ers were intermediate to advanced learners[4].

---

[4]49 participants scored between 50 to 60, 59 between 40 to 49, 44 between 30 to 39 and 24 between 22 to 29.

5.2. Materials

Stimuli for all experiments are available online (https://osf.io/fy2aw/). Stimuli in the GJT consisted of 24 sentences testing S-V agreement as in (5) and 24 sentences testing reflexives as in (6). Participants saw six different items per condition for each dependency. The distractor was always embedded in a PP. For reflexives, an equal number of masculine and feminine reflexives were used across items. The 48 experimental items were pseudorandomized alongside 64 fillers, half of which were grammatical and half ungrammatical.

24 SPR item sets were constructed for each dependency as in (7) and (8), manipulating grammaticality (grammatical versus ungrammatical) and distractor (distractor match versus distractor mismatch). The subject consistently had an embedded distractor within a PP. For agreement, the main verb was always a plural verb be in the past tense ('were'). Some previous studies on agreement have included singular verbs ('was') in grammatical conditions and plural verbs ('were') in ungrammatical conditions, as in (3). We instead manipulated the number properties of the subject whilst keeping the verb identical across grammatical/ungrammatical conditions to ensure any grammaticality effects are not confounded with word length. The verb was always preceded by an adverb to avoid potential reading time differences between singular and plural distractors from influencing reading times at the critical region (Wagers et al., 2009).

For reflexives, singular gender-biased nouns (e.g., schoolboy) were used rather than gender-stereotyped nouns (e.g., surgeon) to avoid any potential cultural differences in stereotypical gender. The gender of the reflexive was identical across conditions, and grammaticality was manipulated based on its match with the subject. Half of the items contained the reflexive *himself* and half contained *herself.*

64 fillers were also created. These included items with various anaphors positioned differently from the reflexive in the experimental items, and others with main verbs not requiring agreement morphology or differing in form or number from the S-V agreement experimental items. To ensure participants read for meaning, all experimental items and fillers were followed by a yes/no comprehension question, half of which required a yes answer and half no. The question never asked about the critical dependencies.

5.3. Procedure

Study 1 was conducted online using IbexFarm (Drummond, 2013) over two sessions. In the first session, participants first filled out a background questionnaire and then completed the SPR experiment. Each sentence was initially presented as a series of dashes that masked the sentence's words, and participants needed to press the space bar to reveal each word. The presentation was noncumulative, such that the previous word was hidden once the next word appeared. After each sentence, participants answered a yes/no comprehension question on a separate screen by pressing the "1" key for yes or "2" for no. Prior to the test items, participants completed three practice items to become familiar with the task. Afterwards, L2ers completed the Quick Placement Test. This session took 30–40 minutes for L1ers and 50–60 minutes for L2ers.

Participants completed the GJT in a second session, at least one week after the first session. Each sentence in the GJT was presented one a time in its entirety by pressing the space bar. Below each sentence, two choices were given (grammatical/ungrammatical) to which participants responded by pressing either "1"

(for grammatical) or "2" (for ungrammatical). Two practice items preceded the test items, and there was no time limit to finish the task[5]. This session took around 30 minutes for L1ers and 30–40 minutes for L2ers.

5.4. Data analysis

Data and analysis code for all experiments is available online (https://osf.io/fy2aw/). Analyses were conducted separately for each dependency using R (R Core Team, 2019). Accuracy rates in the GJT were analysed using generalised mixed-effects models with the lme4 package (Baayen et al., 2008; Bates et al., 2015). The analysis included by-subject and by-item random effects, and sum-coded main effects of group (L1/L2), grammaticality (grammatical/ungrammatical), distractor (distractor/no distractor) and their interactions as fixed effects. Reading times in the SPR task were log-transformed to remove skewness. Analysis was performed at the critical region ("were" in S-V agreement dependencies and "himself/herself" in reflexive dependencies) and a spillover region (the following word) using linear mixed-effects models. Fixed effects included sum-coded main effects of group (L1/L2), grammaticality (grammatical/ungrammatical), interference (distractor match/mismatch) and their interactions. In addition, region (critical/spillover) was included to test for any potential effects that may result from the time course of processing (see Cunnings & Sturt, 2018 for discussion). The main effect of region and the group-by-region interaction will not be discussed unless they interact with another variable of interest, such as grammaticality and/or distractor, since these effects on their own are not of theoretical significance.

For both tasks, we initially ran an analysis comparing the two groups. We also conducted additional analyses of the L2 group separately with proficiency as a continuous predictor. For reasons of space, we report a summary of these additional analyses only when significant L1/L2 differences of theoretical relevance are found in the main analysis (that is interactions between group and other factors of theoretical interest), as this approach is crucial for evaluating whether individual differences elucidate between-groups L1/L2 differences predicted by different models such as the interference account and SSH. A full report of the proficiency analysis is provided as an online supplement (Appendix S1).

We initially fit models with maximal random effects (Barr et al., 2013). If the maximal model failed to converge, we refit the model after removing the correlation parameters. If this model still did not converge, we iteratively removed the random effects that accounted for the least variance until convergence was achieved. If an interaction was observed, follow-up analysis was performed using nested contrasts. The p values for each fixed effect were calculated using the Satterthwaite approximation by the lmerTest package (Kuznetsova et al., 2017).

Before analysing the GJT data, we checked each participant's median reaction time across experimental and filler items as a measure of attention. Only participants with a median reaction time exceeding 1500 ms were included in the analysis, leading to the exclusion of the previously mentioned two L2ers and one L1er.

Before examining the reading time data, we ensured participants had at least 75% accuracy on comprehension questions in the SPR

---

[5]Participants also completed a lexical decision task to investigate if lexical processing speed influences L2 processing (Hopp, 2014). However, analysis including it as a predictor revealed no reliable effects or interactions. As such, we do not discuss this further.

Table 1. Accuracy in percentages for S–V agreement and reflexives in Study 1

	L1ers	L2ers
S-V Agreement		
No distractor, grammatical	96 (0.01)	86 (0.01)
No distractor, ungrammatical	89 (0.01)	73 (0.01)
Distractor, grammatical	83 (0.01)	83 (0.01)
Distractor, ungrammatical	75 (0.01)	74 (0.01)
Reflexives		
No distractor, grammatical	97 (0.01)	88 (0.01)
No distractor, ungrammatical	81 (0.01)	79 (0.01)
Distractor, grammatical	87 (0.01)	82 (0.01)
Distractor, ungrammatical	74 (0.01)	77 (0.01)

Note. Standard errors in parentheses.

task to gauge attention. All participants scored over this threshold. Mean comprehension accuracy rates across the experimental and filler items were 96% for L1ers (range = 86–100%) and 94% for L2ers (range = 78–100%). Reading times shorter than 100 ms or longer than 10,000 ms were excluded since these likely reflect lapses of attention. This affected less than 0.05% of the data.

5.5. Results

5.5.1. Grammaticality judgements

Accuracy rates are reported in Table 1. For S-V agreement, the analysis revealed a significant main effect of group, as L1ers performed more accurately than L2ers (estimate = −0.80 (0.151), $z = −5.31$, $p < .001$). There was also a significant main effect of grammaticality, with more correct responses to grammatical than ungrammatical conditions (estimate = −0.91 (0.196), $z = −4.67$, $p < .001$), and a significant main effect of distractor due to lower accuracy for conditions with distractors than conditions without (estimate = 0.83 (0.100), z = 8.24, $p < .001$).

A significant group by distractor interaction was also found. Nested contrasts showed lower accuracy in distractor than no distractor conditions for L1ers only (L1 estimate = 1.63 (0.173), $z = 9.42$, $p < .001$; L2 estimate = 0.047 (0.110), $z = 0.43$, $p = .66$). There was also a significant interaction between grammaticality

and distractor (estimate = −0.48 (0.18), $z = −2.73$, p = .006), but the three-way interaction with group was not significant (estimate = 0.20 (0.38), $z = 0.53$, $p = .596$). Nested contrasts, collapsed across distractor/no distractor conditions, indicated higher accuracy in grammatical than ungrammatical conditions, with a larger effect in no distractor conditions (estimate = −1.15 (0.226), $z = −5.12$, $p < .001$) than the distractor conditions (estimate = −0.664 (0.205), $z = −3.23$, $p = .001$).

For reflexives, the analysis revealed a significant main effect of group, with the L1ers being overall more accurate than the L2ers (estimate = −0.54 (0.163), $z = −3.31$, $p < .001$), and a significant main effect of grammaticality, with the grammatical conditions having higher accuracy rates (estimate = −1.14 (0.201), $z = −5.67$, $p < .001$). We also observed a significant main effect of distractor, such that conditions without distractors received higher accuracy rates than conditions with distractors (estimate = 0.76 (0.136), $z = 5.60$, $p < .001$).

A significant three-way interaction between group, grammaticality and distractor was also found (estimate = 0.82 (0.38), $z = 2.14$, $p = .033$). Nested contrasts suggest significant grammaticality effects, with higher accuracy in grammatical than ungrammatical conditions, for L1ers in both distractor/no distractor conditions, with larger effects in no distractor conditions (no distractor estimate = −2.29 (0.382), z = −6.01, $p < .001$; distractor estimate = −1.17 (0.269), z = −4.35, $p < .001$). L2ers, however, showed significant grammaticality effects only in the no distractor conditions (no distractor estimate = −0.72 (0.242), $z = −3.00$, $p = .002$; distractor estimate = −0.380 (0.303), $z = −1.25$, $p = .210$).

Considering the significant between-group interactions in the main analysis, we conducted an additional L2 analysis including proficiency. Separate analyses for agreement and reflexives indicated significant main effects of proficiency, with increasing accuracy as proficiency increased. For agreement only, proficiency significantly interacted with the distractor, with higher-proficiency L2ers having higher accuracy in distractor than no distractor conditions (see online supplement for further details).

5.5.2. Self-paced reading

Reading times for S-V agreement and reflexives are shown in Figures 1 and 2 respectively, while a statistical analysis summary is provided in Table 2.
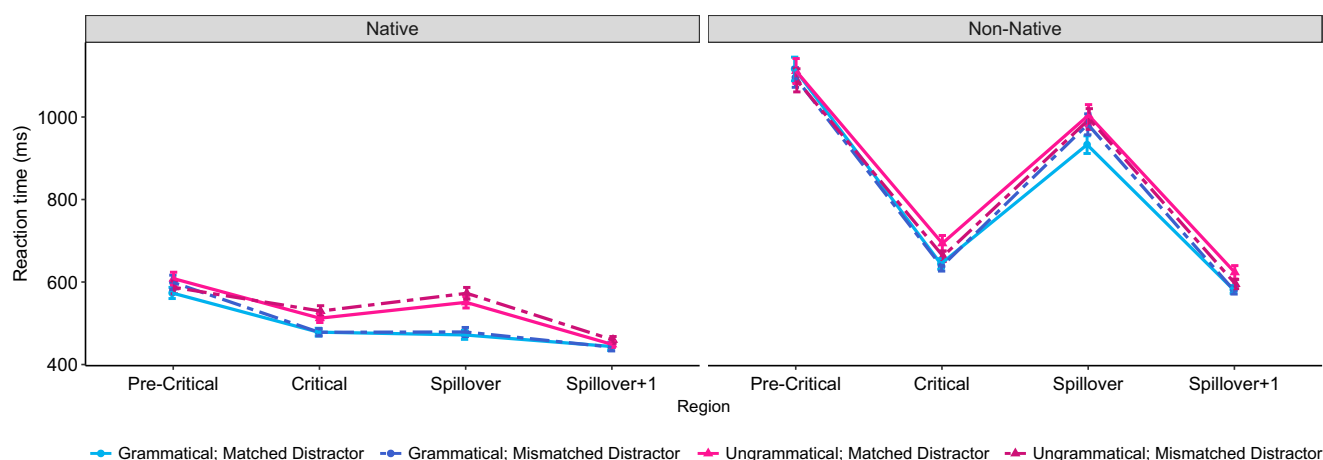
Figure 1. Reading times for S-V agreement dependencies in Study 1. Error bars represent standard errors.
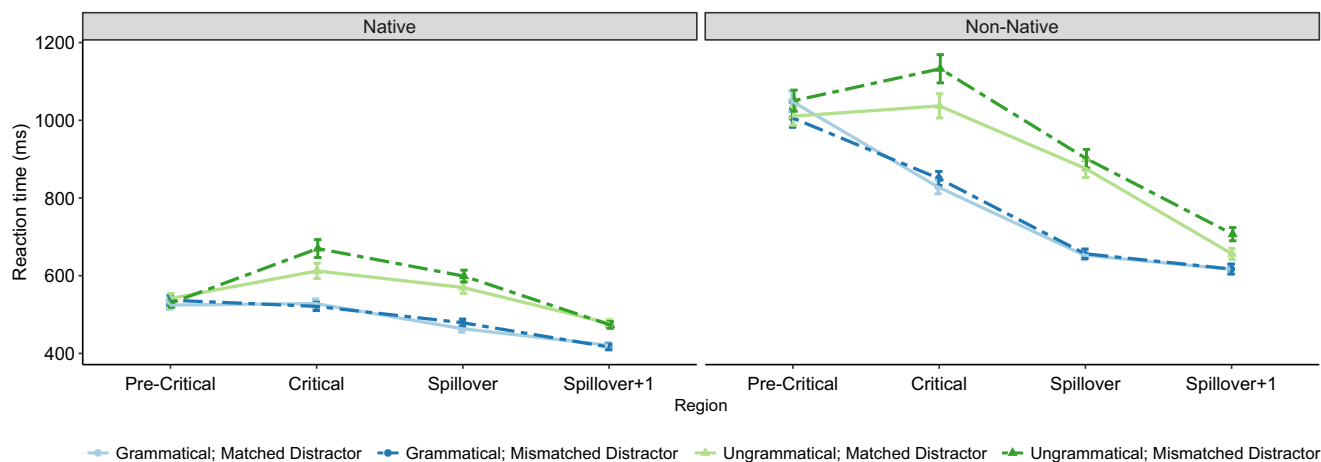
Figure 2. Reading times for reflexive-antecedent dependencies in Study 1. Error bars represent standard errors.

Reading times for S-V agreement revealed a significant main effect of group, as L2ers had longer reading times. There was also a significant main effect of grammaticality, driven by longer reading times in the ungrammatical sentences relative to the grammatical sentences. This ungrammaticality effect was not attenuated by matching distractors as expected, and neither the main effect of the distractor nor any interactions with the distractor were significant.

We also found a significant three-way interaction between group, grammaticality and region. Nested contrasts revealed that grammaticality effects were significant only for the L1ers at the critical region (L1 estimate = 0.049 (0.013), $t = 3.80$, $p < .001$; L2 estimate 0.027 (0.013), $t = 1.93$, $p = .064$), but they were significant for both groups at the spillover region, though the L1 estimate was numerically larger (L1 estimate = 0.126 (0.013), t = 9.37, $p < .001$; L2 estimate = 0.035 (0.015), $t = 2.32$, $p = .029$).

For reflexives, there was a significant main effect of group as L2ers were slower than L1ers. There were also significant main effects of grammaticality and distractor, with longer reading times in ungrammatical than grammatical sentences and in sentences with mismatching distractors than with matching distractors. As shown in Figure 2, the distractor effect is largely restricted to ungrammatical sentences and is suggestive of facilitatory interference in ungrammatical sentences only. However, the interaction between grammaticality and distractor did not reach significance.

The group by grammaticality interaction was significant. Nested contrasts indicated significant grammaticality effects for both groups, with larger effects for the L2ers (L1 estimate = 0.111 (0.014), t = 7.70, $p < .001$; L2 estimate = 0.168 (0.018), $t = 9.37$, $p < .001$). A significant grammaticality by region interaction was also found. Nested contrasts showed that grammaticality effects were significant at both critical and spillover regions, with larger effects at the spillover region (critical region estimate = 0.103 (0.015), t = 6.77, $p < .001$; spillover region estimate = 0.176 (0.013), t = 12.99, $p < .001$).

Given the significant between-group interactions, we conducted additional L2 analyses including proficiency. For agreement, we observed a significant main effect of proficiency, indicating shorter reading times with increased proficiency, and a significant proficiency by grammaticality interaction, where higher proficiency but not lower proficiency L2ers showed clear grammaticality effects. For reflexives, there was also a significant proficiency by grammaticality interaction, with larger grammaticality effects as proficiency increased. These effects are illustrated in Figure 3.

5.6. Discussion

The offline data indicated high accuracy for both groups, suggesting our L2ers have sufficient grammatical knowledge of these dependencies. Both groups were more accurate in grammatical than ungrammatical conditions, as evidenced by the main effects of grammaticality. This might indicate a response bias towards 'grammatical' responses (see Felser et al., 2009; Hammerly et al., 2019). This effect was larger in the 'no distractor' conditions. Conditions containing distractors also had lower accuracy than conditions without, especially in L1ers, which may just indicate lower acceptance rates for longer sentences. Finally, though overall L2ers were less accurate than L1ers, L2 accuracy increased with higher L2 proficiency.

The SPR data showed that L2ers largely patterned with L1ers in that their reading times were guided by structural constraints. Specifically, they showed significant grammaticality effects on both dependencies, but the size of effects differed across the two dependencies. While L2ers' grammaticality effects were relatively smaller for S-V agreement compared to L1ers, grammaticality effects were bigger for the L2ers for reflexives relative to L1ers. The L2 proficiency analysis may explain this relative difference. That is, while L2ers showed grammaticality effects for reflexives across different proficiency levels (with larger effects as proficiency increased), only higher-proficiency L2ers demonstrated grammaticality effects in the expected direction for S-V agreement.

Effects of distractors were only observed in reflexives, such that conditions with matching distractors were read faster than those with mismatching distractors. Figure 2 suggests this effect was mainly driven by the ungrammatical condition. This descriptive finding is indicative of facilitatory interference. The lack of distractor effects for the S-V agreement was surprising, and we return to this in the General Discussion. Importantly for present purposes, however, distractor effects did not significantly interact with group, as such we found no evidence to suggest increased interference in L2ers.

L2ers displaying smaller grammaticality effects than L1ers during S-V agreement processing compared to reflexives may suggest reduced L2 sensitivity to number than gender. However, this reduced sensitivity to number might be related either to L2 difficulty in encoding number features on nouns held in memory during sentence processing, or alternatively, it might be related to difficulty in computing agreement. To assess whether this effect relates to the encoding of number itself we ran Study 2, which examined reflexives using a number, rather than gender, manipulation.

Table 2. SPR statistical analyses for S-V agreement and reflexives in Study 1

	Estimate (SE)	*t*	*p*
S-V Agreement			
Group	0.45 (0.04)	12.45	**<0.001**
Grammaticality	0.06 (0.01)	8.23	**<0.001**
Distractor	0.01 (0.01)	1.17	0.242
Region	0.16 (0.03)	6.11	**<0.001**
Group*grammaticality	−0.06 (0.02)	−3.78	**<0.001**
Group*distractor	−0.01 (0.01)	−0.99	0.323
Grammaticality*distractor	−0.00 (0.01)	−0.25	0.8
Group*region	0.30 (0.04)	7.36	**<0.001**
Grammaticality*region	0.04 (0.01)	3.52	**<0.001**
Distractor*region	0.01 (0.01)	1.26	0.208
Group*grammaticality*distractor	−0.04 (0.03)	−1.20	0.232
Group*grammaticality*region	−0.07 (0.03)	−2.64	**0.008**
Group*distractor*region	−0.00 (0.02)	−0.02	0.988
Grammaticality*distractor*region	−0.00 (0.02)	−0.15	0.882
Group*grammaticality* distractor*region	−0.06 (0.05)	−1.35	0.177
Reflexives			
Group	0.43 (0.04)	11.9	**<0.001**
Grammaticality	0.14 (0.01)	10.8	**<0.001**
Distractor	0.02 (0.01)	3.46	**0.001**
Region	−0.11 (0.03)	−4.26	**<0.001**
Group*grammaticality	0.06 (0.02)	2.89	**0.004**
Group*distractor	−0.01 (0.01)	−0.38	0.707
Grammaticality*distractor	0.02 (0.02)	1.48	0.139
Group*region	−0.12 (0.03)	−4.05	**<0.001**
Grammaticality*region	0.07 (0.01)	5.18	**<0.001**
Distractor*region	0.00 (0.01)	0.26	0.792
Group*grammaticality *distractor	−0.01 (0.03)	−0.33	0.744
Group*grammaticality*region	0.02 (0.03)	0.68	0.495
Group*distractor*region	−0.04 (0.03)	−1.46	0.145
Grammaticality*distractor*region	−0.01 (0.03)	−0.30	0.764
Group*grammaticality *distractor*region	0.02 (0.06)	0.3	0.767

Note. Significant effects (p < .05) are in bold.

6. Study 2

This study tested the extent to which L2ers are sensitive to number in reflexive dependencies and additionally manipulated whether number in reflexive dependencies triggers interference. If L2ers have difficulty encoding number during processing in general, they should show reduced grammaticality effects irrespective of the dependency. If, however, this difficulty is restricted to agreement, L2ers may not show reduced grammaticality effects for number when tested using reflexives.

6.1. Participants

168 Saudi-Arabic L2 English speakers and 171 L1 English speakers were recruited. New L1ers were recruited using the same recruitment method in Study 1. All L2ers in this study also participated in the first study. The interval between their first and second participation was between six to eleven months.

Using the same inclusion criteria as in Study 1, we excluded four L2ers and five L1ers because they scored less than 75% correct on the SPR comprehension questions. Two more L1ers were removed due to fast median reaction times (< 1500 ms) during the GJT.

This led to the inclusion of 164 L2ers (29 males, mean age = 31, range = 19–44) and 164 L1ers (21 males, mean age = 25, range = 18–60).

6.2. Materials

Study 2 utilised a GJT and an SPR task. In the GJT, 24 sets of experimental sentences were created, as in (9). Distractor (no distractor versus distractor) and grammaticality (grammatical versus ungrammatical) were manipulated. The reflexive was plural across all conditions, varying subject number to manipulate grammaticality. The distractor was embedded in a PP and mismatched the reflexive in number in grammatical conditions and matched it in ungrammatical conditions. The 24 experimental items were distributed across four lists in a Latin-square design, alongside 60 fillers, half of which were grammatical.

(9)   **Reflexives**
  a.   *No Distractor, Grammatical*
       The boys hurt themselves yesterday afternoon.
  b.   *No Distractor, Ungrammatical*
       *The boy hurt themselves yesterday afternoon.
  c.   *Distractor, Grammatical*
       The boys near the girl hurt themselves yesterday afternoon.
  d.   *Distractor, Ungrammatical*
       *The boy near the girls hurt themselves yesterday.

In the SPR task, grammaticality (grammatical vs. ungrammatical) and distractor (distractor match versus distractor mismatch) were manipulated in 24 maximally similar configurations to those in Study 1, as in (10). The reflexive was always plural and grammaticality depended on its number match with its antecedent ("the girl/s"). Interference was tested by manipulating distractor number.

12 experimental items in two baseline conditions were also added. In these conditions, the reflexive was also kept plural but there was no intervening distractor between the reflexive and the subject, as in (11). These conditions were included to test grammaticality effects without distractors, especially given that some may treat the reflexive 'themselves' as a non-gendered singular (Foertsch & Gernsbacher, 1997; Speyer & Schleef, 2019). This may affect reading times in the ungrammatical conditions (10c/d), as it may be acceptable to retrieve the singular subject as an antecedent for 'themselves'. The baseline conditions (11a/b) can rule out this possibility, as finding grammaticality effects here would suggest the plural reflexive is indeed treated as plural. The 36 experimental items were mixed with 90 filler items in a pseudorandomised Latin-square design, such that each participant read a total of 126 sentences. All sentences were followed by a
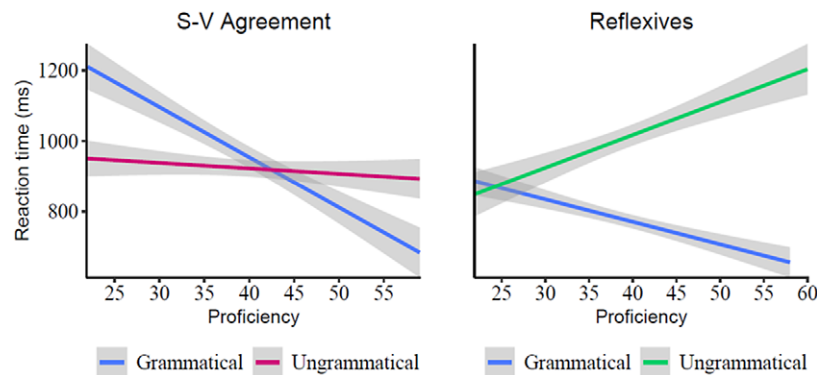
Figure 3. Interaction between proficiency and grammaticality on L2 speakers' reading times in Study 1.

counterbalanced yes/no comprehension question to ensure reading for meaning.

(10)   **Reflexives**
   a.   *Grammatical, distractor match*
         The girls near the boys accidentally hurt themselves
            while playing in the garden.
   b.   *Grammatical, distractor mismatch*
         The girls near the boy accidentally hurt themselves while
            playing in the garden.
   c.   *Ungrammatical, distractor match*
         *The girl near the boys accidentally hurt themselves
            while playing in the garden.
   d.   *Ungrammatical, distractor mismatch*
         *The girl near the boy accidentally hurt themselves while
            playing in the garden.

(11)
   a.   *Grammatical*
         The boys carefully washed themselves using soap before
            having dinner.
   b.   *Ungrammatical*
   c.   *The boy carefully washed themselves using soap before
            having dinner.

6.3. Procedure and data analysis

The procedure was the same as Study 1, except that we used PCIbex (Zehr & Schwarz, 2018). Participants completed the SPR task and a proficiency test for the L2ers in a first session, and the GJT in a second session at least one week later. Data analysis was the same as Study 1.

6.4. Results

6.4.1. Grammaticality judgements

Accuracy rates are provided in Table 3. There was a significant main effect of group (estimate = 0.427 (0.176), $z = 2.42$, $p = .015$), with L2ers having higher accuracy rates than L1ers. Significant main effects of grammaticality (estimate = −3.049 (0.267), $z = 11.38$, $p < .001$) and distractor (estimate = 0.748 (0.102), $z = 7.32$, $p < .001$) were also found, revealing that grammatical conditions received more correct responses than ungrammatical conditions, and responses to conditions without distractors were more accurate compared to conditions with distractors.

There was also a significant three-way interaction between group, grammaticality and distractor. Using nested contrasts, we tested grammaticality effects for each group in the distractor/no distractor conditions. This demonstrated significant grammaticality effects, with lower accuracy in ungrammatical conditions, for both groups in the distractor and no distractor conditions. Grammaticality effects were larger in the L1ers, especially in no distractor conditions (for L1ers, no distractor estimate = −5.26 (0.468), $z = −11.24$, $p < .001$, distractor estimate = −3.64 (0.371), $z = −9.82$, $p < .001$; for L2ers, no distractor estimate = −1.88 (0.292), $z = −6.44$, $p < .001$, distractor estimate = −1.61 (0.269), $z = −5.98$, $p < .001$).

Given the between-group interaction, we conducted an L2 analysis including proficiency. This revealed a significant main effect of proficiency, with higher accuracy as proficiency increased, and a significant proficiency by distractor interaction. Here, lower but not higher-proficiency learners were less accurate in the distractor than no distractor conditions (see online supplement for further details).

6.4.2. Self-paced reading

An average of 94% (range = 82–100%) accuracy on the comprehension questions was achieved by L1ers and 93% (range = 75–99%) by L2ers. We conducted separate analyses for each experimental item set. Reading times are shown in Figure 4. The inferential statistics are presented in Table 4.

As shown in Table 4, there were main effects of group and grammaticality in the baseline conditions, revealing that L2ers were slower than L1ers and reading times were longer in ungrammatical relative to grammatical sentences. There was also a significant interaction between grammaticality and region. Nested contrasts showed that grammaticality effects were significant at the spillover region, but not at the critical region (critical region estimate = 0.025 (0.012), $t = 2.05$, $p = .060$; spillover estimate = 0.082 (0.017), $t = 4.68$, $p < .001$). The group by grammaticality interaction was not significant.

Table 3. Accuracy in percentages for reflexives in Study 2

	L1ers	L2ers
No distractor, grammatical	97 (0.01)	91 (0.01)
No distractor, ungrammatical	46 (0.01)	71 (0.01)
Distractor, grammatical	87 (0.01)	88 (0.01)
Distractor, ungrammatical	43 (0.01)	67 (0.01)

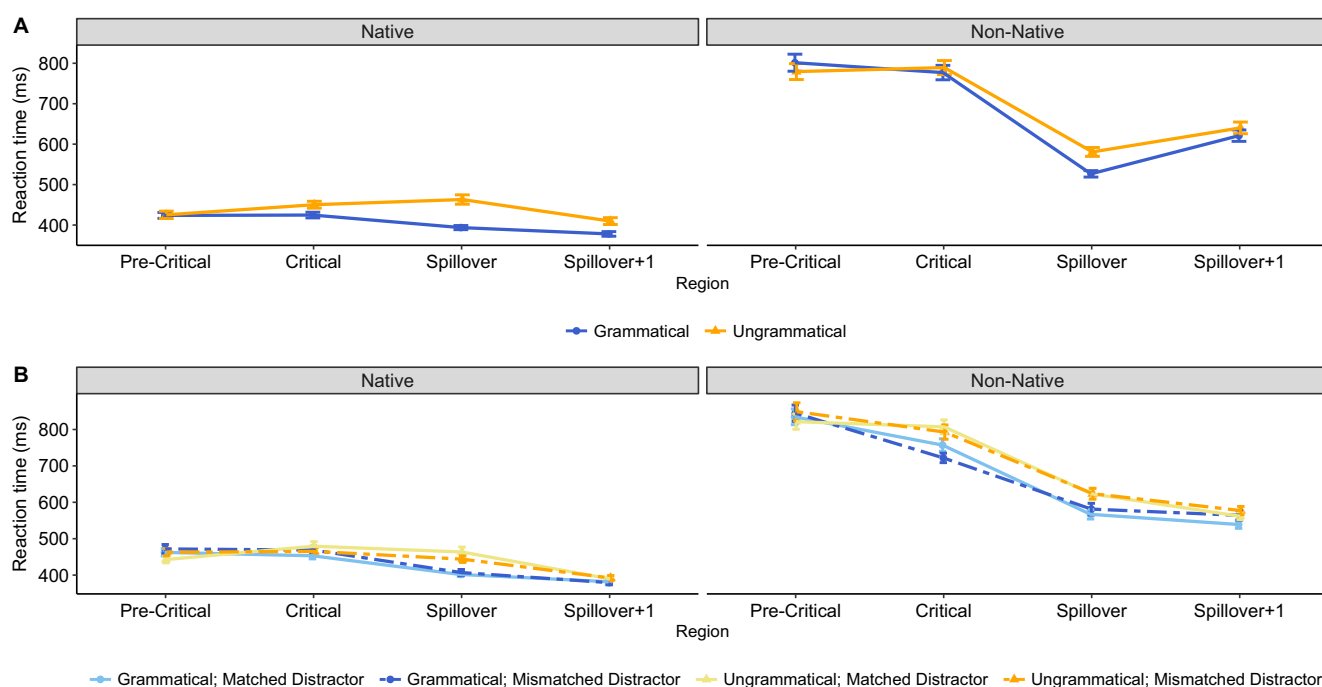*Note. Standard errors in parentheses.*

Figure 4. Reading times for reflexive-antecedent dependencies in (A) baseline and (B) main experimental items in Study 2. Error bars represent standard errors.

Table 4. SPR statistical analysis for reflexives in Study 2

	Estimates (SE)	t	p
Baseline conditions			
Group	0.41 (0.03)	15.08	**<0.001**
Grammaticality	0.05 (0.01)	5.77	**<0.001**
Region	−0.15 (0.02)	−7.20	**<0.001**
Group *grammaticality	−0.03 (0.02)	−1.73	0.084
Group * region	−0.28 (0.01)	−19.61	**<0.001**
Grammaticality *region	0.06 (0.02)	3.04	**0.002**
Group*grammaticality*region	−0.01 (0.04)	−0.36	0.717
Distractor conditions			
Group	0.41 (0.03)	13.59	**<0.001**
Grammaticality	0.05 (0.01)	4.64	**<0.001**
Distractor	−0.01 (0.01)	−0.95	0.344
Region	−0.16 (0.01)	−22.19	**<0.001**
Group*grammaticality	0.02 (0.01)	1.49	0.136
Group*distractor	−0.01 (0.01)	−0.66	0.507
Grammaticality*distractor	−0.01 (0.02)	−0.80	0.423
Group*region	−0.19 (0.02)	−7.73	**<0.001**
Grammaticality*region	0.04 (0.01)	3.86	**<0.001**
Distractor*region	0.01 (0.01)	1.04	0.3
Group*grammaticality *distractor	0.01 (0.03)	0.28	0.78
Group*grammaticality *region	−0.04 (0.02)	−1.72	0.085

(Continued)

Table 4. (*Continued*)

	Estimates (SE)	t	p
Group*distractor*region	0.04 (0.02)	1.96	0.051
Grammaticality *distractor*region	0.01 (0.02)	0.26	0.793
Group*grammaticality *distractor*region	−0.00 (0.04)	−0.10	0.919

Note. Significant effects (*p* < .05) are in bold.

For the main set of experimental items, there were significant main effects of group and grammaticality. Grammaticality and region were also found to interact significantly. Nested contrasts revealed significant grammaticality effects at both regions, with larger effects at the spillover region (critical region estimate = 0.028 (0.008), *t* = 3.28, *p* = .001; spillover estimate = 0.068 (0.009), *t* = 7.26, *p* < .001). Interestingly, no significant interaction was observed between group and grammaticality. We also did not find any significant distractor effects.

As we did not observe significant interactions with group for reading times in Study 2, we do not report the L2 proficiency analysis here, but see the online supplement for a summary.

6.5. Discussion

The offline data replicated Study 1 with higher accuracy rates in grammatical than ungrammatical conditions. This may indicate response bias (Hammerly et al., 2019). Despite this, L2ers generally demonstrated knowledge of binding constraints. Indeed, we found L1ers overall less accurate than L2ers. Numerically, this L1/L2 difference is largely carried by the ungrammatical conditions, which L1ers accepted roughly 50% of the time. This may suggest that L1ers sometimes treated the plural reflexive as a generic,

singular form as has been observed with 'they' (Konnelly & Cowper, 2020; Speyer & Schleef, 2019). L2 accuracy also increased with higher proficiency.

In SPR, L2ers showed grammaticality effects similar to L1ers. This suggests that L2ers reduced grammaticality effects compared to L1ers for S-V agreement in Study 1 relates to the computation of agreement, rather than the encoding of number. Note also that the grammaticality effects we observed in both baseline and distractor conditions suggest that, even if 'themselves' is sometimes treated as generic, it is still also treated as plural. Finally, we did not find any significant interference effects. We discuss these results, along with Study 1, in detail below.

7. General discussion

We examined L2 sensitivity to syntactic constraints on S-V agreement and reflexives, assessing both grammatical knowledge (GJT) and syntactic processing (SPR), to test competing accounts of L2 processing. L2ers' performance across different tasks largely patterned with L1ers, and the differences that we observed were driven by proficiency. Implications of these findings are discussed below.

7.1. S-V agreement

Study 1 revealed L2ers' accurate untimed knowledge of S-V agreement constraints in line with previous findings (Chen et al., 2007; Lim & Christianson, 2015; Tanner et al., 2012). L2ers' reading times showed grammaticality effects at the spillover region which is contrary to some previous studies that showed null effects (Chen et al., 2007; Jiang, 2004). The L2 grammaticality effects were, however, significantly smaller than those in L1ers. However, sensitivity to grammaticality was modulated by L2 proficiency, with higher-proficiency L2ers exhibiting larger grammaticality effects, suggesting proficiency plays a role in L2 attainment during agreement processing (Hopp, 2006; Lim & Christianson, 2015).

Contrary to expectations, reading times did not show any significant similarity-based interference effects typically found in L1 studies (Dillon et al., 2013; Jäger et al., 2020; Wagers et al., 2009) and have been reported among L2ers (Tanner et al., 2012). Given previous findings, we are cautious in drawing strong conclusions about the lack of effects here, though note that null results also exist (Parker & An, 2018, relative clause conditions; Schlueter, 2017, Experiment 11; 2019). Based on our results alone, one might speculate that L1ers and L2ers favoured structural cues rather than number, but we believe this conclusion would however be premature given previous results and we are as such cautious in drawing an erroneous conclusion from a single study. Instead, our results highlight the need for replication in L1 and L2 research (Vasishth & Gelman, 2021).

For present purposes, despite this discrepancy with previous research, our most important finding is that we found no significant evidence that L2ers were more influenced by distractors than L1ers. Thus, in subject-verb agreement, our data did not support the claim that L2ers are more susceptible to interference than L1ers (cf. Cunnings, 2017), nor did we find evidence of shallow L2 processing of complex noun phrases (cf. Reifegerste et al., 2020).

7.2. Reflexives

L2ers showed knowledge of binding Principle A in our judgement tasks, consistent with previous findings (Felser et al., 2009; Felser &

Cunnings, 2012). Our SPR results suggest L2ers apply binding constraints during processing, and indeed our L2ers demonstrated larger grammaticality effects for gender mismatches than L1ers. This finding is contrary to previous studies (Felser et al., 2009; Felser & Cunnings, 2012) that suggested L2 reflexive processing is not initially constrained by Principle A. Across two experiments with gender and number manipulations, our results suggest both L1ers and L2ers rely more heavily on syntactic cues for reflexive resolution than gender, number or linear proximity.

Our results may differ from previous studies for various reasons. One possibility might be that the distractor was more discourse-prominent in Felser et al. and Felser and Cunnings, such that distractors only influence L2 processing more than L1 processing when they are discourse-prominent. Both these studies introduced the distractor in a lead-in sentence before the critical sentence, where it also occupied the discourse-prominent subject position. However, our study placed the distractor within a PP, a non-prominent discourse position. If discourse prominence influences L2 anaphora resolution (e.g., Cunnings, 2017; Felser, 2015, 2019), the different results between the current study and Felser et al. (2009) and Felser and Cunnings (2012), might be a result of the discourse prominence of distractor antecedents across studies. Teasing these issues apart is an avenue for future research. Despite this, our results still nevertheless suggest L2ers are not always more influenced by distractors than L1ers during anaphora resolution. The different findings between our experiments and previous studies could also be a result of the small sample sizes in previous research, which can overestimate effect sizes (Vasishth & Gelman, 2021).

The main effect of distractor for reflexives in Study 1 indicated shorter reading times in conditions with matching distractors. This effect was numerically larger in ungrammatical conditions, suggesting facilitatory interference in both groups. No significant distractor effects were found in Study 2, however. Together, these results suggest that although syntactic cues may be more highly weighted during reflexive resolution than gender/number, reflexive resolution is susceptible to interference (compare Cunnings & Sturt, 2014; Dillon et al., 2013; Jäger et al., 2020; Parker & Phillips, 2017). The fact we observed interference from gender-matching distractors in Study 1 but did not find this effect for number-matching distractors in Study 2, may suggest gender is a more highly weighted retrieval cue for reflexives than number. As this finding has not previously been observed, we are cautious in drawing strong conclusions here without further replication.

In sum, our results for reflexives did not indicate that L2ers are more susceptible to interference than L1ers (cf. Cunnings, 2017), nor did we find evidence consistent with shallow L2 processing (cf. Clahsen & Felser, 2006, 2018). Instead, while we do not rule out the possibility of L1/L2 differences in certain conditions, our results suggest L1 and L2 speakers process reflexives utilising similar memory retrieval mechanisms.

7.3. Number in agreement and reflexives

With regard to the reduced L2 sensitivity to grammaticality effects observed for S-V number agreement in Study 1, we did not observe significantly smaller grammatical effects for number violations during the processing of reflexives in Study 2. This suggests L2ers do not have difficulty with encoding number per se, but instead, the difficulty may lie in the processing of agreement. One possible explanation of this finding is that reflexive-antecedent dependencies generally have a stronger impact on sentence interpretation than agreement, and this may have additionally supported L2

processing. L2ers may focus more on feature (mis)match when it has a larger consequence on interpretation. We do not intend for this to imply that L2ers focus on semantics over syntax, however, as our L2ers generally demonstrated an understanding of structural constraints during processing. Note also, that L2 grammaticality effects in Study 1 interacted with proficiency, indicating more native-like performance in higher-proficiency L2ers, suggesting this L1/L2 difference in grammaticality effects in agreement is influenced by proficiency.

8. Conclusion

We conducted two large-scale studies investigating subject-verb agreement and reflexive resolution during L1 and L2 comprehension. Although some previous research has indicated L1/L2 differences in the application of syntactic constraints on linguistic dependency resolution, our results revealed that applying syntactic constraints during processing is not always more problematic for L2ers than for L1ers. We did not find that L2ers were more susceptible to memory-based interference than L1ers, nor did we find evidence of shallow L2 processing. Though we do not intend to restrict all L1/L2 differences to a single source, the L1/L2 differences in grammaticality effects we did observe were influenced by proficiency. In sum, our results suggest a similar use of syntactic constraints on linguistic dependencies by L1ers and L2ers during real-time sentence processing.

References

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4), 390–412. https://doi.org/10.1016/J.JML.2007.12.005

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3), 255–278. https://doi.org/10.1016/J.JML.2012.11.001

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects Models using lme4. *Journal of Statistical Software*, 67(1), 1–48. https://doi.org/10.18637/JSS.V067.I01

Benmamoun, E. (1992). *Functional and inflectional morphology problems of projection, representation and derivation* (PhD Thesis, University of Southern California).

Chen, L., Shu, H., Liu, Y., Zhao, J., & Li, P. (2007). ERP signatures of subject–verb agreement in L2 learning*. *Bilingualism: Language and Cognition*, 10(2), 161–174. https://doi.org/10.1017/S136672890700291X

Chomsky, N. (1981). *Lectures on government and binding*. Dordrecht, The Netherlands: Fori.

Chomsky, N. (1986). *Knowledge of Language: Its Nature, Origin, and Use*. Westport, CT: Greenwood Publishing Group.

Clahsen, H., & Felser, C. (2006). Grammatical processing in language learners. *Applied Psycholinguistics*, 27(1), 3–42. https://doi.org/10.1017/S0142716406060024

Clahsen, H., & Felser, C. (2018). Some notes on the shallow structure hypothesis. *Studies in Second Language Acquisition*, 40(3), 693–706. https://doi.org/10.1017/S0272263117000250

Cunnings, I. (2017). Parsing and working memory in bilingual sentence processing. *Bilingualism: Language and Cognition*, 20(4), 659–678. https://doi.org/10.1017/S1366728916000675

Cunnings, I., & Sturt, P. (2014). Coargumenthood and the processing of reflexives. *Journal of Memory and Language*, 75, 117–139. https://doi.org/10.1016/j.jml.2014.05.006

Cunnings, I., & Sturt, P. (2018). Coargumenthood and the processing of pronouns. *Language, Cognition and Neuroscience*, 33(10), 1235–1251. https://doi.org/10.1080/23273798.2018.1465188

Dallas, A., & Kaan, E. (2008). Second language processing of filler-gap dependencies by late learners. *Linguistics and Language Compass*, 2(3), 372–388. https://doi.org/10.1111/j.1749-818X.2008.00056.x

Dillon, B., Mishler, A., Sloggett, S., & Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: Experimental and modeling evidence. *Journal of Memory and Language*, 69(2), 85–103. https://doi.org/10.1016/j.jml.2013.04.003

Drummond, A. (2013). *Ibex Farm*. http://spellout.net/ibexfarm/.

Eberhard, K. M. (1997). The marked effect of number on subject-verb agreement. *Journal of Memory and Language*, 36(2), 147–164. https://doi.org/10.1006/JMLA.1996.2484

Felser, C. (2015). Native vs. Non-native processing of discontinuous dependencies. *Second Language*, 14, 5–19. https://doi.org/10.11431/secondlanguage.14.0_5

Felser, C. (2019). Structure-sensitive constraints in non-native sentence processing. *Journal of the European Second Language Association*, 3(1), 12–22. https://doi.org/10.22599/jesla.52

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

Felser, C., Sato, M., & Bertenshaw, N. (2009). The on-line application of binding principle A in english as a second language. *Bilingualism: Language and Cognition*, 12(4), 485–502. https://doi.org/10.1017/S1366728909990228

Foertsch, J., & Gernsbacher, M. A. (1997). In search of gender nertrality: Is singular they a cognitively efficient substitute for generic he? *Psychological Science*, 8(2), 106–111. https://doi.org/10.1111/J.1467-9280.1997.TB00691.X

Hammerly, C., Staub, A., & Dillon, B. (2019). The grammaticality asymmetry in agreement attraction reflects response bias: Experimental and modeling evidence. *Cognitive Psychology*, 110, 70–104. https://doi.org/10.1016/j.cogpsych.2019.01.001

Hopp, H. (2006). Syntactic features and reanalysis in near-native processing. *Second Language Research*, 22(3), 369–397. https://doi.org/10.1191/0267658306sr272oa

Hopp, H. (2014). Working memory effects in the L2 processing of ambiguous relative clauses. *Language Acquisition*, 21(3), 250–278. https://doi.org/10.1080/10489223.2014.892943

Jäger, L. A., Engelmann, F., & Vasishth, S. (2017). Similarity-based interference in sentence comprehension: Literature review and Bayesian meta-analysis. *Journal of Memory and Language*, 94, 316–339. https://doi.org/10.1016/j.jml.2017.01.004

Jäger, L. A., Mertzen, D., Van Dyke, J. A., & Vasishth, S. (2020). Interference patterns in subject-verb agreement and reflexives revisited: A large-sample study. *Journal of Memory and Language*, 111(104063), 1–21. https://doi.org/10.1016/j.jml.2019.104063

Jiang, N. (2004). Morphological insensitivity in second language processing. *Applied Psycholinguistics*, 25(4), 603–634. https://doi.org/10.1017/s0142716404001298

Konnelly, L., & Cowper, E. (2020). Gender diversity and morphosyntax: An account of singular they. *Glossa: A Journal of General Linguistics*, 5(1), 1–19. https://doi.org/10.5334/GJGL.1000

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13), 1–26. https://doi.org/10.18637/JSS.V082.I13

Kush, D. (2013). Respecting Relations: Memory Access and Antecedent Retrieval in Incremental Sentence Processing. *Unpublished doctoral thesis*, University of Maryland.

Lee, E.-K. R., & Phillips, C. (2022). Why non-native speakers sometimes outperform native speakers in agreement processing. *Bilingualism: Language and Cognition*, 26(1), 152–164. https://doi.org/10.1017/s1366728922000414

Lewis, R. L., Vasishth, S., & Van Dyke, J. A. (2006). Computational principles of working memory in sentence comprehension. *Trends in Cognitive Sciences*, **10**(10), 447–454. https://doi.org/10.1016/j.tics.2006.08.007

Lewis, S., & Phillips, C. (2014). Aligning grammatical theories and language processing models. *Journal of Psycholinguistic Research*, **44**(1), 27–46. https://doi.org/10.1007/s10936-014-9329-z

Lim, J. H., & Christianson, K. (2015). Second language sensitivity to agreement errors: Evidence from eye movements during comprehension and translation. *Applied Psycholinguistics*, **36**(6), 1283–1315. https://doi.org/10.1017/S0142716414000290

McDonald, J. L. (2006). Beyond the critical period: Processing-based explanations for poor grammaticality judgment performance by late second language learners. *Journal of Memory and Language*, **55**(3), 381–401. https://doi.org/10.1016/j.jml.2006.06.006

McElree, B. (2000). Sentence comprehension is mediated by content-addressable memory structures. *Journal of Psycholinguistic Research*, **29**(2), 111–123. https://doi.org/10.1023/A:1005184709695

Musabhien, M. (2008). *Case, Agreement and Movement in Arabic: A Minimalist Approach* (PhD Thesis, Newcastle University). Retrieved from https://theses.ncl.ac.uk/jspui/handle/10443/2046

Nicenboim, B., Vasishth, S., Engelmann, F., & Suckow, K. (2018). Exploratory and confirmatory analyses in sentence processing: A case study of number interference in German. *Cognitive Science*, **42**, 1075–1100. https://doi.org/10.1111/cogs.12589

Parker, D., & An, A. (2018). Not all phrases are equally attractive: Experimental evidence for selective agreement attraction effects. *Frontiers in Psychology*, **9** (Article 1566), 1–16. https://doi.org/10.3389/fpsyg.2018.01566

Parker, D., & Phillips, C. (2017). Reflexive attraction in comprehension is selective. *Journal of Memory and Language*, **94**, 272–290. https://doi.org/10.1016/j.jml.2017.01.002

Pearlmutter, N. J., Garnsey, S. M., & Bock, K. (1999). Agreement processes in sentence comprehension. *Journal of Memory and Language*, **41**(3), 427–456. https://doi.org/10.1006/jmla.1999.2653

Quick Placement Test:Version1. (2004). Oxford University Press.

R Development Core Team. (2019). *R: A Language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.r-project.org/

Reifegerste, J., Jarvis, R., & Felser, C. (2020). Effects of chronological age on native and nonnative sentence processing: Evidence from subject-verb agreement in German. *Journal of Memory and Language*, **111**(February 2019), 104083. https://doi.org/10.1016/j.jml.2019.104083

Roberts, L. 2013). Sentence processing in bilinguals. In Roger P. G. van Gompel (Ed.), *Sentence processing* (pp. 221–246). London, UK: Psychology Press. https://doi.org/10.4324/9780203488454

Schlueter, Z. (2017). *Memory retrieval in parsing and interference*. The University of Maryland.

Schlueter, Z. (2019). *No grammatical illusions with L2-specific memory retrieval cues in agreement processing*. Manuscript Sumbitted for Publication.

Speyer, L. G., & Schleef, E. (2019). Processing 'Gender-neutral' pronouns: A Self-paced reading study of learners of english. *Applied Linguistics*, **40**(5), 793–815. https://doi.org/10.1093/APPLIN/AMY022

Sturt, P. (2003). The time-course of the application of binding constraints in reference resolution. *Journal of Memory and Language*, **48**(3), 542–562. https://doi.org/10.1016/S0749-596X(02)00536-3

Tanner, D., Nicol, J., Herschensohn, J., & Osterhout, L. (2012). Electrophysiological markers of interference and structural facilitation in native and nonnative agreement processing. In A. Biller, E. Chung, & A. Kimball (Eds.), *Proceedings of the 36th annual Boston University conference on language development* (pp. 594–606). Somerville, MA: Cascadilla Press.

Vasishth, S., & Gelman, A. (2021). How to embrace variation and accept uncertainty in linguistic and psycholinguistic data analysis. *Linguistics*, **59** (5), 1311–1342. https://doi.org/10.1515/ling-2019-0051

Vasishth, S., Nicenboim, B., Engelmann, F., & Burchert, F. (2019). Computational models of retrieval processes in sentence processing. *Trends in Cognitive Sciences*, **23**(11), 968–982. https://doi.org/10.1016/j.tics.2019.09.003

Wagers, M. W., Lau, E. F., & Phillips, C. (2009). Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language*, **61**(2), 206–237. https://doi.org/10.1016/j.jml.2009.04.002

Zehr, J., & Schwarz, F. (2018). *PennController for internet based experiments (IBEX)*. https://doi.org/10.17605/OSF.IO/MD832