Highlights

- Agreement attraction is used as a diagnostic for the timing of morphosyntactic planning.
- Unaccusative and unergative verbs are tested to probe early versus late agreement planning.
- No verb-type differences are found in attraction errors or pause likelihood across experiments.
- Onset latencies suggest early number encoding for unaccusatives.
- Results support an early assignment of number diacritics but a late use of these diacritics to retrieve the correct form, which appears to be the locus of attraction.
- Attractor subjecthood and the nature of the modifier (restrictive versus non-restrictive) modulate attraction effects.

When do we plan agreement: Evidence from agreement attraction and unaccusativity

Utku Turk

Department of Linguistics, University of Maryland, College Park

Author Note

Utku Turk https://orcid.org/0000-0001-8011-1541

All available data, code, and experimental materials can be found in https://github.com/umd-psycholing/morph_planning.

Utku Turk, Department of Linguistics, University of Maryland, College Park, 1401 Marie Mount Hall, College Park, MD 20742, USA, Email: utkuturk@umd.edu

Abstract

This paper investigates the timing of the planning of verbal agreement in sentence production by examining agreement errors in environments where the syntactic relation between the subject and the verb is manipulated. Recent work in sentence planning has shown that when the syntactic relation between the verb and the subject is tighter (as in unaccusative verbs, where subjects function as their complements), participants diverge from their usual incremental planning and plan the verb before they begin uttering sentences, along with the subject, in sentences like *The octopus below the spoon is boiling* (Momma & Ferreira, 2019). More importantly, planning the verb early does not mean that the modifier NP, *spoon*, is planned early as well. We present two picture description experiments in which we manipulated the number of the modifier noun and the type of intransitive verb in sentences like *The doctor(s) by the wizard is running/boiling*. We argue that if agreement-related features are planned early and eagerly, we should not see agreement errors in unaccusative sentences, since the verb is planned before the modifier noun. However, if agreement is planned later, that is, not when the verb is retrieved, we should observe agreement errors with both unaccusative and unergative verbs; the latter have subjects as modifiers rather than complements. Both experiments ($N = 74$, $N = 59$) showed that agreement errors and agreement-related timing effects are comparable in both unaccusative and unergative verbs. However, our onset latencies (that is, sentence initiation measures) consistently showed that the number of the second noun affected the planning of the verb early on, but only in unaccusatives. We take our results as evidence that agreement is planned late, given the comparable patterns of errors. However, these error patterns might be related to factors other than the assignment of the number feature on the verb, such as the retrieval of the correct auxiliary form.

Wordcount: 46686 total words (42979 in the body & 3127 in the references)

Keywords: planning, agreement, sentence production, morphology

When do we plan agreement: Evidence from agreement attraction and unaccusativity

Table of contents

List of Figures

List of Tables

When do we plan agreement: Evidence from agreement attraction and unaccusativity

1 Introduction

When we speak, we start with a message that we wish to express. In its simplest version, it is assumed that a notion of events or ideas is filtered to form a preverbal message, which is then mapped to a syntactically and semantically specified representation. This representation is subsequently used to access the phonological representation.

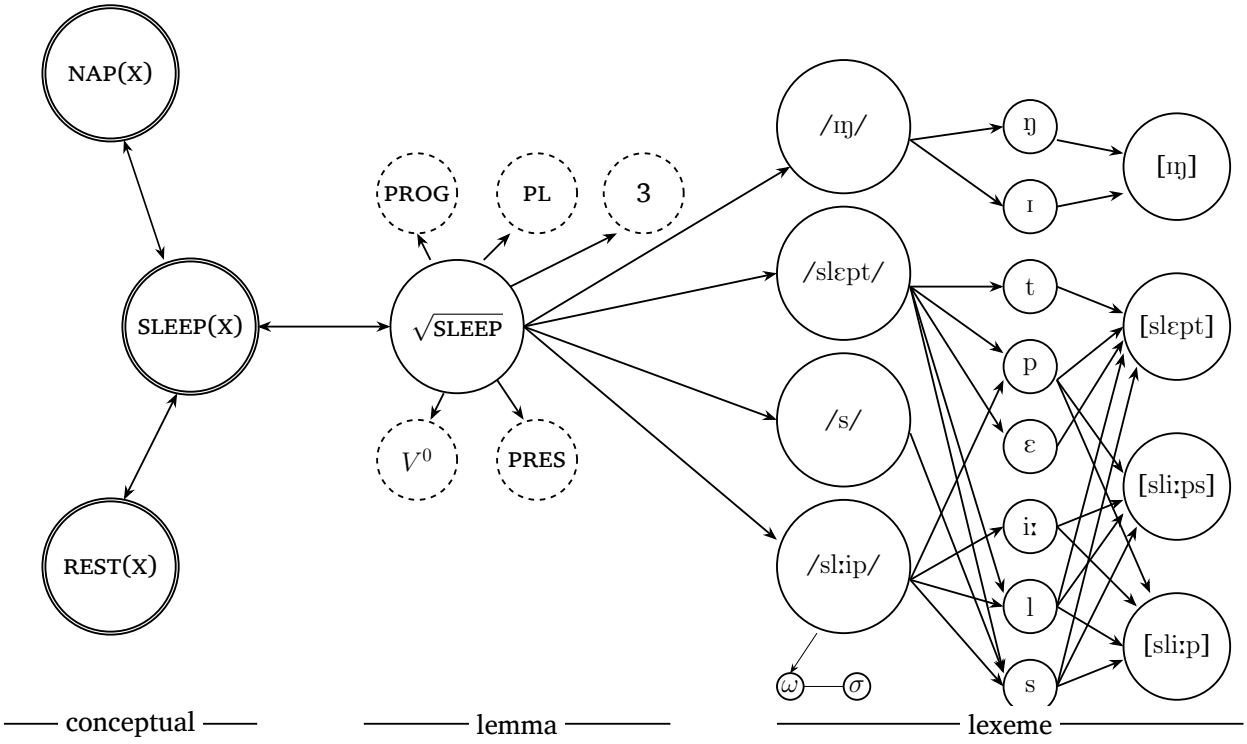
One influential model is the Lemma Model proposed by Levelt (1989) and Levelt et al. (1999). This model assumes that language production operates over a lexical network that certain functions can access. This lexical network consists of three levels: conceptual (notional), lemma (syntactic), and lexeme (form). See Figure 1 for a schematic representation of a fragment of this lexical network. The top level in Figure 1 represents the concept node for the concept SLEEP(X), which is connected to other concepts such as NAP(X) or REST(X). Once the concept node SLEEP(X) is activated, it spreads activation to the lemma node, the second layer in this network. When the lemma $\sqrt{\text{SLEEP}}$ is activated, the relevant information for syntactic computation, such as lexical category, tense, or person, becomes available in the form of diacritics.[1] Activation from the lemma $\sqrt{\text{SLEEP}}$ then spreads to the phonological representation, which is the lexeme /sliːp/, along with metrical and syllable information. This lexeme information is then used to access articulations such as [sliːp], [sliːp], [sliːpɪŋ], or [slɛpt], depending on the diacritics.

In this model, when speakers utter the sentence *the cats are sleeping* in a conversation, they start by boiling the entire frame of reference down to a message that involves the entity CAT and the action SLEEP(X). Even though there might be other

[1] We are aware that the square root ($\sqrt{\ }$) is used in other linguistic theories, such as Distributed Morphology, to explain morphological structure. Here, we use the square root notation solely to indicate the lemma node in Levelt's model. Our usage does not imply any connection between the two theories.

Figure 1

Adaptation of the lexical fragment from Levelt's model of speech production. The top level represents the conceptual node, the second level represents the lemma node, and the bottom level represents the lexeme node.



elements in the reference frame, such as the MAT that the cats are sleeping on, speakers can ignore these additional elements for the time being (Bock & Ferreira, 2014). Conceptual elements that are not ignored are then mapped to representations called lemmas, which are specified with relevant syntactic information for building a hierarchical structure. The lemma $\sqrt{\text{CAT}}$, for example, is specified as a noun (N), along with other features such as definite (DEF) and plural (PL). According to their specifications, these lemmas are placed into a syntactic frame that is compatible with the message, [$_S$ [$_{NP}$ cats] …]. The corresponding phonological representation, or lexeme, is then accessed: /kæts/. Along with $\sqrt{\text{CAT}}$, the determiner $\sqrt{\text{THE}}$ is also accessed at the lemma level and is later pronounced before *cats* as [ðəkʰæts] (Levelt, 1989; Levelt et al., 1999; Schriefers et al., 2002). While the phonological or articulatory information for the lemmas $\sqrt{\text{THE}}$ and $\sqrt{\text{CAT}}$ is accessed, speakers plan the next utterance related to the

concept SLEEP(X). For example, while [ðəkʰæts] is being uttered, the lemma $\sqrt{\text{SLEEP}}$ might be retrieved with its additional specifications, such as present tense and plural agreement, and then placed in the stipulated syntactic frame, [$_S$ [$_{DP}$ the [$_{NP}$ cats]] [$_{VP}$ sleep]]. Its lexeme, according to the additional specifications, is then accessed, /ər slipɪŋ/, and concatenated to the previous [ðəkʰæts] to form [əsliːpɪŋ]. In this simple model, speech production follows a linear process with cascading stages: some of the planning for the immediately upcoming word can be carried out while another process is still being completed for the previous word (Kempen & Hoenkamp, 1987).

The present study investigates when number agreement features are computed during a sentence production process like the one described above. Agreement errors suggest that agreement planning may be more dynamic than previously assumed. The idea that the marking of grammatical features such as number or tense is trivial has been called into question by a large body of evidence on agreement errors in production. A particularly well-known phenomenon is *agreement attraction*, in which speakers produce number agreement errors in complex subject phrases. For instance, given a preamble like *The key to the cabinets…*, speakers often complete the sentence with a plural verb, producing **The key to the cabinets are rusty* (Bock & Miller, 1991). Similarly, in comprehension, participants often find comparable sentences grammatical due to either an erroneous retrieval of *cabinets* as an agreement controller (Wagers et al., 2009) or an erroneous representation of the entire subject phrase (Eberhard et al., 2005). This erroneous production or acceptability, known as agreement attraction, has been replicated in many other languages (Avetisyan et al., 2020; Lago et al., 2015; Ristic et al., 2016; Tucker et al., 2015). However, it is also noted that the magnitude of these attraction effects is affected by the distributive reading of the distractor (Humphreys & Bock, 2005), the syntactic structure (Franck et al., 2006), form syncretism (Slioussar, 2018), and the semantic similarity between the subject and the distractor (Solomon & Pearlmutter, 2004). Such errors and their interaction with other

parts of production and comprehension suggest that agreement computation is not a simple, mechanical process but rather one that is sensitive to structural complexity and potentially susceptible to interference. These findings complicate the idea that feature marking is trivial and instead point to the need for more precise models of how agreement features are computed during production.

We are going to test the timing of agreement by integrating our agreement question with recent findings on advance planning influenced by syntactic factors. The order in which constituents in a sentence are planned is subject to syntactic constraints and does not have to follow a linear sequence (Momma & Ferreira, 2019). For example, verbs are planned before objects are articulated, even if the verb appears after the object in linear order. It is no surprise that syntax influences what speakers choose to say (Bock, 1986; Pickering & Branigan, 1998; Xiang et al., 2019). More recently, it has been shown that sentences identical in their linear arrangement but differing in syntactic structure exhibit different timing patterns in planning (Momma et al., 2016; Momma & Ferreira, 2019; Momma & Yoshida, 2023; Zhao et al., 2024).

For example, in Japanese, speakers can omit subjects or objects when these are recoverable from context. Native speakers of Japanese have been shown to plan the verb before uttering the object in OV sentences where the subject is dropped. However, this early planning of the verb is not observed in SV sentences, even though in both SV and OV structures the verb comes after the initial element (Momma et al., 2016). Similar evidence on timing shows that in sentences like *The cats on the mats are freezing*, participants plan the verb *freeze* together with the noun *cats* before they begin speaking, and do not plan the intervening noun *mats*. This departure from strictly linear planning is evident only for verbs like *freeze*, which have subjects that share syntactic and semantic properties with canonical objects cross-linguistically, compared to verbs like *sleep*, whose subjects do not share these properties (Momma & Ferreira, 2019).

Our study builds on these two observations and aims to address the question of

when morphosyntactic information, such as number or gender, is planned. Manipulating timing differences in verb planning (Momma & Ferreira, 2019; Momma & Yoshida, 2023) allows us to contrast two possible timings for when morphological information is planned. One possibility is that morphological information is immediately available when the lemma is accessed. Another possibility is that morphemes containing morphosyntactic information, such as number, have an independent representation and are planned later in the process, even though their host verb is accessed earlier. These two accounts make different predictions about whether participants might produce agreement errors in sentences that show different planning patterns, such as *The cat on the mats is freezing* and *The cat on the mats is sleeping*.

Spoilers for size of planning units

- Extended picture–word interference tasks are widely used to investigate the timing of planning in word and sentence production.
- The extent of advance planning is often described in terms of the size of the planning unit.
- Although evidence in the nominal domain remains mixed, findings are more robust for verbs.
- Stronger syntactic or conceptual links between the verb and the subject appear to facilitate advance verb planning.

2 Size of planning units in speech

In this study, we aim to investigate the timing of morphosyntactic planning for the verb. To do so, we begin by reviewing previous methods for examining when different parts of a sentence are planned and what these studies have revealed. A growing body of work has shown that sentence planning does not always proceed in a strictly linear or uniform fashion (Griffin & Ferreira, 2006; Hwang & Kaiser, 2014; Momma et al., 2016; Sauppe, 2017; Schriefers et al., 1998).

One method that has been used is a simple scene description task. Kempen and Huijbers (1983) presented participants with line-drawing pictures and asked them to name the action, the actor, or both. The type of target utterance was signaled to participants by displaying “V”, “S”, “VS” or “SV” on the screen. For sentential utterances, they were asked to produce either a simple sentence with SV order (*man greets*) or a slightly more complex sentence with VS order, such as *here greets the man* in Dutch. They aimed to compare how long it takes participants to begin uttering SV or VS sentences compared to single-word utterances.

Their results showed that there was no onset latency difference between the SV and VS conditions, and the latencies for the sentential conditions were similar to the condition in which participants only named the action. They interpreted these results as

evidence for parallel retrieval of the subject and verb lemmas in sentential conditions before sentence onset, given the comparable timings for sentence and verb naming. However, it remains unclear whether the nature of the representation for the verb and the subject was the same. It is also not clear whether some of the processing required to utter the second word can be completed while uttering the first word. In other words, these results are not conclusive regarding the exact timing of verb planning.

2.1 Decomposing planning

To investigate the timing of planning during sentence production, many studies have used the picture–word interference (PWI) paradigm. This methodology has proven especially useful for identifying when different levels of linguistic representation, such as conceptual, syntactic, and phonological information, are activated. In a typical PWI task, participants name a pictured object (e.g., *cat*) while ignoring a distractor word (e.g., *dog* or *cash*) that appears near or on the picture, or is presented auditorily. The latency to initiate speech reflects the interaction between the target and the distractor, revealing which planning processes are active at the moment of naming.

One robust finding from PWI is *semantic interference*: naming is slower when the distractor is semantically related to the picture (*cat–dog*) compared to when it is unrelated (*cat–train*) (Bürki et al., 2020; Rosinski et al., 1975). This effect occurs when the distractor appears around the same time as the picture, typically within a window of –150 ms to +150 ms stimulus onset asynchrony (SOA). This timing suggests that semantic interference arises during conceptual or lemma-level planning, when multiple related candidates compete for selection (Damian & Martin, 1999; Jescheniak & Schriefers, 2001).

In contrast, *phonological facilitation* occurs when the distractor shares sound structure with the target (*cat–cash*), leading to faster naming latencies (Meyer & Schriefers, 1991). Unlike semantic interference, phonological facilitation is reliably observed when the picture precedes the distractor (positive SOA), suggesting that

phonological forms are retrieved later, after lemma selection has occurred (Jescheniak & Schriefers, 2001; Schriefers et al., 1990).

Together, these effects show that the timing and nature of interference depend on the type of relationship between the target and the distractor, as well as on their relative timing. Crucially, this makes the PWI paradigm a powerful tool for probing when different components of language production are activated. In the current study, we build on this logic to investigate whether agreement-related features on the verb, such as number, are computed early, during lemma retrieval, or later, during morphophonological encoding.

2.2 Uttering more than a word

An extended version of the PWI task (ePWI) has been used to study planning in sentence production (Hwang & Kaiser, 2014; Kempen & Huijbers, 1983; Meyer, 1996; Momma et al., 2016; Momma & Ferreira, 2019; Schnur, 2011; Schriefers et al., 1998). In studies using ePWI, participants were either presented with a set of entities and asked to produce a coordinated phrase such as *cats and mats* (Meyer, 1996), or shown a scene and asked to describe it with a simple sentence such as *The cats on the mat are sleeping* (Schriefers et al., 1998). The main aim of these experiments using ePWI was to test the size of the planning units and whether speakers plan words in advance.

For instance, Meyer (1996) tested advance planning in phrasal utterances using ePWI. She presented participants with pairs of objects and asked them to produce sentences using a conjunction, such as *the cat and the mat*, or sentences like *the cat is on the mat*. She tested whether auditory distractors related to the second word affected onset latency, that is, how long it took participants to begin speaking. If the distractor word was semantically related to the second target word, in this case *mat* (target) and *rug* (distractor), participants were expected to show increased onset latency compared to the unrelated condition *mat* (target) and *chair* (distractor). This increased latency would indicate whether participants plan larger units than a single word, meaning they

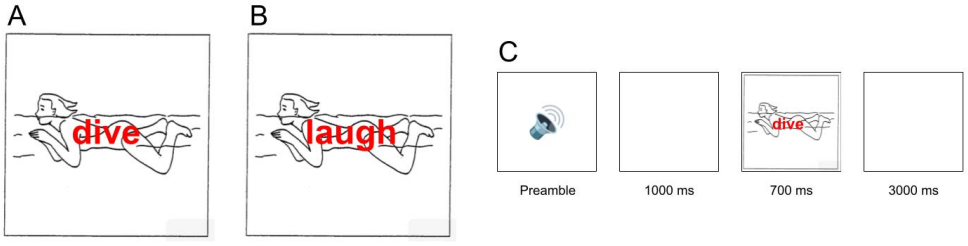
plan both *cat* and *mat* before beginning to speak.

Meyer's (1996) results suggest that speakers do not engage in extensive advance planning in this context. She found that semantic relatedness affected latency only when the distractor was related to the first noun (*cat* and *dog*), but not to the second noun (*mat* and *rug*). Thus, her results suggested that the second noun was not included in the planning scope before sentence onset.

Understanding the timing of verb lemma planning in sentences is crucial for investigating the planning of verbal morphosyntax. Schriefers et al. (1998) used ePWI to investigate advance planning of the verb in SV sentences. Building on earlier work by Kempen and Huijbers (1983) and Meyer (1996), they tested whether the semantic relatedness between a distractor word and a target verb affects the time it takes to initiate a sentence. In their task, native German speakers described scenes using simple sentences. A distractor word, either semantically related (e.g., *dive* for *swim*) or unrelated (e.g., *laugh* for *swim*), was superimposed on the picture. They examined whether utterance onset latency was influenced by the relationship between the distractor and the verb, comparing conditions in which the verb followed the subject (verb-final) to those in which the verb appeared earlier in the sentence (verb-initial). The experimental design and sample stimuli are shown in Figure 2.

Figure 2

*Simplified schematic of the experiments in Schriefers et al. (1998). **(A)** Example of a semantically related condition. **(B)** Example of a semantically unrelated condition. **(C)** Participants first heard the preamble of the sentence, after 1000 ms, they saw the scene with a superimposed word for 700 ms. Participants were asked to describe the scene as soon as they have seen the scene. After the picture disappear, they had additional 3000 ms before the next trial.*



Schriefers et al. (1998) showed that a related distractor verb affected sentence onset latency only in verb-initial follow ups, but not in verb-final follow ups. The absence of semantic interference in the verb-final conditions was taken as evidence that participants did not engage in advance planning. The retrieval of the verb lemma to be produced immediately was influenced by the distractor word; however, in conditions where the verb was to be uttered sentence-finally, the verb was not retrieved before the sentence began. Contra Kempen and Huijbers (1983), they argued that the verb is not obligatorily planned early in the sentence production process and that verb planning is delayed until the verb needs to be produced. However, as we will see next, Schriefers et al.'s (1998) findings on verb planning likely depended on the type of verbs used and the structure of their sentences.

2.3 Structurally guided advance planning

We will now review ePWI work by Momma and colleagues that is most directly relevant to the current study, as they investigated the timing of all the relevant lemmas in the sentence structures most commonly used to elicit agreement attraction in production, namely, sentences with PP-modified subjects. Their results show that the earlier findings of Schriefers et al. (1998) cannot be generalized to all verbs. The timing of verb planning is not uniform across different types of verbs. Momma and his colleagues showed that the planning procedure and the scope of planning are affected by the structural properties of the target sentence. Momma et al. (2016) demonstrated that verbs are planned before the object noun, but not before the subject noun. Momma and Ferreira (2019) showed that verbs are planned early when their subjects are patient-like arguments rather than agent-like arguments.

The first study focused on Japanese sentence production. In a series of three experiments in Japanese, Momma et al. (2016) showed that verbs are planned before the articulation of objects but not subjects. The scenes they used included target sentences with either transitive sentences like (1a) or intransitive sentences like (1b).

Since it is common to drop the subject in Japanese, both possible descriptions have similar linear templates: a noun followed by a verb.

(1) a. *pro* *neko-o* *naderu*
cat-ACC pet
'(She) pets the cat.'

b. *innu-ga* *hoeru*
dog-NOM howl
'The dog howls.'

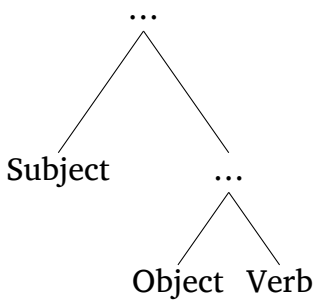
Since we build on ePWI and semantic interference as they did, it is important to note some crucial controls that ensure the task works as intended. One essential control is verifying that the verbal distractors reliably produce semantic interference. To that end, they prompted participants to name verbs in isolation, similar to previous PWI studies. The scenes they used included either a transitive verb such as *pet* or an intransitive verb such as *howl*. Each scene also included a superimposed word, either semantically related or unrelated to the verb. They confirmed that when the superimposed word was related to the target verb (*pet* and *rub*; *howl* and *cry*), participants took longer to name the actions compared to the unrelated distractor condition (*pet* and *sound*; *howl* and *break*).

After confirming that the distractors worked as intended, they tested whether this effect still occurred when participants produced simple Japanese sentences instead of naming verbs in isolation. They asked participants to describe the scenes using only two words. Japanese, a strongly verb-final language, allows sentences without overt subjects. This property enabled Momma et al. (2016) to compare the effect of semantic interference on the verb between SV (intransitive) and OV (transitive) utterances. They found that participants took similar amounts of time to begin sentences in the SV condition, regardless of semantic relatedness. In other words, the semantic interference effect observed when producing verbs in isolation did not appear in sentential utterances, consistent with Schriefers et al. (1998). However, participants took longer

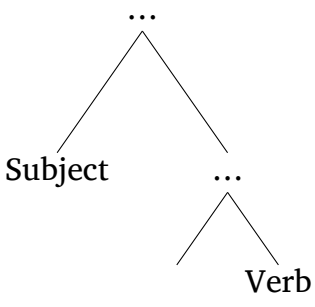
to initiate sentences in the OV condition when the distractor was semantically related. This difference in semantic interference between OV and SV conditions suggests selective early planning of verbs. This initial study by Momma et al. (2016), by showing that syntactic structure determines the scope of sentence planning, set the stage for their subsequent work on unergative and unaccusative verbs, which our study builds upon.

Perlmutter (1978), in his seminal work, proposed that intransitive verbs can be divided into two groups: unergatives and unaccusatives. His primary evidence came from cross-linguistic data on intransitive verb passivization. He demonstrated that certain intransitive verbs in German, such as *walk*, *dance*, or *sleep*, can appear with passive morphology. These verbs, called unergatives (structure in 3), have subjects that are similar to the subjects of transitive verbs (2) in terms of syntactic position and thematic roles. In contrast, unaccusative verbs (4) such as *melt*, *freeze*, or *fall* do not appear with passive morphology. This restriction is argued to stem from the absence of a noun phrase in the canonical subject position. The supposed subjects of unaccusative verbs share properties with the objects of transitive verbs: they are considered the complements of verbs and typically have patient-like thematic roles. In short, it has been theorized that unergative verbs (structure in 3) lack an internal patient-like argument (deep object), while unaccusative verbs (structure in 4) lack an external agent-like argument (deep subject). While the subjects of unergative verbs resemble subjects of other verbs, the subjects of unaccusative verbs show object-like properties.

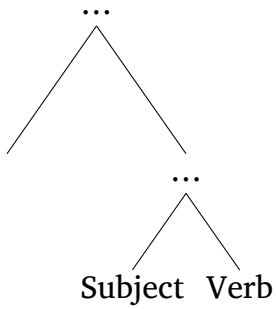
(2) Transitive Verbs



(3) Unergative Verbs



(4) Unaccusative Verbs



Remember Momma et al.'s (2016) findings: verbs are planned before the object,

but not before the subject, even though both were the only elements preceding the verb in their experiment. If Momma et al.'s (2016) findings are related to the difference between internal and external arguments, one might expect a similar distinction in unergative and unaccusative sentences. Their results, together with Perlmutter's (1978) categorization, raise the possibility that verbs may be planned before the subjects of unaccusatives, which resemble objects in transitive sentences. In contrast, verbs in unergative sentences would not be planned before their subjects, which align with the subjects of transitive verbs.

Momma and Ferreira (2019) tested this hypothesis in English using an extended PWI task. In a series of experiments, they manipulated the target verb's category: unergatives, whose subjects were agentive, and unaccusatives, whose subjects were non-agentive, such as themes. The unergative verbs included verbs such as *swim*, *walk*, and *wink*. The unaccusative verbs included verbs such as *boil*, *fall*, and *melt*. Their target sentences, shown in (5), were slightly more complex than in previous studies. In addition to a subject and a verb, the target sentences also included a prepositional phrase. This prepositional phrase was elicited by the presence of another object in the scenes that was either located below or above the subject head.

(5) a. **Unergative:** The octopus below the lemon is swimming.

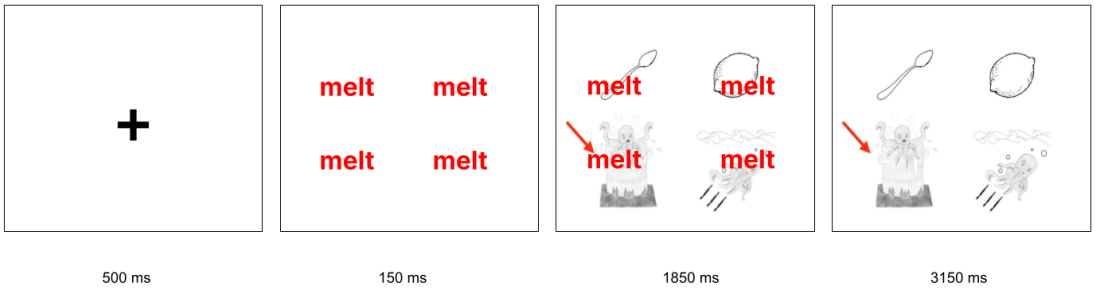
b. **Unaccusative:** The octopus below the spoon is boiling.

Similar to previous ePWI experiments, the scenes included superimposed words; however, Momma and Ferreira (2019) did not use only verbs as distractors, but also included nominal distractors. While the verbal distractors were either semantically related or unrelated to the verb of the sentence, the nouns were manipulated for semantic relatedness with respect to the second noun. An example trial from their experiment and their experimental procedure are shown in Figure 3.

Their results confirmed that, depending on the verb type, the scope of planning included the verb. The utterance onset of sentences with unaccusative verbs (which lack

Figure 3

Momma and Ferreira's (2019) experimental procedure in Experiment 1-4. Each trial starts with a fixation cross for 500 ms. Following the cross, distractors were appear and stays in the screen alone for 150 ms. After 150 ms, the scene appears behind the distractor words. With the scene, participants hear a click and prompted to describe the scene that is marked with a red arrow. Distractor words stay on the screen for 2000 ms, and the scene stays in the screen total of 5000 ms.



an external argument or deep subject) was affected by the semantic relatedness of the verbal distractor. When participants produced sentences like *The octopus below the spoon is boiling*, they were slower when the superimposed word was semantically related, such as *melt*, compared to an unrelated word like *fall*. However, this slowdown occurred only in unaccusative sentences. When participants were prompted to produce sentences like *The octopus below the lemon is swimming*, they took similar amounts of time regardless of whether the distractor was unrelated (*wink*) or related (*run*). These results demonstrate that verb planning is not uniform across different syntactic structures. Schriefers et al.'s (1998) findings on verb timing in German with agentive verbs cannot be generalized to all verbs in sentence production. Verbs are planned in advance in certain syntactic configurations.

Another important aspect of Momma and Ferreira's (2019) study concerns their use of nominal distractors. Unlike previous ePWI studies, Momma and Ferreira's (2019) target sentences included a complex subject phrase. It was important to test whether only the verb was planned in advance, or whether the entire utterance, including the subject, second noun phrase, and verb, was planned before speaking. To test this, they

included nominal distractors in their experiment.

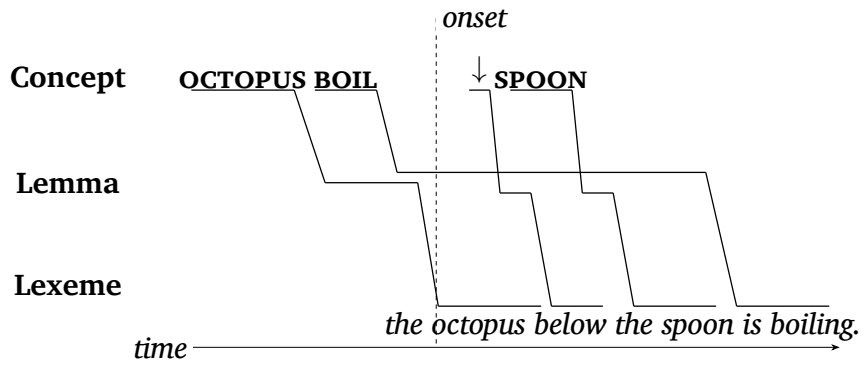
In the nominal distractor conditions, they manipulated the semantic relatedness of the distractor relative to the second noun phrase in the sentence. For example, when participants produced *The octopus below the spoon is boiling*, they saw *knife* as a related distractor and *apple* as an unrelated distractor. They did not observe any onset latency differences based on semantic relatedness in the nominal conditions. This lack of effect was consistent for both unaccusative and unergative scenes. Even though the verb was planned early in unaccusative sentences, the noun within the prepositional phrase was not planned before speech onset. In other words, participants planned the verb and the head noun but excluded the prepositional phrase from this initial planning.

Our work builds on these two findings: (i) certain verbs are planned early and (ii) other intervening nouns between the subject head and the verb are excluded from this initial planning process. Thus, during sentence planning, there is a time window in which only the subject head, *octopus*, and the verb, *boil*, are retrieved, while the noun within the prepositional phrase, *spoon*, is not yet retrieved.

Figure 4 illustrates a simplified version of this planning process. The y-axis represents the levels of planning: Conceptualization; Lemma, which includes syntactic and lexical retrieval; and Lexeme. The x-axis represents time, and the moment of speech onset is marked with a dashed line. The Concept level includes the concepts OCTOPUS, BOIL, the spatial relation below (↓), and SPOON. Momma and Ferreira's (2019) results suggest that only OCTOPUS and BOIL are planned before speech onset. Before participants begin uttering the phrase *the octopus*, they have access only to these two lemmas. While the subject head is being pronounced, participants complete the necessary planning for the prepositional phrase *under the spoon*. After the entire subject phrase is produced, the verb representation that has been maintained is sent to the articulator.

Figure 4

Planning procedure of the sentence with an unaccusative verb, such as 'The octopus under the spoons is boiling,' following Momma and Ferreira's (2019) findings.



Spoilers for functional elements in planning

- We review how functional elements like determiners and auxiliaries are planned, since morphosyntactic agreement surfaces on auxiliaries in our experiments.
- The lemma model distinguishes diacritics based on whether they are conceptual, inherent, or dependent—only some are planned with their host.
- Determiner planning has been more extensively studied; findings are mixed on whether gender diacritics compete at the lemma or phonological level.
- Some evidence suggests determiners are planned at the lemma level; others argue selection competition happens only at the form level.
- Auxiliary verbs are less studied, but tip-of-the-tongue data suggest their selection can occur independently of full verb retrieval.
- Both determiners and auxiliaries may be introduced with or without lemmas, but current models say little about the timing of these planning processes.

3 Planning the functional elements

To understand the timing of agreement computation in our experiments, where number agreement appears on the auxiliary verb, we must first consider how functional elements like determiners and auxiliaries are planned in sentence production. Although classic models such as the Lemma Model specify how lemmas carry syntactic features (e.g., number, definiteness), they are less explicit about how these features are morphologically realized on downstream elements like auxiliaries. For instance, in producing *The cats are sleeping*, speakers must retrieve *are* in agreement with the plural subject *cats*, but it remains unclear when this agreement-marked auxiliary is introduced into the planning process. This question extends to other functional items as well: when are determiners like *the* or *some* selected, and how closely are they tied to the noun's features? In what follows, we first review what is known about the timing of determiner planning, an area where psycholinguistic models and empirical data are more developed, before turning to auxiliary verbs and the implications for inflectional

morphology.

The first point to note is that under the Lemma Model (Levelt, 1989; Levelt et al., 1999), most of the information relevant to our study is implemented through diacritics. It is therefore crucial to distinguish the different types of diacritics present in the system, which are not always clearly described. There are three distinct categories of morphosyntactic diacritics that help clarify possible processes in sentence planning: inherent, conceptual, and dependent morphosyntax. For example, gender information in German is an inherent characteristic of a noun, independent of the noun's conceptual properties. In contrast, number information is not an inherent characteristic of a noun; rather, it is inferred from the numerosity represented at the conceptual level. In addition to these two types, there are morphosyntactic diacritics that depend on other elements in the sentence. For example, the number feature on the verb in English depends on the number feature of the subject noun.

In the case of the verb *sleep*, information such as being an intransitive verb with an agentive subject, as opposed to *fall asleep*, is inherent to the verb. The tense (PRES) and aspectual information (PROG) are provided at the conceptual level, as they are part of the intended message. However, the number feature (PL) results from a syntactic dependency between the subject noun and the verb. Since the subject is *the cats* and its number is plural, the verb is marked with a plural diacritic.

It is also important to note that morphosyntactic features such as number, gender, or definiteness are not always phonologically realized on the lemma that carries them. For example, in German, a noun lemma may be specified for gender, but this feature is expressed phonologically not on the noun itself, but on the accompanying determiner (e.g., *der*, *die*, *das*). In this paper, we are specifically concerned with the planning of number features that are reflected on auxiliary verbs. In English, number is morphophonologically realized in different ways: sometimes on the main verb (*sleeps* vs. *sleep*), and sometimes on an auxiliary verb (*is* vs. *are*), depending on the aspectual

structure of the sentence. Understanding when and how these features are activated and mapped onto their surface forms is crucial for interpreting agreement planning effects.

It is widely assumed that functional elements such as *the* or *are* do not have corresponding nodes at the conceptual level of the production system (Levelt, 1989; Roelofs, 1992). Instead, these elements are introduced at the lemma level, where they carry morphosyntactic diacritic information. While the Lemma Model does not fully specify how functional items are selected, it is generally agreed that they are not planned in competition with other conceptual candidates. However, questions remain about how and when their associated diacritics, such as number, gender, or definiteness, are mapped to surface forms. In the remainder of this section, we first review evidence on the planning of determiners and then discuss auxiliary verbs, with a focus on how each reflects morphosyntactic features during production. It is important to note that the studies reviewed here focus primarily on inherent diacritics rather than dependent ones.

3.1 Determiners and Gender Congruency

Determiners have attracted the most attention in the literature concerning the planning of functional elements. Given that none of the psycholinguistic models of language production assume category-specific planning procedures, the findings for determiners should inform our understanding of the planning of other functional elements. Research on determiner planning has focused primarily on how the features of the lemma are reflected in the determiner: whether they are mapped directly from the lexical representation to the determiner without competition, or whether the available features are selected among competing options.

According to the Lemma Model, when a conceptual node is activated in the network, it sends activation to the corresponding lemma node (Levelt, 1989; Roelofs, 1992). As discussed earlier in 2, semantic interference effects are thought to result from this spreading activation. In the case of semantically related distractors, the activation

of competing lemmas by the superimposed distractor word creates competition with the target lemma, which slows the retrieval of the target lemma. One prediction that follows from this logic is that when two elements are present during the experiment—a word and a picture—both their lemmas will be accessed, and their features will also be activated. If diacritic specification is a competitive process, then the different features of the superimposed word and the picture can both be candidates for insertion into the diacritic representation. This competition should be reflected in the time it takes to initiate the sentence, consistent with findings from other PWI studies.

Schriefers (1993) tested the prediction that sharing features would affect the retrieval of the target lemma using a PWI experiment. In his experiments, he manipulated the gender information of target words in Dutch. Dutch native speakers were prompted to describe the entity on the screen using a simple noun phrase (either Det + Adj + Noun or Adj + Noun). Dutch has two genders, which surface differently only in singular nouns. In phrases with determiners, gender information is expressed as different determiners, *het* or *de*, for neuter and common gender respectively, as shown in (6). In phrases without determiners, the same gender information appears as different suffixes, -Ø or -ə, for neuter and common gender respectively, as illustrated in (7).

(6)

a.	*het*	*rode*	*huis*
	hɛt	rodə	həys
	DET.N	red	house
b.	*de*	*rode*	*tafel*
	də	rodə	ta:fəl
	DET.C	red	table

(7)

a.	*rood*	*huis*	
	rot	həys	
	red[N]	house	
b.	*rod-e*	*tafel*	
	rod-ə	ta:fəl	
	red-C	table	

Schriefers (1993) measured utterance onset latencies in conditions where the superimposed word and the target word shared the same gender, and compared them to conditions where the genders of the superimposed and target words did not match. He tested both types of structures presented in (6–7) and found that participants were slower to begin speaking when the genders of the superimposed and target words

mismatched. This effect was present across both types of structures, regardless of whether a determiner was included. Schriefers (1993) interpreted these results as evidence for lemma-level competition in determining the gender of a noun. When different genders are activated, the specification of the noun lemma is delayed compared to when both the target and the distractor activate the same gender information.

Schiller and Caramazza (2003), on the other hand, argued that the retrieval of features for the diacritic is not a competitive process and that Schriefers' (1993) findings can be explained by competition at the form selection stage. Using PWI experiments in Dutch and German, they replicated the gender congruency effects for singular nouns: when the gender of the superimposed and target word differed, participants were slower to name the pictures. However, the same slowdown did not occur for plural nouns. An important difference between singular and plural nouns is the syncretism among different genders, as shown in Table 1. In German singular forms, different genders have distinct morphological reflexes on the determiner (der_{masc}, die_{fem}, das_{neu}). By contrast, all genders share the same determiner form in plural (die_{masc}, die_{fem}, die_{neu}), as in Dutch.

Table 1
German definite articles

	Masculine	**Feminine**	**Neuter**
Singular	der	die	das
Plural	die	die	die

According to Schiller and Caramazza (2003), abstract features for diacritics become available and are assigned at the lemma level automatically when the lexical node is selected. They argued that the competition reflected in utterance onset latency arises from competition at the level of phonological realization. For instance, when participants see a picture of a table and need to produce the word *Tisch* and the distractor item is *Wand*, both *der* and *die* articles are activated, and the speaker must choose between them. The process of selecting between these articles is reflected in the

utterance onset latency, rather than competition between different gender features during diacritic specification. In cases where both words are plural, only the article *die* is activated, even when the distractor word and the picture have different genders. Therefore, when the utterance is plural, there is no competition between forms and thus no difference in onset latency.

Schriefers et al. (2002) combined these findings with additional evidence from tasks without picture–word interference. They presented participants with one or two identical pictures and prompted them to produce singular or plural noun phrases. They found that participants were generally faster to produce German feminine noun phrases and that the difference between plural and singular forms was smaller (sometimes even negative) for feminine nouns. This difference did not appear in the experiment where participants did not produce determiners.

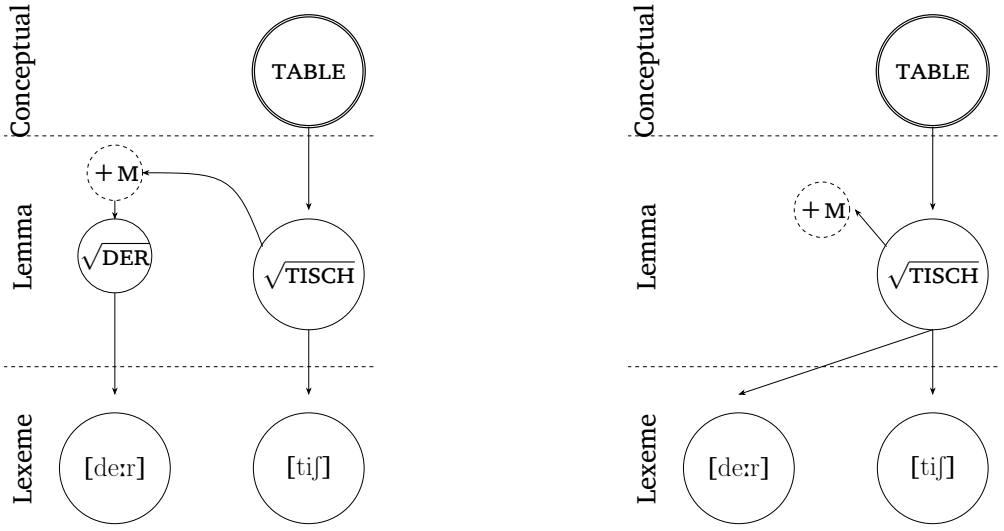
Following Schiller and Caramazza (2003), Schriefers et al. proposed that the diacritics carrying inherent features are mapped from the lexical representation to the lemma without competition. They hypothesized that after nouns are retrieved and their number and gender information is accessed, the determiner is introduced at the lemma level if other syntactic and conceptual constraints require it. Once the determiner is introduced, the features of the noun are copied to the determiner. Figure 5a illustrates the planning procedure of determiners as proposed by Schriefers et al. (2002).

These findings leave open the possibility that determiners are introduced not at the lemma level but at the lexeme level. In this view, diacritics established non-competitively at the lemma level do not require their own separate lexical entry; instead, determiners may emerge at the lexeme level as a function of morphosyntactic specifications on the noun lemma, similar to how bound morphemes are retrieved. Supporting this idea, recent evidence shows that both free-standing and bound morphemes produce similar gender incongruency effects on utterance onset latencies, suggesting comparable processing dynamics. Figure 5b illustrates this alternative

planning architecture.

Figure 5
Two possible planning procedures for determiners.

(a) *Lemma based determiner planning in German.* (b) *No-lemma determiner planning.*



In this section, we have reviewed the assumptions and findings within the determiner planning literature. It is generally assumed that determiners do not have conceptual representations like content words do. Instead, they are introduced at the lexical selection level (Schiller & Caramazza, 2003; Schriefers et al., 2002). The diacritics inherent to determiners are not selected among competing options but are accessed automatically as part of lexical selection. Given more recent findings that bound and free elements are selected by the same mechanisms (Jescheniak et al., 2014), it is not definitive that determiners are planned at the lemma level. Based on these findings, we have outlined two possible planning procedures for determiners. It is important to note that the hypotheses presented here do not directly determine the timing of this process. In scenarios where the noun is retrieved early and other elements remain to be uttered, the determiner can be introduced either at the moment of early planning or just in time, immediately before it is pronounced. This distinction will become important when we discuss auxiliaries.

3.2 Auxiliary Verbs

Except for the agreement attraction literature, which we will cover later in 4, auxiliary verbs have not received the same level of attention as determiners in the literature. One notable exception is Miozzo and Caramazza's (1997a) study on the tip-of-the-tongue phenomenon. In this section, we will review their findings and relate them to the determiner planning literature.

The tip-of-the-tongue phenomenon occurs when native speakers are temporarily unable to retrieve a word that they know (Badecker et al., 1995; Miozzo & Caramazza, 1997b; 1997a; Vigliocco et al., 1997). The significance of this phenomenon lies in cases where the failure to retrieve information about a word is selective. This selectivity has been a central piece of evidence for the existence of distinct levels in the sentence planning process. For example, Vigliocco et al. (1997) showed that Italian native speakers were able to successfully retrieve the gender of a noun during a tip-of-the-tongue state but failed to retrieve its phonological form. This selective failure was interpreted as evidence for the existence of a lemma level separate from the phonological level.

Miozzo and Caramazza (1997a) focused on the auxiliary selection abilities of an Italian anomic patient. Dante, who had spent a month in a comatose state, showed abnormalities in the right fronto-temporal regions (Sartori et al., 1992). Although his speech was fluent, he experienced difficulties in naming objects. In their study, Miozzo and Caramazza (1997a) tested Dante's ability to select the correct auxiliary given a context and to retrieve the correct verb and its initial phoneme. They presented Dante with partial sentences, as in (8), along with pictures depicting the intended sentence. In (8a), given the target verb *cantare* 'to sing', Dante was expected to retrieve the auxiliary *avere* in the form *ha* and the initial phoneme of the verb, [c]. In (8b), given the target verb *sorgere* 'to rise', he was expected to retrieve the auxiliary *essere* in the form *è* and the initial phoneme [s].

(8) a. *Al* *concerto* *di* *ieri,* *il* *coro* ____ ____ *benissimo.*
at the concert of yesterday, the choir ____ ____ very well
Target: 'At yesterday's concert, the choir sang very well.'

b. *Ieri* *mattina* *il* *sole* ____ ____ *alle* *5.30.*
yesterday morning the sun ____ ____ at 5:30
Target: 'Yesterday morning the sun rose at 5:30.'

Miozzo and Caramazza (1997a) specifically measured Dante's ability to retrieve the auxiliary verb and the initial phoneme of the verb in trials where he was unable to retrieve the target verb. They found that in tip-of-the-tongue states, Dante was able to retrieve the correct auxiliary verb with high accuracy (100% for unergatives and 97% for unaccusatives). Even though he could retrieve the auxiliary verb, he was unable to retrieve the initial phoneme of the verb reliably. His success in retrieving the initial phoneme was not above chance (55% for unergatives and 45% for unaccusatives).

For our purposes, the most important aspect of Miozzo and Caramazza's (1997a) study is that the diacritic information inherent to the verb is retrieved independently of its phonological information. This is consistent with the Lemma Model described in our introduction and the two possible procedures we proposed for determiners. In these models, after a conceptual node is activated, the lemma is retrieved via spreading activation, and the inherent diacritic information is accessed automatically, without competition.

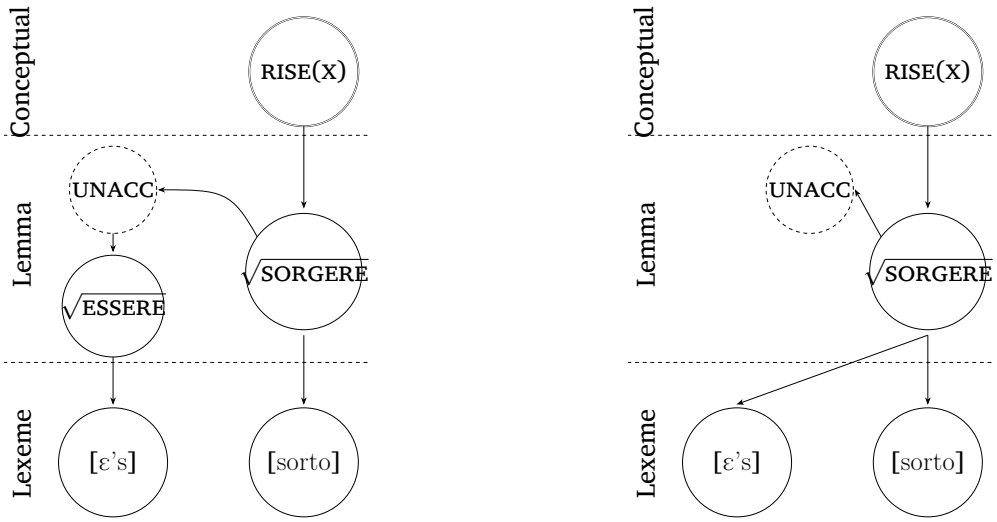
With this inference in mind, we argue that the same planning procedures outlined for determiners can be applied to auxiliary verbs. Figure 6 shows both lemma-based and no-lemma variants. In Figure 6a, after the verb lemma $\sqrt{\text{SORGERE}}$ is retrieved, the inherent unaccusativity diacritic is used to introduce the lemma $\sqrt{\text{ESSERE}}$, and then other morphosyntactic diacritics are copied to the auxiliary lemma from the verb. By contrast, Figure 6b shows the planning procedure in which the auxiliary verb is introduced directly as a dependent of the verb lemma.

Our main question in this study concerns how number diacritics that appear on

Figure 6

Auxiliary verb planning based on the findings of Schriefers et al. (2002) and Miozzo and Caramazza (1997a).

(a) *Lemma based auxiliary verb planning in Italian.* (b) *Auxiliary verb planning in English.*



English auxiliary verbs in our experiments are planned. How do these possible procedures translate to auxiliary planning in English? The comparison between the German determiner and the auxiliary verb is more direct in the case of Italian, since both the gender feature of the noun and the unaccusativity of the verb are inherent features.[2] In the case of English auxiliary planning, we will for now remain agnostic about how the PL or SG diacritics are specified. Given that they are dependent on other elements in the sentence, this process will differ from the specification of inherent features. However, once the number diacritic is specified, both planning procedures outlined in this section can be applied to auxiliary verb planning in English. In Figure 7a, we illustrate the lemma-based planning procedure in which the number diacritic is copied from the main verb to the lemma of the auxiliary verb. In Figure 7b, we show the no-lemma planning

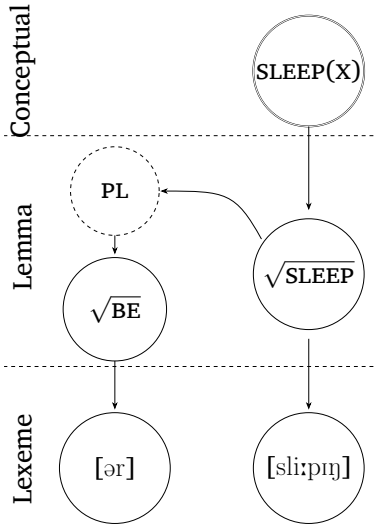
[2] It is important to note here that Italian unaccusativity is not completely predictable from the conceptual meaning of the verb, as argued by Miozzo and Caramazza (1997a). There have been attempts to explain unaccusative and unergative differences entirely in terms of the conceptual meaning of the verb (Perlmutter & Postal, 1984). However, a more in-depth look at cross-linguistic data shows that there are cases where the conceptual meaning does not fully predict the unaccusative behavior of the verb (see Williams, 2015 for discussion).

procedure in which the auxiliary verb is introduced directly as a dependent of the verb lemma. ry verb is introduced directly as a dependent of the verb lemma.

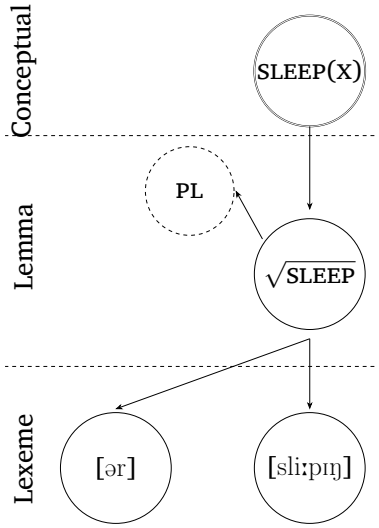
Figure 7

Different planning procedures for auxiliary verbs in English.

(a) *Lemma based auxiliary verb planning in English.*



(b) *No-lemma auxiliary verb planning in English.*



A key issue in understanding the planning of functional elements is the long-distance dependency between a functional item and its controller. In a sentence such as *The cats in the garden are sleeping*, the auxiliary verb *are* reflects the number feature of the distant subject *cats*, raising questions about when that feature is retrieved and mapped onto the auxiliary. Do these diacritics become available early, when the verb lemma is first planned, or are they introduced later, closer to articulation, in a "just-in-time" manner? While existing models help explain some aspects of the mechanism for introducing and selecting functional elements like auxiliaries, they provide little guidance on when their morphosyntactic features are computed. This timing question is especially important for dependencies that span multiple words, where the controller and the agreement target are not adjacent. In the next section, we examine the phenomenon of agreement attraction, cases where agreement fails, as a window into the planning of dependent diacritics. We introduce two competing

accounts that differ in their assumptions about when agreement features are specified.

Spoilers for Dependent functional elements

- Agreement attraction arises from planning errors in dependent morpho-syntactic features, such as number on verbs.
- We test whether agreement is planned early, alongside the verb lemma, or later, just before producing the auxiliary.
- Prior studies use a range of elicitation methods, but few control when the verb is planned.
- We compare two major frameworks for agreement attraction: Marking & Morphing and Cue-Based Retrieval.
- Marking & Morphing attributes attraction to feature assignment and morphological realization; cue-based models trace it to retrieval errors from partial cue overlap.
- While neither model directly predicts when agreement is planned, both provide tools to frame our timing-based manipulation of verb type (unaccusative vs. unergative).

4 Planning Dependent Diacritics

Recall our classification of different types of morphosyntactic diacritics: inherent, conceptual, and dependent. In our discussion of the planning of functional elements, we have covered the planning of inherent diacritics and assumed that once the identity of a dependent diacritic such as PL is known, it follows the same planning procedure as inherent diacritics like UNACC or M.

In this section, we examine the phenomenon of agreement attraction, which is a case where dependent diacritics are planned incorrectly. Agreement attraction is also an area where auxiliary planning has received significant attention, albeit indirectly. After reviewing the agreement attraction phenomenon, we introduce two theories that explain the planning of dependent diacritics. It is important to note, however, that while these theories provide insight into the mechanisms behind planning these diacritics, they do not fully resolve the timing question, which remains open.

4.1 Agreement Attraction Phenomenon

In our study, we are interested in the timing properties of morphosyntactic diacritics that are dependent on other elements in the sentence. We aim to investigate these timing properties by building on the timing differences in verb production observed by Momma and Ferreira (2019). To obtain clearer results, instead of using sentences with a simple subject phrase such as *the cats are sleeping*, we will use target sentences with a complex subject phrase such as *the cat on the mats*, which have previously been shown to induce errors in subject–verb agreement (Bock & Miller, 1991).

These subject–verb agreement errors, called agreement attraction errors, arise when the verb erroneously agrees with a noun that is not the head of the subject. Consider the examples in (9). In English, subject–verb agreement is determined by the number of the subject noun phrase, as shown in (9a–9b). Typically, speakers produce sentences with correct agreement effortlessly. However, in certain instances, speakers fail to produce the correct agreement. These instances usually involve a complex subject noun phrase that contains a noun phrase with mismatching number information, as in (9c). Even though the number of the entire subject noun phrase is syntactically singular, plural morphology on the auxiliary verb is often produced by native speakers in experimental settings, as well as observed in corpus data and comprehension data (Bock & Miller, 1991; Eberhard et al., 2005; Wagers et al., 2009).

(9) a. The cat is sleeping.

b. The cats are sleeping.

c. * The cat on the mats are sleeping.

The most common method for investigating the production of such errors is to use elicitation tasks. In these tasks, participants are asked to read or listen to a preamble consisting of a complex subject noun phrase. After the preamble, they are asked to

repeat the preamble and complete the sentence (Bock & Cutting, 1992; Bock & Eberhard, 1993; Bock & Miller, 1991, among many others). Such errors have been observed in a variety of tasks, including free elicitation (Bock & Miller, 1991), elicitation with prompts (Franck et al., 2006; Vigliocco & Nicol, 1998), and scene description tasks (Kandel & Phillips, 2022; Nozari & Omaki, 2022; Veenstra et al., 2014).

In their seminal work, Bock and Miller (1991) presented participants with recordings of preambles like the one in (10) and asked them to repeat the preamble and complete the sentence. They manipulated the number of the head noun and whether the modifier noun (the attractor) matched in number with the subject. They also manipulated the length of the modifying phrase by adding additional modifiers.

(10) a. The key to the (ornate Victorian) cabinet …

b. The key to the (ornate Victorian) cabinets …

c. The keys to the (ornate Victorian) cabinet …

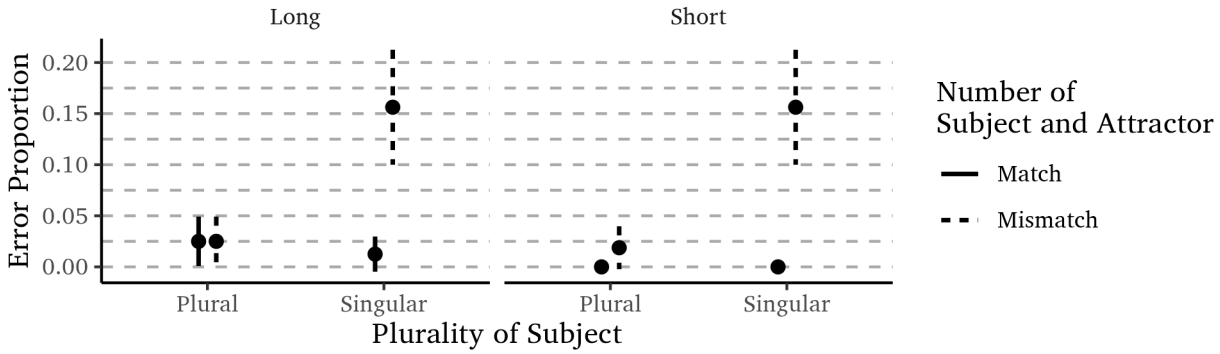
d. The keys to the (ornate Victorian) cabinets …

Their results are shown in Figure 8.[3] They found that participants produced erroneous completions when the number of the attractor and the head noun mismatched. However, this effect was observed only when the subject was singular. The length of the preamble did not have a substantial effect on the overall magnitude of attraction. The increased number of erroneous number markings on the verb in mismatching number conditions is referred to as attraction effects.

In Bock and Miller's (1991) experiments, participants heard the preambles first and then decided on the continuation. Our question concerns the relative timing of the attractor and the agreement planning. It is probable that participants began the planning procedure after hearing both nouns. Therefore, it is crucial for our purposes

[3] In their paper, they do not report standard errors. We calculated these values based on their report of the number of observations and raw number of errors.

Figure 8
Results of Bock & Miller (1991)



that agreement attraction effects are also observed in a paradigm where both nouns are not fully processed prior to sentence production.

Gillespie and Pearlmutter (2011) replicated attraction effects using a picture description paradigm in which this relative timing could be manipulated. They presented participants with pictures of objects on the screen. The head of the subject phrase, which was always singular, was shown with an outline of an object. Participants were asked to produce a complex noun phrase using the objects on the screen. They manipulated the number of the attractor and the preposition used between the head noun and the attractor. They indicated which preposition to use by varying the color of the outline. The experimental conditions are presented in (11).

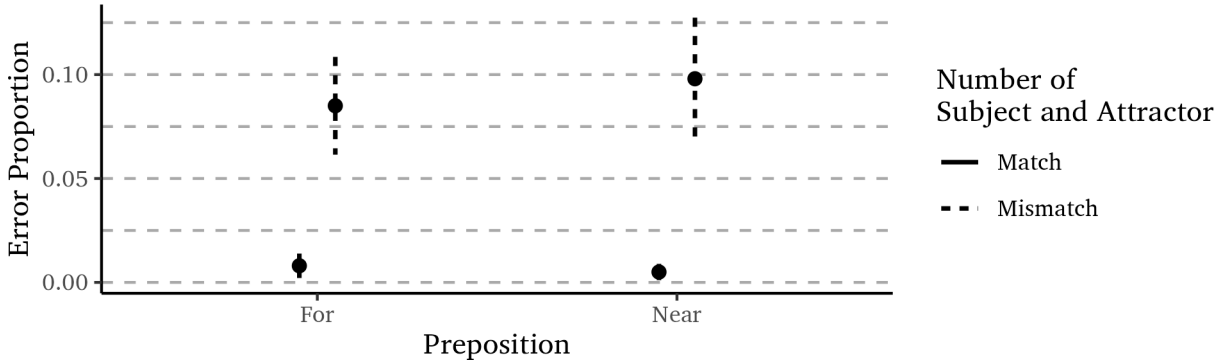
(11) Conditions in Gillespie and Pealmutter (2011)

Image	Target Preamble
	The apple for the pie
	The apple for the pies
	The apple near the pie
	The apple near the pies

Their results are shown in Figure 9. They showed that participants made

agreement errors when the number of the attractor was plural, meaning that the number of the subject head and the attractor mismatched. Their results demonstrated that the picture description paradigm can be reliably used to elicit agreement attraction effects. This aspect of Gillespie and Pearlmutter's (2011) study is crucial for our design because we will need to rely on visual stimuli rather than text or auditory stimuli.

Figure 9
Results of Conditions in Gillespie and Pealmutter (2011)

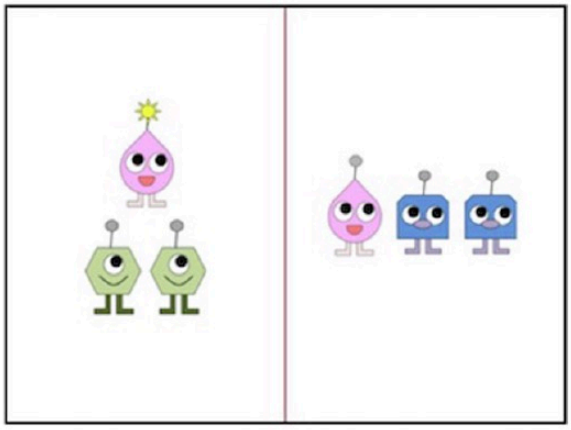


However, both Bock and Miller (1991) and Gillespie and Pearlmutter (2011) employed free-elicitation tasks in which participants were able to complete the sentences with any type of predicate, including ones that do not reflect agreement. Moreover, since our relative timing question depends on manipulating verb types, following Momma and Ferreira (2019), we need to ensure that participants produce errors with a lexically reduced item set. Recent studies have used a picture description paradigm without free elicitation (Kandel & Phillips, 2022; Nozari & Omaki, 2022; Veenstra et al., 2014).

In a series of picture description experiments, Kandel and Phillips (2022) showed that agreement attraction effects occur even with a very limited set of subjects, attractors, and verbs. In their trials, participants saw pictures of three aliens named *the greeny*, *the bluey*, and *the pinky* in various groupings. After a short pause, one or more of the aliens performed an action which participants were asked to describe. In all of their experiments, the target sentence, as shown in (12), included a verb that was a novel

verb *to mim* which was unknown to participants beforehand and explained to them during the instructions. They manipulated the number of the head noun and the attractor. An example trial with the target sentence *the pinky above the greenies is mimming* is shown in Figure 10.

Figure 10
After a 1 second of preview of a scene with no alien had lit up antenna(e), participants saw one or some of the aliens' antenna(e) light up as above for 3 seconds. The target sentence for the scene above is the pinky above the greenies is mimming.



(12) a. The pinky above the bluey is mimming.

b. The pinky above the blueys is mimming.

c. The pinkies above the bluey are mimming.

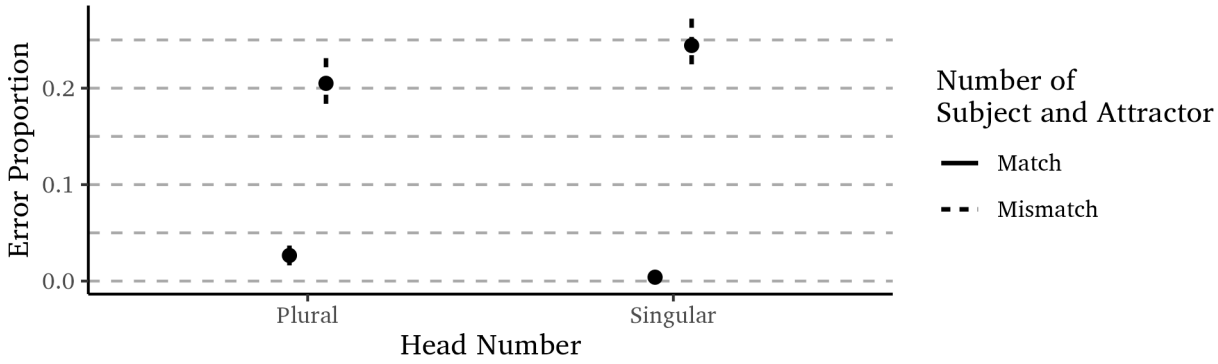
d. The pinkies above the blueys are mimming.

Figure 11 shows the proportion of agreement errors in their Experiment 1 along with the standard error. They found that participants exhibited the well-documented agreement attraction errors and produced sentences such as *The pinky above the blueys are mimming* in the presence of a noun with mismatching number information. They also found that this effect was not limited to cases in which the subject head was singular and the attractor was plural. Kandel and Phillips's (2022) findings demonstrate a crucial point: even in an experimental design with very limited lexical elements,

agreement errors were observed in the picture description paradigm.

Figure 11

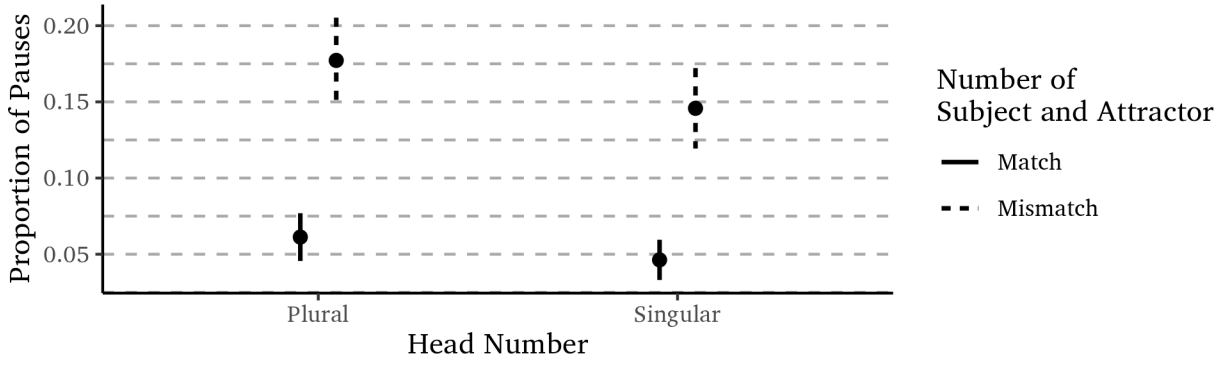
Results of Conditions in Experiment 1 of Kandel and Phillips (2022)



Another contribution of Kandel and Phillips (2022) is their use of time as an index of agreement computation. They showed that even in correct utterances, such as those in (12), participants were more likely to pause before uttering the auxiliary verb when the numbers of the subject head and the attractor mismatched. Figure 12 shows the distribution of pauses immediately before the agreement-bearing auxiliary verb as a function of condition. They interpreted this increased likelihood of pausing in mismatched conditions as evidence for a timing index of agreement during sentence production.

Figure 12

Pause distributions in responses without errors of disfluencies in Experiment 1 of Kandel and Phillips (2022)



However, as we previously mentioned, we are interested in the possible decoupling of verb and auxiliary planning. One way to address this question is to exploit

the syntactic and conceptual differences between unergative and unaccusative verbs, following Momma and Ferreira (2019). Kandel and Phillips's (2022) main aim was to distinguish between number attraction on pronouns and verbs. To that end, they used a novel verb and novel subjects to reduce additional effects of lexical item retrieval. It is not clear whether the novel verb *mim* has an unergative or unaccusative meaning.

Additionally, their research question was orthogonal to verb planning procedures; thus, their experiment does not provide any information regarding when the verb *mim* might be planned. In our study, we aim to manipulate verb type to test whether the timing effects found in Kandel and Phillips's (2022) work are also present in unaccusative environments, which would suggest an independent morphological process and more naturally support marking and morphing theories.

In this paper, we mainly focus on number information and errors related to number marking. However, it is important to note that agreement errors are not limited to number agreement. Similar errors have been observed with other dependent morphosyntactic features, such as gender. Badecker and Kuminiak (2007) used Slovak sentences with nouns that had different genders in the subject and the modifier of the subject, as in (13), and found erroneous gender marking on the verb.[4]

(13) a. *Trest* *za* *zločin* …
punishment.M.NOM for crime.M.ACC …
'The punishment for the crime …'

b. *Trest* *za* *kràdež* …
punishment.M.NOM for theft.F.ACC …
'The punishment for the theft …'

[4] Slovak has three different genders, and in their experiments Badecker and Kuminiak (2007) tested other gender combinations as well. The highest error rates were observed in conditions where the head noun was masculine and the modifier was feminine, with morphophonologically syncretic case markings in their Experiment 2. The reason for this is beyond the scope of this paper; see Dillon and Keshev (n.d.) for further discussion.

4.2 Accounts of Agreement Attraction in Production

What might be the underlying mechanism of sentence production that gives rise to errors like those described in 4? The literature on agreement attraction includes two main families of theories that attempt to answer this question. In this section, I will briefly describe these two models. Both models are compatible with different possible timings, given that sentence production processes can operate in parallel. After covering the basic tenets of these models in relation to sentences with simple subject noun phrases, I will introduce how these models account for errors in sentences with complex subject noun phrases. Lastly, I will discuss how both models can accommodate different timing assumptions.

4.2.1 Marking and Morphing Theory

The first family of accounts is called representational accounts. These accounts assume that the number representation of complex noun phrases can be probabilistically distorted, and this distortion is the reason why agreement attraction errors arise. An influential representational account is the Marking and Morphing model (Eberhard et al., 2005; Hammerly et al., 2019). In this model, Eberhard et al. (2005) propose that number marking on the verb occurs in two steps: marking and morphing. In the marking step, the number information for the relevant nouns is assigned to each lemma as a diacritic using notional, lexical, and morphological information. In the morphing step, a final number for the subject phrase is computed using the spreading activation formula given in (14). In this formula, the notional number of the head noun ($S(n)$) is added to the weighted number information ($S(m)_j$) available in the sentence production state at the moment the number is determined. The weight (w_j) of this information is determined according to its syntactic distance from the subject head. This final morphed number representation is not strictly binary but is instead continuous, with `0` being unambiguously singular and `+1` being unambiguously plural.

$$S(r) = S(n) + \sum_{j} \left(w_j \times S(m)_j\right) \tag{14}$$

As a function of the final number representation on the entire subject node, the probability that the predicate will be marked as plural or singular changes. The mapping between the final subject number and the probability of a plural verb is modeled using a logistic transformation, as shown in (15), where b is a bias term (see Hammerly et al., 2019 for a detailed discussion). According to these accounts, attraction errors arise due to the ambiguous (a value between `0` and `1`) number representation of preambles like *the key to the cabinets*, instead of an unambiguously singular `0`. Non-zero $S(r)$ values from (14) produce higher probabilities of plural marking as $S(r)$ increases, depending on the amount of mismatching number information from non-subject nouns and their syntactic depth.

$$1/\{1 + exp - [S(r) + b]\} \tag{15}$$

While marking the noun lemma with number information, three different types of information must be considered: notional (conceptual) number, grammatical (inherent) number, and morphological number. The notional number is retrieved from the numerosity of the discourse entities. Grammatical number refers to the arbitrary number information that certain nouns are lexically specified with. Lastly, morphological number refers to the number information signaled by morphological markers, such as the presence of the *-s* suffix. These sources of number information do not always align.

Let's take the noun *news* as an example. The noun *news* is notionally underspecified because it can refer to a single event or multiple events reported. Morphologically, the noun *news* obligatorily carries a plural marker *-s*, making it morphologically plural. However, it is lexically specified to have singular number, as the verb that agrees with *news* is singular in English. Another example is the pronoun

they. The pronoun *they* is also notionally underspecified, since it can refer to multiple people or a single person. Morphologically, the pronoun *they* does not have a plural marker, making it morphologically singular. Yet it is lexically specified to have plural number. Even when it refers to a single person in recent usage, the verb that agrees with *they* is plural in English. In simpler cases like *cats*, the noun is both morphologically and notionally plural, as it refers to multiple cats and surfaces with the morphological marker *-s*. However, it is not lexically specified as inherently plural or singular.

How do these different sources of number information converge into a single number marking on the noun lemma? In this model, the lexical specification overrides any other number information. Thus, in cases where there is a lexical specification, as with *news* or *they*, these elements are marked with a SG or PL diacritic, respectively. In other cases, such as *cats*, the numerosity of the discourse entities (the notional number) is used to determine the diacritic information.[5]

This simple case of *marking* in the Marking and Morphing account is illustrated in Figure 13. In this figure, the cascading steps of sentence planning are shown for the sentence *The cats are sleeping*, with planning states divided by lines. In the first state, speakers have only the concept CAT.[6] In the second state, speakers activate the conceptual node for SLEEP while planning the lemma representation and syntactic structure for $\sqrt{\text{CAT}}$.[7] It is also important to note that this model assumes a slightly modified version of Levelt et al.'s (1999) model as proposed by Ferreira et al. (2007). The main difference between these two models is that Ferreira et al. (2007) and

---

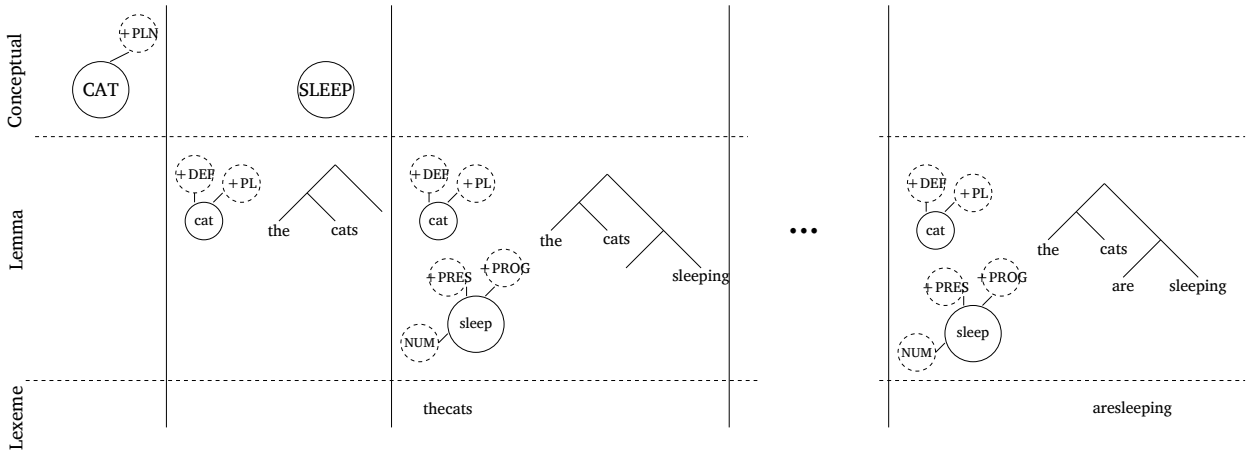
[5] As assumed by Eberhard et al. (2005), we also assume that the notional number of a noun is not directly marked on the verb. Rather, it is first marked on the noun lemma and then copied to the verb lemma.

[6] Even though our schematic presentation of sentence planning suggests that speakers can begin uttering the sentence without having the entire message planned, we remain agnostic about when such message planning occurs. See Gussow and MacDonald (2023) for a detailed discussion. Following their findings, we leave open the possibility that separate conceptual nodes may be activated independently from the preverbal message.

[7] It is important to note that although we use English orthography in the syntactic trees, these do not represent phonological outputs.

Eberhard et al. (2005) assume that structure building and lemma access and specification occur in parallel, as shown in Figure 13. In the third state, the verb lemma $\sqrt{\text{SLEEP}}$ is accessed and its corresponding syntactic structure, given the rest of the sentence, is specified while the subject phrase *the cats* [ðəkʰæts] is pronounced. We also show the final state in which the predicate *are sleeping* [əsliːpɪŋ] is pronounced, along with its lemma and structural representations.[8]

Figure 13
Cascading sentence production with the marking process specified.



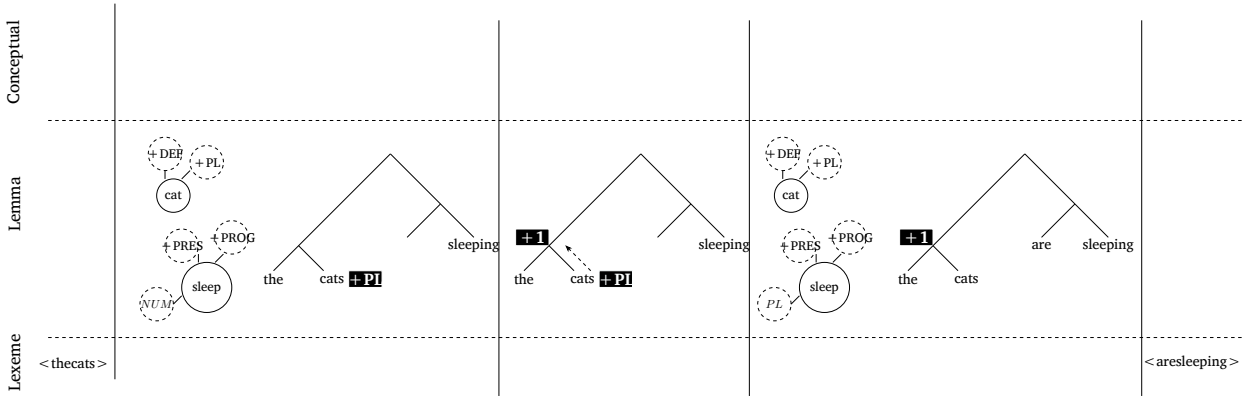
The omitted section in Figure 13 illustrates the morphing part of the Marking and Morphing account. After the noun is marked with a number diacritic, the number features are transmitted to structurally relevant elements, such as verbs, through a process called *morphing*. This process consists of (i) mapping lemmas and diacritics to structural positions, (ii) determining the grammatical number of the entire subject phrase, and (iii) copying this number to the verbal domain. The morphing component of this framework is closely related to other lemma-based models, specifically formulating in the traditional model by Levelt et al. (1999) and constituent assembly in the modal model by Ferreira et al. (2007).

[8] In our scheme, we chose to show the diacritics that will affect the functional elements on the content lemmas, such as $\sqrt{\text{CAT}}$ and $\sqrt{\text{SLEEP}}$. As discussed in Section 3, this choice does not affect the research question posed in this study.

Figure 14 shows the morphing process step by step for a simple sentence like *The cats are sleeping*. In the first state, the diacritics of the terminal nodes are accessed. In the second state, the number of the subject phrase is determined by evaluating the number features of the nouns present in the planning procedure using the formula presented in (14). In addition to the notional number, the other available number information is the number diacritic of $\sqrt{\text{CAT}}$. Given that both are plural and that the noun is overtly marked as plural, the final number of the subject phrase is marked as unambiguously plural (+1). In the final state of the morphing process, this number information is copied to the lemma $\sqrt{\text{SLEEP}}$ and is reflected in the structure as a plural auxiliary verb. Finally, the rest of the sentence is pronounced as [əsliːpɪŋ].

Figure 14

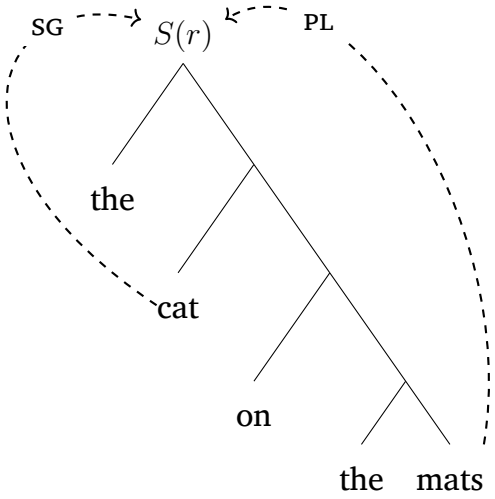
Step-by-step process morphing following Eberhard et al. (2005).



The primary cases that the Marking and Morphing account seeks to explain are agreement attraction errors involving mismatching number information within the subject phrase, such as **The cat on the mats are sleeping* or **The cats on the mat is sleeping*. In these cases, the final subject number does not end up being unambiguously singular or plural but instead takes on a value between 0 and +1. Consider the singular subject case. While *cat* is notionally and morphologically singular, so the lemma $\sqrt{\text{CAT}}$ is marked with the diacritic SG, the noun *mats* is notionally and morphologically plural, and the lemma $\sqrt{\text{MAT}}$ is marked with the diacritic PL. Assuming left-to-right sentence planning, the morphing process is initiated after the number diacritics for both nouns

have been established and situated within the syntactic frame. During morphing, the number information for all relevant nouns is accessed, and then the number of the entire subject phrase is computed. The notional number of the head noun is SG (0), but the number diacritic of $\sqrt{\text{MAT}}$ is PL (1), which yields a final subject number $S(r)$ that is non-zero and ambiguous between singular and plural. This ambiguity increases the likelihood that the lemma $\sqrt{\text{SLEEP}}$ will be marked with a plural diacritic, resulting in a plural auxiliary *are* in a certain proportion of trials.

(16)



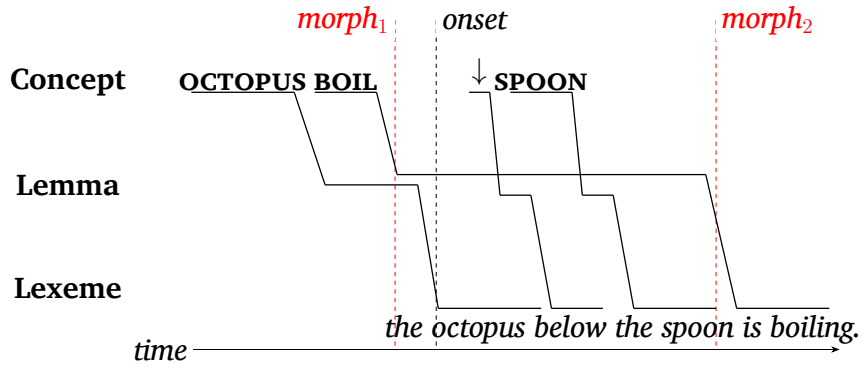
Cases like this, in which sentence planning follows left-to-right structure building, do not help us answer the question of when the auxiliary verb is planned. In these cases, the morphing process is initiated after the last noun is planned, which is immediately followed by the planning and utterance of the auxiliary verb. To disentangle utterance-related pressures from the planning procedures, we need to examine cases in which left-to-right planning is overridden.

Momma and Ferreira's (2019) findings provide relevant cases where the verb and the subject head are planned together, excluding intervening elements. They showed that unaccusative verbs are planned before the utterance of the subject head, as illustrated in Figure 4. The question then arises whether the number diacritic on the verb is also planned prior to the utterance, like the verb itself, or whether this diacritic

is planned only immediately before articulation. Another way to pose this question is whether the morphing process is initiated as soon as the verb lemma is available or whether it is delayed until later in planning. Possible timings for initiating the morphing process are shown in Figure 15.

Figure 15

Possible times for the morphing process given an unaccusative event.



In addition to the absolute timing of the morphing process, Figure 15 also illustrates the relative timing of the morphing process with respect to the intervening attractor NP. In the Marking and Morphing account, number attraction effects depend on the presence of mismatching number information from different nouns during planning. In a lazy agreement system, where the number diacritic on the verb lemma is planned later ($morph_2$), both nouns will be available in the planning procedure for both unaccusative and unergative verbs, so the final number will encode information from both nouns. In an eager agreement system, where the diacritic is set as soon as the verb lemma is accessed ($morph_1$), the number diacritic on the verb lemma will be set before the attractor NP's lemma is planned, meaning the verb's number diacritic will not be influenced by the attractor NP's number. It is important to note that this distinction does not affect sentences with unergative verbs. In these cases, the verb lemma is assumed to be planned while the attractor NP is being uttered (Momma & Ferreira, 2019). Therefore, the syntactic information and diacritics of both NPs will be included in the final number marking on the verb lemma.

In the Marking and Morphing account, the final number representation for the entire subject phrase is the sole driver of attraction effects. We have discussed how the timing of morphing affects this final number evaluation. A natural follow-up, then, is a prediction regarding the relationship between morphing timing and attraction effects. Under this account, if diacritic marking is eager, attraction effects are expected to be reduced or absent in sentences where the verb lemma is planned before the attractor NP is uttered, as in unaccusatives. By contrast, if diacritic marking is lazy, attraction effects are predicted for both verb types. Even if the verb is planned early for a certain class of verbs, the agreement diacritic would be assigned through the morphing process at the same time regardless of verb type.

4.2.2 Cue-Based Production Model

The second family of accounts is called cue-based memory accounts, which are based on the ACT-R cognitive architecture (Anderson, 1996; Lewis & Vasishth, 2005). These accounts provide an overarching framework for understanding various long-distance dependencies, including reflexive pronouns (Jäger et al., 2015; Parker & Phillips, 2017), reciprocal pronouns (Kush & Phillips, 2014), negative polarity items (Drenhaus et al., 2005; Vasishth et al., 2008; Xiang et al., 2009), and subject-verb agreement (Dillon et al., 2013; Wagers et al., 2009). Under cue-based memory accounts, errors in such dependencies are not attributed to an inaccurate linguistic representation but are instead a byproduct of limitations in memory retrieval. Cue-based models have found more success in explaining processes in comprehension than in production. In this section, we will first describe how these models operate in comprehension and then discuss Badecker and Lewis's (2007) production model.

One of the most influential cue-based models is the Lewis and Vasishth model of sentence processing (2005). In this model, structure is built by a left-corner parser following X-bar syntax rules. Syntactic constituents are stored as memory units called chunks. The relations between chunks are specified through feature slots for specifier,

complement, and head, in line with X-bar rules. Chunks also encode additional information such as number, tense, or case. In comprehension, a structure is built incrementally as new words are processed and placed in a lexical buffer. Depending on the current prediction of the existing syntactic chunk and the word in the lexical buffer, the parser retrieves a new chunk matching the cues and attaches it to the syntax.

In the case of erroneous subject-verb agreement in comprehension, Wagers et al. (2009) proposed that when the parser encounters an agreement-carrying element, known as a *probe*, it checks whether the number feature on the probe matches the noun that determines agreement, known as the *agreement controller*. This checking process is guided by cues: the number cue, [PL] or [SG], and the subjecthood cue [SBJ] provided by the probe. These cues are used to retrieve the agreement controller stored in memory as a chunk. However, the retrieval process is not error-free, and the probability of retrieving the correct controller is determined by its activation value. This value changes due to decay over time, reactivation, and noise (Vasishth & Engelmann, 2021). When a retrieval is initiated, a limited amount of activation is spread among all chunks in memory as a function of the match between the retrieval cues and the values stored in each chunk. The activation S_i of a chunk i is calculated as in (17), where all relevant cues (j) are summed by weighting according to their strength of association S_{ji} between cue j and chunk i along with the amount of activation from cue j.

$$S_i = \sum_j W_j S_{ji} \tag{17}$$

The strength of association S_{ji} is calculated by subtracting the number of items stored in memory that match cue j from the theory-internal parameter for maximum associative strength S, as shown in (18).[9]

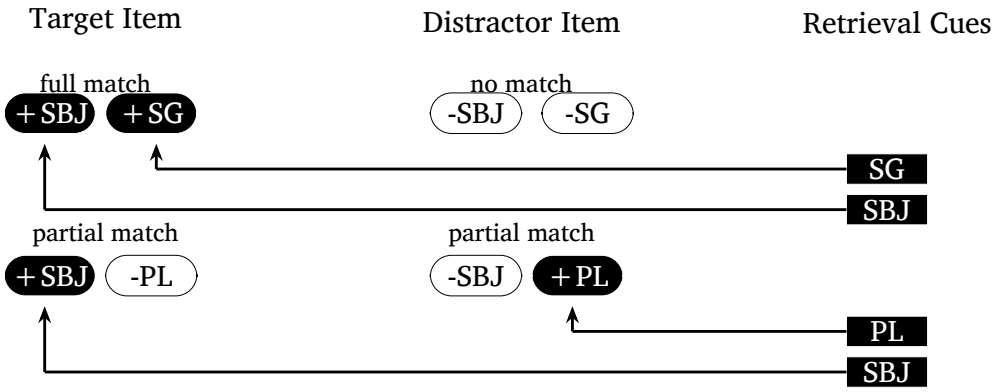
[9] For simplicity, we do not elaborate on the exact values of S and N_{ji}. For our purposes, it is sufficient to understand how changes in the number of chunks that match cue j affect the strength of association and, consequently, the spreading activation.

$$S_{ji} = S - log(N_{ji}) \tag{18}$$

As a result, for each chunk, as the number of cues matching the agreement probe increases, the activation for that chunk also increases (17). However, for each additional chunk that also matches the agreement probe, the activation for any single chunk decreases because the activation is shared among multiple matching chunks (18).

In the comprehension of sentences such as *The cat on the mats is sleeping*, the cues provided by the agreement probe *is* include [SG, SBJ]. Among the relevant chunks stored in memory, only the chunk for *cat* will match either of these cues. For sentences such as *The cat on the mats are sleeping*, the cues provided by the agreement probe *are* are [PL, SBJ]. In this case, no chunk matches both cues, but both *cat* and *mats* match a single cue each. A visualization of these configurations is provided in Figure 16.

Figure 16
Visualization of partial match and full match configurations in cue-based account.



In partial match configurations, the activation of the chunk *mats* will occasionally result in its retrieval as the agreement controller. These occasional retrieval errors lead participants to judge the sentence as grammatical, even though it is not. The important point here is that within this agreement process, the focus is on the retrieval of the agreement controller rather than the representation of the entities.

The production model differs slightly from the comprehension model, primarily

due to the nature of the task. In this paper, we focus on the account proposed by Badecker and Lewis (2007). Similar to the assumptions discussed in Section 2, their production model shares several core ideas with lemma-based theories as a starting point. These include the modular yet cascading and parallel nature of production, with processing levels such as the conceptual level, the formulating (lemma) level, and the articulation (lexeme) level. However, instead of lemma representations, they posit that a syntactic frame drives the production process by using phrasal nodes encoded in a content-addressable memory as bundles of features, called *chunks*, which include information such as number, gender, case, and syntactic function (Smith & Vasishth, 2020).

In their model, retrieval of the syntactic frame is the first step in the production process. Based on communicative intent and other factors that influence message content, such as the salience of entities, the speaker considers multiple syntactic frames that satisfy these constraints. Depending on the overall costs associated with these syntactic frames, one is selected over the others. Once the syntactic frame is determined, the speaker focuses on the nodes of a syntactic tree, proceeding in a left-to-right sequence and retrieving relevant chunks to bind to the focused nodes. Next, the speaker applies procedural rules that govern the planning process. These procedural rules follow an if-then structure, for example: "IF the current goal is to … THEN retrieve … and set the new goal to …."

After retrieving the syntactic frame and placing the first element, such as a subject DP, the syntactic frame is expanded to include the next lexical item. For instance, if the next goal is to integrate a spatial relation between two nouns, the structure is expanded to include a prepositional phrase as a modifier. If the next element is a verb, then the nominal structure is not expanded, and the planning procedure proceeds directly to retrieving the verbal elements.

Let us take the sentence *The cat is sleeping* as an example. After retrieving a

skeleton of an active sentence syntactic tree for the intended utterance, the chunk (19) for *the cat* is retrieved using notional features. This chunk includes the syntactic category (Cat: DP), the dominating syntactic node (DomCat: TP), the case features (NOM/ACC), and the number information (Num: SG).[10] Since there is no additional intended nominal element that modifies *the cat*, the chunk and its syntactic frame $[_{DP}$ the $[_{NP}$ cat$]$ $]$ are connected to the dominating node as in $[_{TP}$ $[_{DP}$ the $[_{NP}$ cat$]]$ $\ldots]$.

(19) *the cat*

$$\left[\begin{array}{lll|lll} Cat: & DP & & DomCat: & TP & \\ \textbf{NOM}: & + & & \textbf{ACC}: & + & \\ Num: & \textbf{SG} & & & & \end{array}\right]$$

The next element is the verb *sleep*. The chunk that matches the notional features of the intended message is retrieved along with its syntactic structure and chunk specifications. The retrieved chunk for *sleep* (20a) includes the category (VP), the dominating syntactic node (TP), and message-related information such as tense (PRS) and aspectual information (PROG). According to these features and the selected syntactic frame, a BE auxiliary verb is inserted into the head of the TP projection. However, the BE auxiliary verb is not fully specified (20b). Although features such as tense and aspect are already available and shared with the main verb *sleep*, both person and number features remain unspecified.[11]

(20) a. *sleep*

$$\left[\begin{array}{lll|lll} Cat: & VP & & DomCat: & TP & \\ Tense: & \textbf{PRS} & & Aspect: & \textbf{PROG} & \\ Num: & & & Person: & & \end{array}\right]$$

[10] Badecker and Lewis (2007) assume that case features are specified through morphological output. In English, non-pronominal elements are syncretic across various cases, all of which are specified within the chunk.

[11] Badecker and Lewis (2007) are not explicit about whether these features are represented in both main verb and auxiliary chunks. Recall that lemma-based models are also not explicit on this question. We assume that these features are represented on both the main verb and the auxiliary verb. We believe that this assumption does not affect our subsequent sentence production processes.

b. *be*

$$\left[\begin{array}{lll|lll} Cat: & & T & DomCat: & & TP \\ Tense: & & \textbf{PRS} & Aspect: & & \textbf{PROG} \\ Num: & & & Person: & & \end{array}\right]$$

In order to set these values, a retrieval process is initiated using procedural rules. (21a) shows the syntactically driven search procedure for retrieving the subject. The cues used in this search are shown in (21b). These retrieval cues are determined based on the nature of the linguistic dependency and the properties of the language. In English subject-verb agreement, the cues include syntactic position, number, aspect, and case marking.

(21) a. **IF** the current goal is to set the agreement features of a tensed form of BE in TP, and the subject of TP can be retrieved based on the cues (21b),
THEN retrieve the subject, copy the person and number feature values from the subject to the verb BE.

b. Retrieval Cues

$$\left[\begin{array}{lll|lll} Cat: & & DP & DomCat: & & TP \\ \textbf{NOM}: & & + & \textbf{ACC}: & & - \\ Num: & & var & Person: & & var \end{array}\right]$$

It is important to note that the non-trivial aspect in this process is the retrieval. The retrieval is subject to the same constraints that apply in comprehension. That is, the match is established probabilistically as a function of how many chunks match each feature–value pair specified in the retrieval cues. In our simple example sentence, *The cat is sleeping*, the only chunk that can be retrieved as the subject is *the cat* (19). It matches the values for the Cat, $DomCat$, and NOM features. After this chunk is retrieved, its person and number features are copied to the verb BE, resulting in the output *is*.

Similar to the Marking and Morphing account, this account is also designed to

explain more complex sentences that can give rise to agreement attraction effects. Consider the sentence *The cat on the mats is sleeping*. The first difference between this example and the previous one is the addition of a prepositional phrase. Following the sentence plan, the PP is bound to the subject noun phrase, and the chunk for the noun *the mats* (22) is retrieved and bound to the complement position of the prepositional head.

(22) *the mats*

$$\left[\begin{array}{ll|ll} Cat: & DP & DomCat: & PP \\ \textsc{nom}: & + & \textsc{acc}: & + \\ Num: & \textsc{pl} & & \end{array}\right]$$

Integrating the verbal predicate and auxiliary follows the same steps discussed for the simpler case. The retrieval cues remain unchanged as well. Under optimal conditions, the retrieval cues match better with the chunk *the cat* than with *the mats*, because *the cat* matches the $DomCat: TP$ feature–value pair, whereas *the mats* does not. However, the retrieval process is not always flawless. Given the assumed dynamics of activation spread discussed previously, the fact that other feature–value pairs such as $Cat: DP$ and $\textsc{nom}: +$ are shared by both chunks increases the probability of an erroneous retrieval of *the mats* as the agreement controller. When another chunk is mistakenly retrieved, its number and person information is copied into the unspecified slots, leading speakers to produce sentences such as *The cat on the mats **are** sleeping*.

What happens when left-to-right production is overridden by syntactic constraints, as suggested by Momma and Ferreira (2019)? First, it is important to note that this scenario is less straightforward than in the lemma model due to the syntactic assumptions within the cue-based retrieval production model. Two issues contribute to this complexity: (i) the syntactic assumptions and (ii) the level of representation required for semantic interference.

Momma and Ferreira (2019) argue that speakers plan unaccusative verbs, whose

subjects are patient-like, before sentence onset and before the intervening prepositional phrase. Their main evidence comes from patterns of semantic interference. What does semantic interference imply for a cue-based production model? One possibility is that semantic interference arises from competing chunks during retrieval. In the case of the target verb *boil* and its semantically related distractor *melt*, both verbs are sufficiently activated due to shared semantic features, making them candidates for placement in the syntactic tree (Logačev & Vasishth, 2011 for similar effects in comprehension; see Van Dyke & McElree, 2011). This suggests that semantic interference may be a byproduct of syntactic structure building. However, as described by Badecker and Lewis (2007) and Lewis and Vasishth (2005), syntactic parsing and incremental structure building follow a left-corner parser using X-bar syntax. Moreover, the syntactic structure is constructed and established before sentence production begins. To account for Momma and Ferreira's (2019) findings within this framework, we must either hypothesize that the speaker reanalyzes the syntactic frame or assume a Tree-Adjoining Grammar, which allows for the manipulation of established syntactic trees (Frank, 2004; Joshi et al., 1975).

Another possibility is to interpret the semantic interference effects as a byproduct of chunk activation. In unaccusative scenes, the verb chunk may be activated not due to syntactic reasons but due to conceptual factors. Momma and Yoshida (2023) suggest that early semantic interference results might reflect the activation of high-level conceptual representations linked to the theme or patient role of the constituent.[12] According to this hypothesis, the increased utterance onset latency reflects a momentary activation of the chunk to establish conceptual relations, not syntactic ones. Under this account, there is no need to assume a different syntactic framework for the cue-based production model.

---

[12] Even though Momma and Yoshida (2023) argue that their results support a syntactic account rather than a conceptual one, the effect sizes they report are relatively small (approximately 10–25 ms).

Different interpretations of semantic interference do not affect scenes in which the target sentence includes an unergative verb. In these scenes, the speaker follows procedures similar to those used for sentences with complex subject phrases, such as *The cat on the mats is sleeping*. The retrieval of elements and syntactic expansion occur through left-to-right structure building. Because the subject does not have a theme or patient role, it is not activated early. The activation of the verb and its syntactic integration occur within the same time frame. Therefore, either interpretation of semantic interference results in the same outcome. Regarding agreement, the features of the BE auxiliary or the main verb are set after both noun phrases are activated and integrated into the structure.

On the other hand, assumptions about semantic interference have consequences for unaccusative scenes. If semantic interference arises from syntactic integration, then the production of unaccusative scenes proceeds initially in the same way as for simple sentences like *The cat is sleeping*. A simple subject is planned, and the verbal predicate is integrated without the modifier. The subject phrase is then expanded to include the modifier *on the mats*. Because the verb is integrated into the structure early, the procedural rules that trigger retrieval of the subject and its features are also activated early, before the modifier is included. Due to this early activation, the only relevant chunk available for retrieval is *the cat*. This is similar to the "eager agreement" scenario described in the Marking and Morphing section. We would expect fewer agreement attraction effects because there are fewer candidate subjects, and activation is not spread across competing chunks.

If, however, semantic interference results from early chunk activation without syntactic integration, the procedural rules are not triggered early. In this case, semantic interference reflects the establishment of conceptual meaning, that is, deciding whether the action is *boiling* or *melting*. Once the conceptual representation is determined, sentence production proceeds similarly to that for unergative sentences. Left-to-right

incremental syntactic integration occurs, and the procedural rules that initiate subject retrieval for determining number and person agreement are triggered only after both nouns have been processed. This corresponds to the “lazy agreement” scenario in the Marking and Morphing section. Because activation would be distributed across existing nouns, we would expect agreement attraction effects to be similar to those observed with unergative sentences.

In our study, we are interested in the planning of the dependent number diacritic, or the number feature more generally. We aim to investigate the timing of number information planning by taking advantage of the different planning timeframes associated with unaccusative verbs. Our prediction is that if this information is planned as early as the initial retrieval of unaccusative verbs, then the patterns of agreement errors in sentences with unaccusative verbs and those with unergative verbs will differ. By contrast, if agreement is planned just before it needs to be produced, regardless of when the verb is planned or activated, then the attraction patterns will be similar for both verb types. In this section, we have shown that existing theories of agreement attraction do not make explicit predictions about the timing of this information. Both models discussed here can account for either hypothesis. In this paper, we do not aim to distinguish between these theories. However, it is important to lay out the predictions and the processes within these models in order to test additional hypotheses that may arise from our results.

Spoilers for Experiment 1

- We hypothesized that unaccusative verbs, planned early, would show reduced agreement attraction compared to unergatives if the agreement planned early.
- Using an extended picture–word interference task, we manipulated verb type, attractor number, and semantic relatedness of the superimposed word.
- Attraction effects were observed for both verb types but were attenuated relative to prior studies.
- This reduction likely stems from design features such as persistent cueing of the subject, non-restrictive modifiers, and attractors not being viable agreement controllers.
- Pause likelihood aligned with attraction patterns, supporting its use as a timing-sensitive measure of agreement computation.
- Unexpectedly, unaccusatives showed prolonged planning rather than clean early planning, as a function of distractor number.
- Results suggest that agreement attraction might be due to planning in late stages.

5 Experiment 1: Using Verb Planning and ePWI to Probe Agreement Timing

This experiment investigates the timing of agreement computation during sentence production, using agreement attraction errors as a way to bring the agreement computation to the surface. The central question is whether the timing of agreement planning depends on the planning of its host. Following Momma and Ferreira (2019), we assume that unaccusatives are planned early, at utterance onset, and unergatives are planned late. This difference creates a prediction: if agreement is planned eagerly alongside the verb, we expect fewer or no agreement attraction errors in unaccusatives. In contrast, for unergatives, where verb planning occurs later, we anticipate the presence of attraction effects, reflecting delayed or more vulnerable agreement computation. This is due to the fact that when an unaccusative verb is planned with the subject at utterance onset, the only other planned noun phrase is the subject itself. However, at

the time of unergative verb planning, both noun phrases are likely already activated.

To test this hypothesis, I contrast unaccusative and unergative verbs in a picture–word interference paradigm, while manipulating two additional factors: the number of the attractor noun (singular vs. plural) and the semantic relatedness of a superimposed distractor word (related vs. unrelated). These manipulations allow me to examine whether agreement attraction errors are modulated by verb type. Ultimately, this study aims to shed light on the timing of agreement computation in real-time sentence production.

All anonymized data (except for participants' voice recordings) and R code used in the data analysis is available at https://go.umd.edu/turk888. Reader can go through the experiment at https://go.umd.edu/turk888-exp1.

5.1 Methods

5.1.1 Participants

We recruited 74 monolingual English speakers located in US at the time of the experiment to participate in the experiment in exchange for monetary compensation. Participants had a mean age of 34.39, ranging between 19 and 50. They were recruited through the online platform Prolific Academy (www.prolific.com) and screened using Prolific Academy's internal screening system. An additional 14 participants completed the study but were excluded from the analysis due to poor sound quality in their recordings, technical difficulties preventing completion of the experiment, or producing the correct sentence structure in less than 30% of their responses. Every participant was informed about the experimental details prior to their consent. Informed consent in compliance with the Institutional Review Board of the University of Maryland, College Park was obtained for all participants. Each session took roughly 30-40 minutes.

5.1.2 Materials

Our experimental items were based on the ones used in Momma and Ferreira (2019). Throughout the experiment, we utilized 12 event participants (each presented

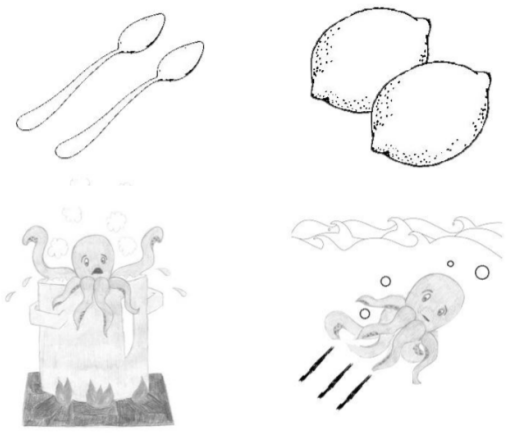
with 2 events), 2 unique objects for each participants (total 24 objects), and a unique pair of superimposed semantic distractors for each scene (48 total: 24 related, 24 unrelated). We manipulated the verb type (unaccusative vs. unergatives), the number of the object (plural vs. singular), the semantic relatedness of the superimposed distractors (related vs. unrelated), resulting in 96 total experimental trials.

Momma and Ferreira's (2019) 24 images (2 for each event participant) that depicts humans or animals doing an action were combined horizontally with by the event participant. Each pair of pictures in these scenes had the same event participant being part of two different events, one of which is an unaccusative event, the other one being an unergative. The unaccusativity and unergativity of the pictures were determined using grammatical tests, such as transitivity alternation tests. (Momma et al., 2016). The frequency of these verbs were also well matched between unaccusative and unergatives ($M_{unacc} = 8.90, SD_{unacc} = 1.03, M_{unerg} = 8.61, SD_{unacc} = 1.68, p = .61$ as reported in Momma and Ferreira (2019)).

These combined action pictures were accompanied by two different objects either located below or above the actions. The positioning of these objects were controlled in the experiment. Participants saw equal number below and above positioning for each event participant, action, and each experimental condition. Both action and the object pictures were all taken from the UCSD International Picture Naming (Szekely et al., 2004). As a result, we ended up with 12 base scenes presented in a 2x2 grid that consisted of an unaccusative action, an unergative action whose event participant is the same as the unaccusative one, and two different objects. An example scene shown in Figure 17.

Following Momma and Ferreira (2019), a red arrow was utilized to signal participants which scene (unaccusative vs. unergative) was to be described. The arrow was always presented on the actions and never on the objects. Unlike Momma and Ferreira (2019), we introduced a manipulation of the number of the object (plural

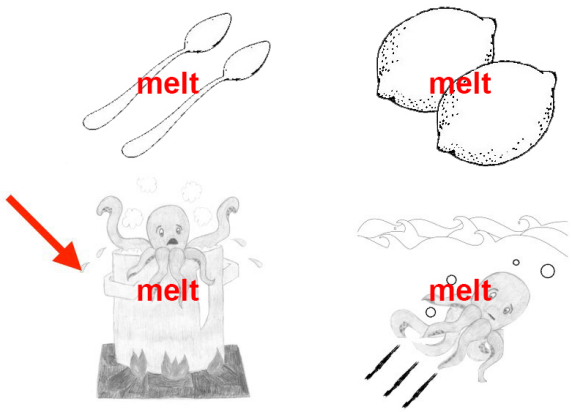
Figure 17
Base scenes adapted by Momma and Ferreira (2019).



vs. singular) to induce agreement attraction errors. In addition to visual manipulations of the scenes, we also included stroop-like semantic relatedness manipulation similar to previous picture word interference tasks. Unlike Momma and Ferreira (2019), we only included verbs as superimposed words. We have not modified previous combinations of picture-superimposed words used in Momma and Ferreira (2019). These picture-word pairs initially was chosen based on intuitiuon, and their semantic relation was later verified using a cosine similarity measure from Latent Semantic Analysis following previous studies (Momma et al., 2016). As a result, every event participants (and base scene) were utilized in 8 different conditions, resulting in 96 trials. An example condition with an arrow present is shown in Figure 18.

Within experimental items, the head of the subject phrase for the target sentence was always singular. To preclude participants from overrelying on the singular verbal forms, 6 more participants were added following the same manipulations above, resulting in 48 additional control trials. Differently from the experimental items, these control items all included a plural subject. Following Momma and Ferreira (2019), the pictures were selected from IPNP (Szekely et al., 2004). The semantically related and

Figure 18
An example condition from Experiment 1.



unrelated superimposed words were selected following intuition first, and then verified using ….

One example set of target sentences is shown in Table 2. Participants saw trials in a randomized order. We made sure that participants did not see the same event participant again without seeing all other event participants. The number manipulation is provided with slashes within each verb-type related condition.

Table 2
Experimental Conditions

Condition	Target Sentence	Related	Unrelated
Unaccusative	The octopus under the spoon/spoons is boiling.	melt	fall
Unergative	The octopus under the lemon/lemons is boiling.	run	smile
Control	The babies under the waffle/waffles are hiding.	find	consider

5.1.3 Procedure

All experiments reported in this paper were conducted using PCIbex, a platform that facilitated online data collection (Drummond, 2013; Zehr & Schwarz, 2018). Each experiment incorporated a voice recording function, allowing participants to provide

spoken responses. These recordings were transmitted to the server in real-time during the experimental session. Critically, only recordings from participants who successfully completed all trials were included in the final data analysis; recordings from participants who did not finalize the experiment were excluded.

Prior to commencing the experimental trials, participants underwent a familiarization period designed to acclimate them to the task. During this phase, participants were presented with action scenes depicting scenarios similar to those in the main experiment, but featuring different objects. Simultaneously, they heard the target sentence associated with the scene, and the sentence was also displayed in written form. Following this exposure, participants were instructed to verbally describe the same scene. Crucially, during their spoken description, the written text was no longer visible, ensuring they relied on their auditory memory and comprehension of the presented material.

Following familiarization, participants completed three practice sessions designed to increase their comfort with the experimental task. These practice sessions utilized a consistent set of images, presented in randomized order across participants. The first session consisted of experimental trials without the semantic interference task or time constraints. After each scene, participants were provided with the target sentence. The second session introduced the semantic interference task with a 0 millisecond Stimulus Onset Asynchrony (SOA). Finally, the third session incorporated the time limit, along with a -150ms SOA for the semantic interference task. Additionally, an auditory cue, a click sound, was introduced to signal the start of the participant's spoken response.

Each trial began with a 500ms fixation cross, followed immediately by the super-imposed distractor. 150ms later, the target picture appeared, while the distractors remained visible. Along with the picture, participants heard an auditory click, signaling the start of their response. 1850 ms after the click, the distractors disappeared. Participants were allotted an additional 3150 ms to complete their spoken description.

The target picture remained on screen for a total duration of 5000 ms, and the superimposed distractor word was visible for 2000 ms.

Throughout the entire experiment, participants were provided with clear visual feedback indicating when their speech was being recorded. This feedback was presented as a small red square, accompanied by the word 'Recording,' located at the top of the screen. The recording period commenced with the auditory click and concluded precisely when the target picture disappeared from the screen.

5.2 Pre-treatment

Our data collection yielded a total of 5613 recordings. Prior to statistical analysis, we performed automatic transcription and alignment of all recordings. Subsequently, each recording was manually reviewed by the researchers. During this process, we hand-coded any errors in the transcriptions or alignments, ensuring the accuracy of our dataset.

Recordings were transcribed using AssemblyAI's speech-to-text model, Universal-2, which is trained on a massive dataset of 12.5 million hours of audio, including 300,000 hours of supervised training data. Data encryption follows stricter EU laws and the data sent to AssemblyAI servers are immediately deleted after the transcription. Moreover, no customer data is shared with any other companies.

Their model employs a Recurrent Neural Network Transducer (RNN-T) architecture with 600 million parameters (Ghodsi et al., 2020). It has demonstrated a 7% word error rate on the CommonVoice v5.1 dataset, outperforming other existing models (Chkhetiani et al., n.d.). To enhance accuracy for our specific stimuli, we implemented a high word boost for every possible word present in our target sentences. The resulting transcriptions were then transferred into TextGrid format, without initial timestamps, for subsequent alignment.

The initial automatic transcriptions were then manually reviewed for accuracy, with any discrepancies corrected. Additionally, each recording was coded for the

presence of disfluencies. Table Table 3 presents the proportion of disfluencies observed within each experimental condition. It is important to note that individual trials could contain multiple disfluencies, and therefore, the summed percentages reported in Table 3 do not represent the overall percentage of trials excluded solely due to disfluencies.

Some participants used “gets boiled” instead of the target “is boiling”. Such errors were categorized as part of speech errors (POS). Stutters (St) at certain elements or interjections like *umm* were also observed. Some participants described the wrong picture (wpt). Incomplete verbs (iv), like producing only “boi-” before the recording ended, and incomplete utterances (i), where speech was cut off before the verb, were also observed. Certain responses were passivized (pass), deviating from the active voice target, or featured a missing determiner (nodet), such as omitting “the.” Errors also included starting the recording too late (toolate), using the distractor verb (d) instead of the target, omitting the auxiliary verb (not_is), or using incorrect noun (bad_n) or verb (bad_v)

Table 3

Counts of disfluencies by condition. Conditions are shortened into three character abbreviations. P = plural, S = singular, R = related, U = unrelated, A = unaccusative, E = unergative.

name	PRA	PRE	PUA	PUE	SRA	SRE	SUA	SUE	rowCount
Pos	14 (18%)	10 (13%)	10 (13%)	9 (11%)	11 (14%)	5 (6%)	10 (13%)	10 (13%)	79 (1.41%)
St	32 (19%)	20 (12%)	30 (18%)	18 (11%)	17 (10%)	24 (14%)	15 (9%)	12 (7%)	168 (2.99%)
Wpt	15 (17%)	5 (6%)	13 (15%)	7 (8%)	15 (17%)	6 (7%)	16 (19%)	9 (10%)	86 (1.53%)
Iv	13 (27%)	3 (6%)	6 (12%)	6 (12%)	6 (12%)	3 (6%)	6 (12%)	6 (12%)	49 (0.87%)
i	45 (16%)	24 (9%)	55 (20%)	24 (9%)	40 (15%)	27 (10%)	35 (13%)	23 (8%)	273 (4.86%)
Pass	1 (17%)	1 (17%)	3 (50%)	1 (17%)	0	0	0	0	6 (0.11%)
Nodet	1 (20%)	1 (20%)	1 (20%)	1 (20%)	1 (20%)	0	0	0	5 (0.09%)
Toolate	1 (25%)	1 (25%)	1 (25%)	1 (25%)	0	0	0	0	4 (0.07%)
d	21 (13%)	24 (15%)	21 (13%)	15 (9%)	22 (13%)	31 (19%)	13 (8%)	16 (10%)	163 (2.9%)
Not_is	29 (19%)	11 (7%)	27 (18%)	10 (7%)	28 (19%)	15 (10%)	20 (13%)	11 (7%)	151 (2.69%)
Bad_v	158 (14%)	132 (12%)	166 (15%)	104 (9%)	160 (14%)	142 (13%)	149 (13%)	116 (10%)	1127 (20.08%)
Bad_n	236 (15%)	177 (11%)	237 (15%)	160 (10%)	232 (14%)	184 (11%)	214 (13%)	166 (10%)	1606 (28.61%)

For our timing analyses, including pause likelihood, we employed a strict exclusion criterion, removing all trials that deviated from the target sentence. Specifically, we excluded trials where (1) the verb form was unidentifiable, (2) the

response did not conform to the predetermined sentence structure, or (3) the response contained any disfluencies detailed in Table 3. In contrast, our error analyses, focusing on agreement attraction and disfluencies, utilized the complete dataset without exclusions. The total percentage of trials excluded from the timing analyses based on these criteria was 37.77%. This percentage of exclusions is significantly higher than what was excluded in both Momma and Ferreira (2019) (11.96%) and Kandel and Phillips (2022) (22.98%, including initial omissions, disfluencies, and agreement errors). However, both of these studies were conducted in person, and our exclusion rate is comparable to the recent online production study by Momma and Yoshida (2023) (35.4% and 34.8% in their Experiment 3a and 3b).

Responses that contained neither agreement errors nor disfluencies were subjected to forced alignment with their corresponding transcriptions, resulting in precise onset and offset times for each word. For all experiments in this study, we employed the pre-trained English (US) models from Montreal Forced Aligner (MFA) version 3.1.0 (McAuliffe & Sonderegger, 2023, 2024a, 2024b). MFA reports an average 24ms difference between its automatically generated boundaries and manually annotated boundaries, which is comparable to the 26ms inter-transcriber reliability. To account for between-participant variability, we ensured that MFA was configured to consider individual speaker characteristics during the alignment process. Following the alignment procedure, a random sample of 100 responses was manually checked to verify the accuracy of the word onset and offset times.

5.3 Analysis

Prior to modeling, we cleaned the data and visualized general tendencies present in the data as summary plots using the tidyverse package system in R (Wickham et al., 2019). We did not include missing data points or exclusions in our analysis and assumed that data were missing completely at random (Van Buuren, 2018).

In summary plots, we are mainly interested in showing whether error bars

overlap for certain conditions. Our confidence intervals were computed following Morey (2008) and his correction of Cousineau (2005). We reported summaries in two formats. One format followed the classical (M = value, SE = value) format, where the mean and the standard error for a condition are presented. We also presented differences of means between conditions and the confidence intervals for those mean differences. These followed the format of ($\Delta_{A\text{-}B}$M [UpperCI, LowerCI]), where A and B are the conditions we contrasted.

Statistical analysis was carried out by fitting Bayesian GLMs with Stan (Carpenter et al., 2017; Stan Development Team, 2024). Models were fitted using the brms package in R (Bürkner, 2017a, 2018a), with weakly-informative priors, random intercepts and slopes for subject and items, and sum-coded predictors. Instead of having a maximal model, we included all manipulations as predictors and only the interactions we think plausible: the two-way interactions between the verb type and attractor number (our main question), the verb type and the semantic relatedness (different planning patterns for different verbs), and the three-way interactions. Measure specific details can be found in their own sections below.

In posterior plots, we visualized the mean of posterior distributions of Bayesian models for population-level coefficients and their interactions. We included %89 posterior intervals, and the probability of each coefficient to be smaller than 0, i.e. evidence rate. If a distribution does not include the value 0, we can say we have strong evidence for an effect. If the distribution consists of both negative and positive values, we can say that according to our data, there seems to be no evidence for an effect. On occasions in which only a small part of the distribution resides in a different sign area, we explicitly quantify our degree of belief towards an effect.

We used R version 4.4.1 (R Core Team, 2024a) and the following R packages: bayestestR v. 0.15.2 (Makowski et al., 2019), brms v. 2.22.2 (Bürkner, 2017b, 2018b, 2021), cmdstanr v. 0.8.1.9000 (Gabry et al., 2024), cowplot v. 1.1.3 (Wilke, 2024),

ggrepel v. 0.9.6 (Slowikowski, 2024), ggsci v. 3.2.0 (Xiao, 2024), kableExtra v. 1.4.0 (Zhu, 2024), knitr v. 1.49 (Xie, 2014, 2015, 2024), rmarkdown v. 2.29 (Allaire et al., 2024; Xie et al., 2018, 2020), Rmisc v. 1.5.1 (Hope, 2022), rstan v. 2.35.0.9000 (Stan Development Team, n.d.), scales v. 1.3.0 (Wickham et al., 2023), shiny v. 1.9.1 (Chang et al., 2024), systemfonts v. 1.2.1 (Pedersen et al., 2025), tidyverse v. 2.0.0.9000 (Wickham et al., 2019), tools v. 4.4.1 (R Core Team, 2024b).

5.3.1 Disfluency Analysis

The disfluency analysis in this study aimed to determine if any experimental condition posed a significantly greater challenge to speech production. We investigated the likelihood of producing a scene description with a disfluency within each trial using Bayesian Generalized Linear Models. Trials containing agreement attraction errors were excluded, while all other trials were included in the analysis. The presence of any disfluency, as reported in Table 3, was coded as 1.

We assumed a Bernoulli distribution with a probit link function, modeling the probability of disfluencies. The model-fitting specifications used in brms are reported in Table 4, along with the priors, and the contrasts of factors are reported in Table 6.

Table Table 4 details the Bayesian model specifications. The `Family` parameter, `bernoulli("probit")`, indicates that we assumed a binary outcome (presence or absence of disfluency) using a Bernoulli distribution with a probit link function, which transforms the probability scale to a standard normal scale. The `Formula` specifies the fixed and random effects in the model. The fixed effects include the intercept, the main effects of `verb_type`, `sem_type`, and `dist_num`, as well as their interactions. Random effects are included for `subject` and `head` (item), capturing variability across participants and items, respectively. These random effects include random intercepts and slopes for all fixed effects. The `Intercept Prior` is a Student's t-distribution with 3 degrees of freedom, a mean of 0, and a scale of 2.5, providing a weakly informative prior. The `Coefficient Prior` for fixed effects is a Normal distribution with a mean of

0 and a standard deviation of 1, also weakly informative. The σ Prior for the standard deviations of the random effects is a half-Cauchy distribution with a location of 0 and a scale of 1, ensuring positive standard deviations. The ρ Prior for correlations among random effects is an LKJ distribution with a shape parameter of 2, favoring lower correlations. We utilized 12000 iterations per chain, with the first 2000 iterations used as warmup.

Table 4
Bayesian Model specifications for Disfluency Errors and Agreement Errors in Experiment 1.

Parameter	Specification
Family	`bernoulli("probit")`
Formula	error ~ 1 + verb_type * sem_type * dist_num + (1 + verb_type * sem_type * dist_num | subject) + (1 + verb_type * sem_type * dist_num | head)
Intercept Prior	Student's t(3, 0, 2.5)
Coefficient Prior	Normal(0,1)
σ **Prior** (Random Effects)	Cauchy^{+}(0,1)
ρ **Prior**(Correlations)	LKJ(2)
Chains	12000 (2000 warmup)

Table Table 6 outlines the contrast coding used for the categorical predictors. `Verb-Type` is coded as +0.5 for Unaccusative and -0.5 for Unergative, allowing us to directly compare these verb types. `Semantic Relatedness` is coded as +0.5 for Related and -0.5 for Unrelated, facilitating the comparison of semantic relatedness conditions. `Attractor Number` is coded as +0.5 for Plural and -0.5 for Singular, enabling the analysis of attractor number effects.

Table 6
Contrasts used in the Bayesian model.

Predictors	+0.5	−0.5
Verb-Type	Unaccusative	Unergative
Semantic Relatedness	Related	Unrelated
Attractor Number	Plural	Singular

5.3.2 Attraction Analysis

For each trial that did not contain disfluencies, as previously detailed in Table 3, we coded the presence or absence of an agreement error. This coding included both unrevised errors (e.g., '*the octopus below the spoons are swimming*') and revised errors (e.g., '*the octopus below the spoons are is swimming,*' or '*the octopus below the spoons are swimming is swimming*'). Critically, incomplete productions of an agreement error (e.g., '*the octopus below the spoons a- is swimming*') were excluded from the analysis.

Following the completion of the experiment, we identified that two specific conditions, 'shrinking' and 'growing,' exhibited picture ambiguity that likely contributed disproportionately to the observed attraction effects. Consequently, we excluded these picture conditions from our primary analyses. To mitigate this issue in future studies, we ensured the use of clearer and less ambiguous picture stimuli in subsequent experiments.

Agreement attraction errors were analyzed using Bayesian Generalized Linear Model assuming a Bernoulli distribution and a probit link function. Predictors, priors, and contrast coding schemes are the same with the disfluency analysis shown in Table 4 and Table 6. We included interaction terms between predictors, as marked in the table, and maximal random slopes and intercepts for both participants and items to account for variability.

5.3.3 Time Analysis

For all time-related analyses, we included only trials that perfectly matched the target sentence, excluding trials with disfluencies or agreement attraction errors. However, we maintained trials with erroneous spatial marking or non-target modifier nouns, provided these were not the other noun in the scene. For example, 'lime' was accepted as a substitute, but 'spoon' was not, due to its co-presence with 'lemons.' This tolerance did not extend to the head noun or verb. Following data cleaning, we analyzed two key production time measurements: utterance onset time and preverbal onset time. Utterance onset time was defined as the interval between the auditory click and the onset of the second noun, including the definite article 'the,' following Momma and Ferreira (2019). This inclusion accounted for the potential use of 'the' as a filler, allowing participants to begin speaking while formulating the rest of the sentence. Preverbal time was measured as the interval between the onset of the second noun and the onset of the auxiliary verb.

We analyzed both utterance onset time and preverbal time using Bayesian Generalized Linear Models assuming an ex-Gaussian distribution and an identity link function. Production times typically exhibit a skewed distribution, characterized by a short left tail and a long right tail. In our experiment, we expected unaccusative trials and the presence of plural distractors to increase production times, and that these increases would not be present in all participants or conditions uniformly. Additionally, following Momma and Ferreira (2019), we expected semantic interference effects to manifest in trials with longer production times. Consequently, we anticipated that crucial data points would disproportionately contribute to the right-skewed tail, being less likely to fall within the main body of the distribution. Given the importance of the long tail for our research questions, using an ex-Gaussian distribution allows us to accurately model the observed data while preserving the potential for capturing subtle but important effects in the right tail of the distribution.

The ex-Gaussian distribution is particularly well-suited for reaction time and production time data because it combines a Gaussian component with an exponential component, effectively capturing both the main body of the data and the right-skewed tail characteristic of such data. This distribution's sensitivity to variations in the right tail allows us to accurately model differences in the distribution of late trials across conditions, which is crucial for investigating semantic interference effects. Furthermore, the ex-Gaussian allows us to model the data in its original scale, avoiding issues with data transformations that can disproportionately shrink the right tail and obscure important effects. By using the ex-Gaussian, we can directly interpret its parameters (mu, sigma, tau), with tau being particularly sensitive to changes in the exponential component and thus reflecting changes in the frequency or magnitude of late trials. This approach enhances our ability to detect potential semantic interference effects in the slower trials, without losing information due to data transformations.

Predictors and priors for the time analysis are detailed in Table 8. We included interaction terms between predictors, as marked in the table, and maximal random slopes and intercepts for both participants and items to account for variability. If the model did not converge, we successively removed by-item slopes. Given our repeated-measures design, where participants saw the same items across different conditions, we anticipated adaptation effects and systematic changes in estimates. To address this, we included the log of recurrence number (`l_recur`), representing the number of times an item was presented, as a predictor. Trial order was excluded due to its high correlation with recurrence number. We used the same contrasts presented in Table 6 for disfluency errors.

The `Family` parameter, `exgaussian()`, indicates that the model assumes an Exponential Gaussian distribution, which is suitable for modeling duration data. The Exponential component models the time until an event occurs, while the Gaussian component accounts for additional variability in the response times. The `Formula`

specifies the fixed and random effects in the model. The fixed effects include the intercept, the main effects of `verb_type`, `dist_num`, and `sem_type`, as well as their interactions. Additionally, `l_pres` is included as a fixed effect, representing a predictor that may influence the response time. Random effects are included for `subject_id` and `head` (item), capturing variability across participants and items, respectively. The random effects include random intercepts and slopes for the interaction terms of `verb_type`, `dist_num`, and `sem_type`, while the random intercept for `head` captures item-level variability.

The `Intercept Prior` is a Normal distribution with a mean of 1000 and a standard deviation of 50, reflecting a belief that the intercept should be large, but allowing for some variability. The `Coefficient Prior` for the fixed effects is a Normal distribution with a mean of 50 and a standard deviation of 10, indicating that the coefficients for the main effects and interactions are expected to be reasonably large but with some degree of uncertainty. The σ Prior for both random effects and residuals is a half-Cauchy distribution with a location of 50 and a scale of 10, which ensures that the model allows for potentially large but unlikely values for the standard deviations of the random effects and residuals.

We used 12000 iterations per chain, with the first 2000 iterations allocated as warmup. The model runs on 8 cores using the `cmdstanr` backend, ensuring efficient computation. This setup is designed to capture both participant-level and item-level variability in the duration data, allowing for more accurate and robust model estimation.

5.3.4 Pause Likelihood Analysis

Kandel and Phillips (2022) observed an increased likelihood of pauses (gaps in their paper) occurring between the offset of the second noun and the onset of the auxiliary verb (‘is’/‘are’) in sentences, specifically when the nouns mismatched in number (e.g., *the pinky above the greenies were mimming*). They hypothesized that this pause serves as a time index of agreement computation.

Table 8
Bayesian Model specifications for Time Duration Analysis in Experiment 1.

Parameter	Specification
Family	`exgaussian()`
Formula	`duration ~ 1 + verb_type * dist_num * sem_type + l_pres + (1 + verb_type * dist_num * sem_type | subject_id) + (1 | head)`
Intercept Prior	Normal(1000, 50)
Coefficient Prior	Normal(50, 10)
σ **Prior (Random Effects)**	Cauchy^{+}(50, 10)
σ **Prior (Residual)**	Cauchy^{+}(50, 10)
Chains	12000 (2000 warmup)
Backend	cmdstanr
Cores	8

Instead of strictly adhering to their methodology, which coded all non-zero differences between the offset of the second noun and the onset of the auxiliary verb as pauses, we adopted a modified approach. We classified gaps below 50ms as not containing a pause. This adjustment was made because we found that over thirty percent of the gaps fell within the 0-50ms range, and we observed a generally higher incidence of pauses in our data. We attribute this difference to the fact that Kandel and Phillips (2022) used only one verb, whereas our experiment involved more than 30 verbs, likely increasing the complexity of sentence planning. Therefore, we coded sentences with a difference of 50ms or greater between these points as containing a pause.

We hypothesize that these slowdowns should, at a minimum, occur in unergative sentences. Furthermore, we predict a similar increase in pause likelihood in unaccusative sentences if agreement computation occurs at the same time for both verb types. Given that our experiment required participants to recall a verb in addition to processing agreement, we hypothesize that the observed effects will be more pronounced than those found in previous agreement attraction experiments.

Similar to the disfluency and agreement attraction analyses, trials containing pauses were coded as 1 in the dataset. We then employed a Bayesian Generalized Linear Model, assuming a Bernoulli distribution with a probit link function, to analyze the presence of pauses. The predictors used in the models, along with their respective priors and contrast coding schemes, are the same with the ones in the disfluency and attractor analysis given in Table 4 and in Table 6. As in every model throughout the paper, models included random intercepts and slopes for participants and items.

5.4 Results

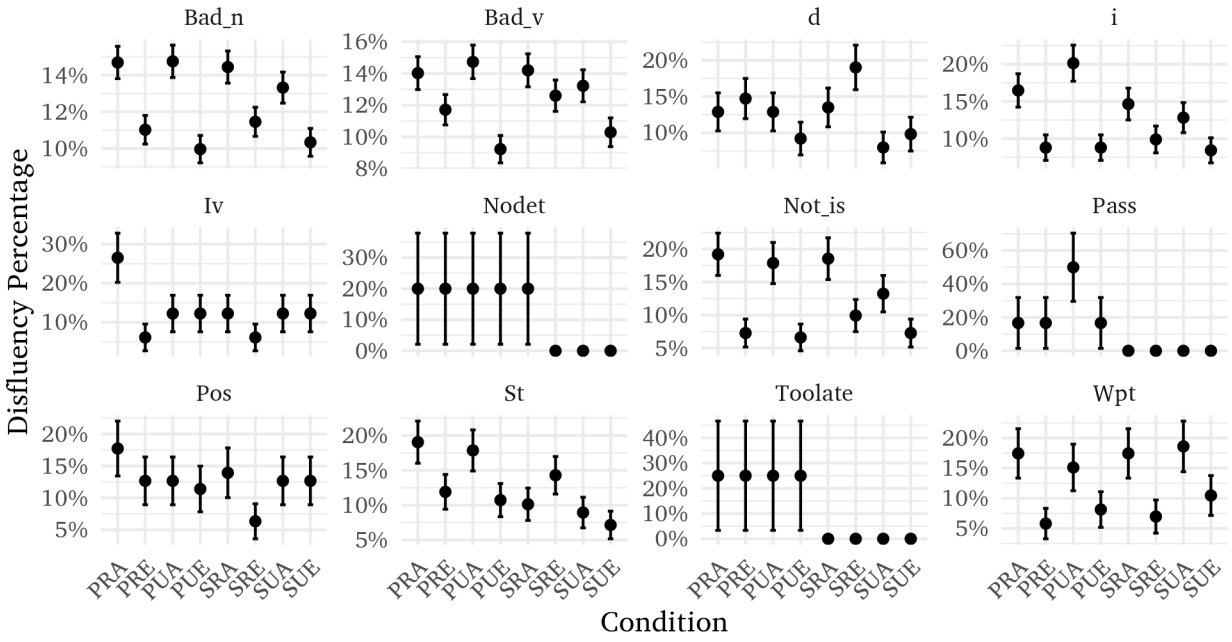
5.4.1 Disfluencies

To simplify the presentation of experimental conditions, we used three-letter abbreviations: 'P' for plural, 'S' for singular, 'R' for related, 'U' for unrelated, 'A' for unaccusative, and 'E' for unergative. Figure 19 illustrates the mean percentage of each disfluency type across these conditions. We observed that certain error types—stutters (at), incomplete utterances (i), incorrect nouns (bad_n), omitted auxiliary verbs (not_is), wrong picture descriptions (wpt), and incorrect verbs (bad_v)—were more prevalent in unaccusative ('A') sentences. This increased difficulty was evident in both plural ('P') and singular ('S') trials, but was more consistently pronounced in plural trials, particularly for stutters (at) and incomplete utterances (i). Interestingly, the use of the distractor verb (d) did not show a strong dependence on semantic relatedness ('R' vs. 'U'). Specifically, in singular distractor conditions, we found a clear effect of semantic relatedness when comparing related singular unergative (SRE) to related singular unaccusative (SRA) conditions, but this was not the case when comparing unrelated singular unaccusative (SUA) to unrelated singular unergative (SUE) conditions. This pattern was also absent in plural conditions. Overall, the data suggests that describing unaccusative pictures induced more difficulty and presented a slightly greater cognitive burden for participants during trials.

In Figure 20, we present the posterior distribution for the predictors of the probit

Figure 19

Mean Percentages with Standard Error for disfluencies in Experiment 1 for each condition.



regression coefficients of the model analyzing attraction errors in Experiment 1. In addition to the posterior distribution, we report the posterior probability of the effect of coefficient β being greater or smaller than zero, which we refer to as the degree of belief for a positive or negative effect.

Unaccusativity had a positive effect on disfluency error presence ($\hat{\beta} = 0.28$; $CI = [0.16; 0.39]$; $P(\beta > 0) > .999$) suggesting a very strong likelihood that unaccusative verbs caused more disfluencies. We also saw a relatively smaller positive effect for semantically related superimposed verbs ($\hat{\beta} = 0.08$; $CI = [0.01; 0.15]$; $P(\beta > 0) = .98$), indicating a high probability but weaker effect of semantic relatedness. Plural attractor also had a positive effect ($\hat{\beta} = 0.09$; $CI = [0.02; 0.17]$; $P(\beta > 0) = .993$), with a high degree of belief in its positive influence on disfluency errors.

The unaccusativity $\times$ semantic relatedness interaction showed a non-significant negative effect ($\hat{\beta} = -0.06$; $CI = [-0.21; 0.09]$; $P(\beta > 0) = .21$), indicating that the effect of unaccusativity was not significantly amplified when the distractor verb was semantically related to the target verb. Similarly, the semantic relatedness $\times$ plural

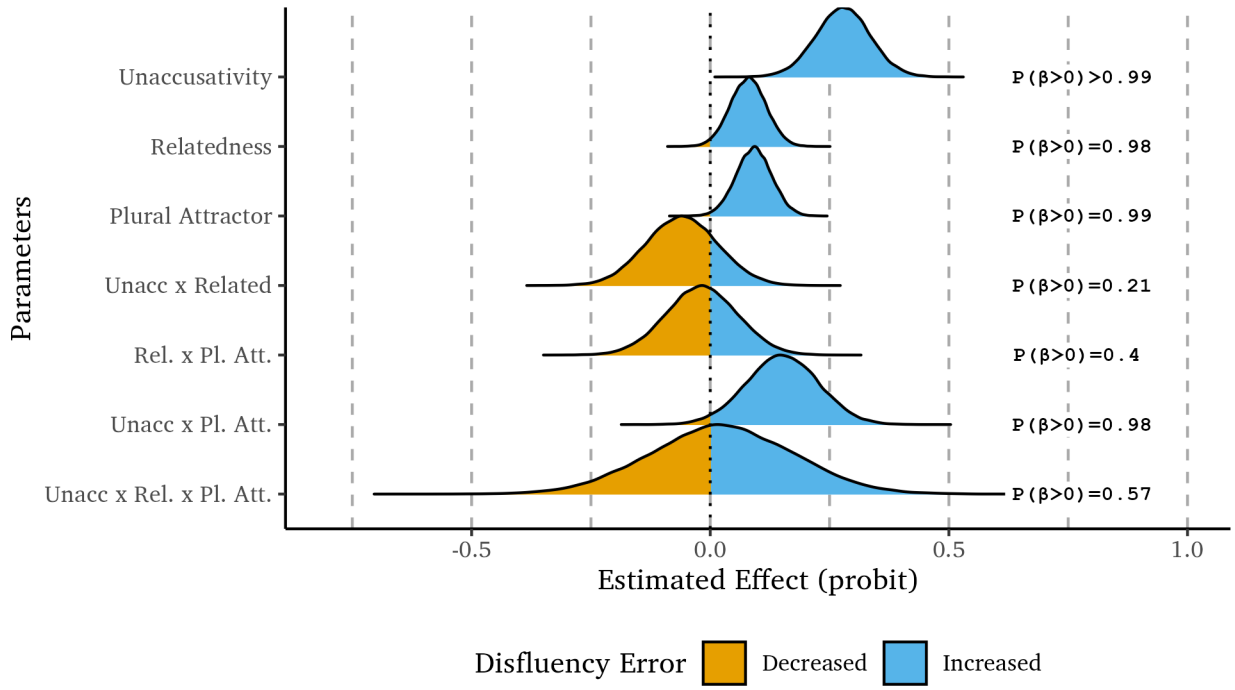
attractor interaction was non-significant ($\hat{\beta} = -0.02$; $CI = [-0.17; 0.13]$; $P(\beta > 0) = .40$), indicating that the effect of semantic relatedness was not significantly influenced by the plural attractor. The three-way interaction was also not significant ($\hat{\beta} = 0.02$; $CI = [-0.27; 0.32]$; $P(\beta > 0) = .57$)

However, the unaccusativity × plural attractor interaction was positive and significant ($\hat{\beta} = 0.15$; $CI = [0.00; 0.30]$; $P(\beta > 0) = .98$), suggesting that the effect of unaccusativity on disfluency errors was amplified when the attractor was plural. The difficulty of producing scene descriptions was significantly increased when the attractor noun was plural and the verb was unaccusative.

These results indicate that individual predictors such as unaccusativity, semantic relatedness, and plural attractor have significantly increased the difficulty of the task, which verifies our interpretation of the descriptive statistics. On the other hand, the only real difficulty inducing interaction was between unaccusativity and plural attractor.

Figure 20

Posterior distribution and the degree of belief for the probit regression coefficients for the model of disfluency errors in Experiment 1.



5.4.2 Agreement Attraction

Figure 21 presents the average proportions of agreement attraction errors as a function of verb type and attractor number. To focus on the effect of verb type, and given we had no specific predictions for semantic relatedness, we initially grouped the data into four core conditions, omitting semantic relatedness from the initial plot. The x-axis represents verb type (unaccusative “boil” vs. unergative “swim”), and line type indicates attractor number.

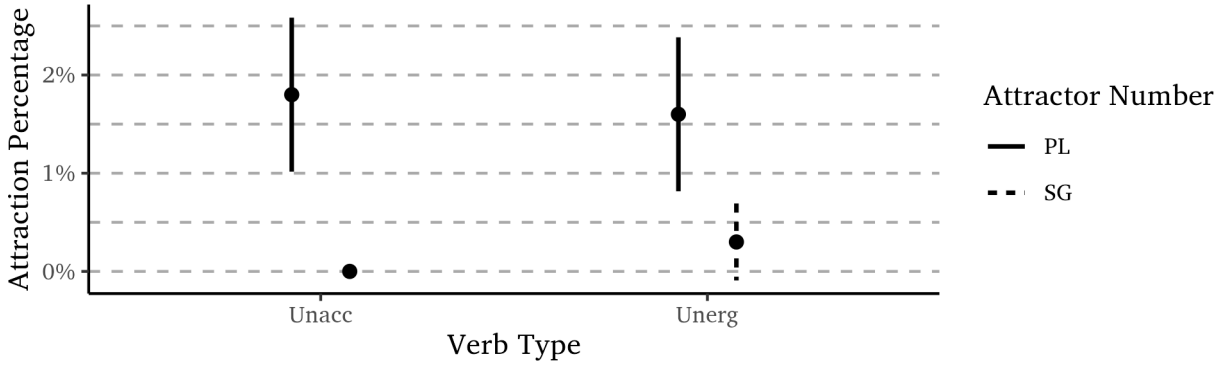
As predicted by Kandel and Phillips (2022), who used nonce verbs resembling unergatives, participants made significantly more errors when the attractor object was plural in unergative sentences (M = 0.02, SE = 0.00) compared to singular attractors (M = 0.00, SE = 0.00). However, the error rate was lower than the approximately 20% reported by Kandel and Phillips (2022). Crucially, we also observed a significant effect of plural attractors in unaccusative sentences. Participants produced more erroneous “are” markings with plural attractors (M = 0.02, SE = 0.00) than with singular attractors (M = 0.00, SE = 0.00). Notably, the magnitude of the plural attractor effect was comparable across verb types (Unaccusatives: $\Delta_{*\text{PL-SG}*}\text{M} = 0.02$ [0.01, 0.03], Unergatives: $\Delta_{*\text{PL-SG}*}\text{M} = 0.01$ [0.00, 0.02]).

It is crucial to note that, even though there was an observed effect of attraction in this experiment, the magnitude is extremely small compared to other production studies with picture description Veenstra et al. (2014). In fact, it is so low that even the experiment reporting the weakest attraction effects, Nozari and Omaki (2022) which found a 3% difference between the singular head and plural attractor conditions, and a 2% difference between the plural head and singular attractor conditions, showed more pronounced attraction errors.

Adding semantic relatedness to the analysis reveals a more complex picture. As depicted in Figure Figure 22, which presents the means and credible intervals for all eight experimental conditions, the attraction effect in unergative sentences manifests

Figure 21

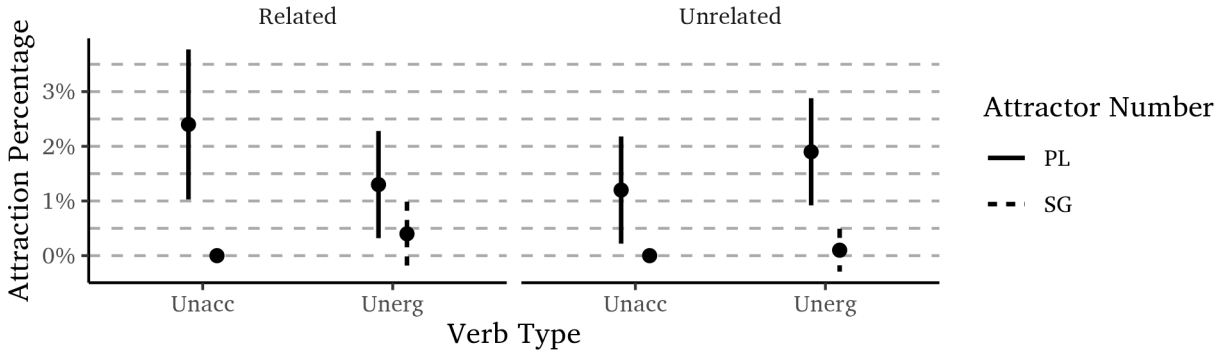
*The average percentages of agreement errors according to the experimental conditions (excluding semantic relatedness) in our Experiment 1 Error bars signal Morey-Cousineau 95% confidence intervals (1.96*Standard Error) corrected for by-subject variance and not by-item variance.*



exclusively when participants encountered an unrelated distractor ($\Delta_{*PL\text{-}SG*}M$ = 0.02 [0.01, 0.03]). Notably, this effect dissipates in the presence of related distractors ($\Delta_{*PL\text{-}SG*}M$ = 0.01 [-0.00, 0.02]). Conversely, within unaccusative sentences, the attraction effect emerges both in related conditions ($\Delta_{*PL\text{-}SG*}M$ = 0.02 [0.01, 0.04]) and in unrelated contexts ($\Delta_{*PL\text{-}SG*}M$ = 0.01 [0.00, 0.02]). However, the magnitude in unrelated contexts is as little as 1%, suggesting a reverse picture of the unergative scenes.

Figure 22

*The average percentages of agreement errors according to the experimental conditions (including semantic relatedness) in our Experiment 1. Error bars signal Morey-Cousineau 95% confidence intervals (1.96*Standard Error) corrected for by-subject variance and not by-item variance.*



In Figure 23, we present the posterior distribution for the predictors of the probit

regression coefficients of the model analyzing attraction errors in Experiment 1. In addition to the posterior distribution, we report the posterior probability of the effect of coefficient β being greater or smaller than zero, which we refer to as the degree of belief for a positive or negative effect on the presence of an attraction effect, an erroneous production of agreement. In this model, we included all predictor to our analysis.

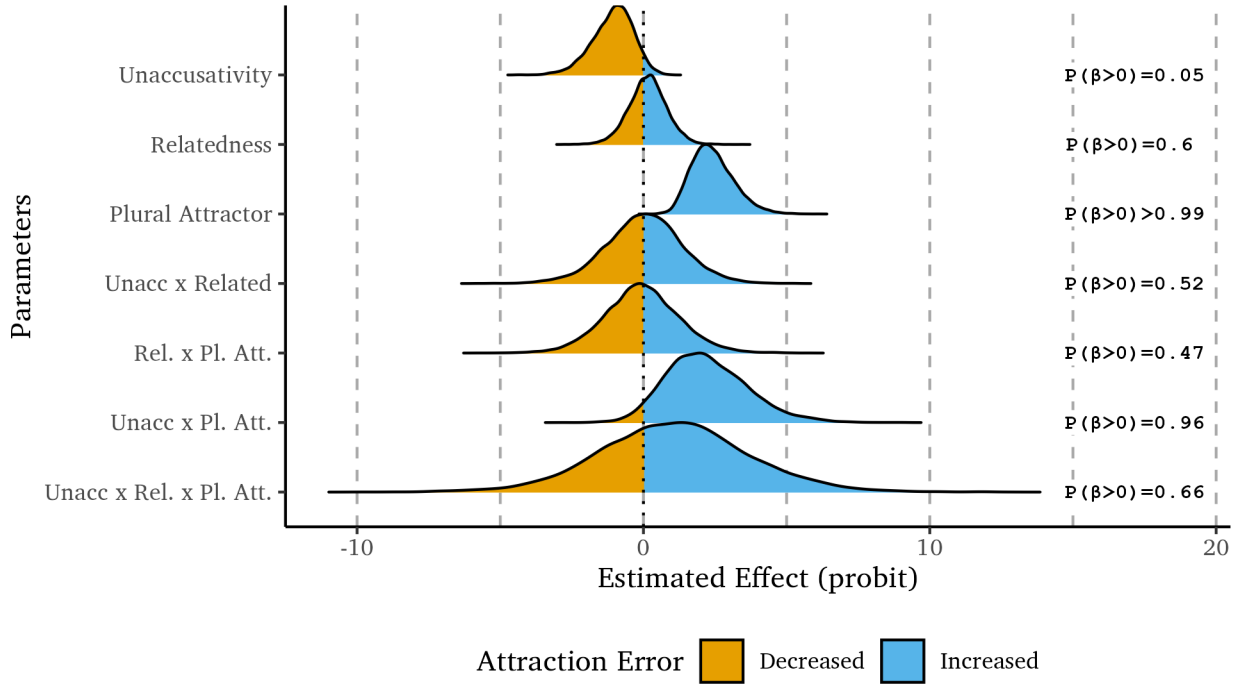
Unaccusativity had a main negative effect on attraction error presence ($\hat{\beta} = -1.07$; $CI = [-2.63; 0.16]$; $P(\beta < 0) = .95$), suggesting a small but significant likelihood that unaccusative scenes had overall less attraction errors, pooling over both semantic relatedness and number. This is not surprising, as in our descriptive plots, we observed that while unaccusatives with plural attractor had more attraction errors, participant did not do any attraction errors with singular attractor. Which was not the case for unergatives, where participants did attraction errors with both singular and plural attractors. As expected, the presence of a plural attractor had a significant positive effect with a very strong degree of belief in its influence ($\hat{\beta} = 2.42$; $CI = [1.20; 4.07]$; $P(\beta > 0) > .999$), however this effect is rather small in magnitude. Lastly, semantic relatedness show a high probability of no significant effect on attraction errors ($\hat{\beta} = 0.15$; $CI = [-1.18; 1.44]$; $P(\beta > 0) = .60$).

The only significant interaction was between unaccusativity and plural attractor ($\hat{\beta} = 2.22$; $CI = [-0.20; 5.39]$; $P(\beta > 0) = .96$), suggesting that the effect of plural attractor was significantly amplified by the unaccusative scenes. The unaccusativity $\times$ semantic relatedness interaction ($\hat{\beta} = 0.03$; $CI = [-2.67; 2.63]$; $P(\beta > 0) = .52$), the semantic relatedness $\times$ plural attractor interaction ($\hat{\beta} = -0.06$; $CI = [-2.61; 2.58]$; $P(\beta > 0) = .47$), and the three-way interaction ($\hat{\beta} = 1.08$; $CI = [-4.05; 6.34]$; $P(\beta > 0) = .66$) were all non-significant. These findings are surprising, given that our by-subject corrected descriptive statistics suggested that the difference between singular and plural attractors within unaccusatives and unergatives changed as a function of semantic relatedness, suggesting a three-way interaction. However, it is possible that

this effect was clouded by the small magnitude of the attraction errors in our data, as well as the no attraction effects in singular unaccusatives.

Figure 23

Posterior distribution and the degree of belief for the probit regression coefficients for the model of attraction errors in Experiment 1.



To uncouple this interaction, we fitted another model where we only included the plural attractors in our analysis. Figure 24 presents the posterior distribution for the predictors of the probit regression coefficients of the model analyzing attraction errors in Experiment 1, focusing on the plural attractors.

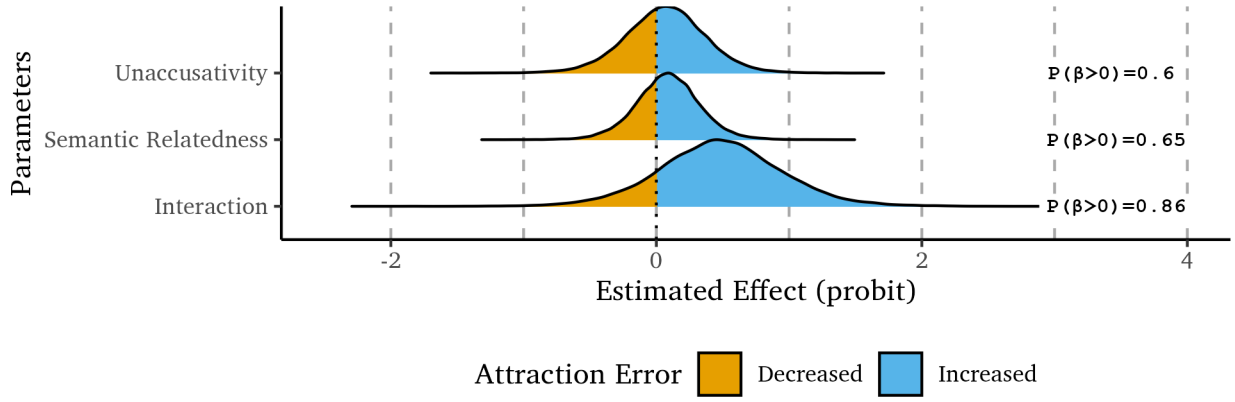
Within the plural attractors, we observed that the presence of the unaccusatives did not have any significant effect on attraction error presence ($\hat{\beta} = 0.07$; $CI = [-0.55; 0.67]$; $P(\beta > 0) = .60$). Similarly, semantic relatedness effect was not definitively positive or negative, suggesting no significant effect ($\hat{\beta} = 0.09$; $CI = [-0.40; 0.57]$; $P(\beta > 0) = .65$).

More importantly, we observed small to moderate effect for the unaccusativity × semantic relatedness interaction ($\hat{\beta} = 0.48$; $CI = [-0.45; 1.43]$; $P(\beta > 0) = .86$),

suggesting that the effect of unaccusativity was amplified by semantic relatedness on the occurrence of attraction errors, but the evidence for this is quite low. This verified our previous doubt that this interaction was clouded by the stronger main effects due to the 0 attraction errors in singulars.

Figure 24

Posterior distribution and the degree of belief for the probit regression coefficients for the model of attraction errors in conditions with plural attractors of Experiment 1.



Since our main question can also be operationalized as an effect of a plural attractor in unaccusative and unergative scenes, we used two additional models with only unaccusative and unergative scenes, respectively. These models included the same model-fitting details, except for the exclusion of verb type predictor and its interaction.

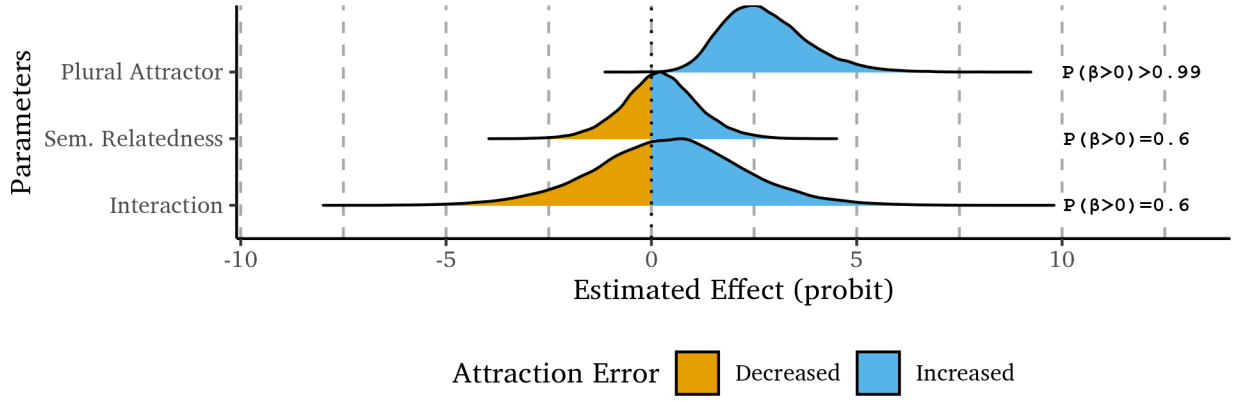
Figure 25 presents the posterior distribution for the predictors of the probit regression coefficients of the model analyzing attraction errors in Experiment 1, focusing on the unaccusative scenes.

The presence of plural attractor showed a strong effect on attraction errors ($\hat{\beta} = 2.77$; $CI = [1.09; 5.11]$; $P(\beta > 0) > .999$), verifying our descriptive results that the attraction effects was present in unaccusative sentences. On the other hand, semantic relatedness ($\hat{\beta} = 0.22$; $CI = [-1.61; 2.09]$; $P(\beta > 0) = .60$) and its interaction with the plural attractor ($\hat{\beta} = 0.43$; $CI = [-3.24; 4.09]$; $P(\beta > 0) = .60$) did not show a strong effect on attraction errors.

Figure 26 presents the posterior distribution for the predictors of the probit

Figure 25

Posterior distribution and the degree of belief for the probit regression coefficients for the model of attraction errors in conditions with unaccusative scenes of Experiment 1.

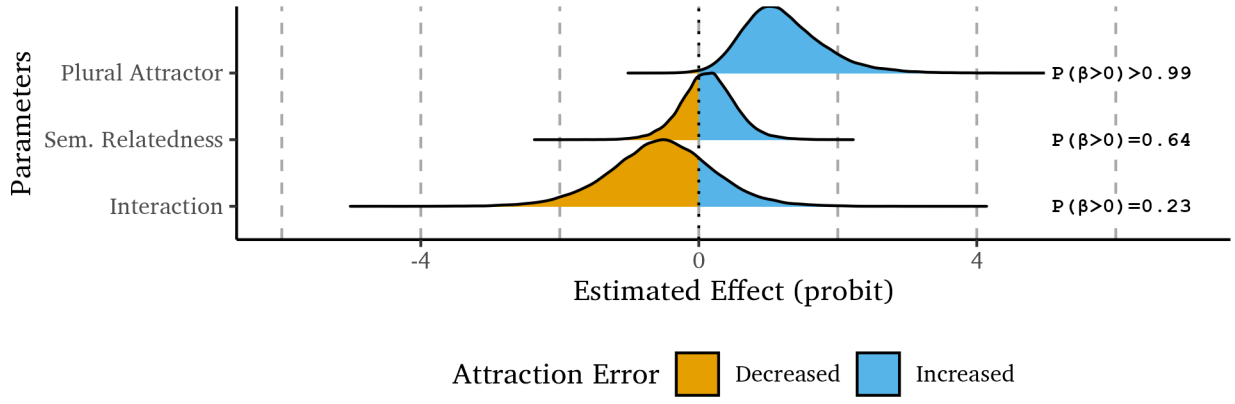


regression coefficients of the model analyzing attraction errors in Experiment 1, focusing on the unergative scenes.

Our results in unergative sentences showed a similar picture. Our model showed a strong evidence for a positive effect of plural attractor on attraction errors ($\hat{\beta} = 1.22$; $CI = [0.28; 2.56]$; $P(\beta > 0) = .995$) and verified the presence of attraction effects in unergative sentences. On the other hand, we did not have sufficient evidence for an effect of semantic relatedness ($\hat{\beta} = 0.22$; $CI = [-1.61; 2.09]$; $P(\beta > 0) = .60$) or its interaction with the plural attractor ($\hat{\beta} = 0.43$; $CI = [-3.24; 4.09]$; $P(\beta > 0) = .60$).

Figure 26

Posterior distribution and the degree of belief for the probit regression coefficients for the model of attraction errors in conditions with unergative scenes of Experiment 1.



Taken together, we saw that participants did attraction errors in both

unaccusative and unergative scenes, suggesting that even thought the planning of the verb might be different, the planning of the agreement is quite similar in these environments. However, this effect was particularly attenuated. The possible reasons for this attenuations is discussed in our discussion session. Moreover, we observed rather an interesting picture: the attraction effects and semantic interference interacted in a way that participants did more errors in unaccusative scenes with semantically-related superimposed verbs and in unergative scenes with semantically-unrelated superimposed verbs.

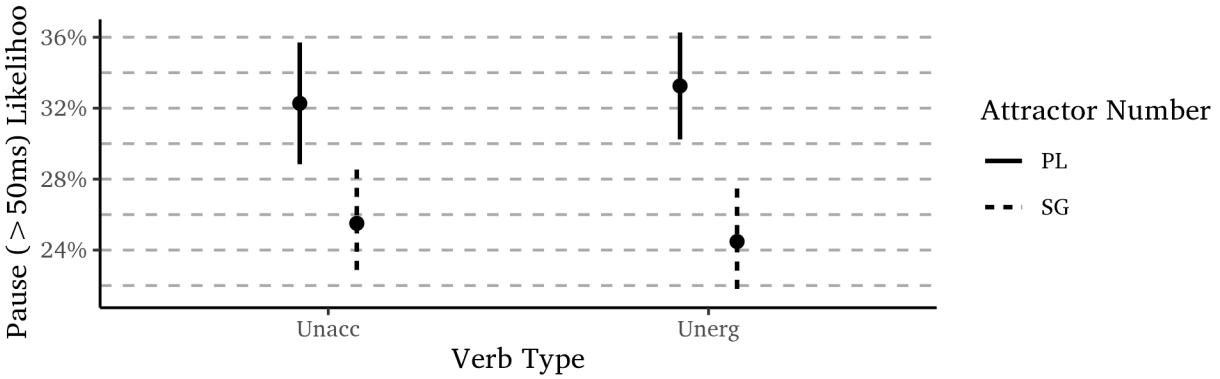
5.4.3 Pause Likelihood

Figure 27 displays the average likelihood of pauses occurring between the second noun (e.g., “spoon”) and the auxiliary verb (e.g., “is”). This analysis was restricted to trials with utterances that perfectly matched the target sentence. Compared to Kandel and Phillips (2022) (M = 0.11, SE = 0.00), our participants exhibited a higher overall pause frequency (M = 0.29, SE = 0.01). We attribute this to the increased number of verbs in our experiment; even with practice, retrieving a verb from a pool of 30 likely required more cognitive effort than in Kandel and Phillips (2022), where only one verb was used. Consistent with Kandel and Phillips (2022), unergative sentences showed a clear difference between plural and singular attractors. Participants were significantly more likely to pause in unergative sentences with plural attractors (M = 0.33, SE = 0.02) compared to singular attractors (M = 0.24, SE = 0.02). This plural attractor effect was also observed in unaccusative sentences. Participants describing unaccusative scenes showed a greater likelihood of pausing with plural attractors (M = 0.32, SE = 0.02) compared to singular attractors (M = 0.26, SE = 0.02). This is not surprising, given that agreement attraction error rates were comparable across both verb types.

Adding semantic relatedness to the analysis reveals an interesting pattern as in Figure 28. The effect of the plural attractor that was observed in unergative sentences was limited solely to the unrelated distractor condition. Specifically, when participants

Figure 27

*The average likelihood of pause between the second noun and the auxilliary according to the experimental conditions (excluding the semantic relatedness) in our Experiment 1. Error bars signal Morey-Cousineau 95% confidence intervals (1.96*Standard Error) corrected for by-subject variance and not by-item variance.*



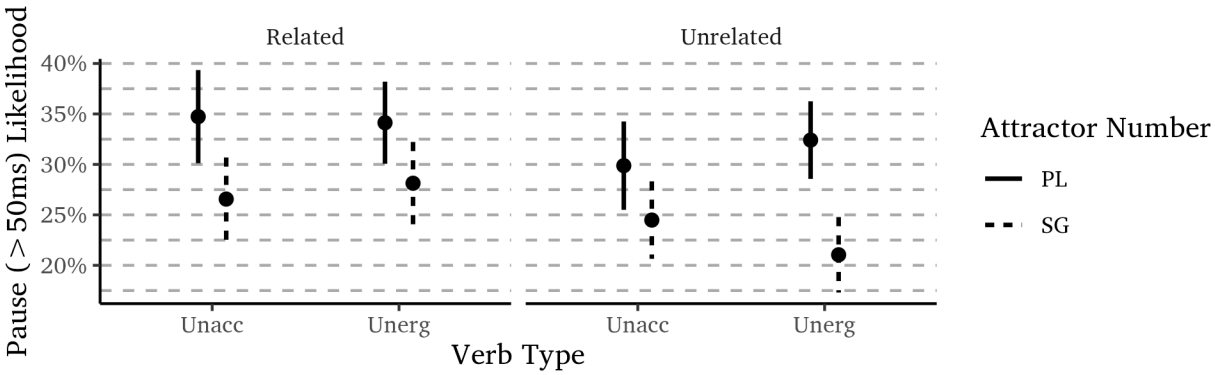
encountered unrelated superimposed verbs for unergative scenes, they exhibited significantly more pauses with plural attractors compared to singular attractors ($\Delta_{*PL\text{-}SG*}M = 0.11$ [0.06, 0.17]). However, when the superimposed verbs for unergative scenes were semantically related to the scene, participants showed comparable amounts of pauses with singular and plural attractors ("spoon" vs. "spoons") ($\Delta_{*PL\text{-}SG*}M = 0.06$ [0.00, 0.12]). Although there is a difference in means, the standard errors and confidence intervals suggest that their distributions likely overlap.

Decoupling semantic relatedness also altered the pattern within unaccusative sentences. While the semantically related superimposed induced a difference in pause likelihood between plural and singular attractor conditions ($\Delta_{*PL\text{-}SG*}M = 0.08$ [0.02, 0.14]), this effect seems to be barely substantial. With unrelated semantic distractors, there was only a difference in means of pause likelihood between the plural and the singular attractor, however, the confidence interval heavily overlap ($\Delta_{*PL\text{-}SG*}M = 0.05$ [-0.00, 0.11]).

In Figure 29, we present the posterior distribution for the predictors of the probit regression coefficients of the model analyzing pause likelihood between the second noun and the auxiliary verb in Experiment 1. In addition to the posterior distribution,

Figure 28

*The average likelihood of pause between the second noun and the auxilliary according to the experimental conditions (including the semantic relatedness) in our Experiment 1. Error bars signal Morey-Cousineau 95% confidence intervals (1.96*Standard Error) corrected for by-subject variance and not by-item variance.*



we report the posterior probability of the effect of coefficient β being greater or smaller than zero, which we refer to as the degree of belief or evidence for a positive or negative effect on the presence of an overall increased or decreased pause likelihood.

Our model showed that plural attractor significantly increased pause likelihood ($\hat{\beta} = 0.27$; $CI = [0.06; 0.49]$; $P(\beta > 0) = .991$). Semantic relatedness also showed a significant positive effect ($\hat{\beta} = 0.15$; $CI = [0.03; 0.27]$; $P(\beta > 0) = .991$). We expected both of these manipulation to induce a pause in speech during the sentence production.

However, we did not find enough evidence for a consistent effect of unaccusativity on pause likelihood ($\hat{\beta} = 0.01$; $CI = [-0.25; 0.26]$; $P(\beta > 0) = .52$). The lack of an effect is expected given that attraction happens both in unaccusative and unergative sentences, thus the timing index of the agreement should be visible in both environments.

There was no evidence for an interaction between unaccusativity $\times$ plural attractor ($\hat{\beta} = -0.02$; $CI = [-0.28; 0.24]$; $P(\beta > 0) = .43$) or plural attractor $\times$ semantic relatedness ($\hat{\beta} = -0.08$; $CI = [-0.31; 0.16]$; $P(\beta > 0) = .26$). Similarly, the unaccusativity $\times$ semantic relatedness interaction ($\hat{\beta} = -0.04$; $CI = [-0.29; 0.20]$; $P(\beta > 0) = .36$) also did not show an effect.

However, we found a weak evidence for the three way interaction between unaccusativity × plural attractor × semantic relatedness ($\hat{\beta} = 0.28$; $CI = [-0.28; 0.86]$; $P(\beta > 0) = .85$). Given that we do not have zero magnitude of an effect in pause likelihood, unlike attraction, this three way interaction surfaced instead of being clouded by other interactions. Even though the evidence is moderate, the credible interval remains wide. This potentially suggest that the effect of the plural attractor as a function of the verb type changed as a function of semantic relatedness. The consequence of this three way interaction is our surprising effect in which we see attraction in unaccusative with related superimposed words but with unrelated superimposed words with unergatives. This dichotomy is also reflected in the pause likelihoods.

Figure 29

Posterior distribution and the degree of belief for the probit regression coefficients for the model of pause likelihoods in Experiment 1.

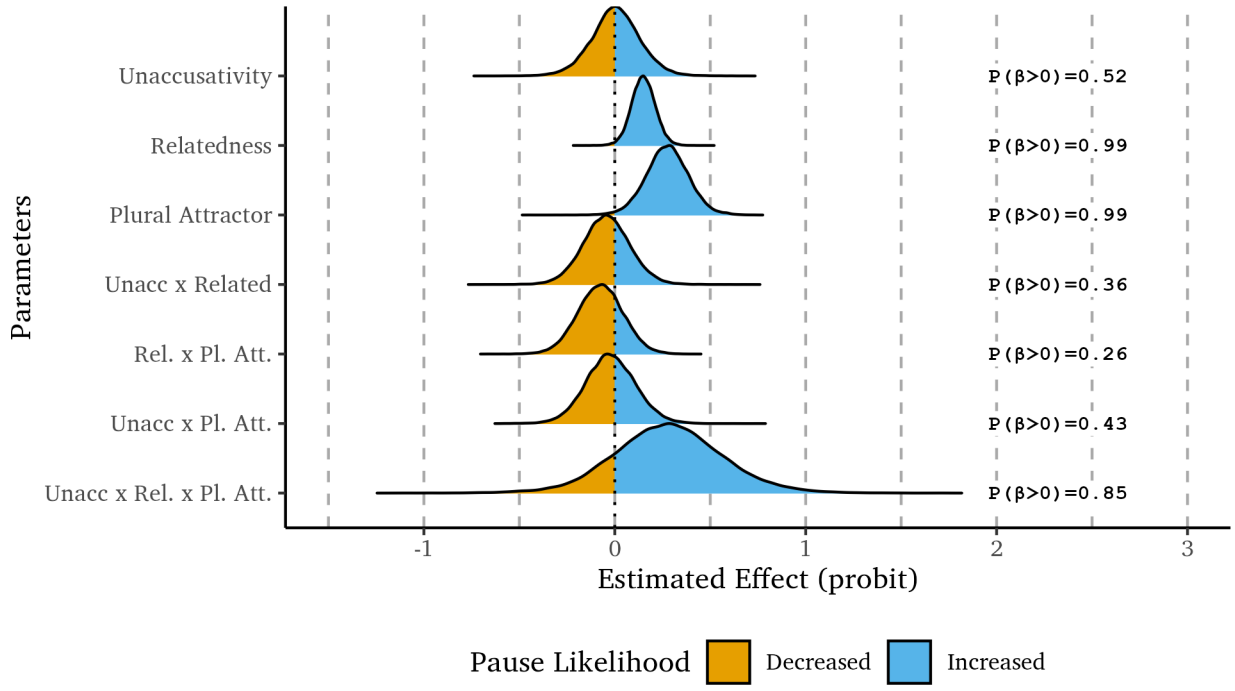


Figure 30 presents the posterior distribution for the predictors of the probit regression coefficients of the model analyzing pause likelihoods only in unaccusative

scenes in Experiment 1.

Our model verified the positive effect of plural attractor on pause likelihood ($\hat{\beta} = 0.26$; $CI = [0.04; 0.47]$; $P(\beta > 0) = .99$). We also found a significant positive effect of semantic relatedness ($\hat{\beta} = 0.13$; $CI = [-0.05; 0.30]$; $P(\beta > 0) = .92$), showing a strong trend toward increasing pause likelihood. However, surprisingly, we did not find an interaction between the plural attractor × semantic relatedness interaction ($\hat{\beta} = 0.04$; $CI = [-0.31; 0.40]$; $P(\beta > 0) = .60$). Our previous model and the descriptive results suggested that the effect of plural attractor on pause likelihood and attraction within unaccusative scenes changed as a function of semantic relatedness, meaning in unrelated superimposed words there was no or reduced effect, but there was an effect in related superimposed words. However, our model did not find enough evidence for this interaction.

Figure 30

Posterior distribution and the degree of belief for the probit regression coefficients for the model of pause likelihoods in conditions with unaccusative scenes of Experiment 1.

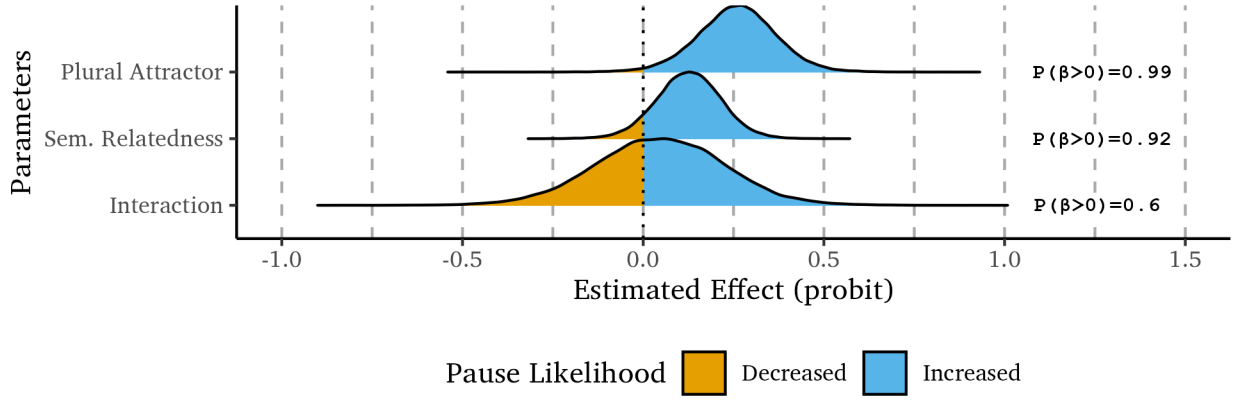


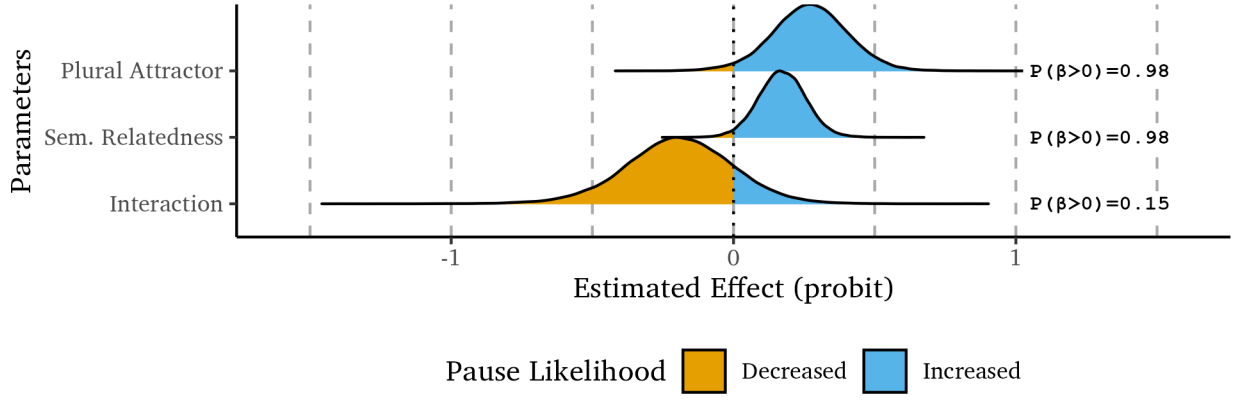
Figure 31 presents the posterior distribution for the predictors of the probit regression coefficients of the model analyzing pause likelihoods only in unergative scenes in Experiment 1.

Similar to unaccusative scenes, we found strong evidence for a positive effect of plural attractor on pause likelihood ($\hat{\beta} = 0.27$; $CI = [0.00; 0.53]$; $P(\beta > 0) = .98$). Semantic relatedness also increased pause likelihood within unergatives ($\hat{\beta} = 0.17$;

$CI = [0.01; 0.34]$; $P(\beta > 0) = .98$). More importantly, we find a weak evidence for an negative effect pf plural attractor × semantic relatedness interaction ($\hat{\beta} = -0.19$; $CI = [-0.59; 0.18]$; $P(\beta > 0) = .15$). This means that even though relatedness did not amplify the effect of plural attractors in unaccusative scenes, it did reduce the effect of plural attractors in unergative scenes. This is consistent with our descriptive results and the three way interaction in the main model.

Figure 31

Posterior distribution and the degree of belief for the probit regression coefficients for the model of pause likelihoods in conditions with unergative scenes of Experiment 1.



Similar to Kandel and Phillips (2022), we found that pause likelihood reflected the difficulty in computing agreement. The plural attractor overall increased the pause likelihood. We also found that semantic relatedness effected the pause likelihood. This effect of semantic relatedness was not affected by the verb type, suggesting that the semantic interference effected the time between the second NP and the auxiliary verb independent of the verb type. This is not consistent with the early-planning idea proposed by Momma and Ferreira (2019). Moreover, we also see the interesting dichotomy present in attraction effects. Similar to attraction effects, the unaccusative scenes with related superimposed word enabled the plural attractor to increase the pause likelihood, whereas the unergative scenes with unrelated superimposed word enabled the plural attractor to increase the pause likelihood. This suggests that the planning of the agreement is not only affected by the verb type but also by the semantic

relatedness of the superimposed word.

5.4.4 Utterance Onset Latency

We are interested in the utterance onset timing as a function of semantic relatedness to verify early planning of the unaccusative verb. To that end, we first plot our onset latency data by collapsing the attractor number manipulation in Figure 32. Unlike previous plots, the immediate comparisons are between semantic relatedness conditions; therefore, the linetypes signal the semantic relatedness (related vs. unrelated).

We see that, as expected from Momma and Ferreira (2019), the unergative scenes are not affected by semantic relatedness. Participants spend similar times before starting to utter the sentence in conditions with a semantically related superimposed word (M = 1015.58, SE = 11.10) and a semantically unrelated superimposed word (M = 1025.19, SE = 11.76).

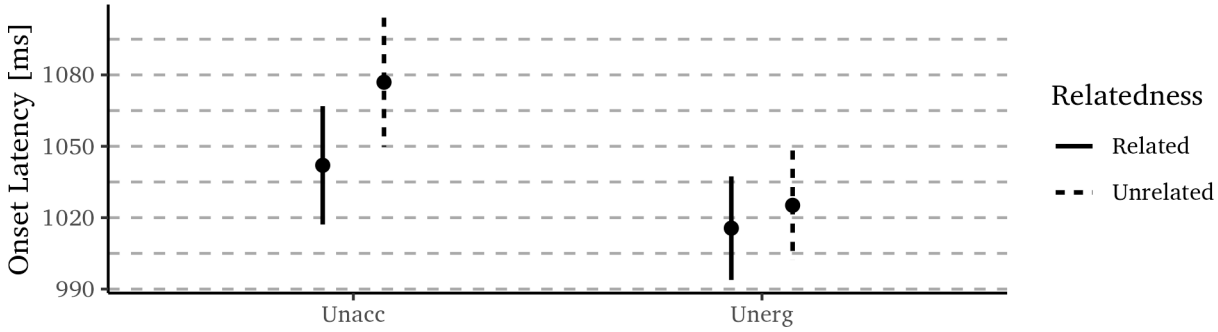
However, the onset latency in unaccusatives does not align with the findings of Momma and Ferreira (2019). Our descriptive results suggest that unaccusative scenes with an unrelated distractor (M = 1076.93, SE = 13.84) take more time to start uttering compared to their related condition counterparts (M = 1042.01, SE = 12.69). This pattern is the reverse of the findings of Momma and Ferreira (2019), in which the related ones took more time to start uttering. Additionally, even though there was a difference in means, the confidence intervals for the difference span across 0 and confidence intervals of the effects are overlapping (Morey, 2008).

The overall pattern did not change drastically when we included attractor number conditions, as shown in Figure 33. Independent of the attractor number, unergatives are not affected by the relatedness of the semantic distractor (SG: $\Delta_{\text{*Related-Unrelated*}}$M = -18.09 [-59.35, 23.16], PL: $\Delta_{\text{*Related-Unrelated*}}$M = -1.00 [-42.74, 40.74]).

As for unaccusatives, the increased onset latency we previously observed in our pooled graph was only present in scenes with a singular distractor. Participants seemed

Figure 32

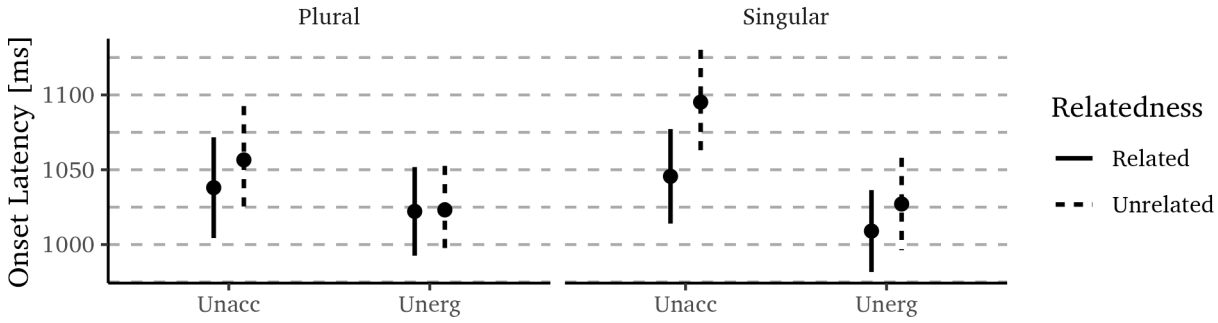
*The average onset times according to the experimental conditions (excluding the attractor number) in our Experiment 1. Error bars signal Morey-Cousineau 95% confidence intervals (1.96*Standard Error) corrected for by-subject variance and not by-item variance.*



to be spending more time to start uttering sentences while describing unaccusative scenes with a singular attractor ($\Delta_{*\text{Related-Unrelated}*}$M = -49.58 [-96.71, -2.45]), but not with a plural attractor ($\Delta_{*\text{Related-Unrelated}*}$M = -18.55 [-67.84, 30.73]). This increase in onset latency appear to be only suggestive, given the slightly overlapping confidence intervals and almost entirely negative confidence intervals of the difference.

Figure 33

*The average onset times according to the experimental conditions (including the attractor number) in our Experiment 1. Error bars signal Morey-Cousineau 95% confidence intervals (1.96*Standard Error) corrected for by-subject variance and not by-item variance.*



In Figure 34, we present the posterior distribution for the predictors of an Gaussian-exponential mixture regression coefficients of the model analyzing onset latencies in Experiment 1, i.e. how much time it takes to start uttering the sentence. In addition to the posterior distribution, we report the posterior probability of the effect of coefficient β being greater or smaller than zero, which we refer to as the degree of belief

or evidence for a positive or negative effect on the presence of an overall increased or decreased onset latency.

Our model showed a strong effect for a positive effect of unaccusativity ($\hat{\beta} = 14.83$; $CI = [2.17; 27.34]$; $P(\beta > 0) = .99$), increasing onset latencies. The presence of a plural attractor ($\hat{\beta} = -2.71$; $CI = [-13.94; 8.40]$; $P(\beta > 0) = .32$) or semantic relatedness ($\hat{\beta} = -3.39$; $CI = [-14.51; 7.71]$; $P(\beta > 0) = .27$) were not associated with any evidence for either positive or negative effect.

We were not able to find any evidence for any interaction in our model. The unaccusativity $\times$ plural attractor interaction did not show any evidence for an effect ($\hat{\beta} = -4.17$; $CI = [-27.86; 19.65]$; $P(\beta > 0) = .36$). Similarly, we did not find any evidence for the plural attractor $\times$ semantic relatedness interaction ($\hat{\beta} = -6.67$; $CI = [-31.65; 17.96]$; $P(\beta > 0) = .30$).

One theoretically expected interaction was the unaccusativity $\times$ semantic relatedness interaction, pointing out people's specific slowdown with unaccusative verbs with a semantically related super imposed words. We found no evidence for such an interaction ($\hat{\beta} = 2.14$; $CI = [-22.30; 26.31]$; $P(\beta > 0) = .57$).

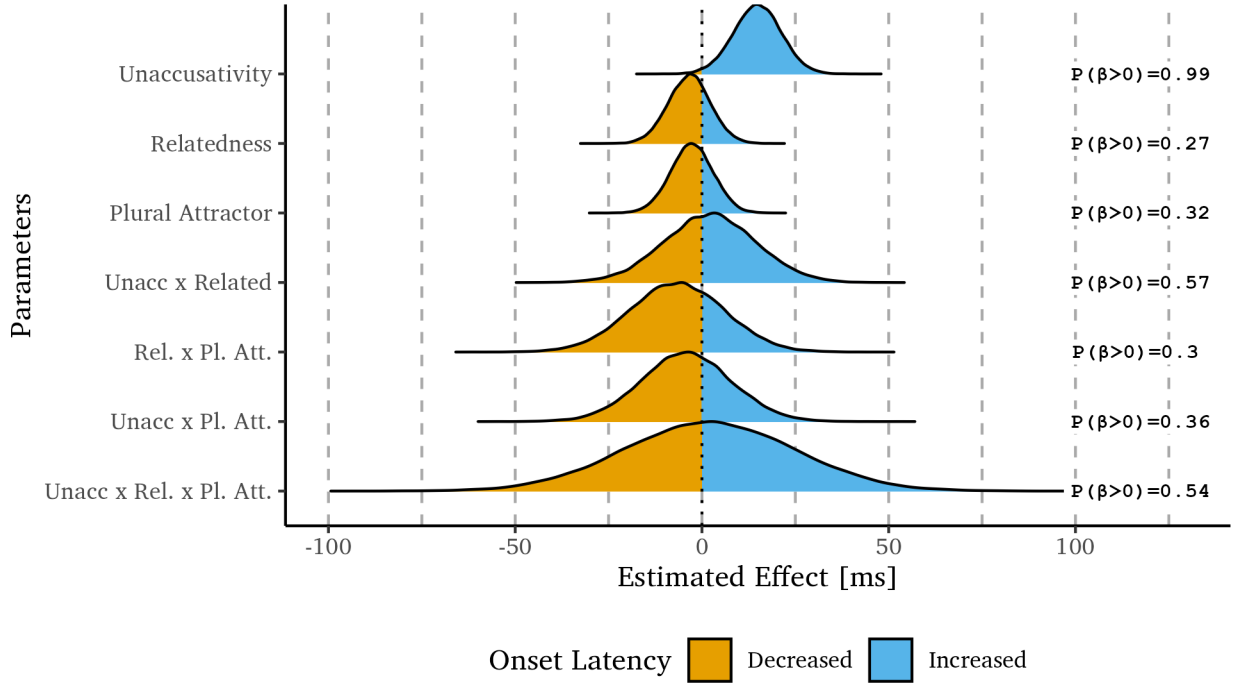
Another expected interaction was the three way interaction. This interaction was borne out of our descriptive plots, we see that there was a suggestion of a negative effect of semantic relatedness on unaccusative within singular attractors. However, this suggestion was not verified by our models ($\hat{\beta} = 2.00$; $CI = [-42.80; 46.43]$; $P(\beta > 0) = .54$). There can be couple of reasons for this. Firstly, it might be the case that this effect is driven by certain outliers in the data, which is shrinked by the model, as intended. Our descriptive results were by-subject corrected and the model integrates by-subject and by-item slopes and intercepts. Item-based outliers might have been shrunk by the model.

Secondly, it might be the case that the effect is not as strong as we thought it was. The confidence intervals of the differences in means were overlapping, suggesting that

the effect might be small and we did not have enough participants to detect this effect.

Figure 34

Posterior distribution and the degree of belief for the exGaussian regression coefficients for the model of onset latencies in Experiment 1.



We again fitted two additional models according to the questions that are of importance for us. For the onset latency, one important question was that whether or not the unaccusative verbs were planned early. To answer this question, we fitted a model to conditions that has only related or unrelated semantically related superimposed words. If unaccusative scenes are planned early and unergative ones are not, we would see a main effect of the verb type in these models.

Figure 35 presents the posterior distribution for the predictors of the Gaussian-exponential mixture regression coefficients of the model analyzing onset latencies only in conditions with related superimposed words in Experiment 1.

Our model showed no evidence for a main effect of the plural attractor ($\hat{\beta} = -7.50$; $CI = [-28.20; 13.74]$; $P(\beta > 0) = .23$) or the interaction between the attractor and the unaccusativity (verb type) ($\hat{\beta} = -9.05$; $CI = [-47.02; 29.08]$;

$P(\beta > 0) = .32$). However, we found a very weak evidence for a positive effect of unaccusativity ($\hat{\beta} = 11.59$; $CI = [-12.70; 35.08]$; $P(\beta > 0) = .83$), suggesting that unaccusative scenes took more time to start the utterance compared to unergative scenes within related superimposed word conditions. The effect is very weak and the confidence intervals are overlapping, suggesting that we might not have enough participants to find enough evidence.

Figure 35

Posterior distribution and the degree of belief for the exGaussian regression coefficients for the model of onset latencies of conditions with semantically related superimposed word in Experiment 1.

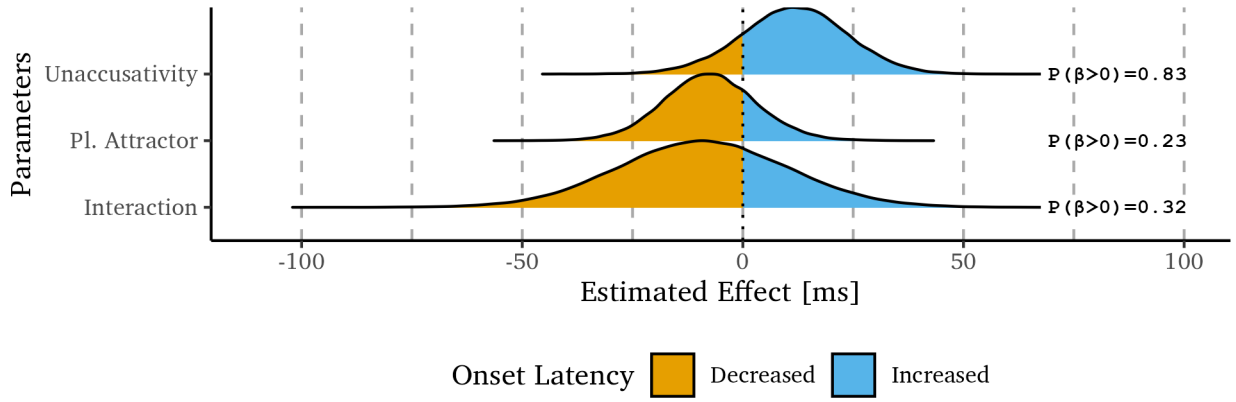


Figure 36 presents the posterior distribution for the predictors of the Gaussian-exponential mixture regression coefficients of the model analyzing onset latencies only in conditions with unrelated superimposed words in Experiment 1.

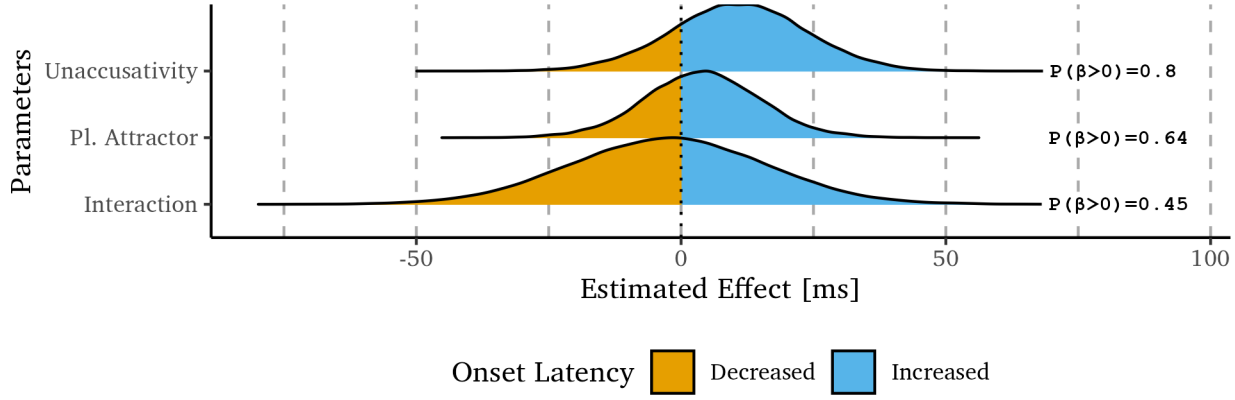
Our model for unrelated superimposed word conditions also showed not effect for the plural attractor ($\hat{\beta} = 4.03$; $CI = [-17.94; 25.96]$; $P(\beta > 0) = .64$) or the interaction between the attractor and the unaccusativity (verb type) ($\hat{\beta} = -2.48$; $CI = [-38.92; 33.75]$; $P(\beta > 0) = .45$). The lack of interaction is surprising given that in our descriptive results we observed that unrelated superimposed words had a different pattern of onset latencies of unaccusatives compared to unergatives.

Following Momma and Ferreira (2019), we would not expect any effect of the verb type in the unrelated superimposed word conditions. We were not able to find any

evidence for the effect of unaccusativity ($\hat{\beta} = 10.82$; $CI = [-14.79; 36.02]$; $P(\beta > 0) = .80$). However, it is important to note that the difference between our degree of beliefs in these two models is quite small. It is quite possible that unaccusatives were overall slower to start utterances or there is no overall difference.

Figure 36

Posterior distribution and the degree of belief for the exGaussian regression coefficients for the model of onset latencies of conditions with semantically unrelated superimposed word in Experiment 1.



5.4.5 Preverbal Production Time

Similar to our analysis of onset latency, our primary interest here is whether the semantic relatedness of the superimposed word influenced speech timing. Consequently, we again plotted semantic relatedness as a line type (dashed: unrelated, solid: related), collapsing across the attractor number manipulation in Figure 37.

Our results for unergative sentences align with Momma and Ferreira's (2019) findings. Semantic relatedness had no discernible effect on preverbal production times. Participants took comparable amounts of time to produce "*spoon is*" when the superimposed word was related (M = 843.76, SE = 10.18) and unrelated (M = 839.73, SE = 10.01).

However, the pattern observed in unaccusative sentences is surprising. Participants spent more time producing the preverbal region with semantically related superimposed words (M = 870.25, SE = 11.65) compared to unrelated superimposed

words (M = 833.13, SE = 10.00). This is unexpected, given that this effect is typically associated with the selection of syntactic lemmas, which has been argued to occur during sentence onset. Our descriptive results, pooled across attractor numbers, suggest that this planning process occurs both preverbally and during utterance onset, as indicated by the differences in onset times shown in Figure 32.

Figure 37

*The average preverbal times according to the experimental conditions (excluding the attractor number) in our Experiment 1. Error bars signal Morey-Cousineau 95% confidence intervals (1.96*Standard Error) corrected for by-subject variance and not by-item variance.*

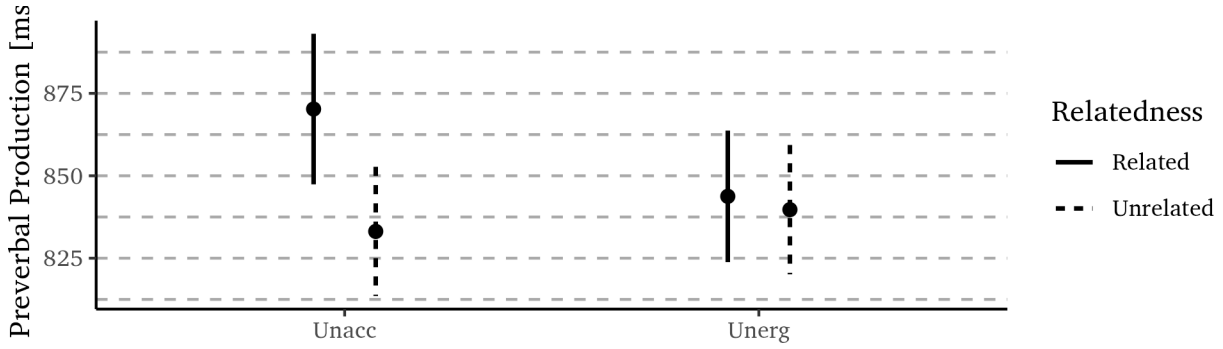


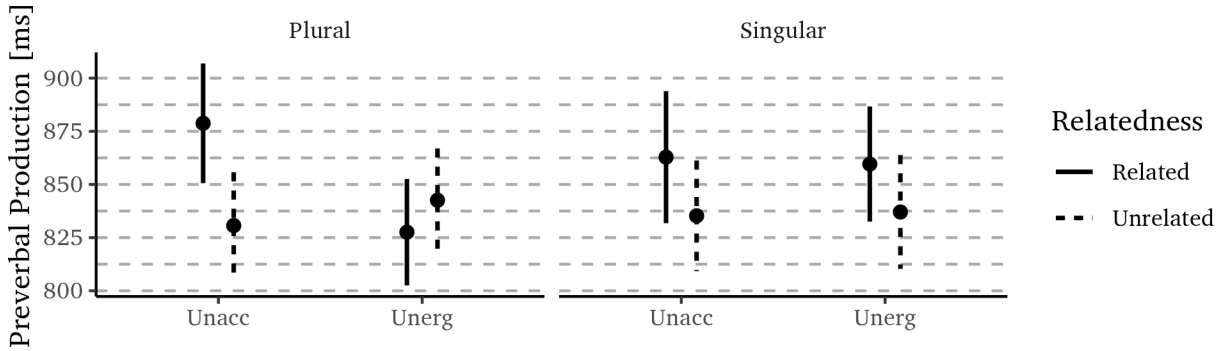
Figure 38 presents the preverbal production times, disaggregated by attractor number conditions. We observe that the effect present in unaccusative sentences was solely driven by conditions with a plural attractor. Specifically, it took significantly more time to produce "*spoons is*" when the superimposed word was related (e.g., *melt* for a *boil* scene) (M = 878.76, SE = 14.36) compared to unrelated superimposed distractors (e.g., *fall* for a *boil* scene) (M = 830.67, SE = 12.79). While a difference in means was also observed with singular attractors ($\Delta_{*\text{Related-Unrelated}*}$M = 27.59 [-12.91, 68.09]), the difference as a function of semantic relatedness is only suggestive, given the overlapping confidence intervals and the zero-containing confidence interval of the differences.

As for unergatives, there was a very minute difference in means and highly overlapping confidence intervals within plural attractors ($\Delta_{*\text{Related-Unrelated}*}$M = -15.00 [-49.94, 19.93]). The pattern with singular attractors was similar to unaccusatives: a suggestive difference in means, but no significant difference ($\Delta_{*\text{Related-Unrelated}*}$M = 22.54

[-15.53, 60.62]).

Figure 38

*The average preverbal times according to the experimental conditions (including the attractor number) in our Experiment 1. Error bars signal Morey-Cousineau 95% confidence intervals (1.96*Standard Error) corrected for by-subject variance and not by-item variance.*



In Figure 39, we present the posterior distribution for the predictors of an Gaussian-exponential mixture regression coefficients of the model analyzing preverbal times in Experiment 1, i.e. how much time it takes to start uttering the sentence. In addition to the posterior distribution, we report the posterior probability of the effect of coefficient β being greater or smaller than zero, which we refer to as the degree of belief or evidence for a positive or negative effect on the presence of an overall increased or decreased preverbal production time, i.e. time to produce the second NP and the auxiliary.

We did not find any evidence for a main effect of unaccusativity ($\hat{\beta} = 4.57$; $CI = [-25.95; 34.75]$; $P(\beta > 0) = .62$). There was a strong evidence for a positive effect of plural attractor on preverbal production time ($\hat{\beta} = 11.78$; $CI = [-4.19; 27.80]$; $P(\beta > 0) = .93$), which is expected given that we also found an effect in pause likelihood. As a final main effect, we also had moderate evidence for the semantic relatedness, meaning that people took more time to utter the second noun and the auxiliary verb when the superimposed word was related to the scene ($\hat{\beta} = 9.45$; $CI = [-4.64; 23.39]$; $P(\beta > 0) = .91$).

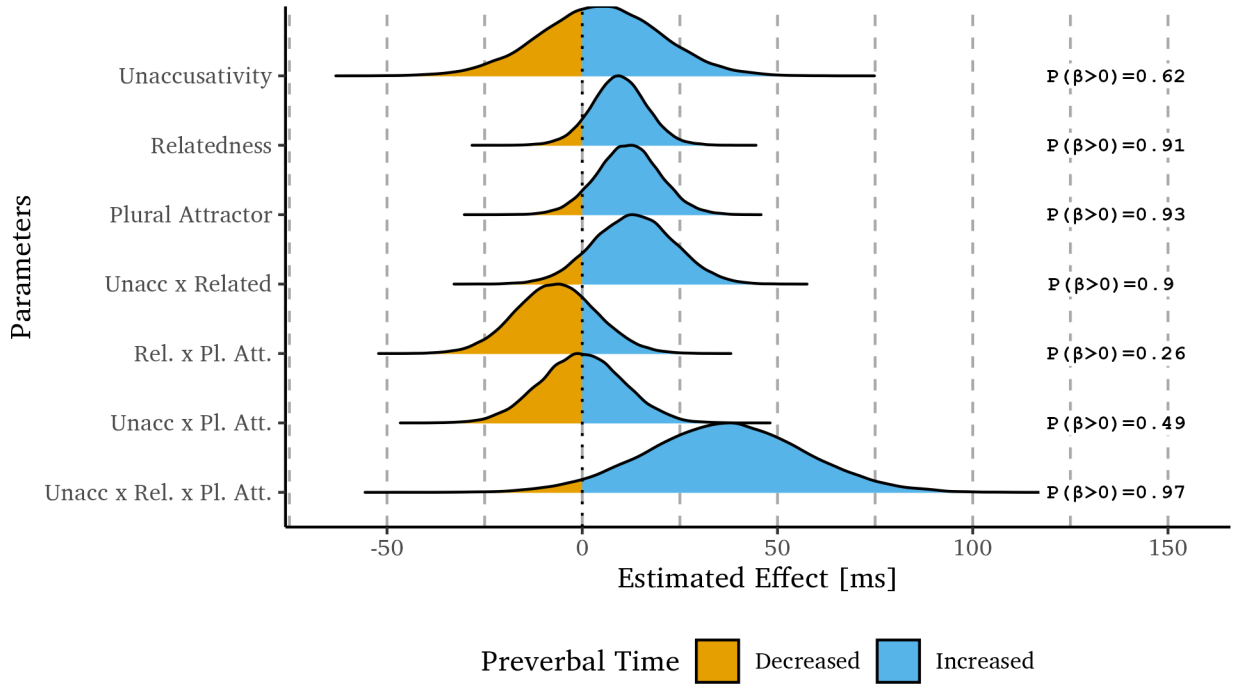
More importantly, there was an evidence for an interaction between the

unaccusativity and the semantic relatedness ($\hat{\beta} = 13.24$; $CI = [-6.94; 33.49]$; $P(\beta > 0) = .90$). This interaction suggests that the effect of the semantic relatedness on the preverbal production time was different for unaccusatives and unergatives. This is suprising given that this interaction is taken as an indication of verb planning in Momma and Ferreira (2019). Differently from them, we found this effect preverbally, suggesting the unaccusative verbs were not planned early.

This interaction was even more amplified as suggested by our three way interaction between unaccusativity × plural attractor × semantic relatedness ($\hat{\beta} = 36.79$; $CI = [-2.69; 76.43]$; $P(\beta > 0) = .97$). This interaction suggests that the difficulty induced by the interaction of unaccusativity and semantic relatedness, i.e. verb planning, was even more difficult when the attractor was plural, suggesting that the interference is not only on the conceptual level, but also on the syntactic level in which participants might be planning the subject-verb agreement for the distractor as well. However, this is a speculative interpretation of our results.

Figure 39

Posterior distribution and the degree of belief for the exGaussian regression coefficients for the model of preverbal time in Experiment 1.



Since the effects of the preverbal timing were easier to interpret, we did not fit any additional models. Our results suggest that in addition to an initial utterance onset difficulty with unaccusatives, the planning of these verbs might spill over to immediately preverbal sites.

5.5 Discussion

Experiment 1 yielded three main findings. First, despite being relatively small, an agreement attraction effect was observed for both unaccusative and unergative verb types. This result suggests that the verb type may not modulate attraction. Second, we found that the likelihood of pauses in correctly produced trials—-used here as a timing index for agreement computation—-was roughly correlated with the magnitude of the attraction effect. This correlation provides converging evidence that pause likelihoods in speech can be a good indicator of the agreement computation. Third, the verb planning pattern deviated from previous findings, particularly those reported in Momma and Ferreira (2019). Specifically, for unaccusative sentences, we observed evidence of unaccusative verb planning both at utterance onset and in the preverbal region. This is a notable departure from the standard view that unaccusative verbs are planned early, along with their subjects, and suggests that unaccusative verb planning can span from sentence onset to preverbal reason. More specifically, even though unaccusatives led to increased onset latencies, this effect was independent of the semantic relatedness manipulation. However, an interaction between unaccusativity and semantic distractor relatedness emerged in the preverbal region.

Together, these findings challenge a strict eager vs. lazy planning dichotomy we hoped to question for agreement computation. Instead, they suggest a more nuanced picture in which the presence of a harder task may change the verb planning, rendering our takeaways from this experiment only suggestive. Further work is needed to refine our understanding of how task difficulty and verb class interact. After discussing our results in this section, we will conduct a simpler experiment.

5.5.1 Attenuated Attraction Errors

While we observed agreement attraction effects in both unaccusative and unergative sentences, the magnitude of these effects was markedly attenuated in our data. Compared to previous studies, our attraction rates were lower: for instance, Kandel and Phillips (2022) report an error rate of approximately 20%, Veenstra et al. (2014) find around 10%, and Nozari and Omaki (2022) report effects in the range of 3–5%. Although our observed effect is broadly comparable to that of Nozari and Omaki (2022), it remains somewhat smaller, raising questions about the underlying cause of this attenuation. In this section, we discuss several experimental factors that may have contributed to the relatively low attraction error rates observed in our study. These considerations are important not only for interpreting our findings, but also for refining the methodological assumptions underlying agreement production research more broadly.

5.5.1.1 Attractors must be possible controllers. One possible reason for the reduced attraction effect in our study concerns the experimental status of the attractor. A key distinction between our materials and those used in prior picture description experiments lies in the possible syntactic function of the attractor noun. In our experiment, attractors were never plausible agreement controllers; that is, nouns like “spoons” did not grammatically function as the subject of the sentence in any trial. This contrasts with the design of previous studies—such as Kandel and Phillips (2022), Veenstra et al. (2014), and Nozari and Omaki (2022)—where the attractors were possible subjects within the experiment and could conceivably serve as agreement controllers. This difference likely reduced the level of competition during agreement computation in our task, leading to a smaller or no observed attraction effect.

This interpretation is further supported by recent findings in agreement attraction during comprehension. Specifically, Bhatia and Dillon (2022) demonstrated that attraction effects emerge predominantly in sentences where the attractor is a viable

syntactic controller given a sentence, i.e. if the agreement controller in a sentence is object, only other objects serve as a attractor. More importantly, the effect of possible subjecthood within the experiment from other sentences is also shown to affect overall attraction effects. Türk (2022) found that the overall magnitude of attraction was attenuated when participants were exposed to trials in which the attractor was not a possible controller—suggesting that within-experiment statistics shape agreement processing. These findings align with our proposal that both within-sentence and within-experiment competition among potential controllers plays a key role in modulating attraction effects. In our study, the attractors were consistently ruled out as possible controllers, which may have globally suppressed attraction. Nevertheless, we acknowledge that additional design-specific factors likely contributed to the reduced error rates observed in our results.

5.5.1.2 Visual cue matters. Another important factor that may have contributed to the attenuated attraction effect in our study concerns the visual cueing of the subject head. Both our experiment and that of Nozari and Omaki (2022) included an explicit cue to direct participants' attention to the subject head. In their study, the subject head was highlighted with a yellow-radiant circle, while in our experiment, we used a red arrow inherited from the design of Momma and Ferreira (2019). Crucially, these visual cues remained on the screen for the duration of the trial, potentially enhancing the salience of the subject head. It is plausible that the continuous visual prominence of the head noun—especially through a direct and unambiguous cue like a red arrow—encouraged participants to maintain focused attention on the correct agreement controller, thereby diminishing the likelihood of attraction errors. Supporting this interpretation, Nozari and Omaki (2022) found that when they instead cued a halo around the attractor (rather than the subject head), attraction errors increased. These findings suggest that visual attention plays a modulatory role in agreement computation, and that emphasizing the head noun may serve to shield

against interference from potential attractors.

5.5.1.3 Non-restrictive modifiers induce less attraction. A further consideration is the communicative context and the role of the attractor within the visual scene. In our experiment, the attractor did not carry a central communicative function and was unlikely to be used by participants as part of their intended message in which they differentiated the intended scene from the otherscene. Because each scene was unique and the attractor was not necessary for disambiguating the subject head, participants' interaction with the attractor was likely constrained to the demands of the experimental task. This stands in contrast to prior studies, where the attractor often played a meaningful role in the scene and served to clarify or restrict the reference of the head noun. In such contexts, the attractor was not only experimentally but also pragmatically relevant, increasing its salience and potential to interfere during agreement computation. In our study, however, modifiers such as "above the spoons" may have been interpreted as non-restrictive adjective, further minimizing the attractor's influence on the production process and contributing to the overall reduction in attraction errors.

This interpretation suggests that participants may have internally represented the modifier as non-essential to the message of the sentence. In other words, modifiers such as "above the spoons" may have been processed in a manner similar to non-restrictive relative clauses—structures or parentheticals that do not limit or define the referent of the head noun, but rather provide supplementary information. Both non-restrictive prepositional modifiers and non-restrictive relative clauses have been argued to be syntactically more opaque, making them less accessible for syntactic operations such as agreement computation (see Lasnik & Uriagereka, 2022). One line of evidence for this increased opaqueness comes from binding theory: while restrictive relative clauses are constrained by binding principles, non-restrictive relatives are not. For example, in the sentence **John$_i$ will fire the person who criticized the bastard$_i$*, the co-reference is ruled out

by Binding Principle C. However, in *John$_i$ will fire Mary, who criticized the bastard$_i$*, the co-reference is permissible due to the non-restrictive status of the relative clause and possibly a different adjunction position.

Moreover, recent findings by Kim and Xiang (2024) further support the syntactic distinction between restrictive and non-restrictive structures and their effect in parsing. They showed that agreement attraction errors are less frequent and attenuated in non-restrictive relative clauses. They provide an explanation for their results via discourse relevance of the attractors. Although our modifiers are not syntactically identical to non-restrictive relatives, it is conceivable that their non-restrictive, parenthetical nature in the discourse led them to function similarly in terms of processing. This, in turn, may have contributed to the reduced likelihood and size of attraction effects observed in our experiment.

5.5.1.4 The response set size changes the planning dynamics. Another important factor that may have influenced the magnitude of attraction effects in our experiment is the level of uncertainty associated with verb retrieval. Unlike other picture description studies that have employed simplified paradigms such as “auxiliary verb + color” (Nozari & Omaki, 2022; Veenstra et al., 2014) or “auxiliary verb + nonce verb” (Kandel & Phillips, 2022), our study required the use of specific lexical verbs. This design choice was motivated by our central research question: to investigate planning differences between unaccusative and unergative verbs. To ensure that participants could not rely on automatic retrieval processes and had a planning procedure that is different depending on the verb, we included multiple lexical verbs, thereby increasing the cognitive load associated with verb access and encouraging more naturalistic sentence planning.

This increased verb variability likely altered the way participants approached scene description compared to earlier attraction studies. For example, it is plausible that participants needed to saccade back to the target scene—for instance, to the image of

the octopus boiling—as a means of retrieving the appropriate verb. Given the lexical specificity and variability of the verbs in our experiment, sentence planning may have unfolded in a more distributed and dynamic manner. One possible sequence of attentional focus might involve an initial fixation on the target scene to retrieve the subject (and potentially initiate planning of the unaccusative verb), followed by a shift to the attractor, and then a return to the target scene to retrieve or confirm the verb. This stands in contrast to other picture-word interference studies of agreement attraction, where participants may have engaged in a more simple pattern without saccading back: first looking at the target scene to retrieve the subject, then shifting to the attractor to plan its mention, and possibly proceeding directly to articulation without needing to saccade back—thanks to the use of repeated verbs or easily accessible features like color.

In our case, the necessity to re-engage with the target visual scene likely encouraged re-fixation on the subject head during verb retrieval, reinforcing its activation at critical moments in agreement computation. This concurrent attention to both the subject and the verb may have contributed to a reduction in attraction errors. Moreover, this shift in the dynamics of sentence planning may also help explain the distinctive pattern of verb planning we observed—particularly for unaccusatives—which departs from the patterns reported in Momma and Ferreira (2019) and will be further discussed in a subsequent section.

5.5.1.5 Why bother with the quest for more attraction? A natural question that arises from these findings is whether stronger attraction effects are necessary to draw meaningful conclusions about agreement computation and its timing. One possibility is to set aside the attenuated nature of the attraction effects and argue that even small effects can yield informative insights into the architecture of sentence production. In our study, we complemented attraction errors with an recent measure—pause likelihood in correctly produced trials—which we interpret as a proxy

for the timing of agreement computation. The fact that this measure correlated with attraction effects across conditions suggests that planning processes related to agreement are still detectable, even when overt errors are rare. Thus, while stronger attraction rates might provide a more robust empirical signal, our findings indicate that subtle fluctuations in agreement planning can be meaningfully captured through alternative indices, allowing us to continue testing theoretical claims about the timing and locus of syntactic encoding in production.

However, there are two important reasons to be cautious about dismissing the attenuation of attraction effects. First, while pause likelihood shows promise as an index of agreement computation timing, it remains a relatively new and unvalidated measure. Its interpretive strength is currently limited by a lack of systematic testing across varied experimental contexts and structures. Even in studies where this measure has been employed—such as Kandel and Phillips (2022)—it was interpreted in conjunction with robust attraction effects, serving to enrich rather than substitute the primary error-based findings. Without a strong baseline of attraction errors, it is difficult to fully disentangle whether pauses reflect syntactic computation, lexical access, or other forms of processing difficulty. As such, relying exclusively on pause likelihood to draw conclusions about the dynamics of agreement planning may overstate the diagnostic precision of this emerging metric.

The second reason to remain cautious about interpreting attenuated attraction effects through pause-based measures alone concerns the theoretical connection between timing and error probability. A well-articulated framework that explicitly links these two dimensions is the drift diffusion model (DDM) of decision making (Ratcliff, 1978; Ratcliff et al., 2016; Ratcliff & Rouder, 1998). Originally developed to explain binary decision tasks, this model has recently been adapted to account for decision-making processes in agreement computation (Hammerly et al., 2019 for two forced-choice completion; Türk, 2022 for speeded acceptability judgments). As illustrated in

Figure 40, the DDM posits that noisy information accumulates over time according to a Gaussian distribution, with the mean drift rate linearly tied to the strength of the stimulus or evidence. A decision is triggered when the accumulating evidence reaches one of two predefined thresholds, each corresponding to a possible response.

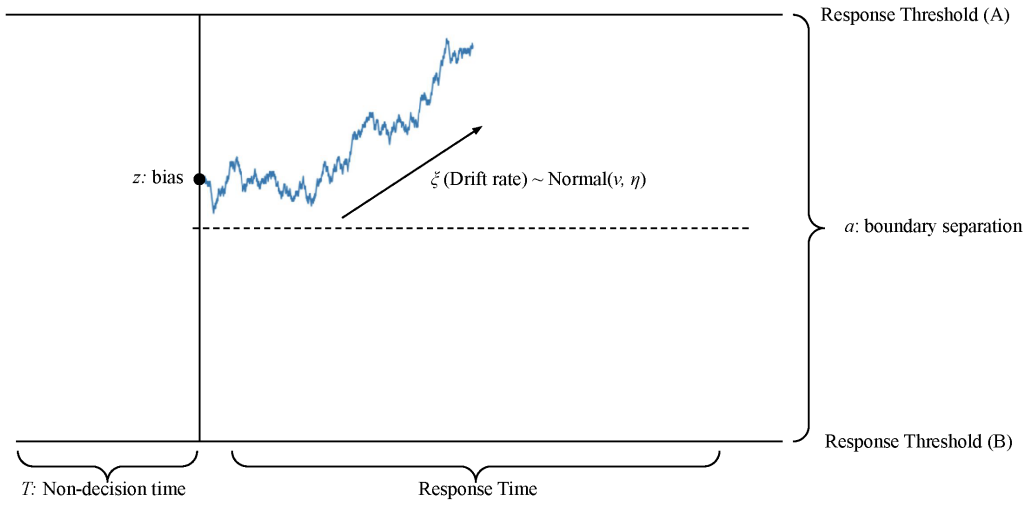
In this framework, both the speed of the decision (i.e., reaction time or pause duration) and the probability of an error are jointly determined by the same underlying process: the rate and variability of information accumulation. As such, reduced error rates are expected to be accompanied by faster decisions—or at least fewer pauses—when processing is efficient and unambiguous. Conversely, if errors are near floor, the timing measure may lose sensitivity, making it harder to detect reliable patterns. This tight coupling between errors and timing in the DDM underscores the value of having both behavioral outputs present to robustly infer the dynamics of syntactic planning. Thus, while pause likelihood offers a promising complementary measure, interpreting it in isolation—especially in contexts of low error—requires caution and further validation.

Among the five core parameters of the drift diffusion model shown in Figure 40, the drift rate (ξ)—which reflects the average rate of evidence accumulation—is of particular interest in the context of agreement computation. For the purposes of this discussion and following Hammerly et al. (2019), we assume that selecting the appropriate auxiliary verb in English (e.g., choosing between *is* and *are*) can be modeled as a two-alternative forced choice process. Each trial presents participants with a certain amount of evidence in favor of the correct choice, and this evidence strength is captured by the drift rate parameter. Crucially, we propose that the drift rate varies as a function of both the experimental condition and the specific processing path participants adopt during the trial.

Given the factors outlined above—including the non-competitor status of the attractor, the presence of a strong head cue, and the possibility of visual re-fixation on

Figure 40

Simplified illustration of the drift diffusion model. In every trial, after a period of time (T), an evidence accumulation process is initiated from the specified position (z) relative to the whole boundary separation (a). Evidence is accumulated stochasticcally according to the drift rate (ξ) that follows a normal distribution with the mean v and the standard deviation η. When enough information is accumulated to cross one of the thresholds, a decision is made.



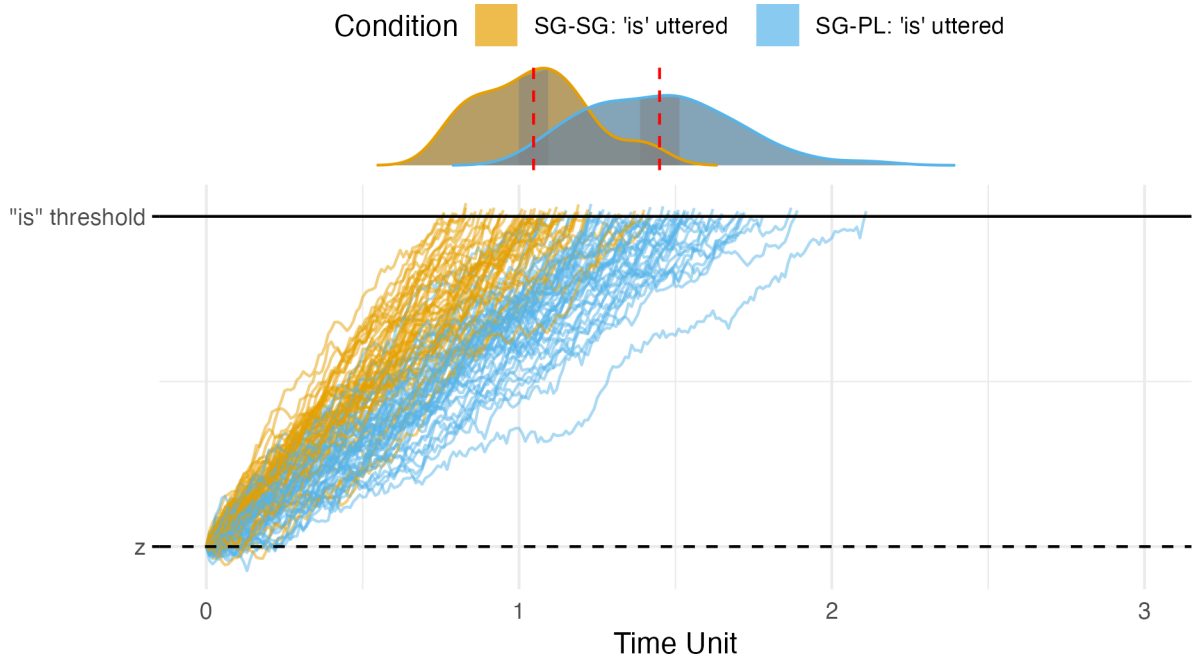
the subject—we hypothesize that participants accumulated similar amounts of evidence in both the singular and plural attractor conditions. That is, the strength of the syntactic signal favoring the correct auxiliary (*is*) remained consistent across conditions, resulting in comparable drift rates. Under the DDM framework, this convergence in drift rate would naturally lead to both reduced error rates and minimal differences in timing measures, offering a principled explanation for the attenuated attraction effects observed in our study.

To illustrate how differences in drift rate can affect decision dynamics under the drift diffusion model, we conducted a simple simulation, presented in Figure 41. This figure mirrors the structure of Figure 40, but instead of displaying both possible decisions (e.g., *is* vs. *are*), we focused exclusively on simulated *is* decisions under two conditions: SG-SG (singular head with singular attractor) and SG-PL (singular head with plural attractor). All model parameters were held constant across conditions except for the drift rate, which was sampled from normal distributions with means of `0.93` and

`0.70` for SG-SG and SG-PL conditions, respectively, both with a standard deviation of `0.15`. These values were chosen arbitrarily for illustrative purposes, with the goal of demonstrating how even modest differences in drift rate can yield measurable changes in decision timing. For now, let's assume that 0.93 signifies that the strength of the visual scene to trigger the response "is" in SG-SG, and 0.70 signifies the strength of the visual scene to trigger the response "is" in SG-PL, which is reduced.

As shown in the figure, the divergent drift rates led to a clear separation in the mean decision times between the two conditions, indicated by the red dashed lines. The SG-SG condition, with a higher positive drift rate, resulted in faster decisions favoring the correct auxiliary (*is*), whereas the negative drift in the SG-PL condition slowed down accumulation. Even though the distribution overlap, their mean is clearly different and the confidence intervals shown by the darkened ribbons do not overlap.

Figure 41
Drift Diffusion Model simulation with considerably different drift rates.



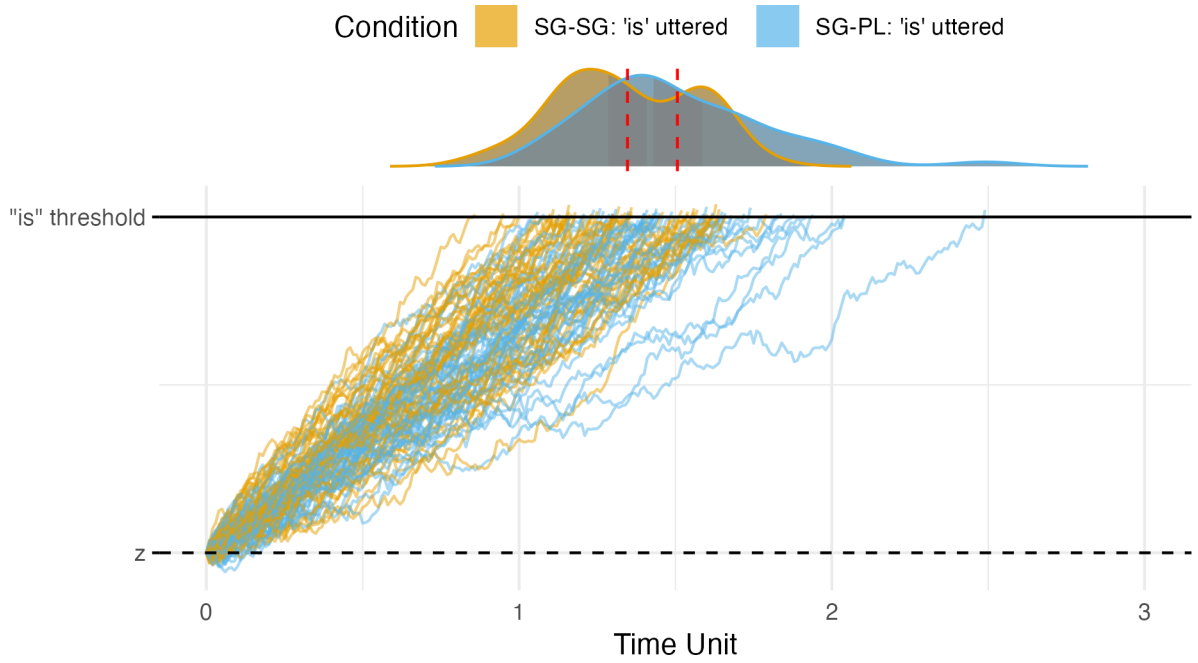
If we instead assume that the drift rates for the two conditions are closer in magnitude, the separation in decision times becomes substantially less pronounced.

This is demonstrated in a second simulation, shown in Figure 42. In this simulation, we assigned the drift rate for the SG-SG condition as N(0.75, 0.15) and for the SG-PL condition as N(0.65, 0.15), values intended to more closely reflect the likely distribution of evidence strength in our experimental data. These narrower differences in drift rate capture the overall similarity in syntactic strength between the two conditions—an outcome we argue results from several of the experimental design choices discussed earlier, including the non-competitor role of the attractor, consistent visual cueing of the subject, and the potential re-fixation on the subject head during verb retrieval.

As illustrated in the figure, the resulting decision times for producing the correct auxiliary *is* are largely overlapping across the two conditions. The mean difference in response latency is minimal, and the difference between the confidence intervals is quite small, suggesting that the reduced drift rate difference is insufficient to generate a reliable difference in decision timing and will probably affected by the number of participants and trials. This simulation supports the interpretation that when the overall evidence between conditions is similar across conditions, assuming it is also reflected in error rates, detecting differences in timing measures will be more challenging to detect.

Taken together, if one assumes a model akin to the drift diffusion framework for decision-making between auxiliary forms, an important prediction follows: as attraction errors decrease, the corresponding differences in timing measures—such as pause likelihood or decision latency—are also expected to diminish. In other words, when the drift rates between different conditions (e.g., SG-SG vs. SG-PL) converge for any reason, both the time it takes to decide on the auxiliary “is” will become less differentiated across conditions. Consequently, the detection of agreement planning time-differences becomes more difficult when the attraction effect is reduced or experimentally minimized.

Figure 42
Drift Diffusion Model simulation with similar drift rates.



5.5.2 Correlated Pause Likelihood and Attraction Errors

One recently proposed measure of agreement computation is pause likelihood, defined as the presence of non-zero gaps—in our study, those longer than 50 milliseconds—between the second noun (e.g., the attractor) and the auxiliary verb. This measure captures subtle disruptions or delays in speech planning that may not manifest as overt errors. In support of its utility, Kandel and Phillips (2022) demonstrated that pause likelihood reliably correlates with attractor number manipulation in correctly produced trials. Their findings suggest that pause likelihood is sensitive to the interference introduced by plural attractors in computing the agreement on the verb, even when speakers successfully produce the correct auxiliary verb form. Thus, pause likelihood offers a promising, gradient index of agreement computation and may serve as a valuable complement to categorical error measures in investigating the temporal dynamics of sentence production.

In our experiment, the overall pause likelihood was notably higher than what

was reported in Kandel and Phillips (2022), a difference we attribute to the substantially greater complexity and variability of our materials. While their study employed a highly constrained stimulus set—using only three recurring entities tied to distinct colors and a single verb (*to mim*) across all trials—our design required participants to retrieve 18 unique characters, 36 distinct objects, and 36 different verbs. This increase in lexical and conceptual variability likely imposed a much heavier cognitive load, especially at the stages of lexical access and message formulation. Participants in our study had to retrieve and integrate more novel information on each trial, which plausibly led to increased planning difficulty and longer pauses in speech, even on correctly produced sentences. In contrast, the regular and predictable structure of Kandel and Phillips (2022)'s materials likely facilitated smoother planning, contributing to the lower overall pause likelihood observed in their results.

Nevertheless, we maintain that pause likelihood remains a valuable indicator of agreement computation. Despite the overall low rate of attraction errors and the increased memory demands imposed by our more complex stimulus set, we observed a meaningful alignment between the two measures. Specifically, the conditions in which attractor number influenced pause likelihood were the same conditions in which we observed attraction errors, albeit attenuated. This convergence suggests that pause likelihood is sensitive to the interference introduced by plural attractors, even in the absence of overt agreement errors. As such, it provides an important window into the real-time processes of sentence planning and agreement computation, particularly in contexts where errors alone may be too infrequent to yield robust insights.

Our results further support the utility of pause likelihood as a measure of agreement computation through its alignment with condition-specific attraction effects. We found that attraction effects were most pronounced in unaccusative sentences when the superimposed distractor was semantically related, and in unergative sentences when the distractor was semantically unrelated. Strikingly, this same pattern was observed in

the pause likelihood data: a clear divergence between plural and singular attractor conditions emerged only in the semantically related–unaccusative and semantically unrelated–unergative conditions. These parallel patterns suggest that, even in the face of overall attenuation, pause likelihood reliably tracks the syntactic processes involved in agreement computation.

5.5.3 Unaccusatives were harder to initiate the sentence

One particularly notable finding in our results concerned the timing of verb planning. We observed that participants generally exhibited longer utterance onset latencies when producing sentences with unaccusative verbs compared to unergative verbs. This aligns with prior claims that unaccusative verbs may be planned earlier, in conjunction with the subject, due to their syntactic configuration. However, unlike previous studies that have used semantic interference as a diagnostic for early verb planning, we did not find that semantically related superimposed verbs specifically delayed utterance onset. This challenges the assumption that onset latency alone reliably captures early lexical access to the verb, at least for our experiment.

Instead, we observed an interaction between verb type and semantic relatedness in the preverbal region. While unaccusative verbs were associated with slower utterance onset overall, semantic interference effects only emerged immediately prior to verb articulation. This pattern suggests that although participants may have initiated sentence planning with some commitment to the unaccusative verb, the planning process remained insusceptible to interference until late in the production timeline. In other words, even when unaccusative verbs are planned early, their lexical access and selection may still occur closer to the point of articulation, particularly when semantically competing information is present.

One possible interpretation of the increased onset latency for unaccusative sentences without semantic interference is that the actions depicted in those scenes were simply harder to discern at first glance compared to unergative scenes. To explore

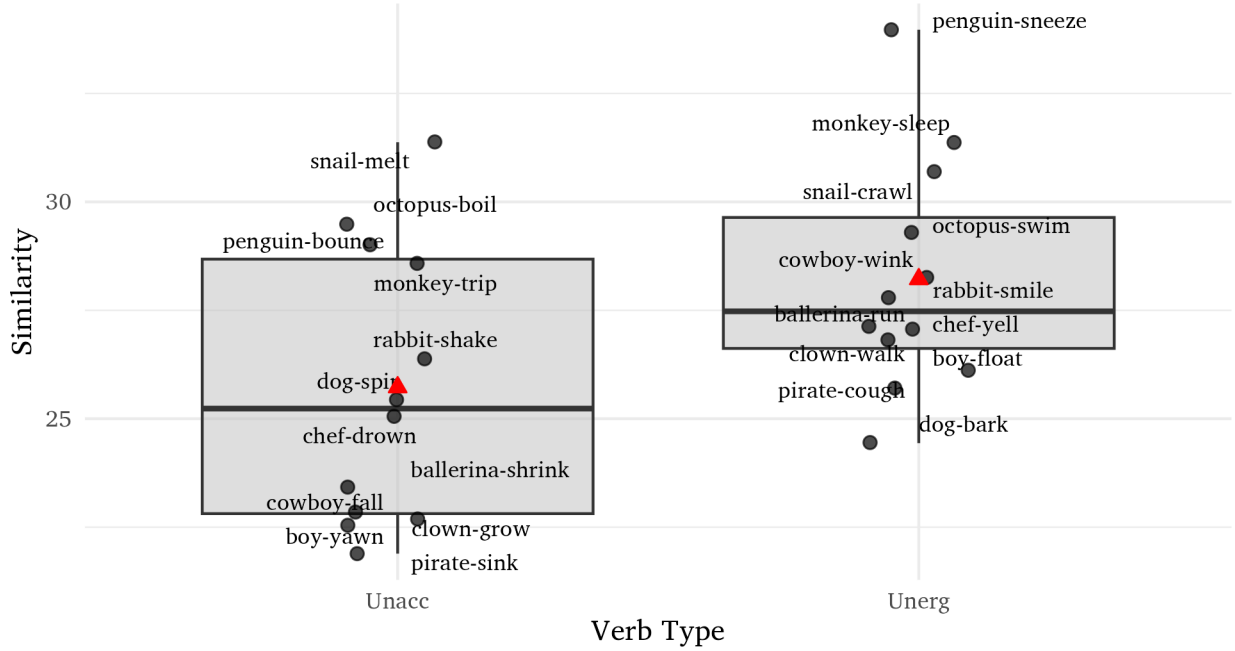
this possibility, we employed the CLIP Vision-and-Language model (Radford et al., 2021), which includes a visual encoder and a text encoder trained jointly on over 400 million noisy image–text pairs from the internet. Using this model, we computed similarity scores between the images in our experiment and their corresponding target sentences (excluding the PP modifier). While this approach does not directly quantify perceptual difficulty, it offers a useful proxy for how easily a sentence might be grounded in the visual input.

As shown in Figure 43, the similarity scores revealed that unaccusative sentences were, if anything, more similar to their associated images than unergative sentences. This suggests that the delay in onset latency for unaccusatives was not due to difficulty in interpreting or identifying the scene itself. If unaccusative actions were harder to visually access, we would expect lower similarity scores in the CLIP analysis. The fact that the opposite pattern emerged provides preliminary evidence against a purely perceptual explanation for the timing differences observed in our data. A more direct test of this hypothesis would involve a separate production task in which participants describe the pictures without prior habituation or task constraints, which could better isolate scene comprehensibility from syntactic planning demands.

Another way to interpret our findings is that verb planning in our task began early—possibly as early as utterance onset, consistent with the findings of Momma and Ferreira (2019)—but that this planning process was prolonged due to the increased number of unique verbs and objects in our experiment. This extended planning window may have diluted or masked the typical “signature” of early verb planning, such as semantic interference effects on onset latencies. This interpretation is further supported by findings from the word production literature, particularly the work of Roelofs (2001). In his review, Roelofs demonstrated a strong relationship between semantic interference effects and response set size: interference was reliably reduced when the number of possible responses was small and item repetition was high. By contrast, both

Figure 43

Boxplot of CLIP similarity scores grouped as a verb type of the scenes in our Experiment 1. Red triangle represents the mean of each verb type. The round points represents individual scene similarity scores. The lower and upper hinges corresponds to the first and third quartiles. The middle line within the boxplot represents the median. The whiskers extend from the hinge to show the range of the data.



our study and that of Momma and Ferreira (2019) featured larger item sets and fewer repetitions than the paradigms discussed in Roelofs' work.

It is possible that the moderate item set in Momma and Ferreira (2019) allowed participants to build conceptual associations that supported more efficient verb retrieval, even if explicit picture–word mappings were not retained–as hypothesized by Roelofs (2001). In our study, however, the number of unique items was even greater, potentially degrading even these more abstract associations. As a result, verb retrieval may have become more effortful and more prolonged, leading to increased onset latencies for unaccusatives and the absence of a clear semantic interference effect at utterance onset. This may also explain the initial facilitatory effects of semantically related superimposed words: participants may have briefly used the related word as a retrieval cue for the target verb concept, before needing to suppress its activation as

they prepared to articulate the correct verb. This extended retrieval process could account for the lack of an early semantic interference effect and the emergence of an interaction only in the preverbal region.

Ultimately, while our descriptive results suggested an interaction between verb type and semantic relatedness, this interaction was supported by weak statistical significance in our model. Nevertheless, the pattern of results—particularly the overall delay in unaccusative utterance onset—suggests that verb planning was initiated early but unfolded more slowly under the increased lexical demands of our paradigm. This interpretation preserves consistency with the findings of Momma and Ferreira (2019) while offering a more nuanced account of how lexical set size and retrieval difficulty can shape the timing and detectability of verb planning effects.

However, this "prolonged" planning interpretation also highlights an important limitation: our data may not be fully suited to answer the precise question we set out to investigate regarding the timing of agreement computation and its dependency on verb type. If much of the utterance onset time was spent recognizing the verb and forming a conceptual representation—rather than generating a syntactic frame with number diacritics—then our measure of early planning may have been confounded by lexical and conceptual retrieval demands. In such a scenario, it becomes unclear whether the planning processes captured at utterance onset include any computation related to agreement. While it remains an open and intriguing question whether early conceptual planning or verb recognition involves number computation, our current data cannot conclusively address it. One important takeaway from these findings is that future experiments should aim to reduce the lexical and conceptual difficulty associated with verb retrieval. By simplifying the verb set and optimizing the task to support earlier and more syntactic-level planning, we may be better positioned to isolate and examine the temporal dynamics of agreement computation.

5.5.4 Facilitation of Semantically Related Distractors

Previous work has shown that semantically related superimposed distractors can slow down the production of the verb, typically by interfering with lexical access. In our study, we attempted to replicate the semantic interference effect reported by Momma and Ferreira (2019), who found that semantically related distractors delayed the onset of sentences with unaccusative verbs. Their results were interpreted as evidence of early verb planning for unaccusatives, with semantic interference disrupting this early retrieval process.

In contrast, our results suggest that semantically related superimposed distractors actually facilitated sentence onset—an effect observed only in unaccusative sentences. We interpret this facilitatory effect as a byproduct of the increased number of verbs used in our experiment. The expanded verb set likely made lexical retrieval more demanding, encouraging participants to use any available cue—such as the semantically related distractor—to assist with verb access early in production.

Interestingly, this facilitatory onset effect was absent in conditions with plural attractors. Within these mismatch conditions, we observed no significant differences in onset latency between related and unrelated distractor trials. However, the semantic interference pattern reported by Momma and Ferreira (2019) did appear in preverbal time, but only for unaccusative sentences with plural attractors. We interpret this as evidence that number mismatch (singular head – plural attractor) may have delayed verb planning, shifting it closer to articulation, where semantic interference from the distractor had a more typical disruptive effect.

When there was no number mismatch—i.e., in singular attractor conditions—we replicated the theoretically significant pattern reported by Momma and Ferreira (2019): the semantic relatedness manipulation had a effect on onset latencies, but only in unaccusatives. In our case, this effect surfaced as a facilitation due to the experimental task demands. We suggest that in conditions where verb planning occurred early,

semantically related distractors helped retrieval. However, in conditions where number mismatch delayed verb planning, this benefit was reduced or eliminated. Instead, participants had more time to visually engage with the scene during verb retrieval, and we believe this increased visual exposure may have functioned similarly to the support provided by a related distractor—thereby reducing the need for such a cue and explaining the absence of further facilitation effects.

5.5.5 Interaction of Semantic Relatedness and Verb Type on Attraction

Another noteworthy finding from our study was the interaction between semantic relatedness and attractor number, which emerged consistently across several dependent measures: the rate of attraction errors, pause likelihood, and preverbal production time. This convergence across distinct metrics suggests a robust underlying effect. Specifically, the influence of a plural attractor—typically expected to increase agreement errors and planning difficulty—was most evident positively in semantically related conditions for unaccusative verbs and negatively in semantically unrelated conditions for unergative verbs. This asymmetry indicates that the role of semantic interference in agreement computation is modulated not only by the lexical similarity between nouns but also by the structural and thematic properties of the sentence, such as verb type.

A closer examination of this interaction across conditions reveals additional nuances, particularly in the pause likelihood data. In unaccusative sentences, pause likelihood did not vary systematically with semantic relatedness alone. However, in the semantically related condition, the presence of a plural attractor led to a noticeable increase in pause likelihood. This suggests that the observed interaction in unaccusatives was primarily driven by the increased planning difficulty introduced by plural attractors when combined with semantically related distractors. Notably, the other unaccusative conditions—including those with singular attractors or unrelated distractors—showed relatively stable pause likelihoods, indicating that the cognitive

load associated with agreement computation was not uniformly elevated across all unaccusative contexts.

In unergative sentences, the pause likelihood was lowest in the condition with a singular attractor and an unrelated distractor—what we take to be the baseline configuration. Interestingly, any deviation from this baseline—whether by introducing a plural attractor, a semantically related distractor, or both—resulted in a moderate increase in pause likelihood. Crucially, these increases appeared relatively uniform regardless of whether one or two features changed, suggesting a kind of thresholded response to added complexity. However, when all three properties differed from the baseline, as in the Unaccusative–Plural–Related condition, the increase in pause likelihood was notably larger. This pattern suggests that the planning difficulty did not simply accumulate linearly across features; rather, when the configuration crossed a certain threshold—combining verb type, attractor number, and semantic relatedness—the cognitive load increased disproportionately, pointing to a nonlinear interaction among these factors.

This three-way interaction was also reflected—albeit in a slightly different pattern—in the preverbal time measure, defined as the duration between the onset of the second noun and the onset of the auxiliary verb. In this measure, conditions with unrelated semantic distractors were relatively uniform across verb types and attractor numbers, showing little variation in preverbal timing. However, a different picture emerged for conditions with semantically related distractors. Specifically, within the related distractor conditions, preverbal time varied as a function of verb type, but only when the attractor was plural. In these cases, unaccusative scenes elicited longer preverbal times than unergative ones, suggesting that the combination of a related distractor and a plural attractor affected the ease of producing the auxiliary more strongly when the underlying structure was unaccusative. This pattern complements the pause likelihood findings, indicating that different measures capture overlapping but

non-identical aspects of the interaction between verb type, attractor number, and semantic relatedness.

At present, we do not have a comprehensive theoretical account for the observed interactions beyond a general cognitive load explanation. We propose that unaccusative structures, plural attractors, and semantically related superimposed words each independently increased the processing difficulty of individual trials. Importantly, this increase in difficulty manifested not as a simple additive effect on our dependent measures, but rather as a non-additive interaction. That is, the combination of these factors had a disproportionately large impact on pause likelihood and preverbal time, suggesting that their joint presence imposed greater demands than the sum of their individual effects would predict.

This interpretation, however, raises concerns for the original theoretical aim of our study. Our goal was to establish early verb planning for unaccusatives and use this as a foundation to investigate whether agreement computation is similarly timed early in production. However, it is possible that by combining two complex manipulations into a single task—one targeting agreement attraction and another targeting verb planning—we inadvertently made the experiment too demanding. This increased task complexity may have prevented participants from engaging in the kind of anticipatory, structured planning we sought to observe. As a result, we did not find the expected signature of early verb planning, namely, the interaction between unaccusative verbs and semantic interference in onset latency.

5.6 Conclusion

In this experiment, we aimed to investigate the timing of agreement computation using an extended picture–word interference paradigm. Specifically, we sought to leverage the established planning differences between unaccusative and unergative verbs to ask whether agreement is planned alongside the verb. To this end, we adapted the experimental materials from Momma and Ferreira (2019) by introducing a

manipulation of the number of the second noun phrase, creating conditions that could elicit agreement attraction.

Our findings revealed that attraction errors occurred in both unaccusative and unergative structures, but the overall effect was attenuated compared to previous studies. This reduction in effect size limited our ability to confidently interpret the temporal dynamics of agreement planning. Furthermore, the increased number of lexical items and overall task complexity appeared to obscure the expected verb planning differences, especially at utterance onset.

Taken together, these results suggest that while our theoretical framework remains viable, the current design introduced too much cognitive load for participants to engage in the type of structured, anticipatory planning needed to isolate early agreement computation. Future research will benefit from simplifying the experimental materials and reducing lexical variability in order to elicit more robust attraction effects and clearer verb planning signatures.

Spoilers for Experiment 2

- This experiment tested the same question–whether the timing of agreement planning depends on verb type–using a simplified design to elicit clearer attraction effects.
- Attraction errors were successfully elicited after we made attractors plausible subjects and part of the core communicative message—supporting the role of subjecthood and structural relevance.
- No differences in attraction errors or pause likelihood were found between unaccusative and unergative verbs, suggesting agreement attraction is due a planning procedure that is later during articulation, and during the same time as verb retrieval.
- An interaction between verb type and number mismatch emerged in onset latency (only for unaccusatives), suggesting early encoding of number.
- Results support a two-stage model of agreement: early morpho-syntactic diacritic assignment, followed by late auxiliary retrieval, which happens to be the locus of agreement attraction.
- The attractor's subject-like status supports cue-based retrieval models, though representational accounts with added assumptions remain viable.
- Exploratory entropy-based codability measures suggest response uncertainty may influence verb planning, but further work is needed.

6 Experiment 2: Simple Picture Description Paradigm

In this simple picture description experiment, we aim to answer the same theoretical question as in Experiment 1: whether the timing of agreement planning depends on the timing of the planning of its host verb. Although Experiment 1 provided valuable insights into the interaction between verb planning and agreement, we were unable to fully answer this question due to attenuated attraction effects and the increased demand on verb planning.

To obtain a clearer effect of agreement attraction, we modified several aspects of the experimental design. Specifically, to enhance the likelihood of agreement attraction effects, we altered the subject-modifier noun combinations. For example, instead of using a subject like “the clown by the apples,” we used “the clown by the pirates.” This change was made because both “pirates” and “clowns” can serve as the controller head, whereas “apples” cannot. This manipulation is expected to increase the possibility of attraction errors.

To reduce the potential confounding effects related to the increased number of entities and verbs influencing verb planning, we limited the experiment to six entities and twenty-four verbs (12 unergative and 12 unaccusative). Unlike Experiment 1, we did not use control trials, but we did manipulate the number of subjects in each trial to explore any potential effects of subject number on agreement attraction. Furthermore, we did not incorporate the semantic interference task used in Experiment 1, as we sought to focus more directly on verb planning without additional semantic distractions.

All anonymized data, excluding participants’ voice recordings, and the R code used for data analysis are available at https://go.umd.edu/turk888. Readers are encouraged to explore the full experiment at https://go.umd.edu/turk888-exp2.

6.1 Methods

6.1.1 Participants

We recruited 59 English speakers from the undergraduate students of University of Maryland in exchange for course credit. Participants had a mean age of 19.8135593, ranging between 18 and 41. They were recruited through the SONA platform. An additional 14 participants completed the study but were excluded from the analysis due to poor sound quality in their recordings, technical difficulties preventing completion of the experiment, or producing the correct sentence structure in less than 30% of their responses. Every participant was informed about the experimental details prior to their consent. Informed consent in compliance with the Institutional Review Board of the

University of Maryland, College Park was obtained for all participants. Each session took roughly 30-40 minutes.

6.1.2 Materials

In Experiment 2, we made significant changes to the materials used in comparison to Experiment 1. We limited the entities to six total event participants, ensuring that there were no additional entities or objects in the experiment. A total of 24 verbs were used, with half being unaccusative and the other half unergative.

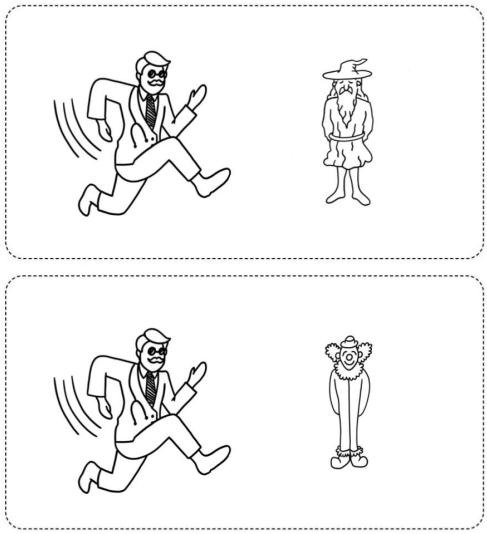
Rather than pairing each entity with two verbs, one for unaccusatives and one for unergatives, we created a unique picture set for each entity-verb pair. This resulted in 144 entity-event pairs in total. These pairs were then divided into two sets, each containing 72 pairs, by reducing the number of verbs for each event participant by half.

We combined each entity performing an action with two other entities standing by. These two standing entities were randomly selected from the list of the other five entities. Each action was combined with itself vertically and with two other entities horizontally. The relational positions of the standing and event participant entities were randomized in the experiment code.

Within the experiment, similar to Experiment 1, we manipulated (i) the number of the attractor (standing entities) and (ii) the verb type (unaccusative vs. unergative). Unlike Experiment 1, we also manipulated the number of the head noun (event participant). We also diverged from Experiment 1 by not including a semantic interference manipulation. In this experiment, participants only saw the pictures and did not see superimposed words. We present an example picture from Experiment 2 in Figure 44.

We also did not include any control trials with the plural subject in this experiment, as we included the plural head conditions. Unlike Experiment 1, we did not use a repeated measures design in this experiment. Instead, participants saw each picture only in one condition. We employed a full Latin square design for our 8

Figure 44
An example experimental scene in our experiment 2. The doctor is a participant of an unergative event swimming and the pirates and the the wizards were randomly selected as a standing entities, attractors.



conditions, and each participant saw a total of 72 scenes, with 9 scenes for each condition.

Target sentences for an item are shown in Table 10.

Table 10
Experimental Conditions

Verb Type	Head	Attractor	Target Sentence
Unaccusative	Singular	Singular	The doctor by the pirate is boiling.
Unaccusative	Singular	Plural	The doctor by the pirates is boiling.
Unaccusative	Plural	Singular	The doctors by the pirate are boiling.
Unaccusative	Plural	Plural	The doctors by the pirates are boiling.
Unergative	Singular	Singular	The doctor by the pirate is running.
Unergative	Singular	Plural	The doctor by the pirates is running.
Unergative	Plural	Singular	The doctors by the pirate are running.
Unergative	Plural	Plural	The doctors by the pirates are running.

6.1.3 Procedure

All experiments reported in this paper were conducted using PCIbex, a platform that facilitated online data collection (Drummond, 2013; Zehr & Schwarz, 2018). Each

experiment incorporated a voice recording function, allowing participants to provide spoken responses. These recordings were transmitted to the server in real-time during the experimental session. Critically, only recordings from participants who successfully completed all trials were included in the final data analysis; recordings from participants who did not finalize the experiment were excluded.

We carried out a different familiarization section for Experiment 2. The familiarization was divided into two blocks: one for entities and one for verbs. Both blocks consisted of two tasks: repeating the name and recalling the name. In the first part of each block of familiarization, participants saw the entity or verb along with its name. After 2500ms, the text disappeared, and participants were asked to name the entity or verb. Once they named every entity or verb, the second part of the block began, in which they were asked to recall the names of the entities or verbs they saw in the first part. Their answers were gathered by asking them to type the names under 7000ms. They were not provided with any help during their production; however, they were given feedback after each recall. If they made a mistake, they were shown the correct answer.

The familiarization section was followed by the practice trials. Participants were shown 5 practice trials that strictly mimicked the experimental trials. The only difference was that after every practice trial, their voice recording was automatically transcribed and checked to see whether the transcription included the target verb or the target head noun. If the transcription did not include the target verb or the target head noun, the correct answer was displayed in red font, and participants were asked to be more careful.

Each trial began with a 500ms fixation cross. The fixation cross was followed by a 2-by-2 grid of entities. The standing entities were randomly selected from the list of all possible entities, excluding the event participant. The side of the grid on which the standing entities appeared was also randomized.

After 1500ms, a square appeared around the head and the modifier attractor. At the same time as the square, participants heard a click noise and their response began to be recorded. Participants were asked to describe the scene as soon as they saw the disambiguating square. They were allotted 4000ms to describe the scene.

We utilized a more holistic approach to signal participants which scene they should describe. Instead of using a red arrow to highlight the specific event, we employed a thick frame around the entire horizontal grouping of the event participant and the standing entities. Participants were asked to infer the event participant from the scene themselves.

The square stayed on the screen for 4000ms. After this, the picture disappeared, and the recording stopped. However, the recording did not actually stop immediately; we continued recording for an additional 500ms, but this was not used in the analysis. This extra time ensured that no participants' responses were cut off prematurely. After the picture disappeared, participants saw a blank screen for 1000ms before the next trial began.

6.2 Pre-treatment

Our data collection yileded a total of 3945 recordings. These recordings were automatically transcribed and words were forced-aligned with their timestamps. The details for the forced alignment with MFA and transcription with AssemblyAI were the same as Experiment 1.

We did not hand-code the data for errors in this experiment. Instead, we used certain heuristics to filter through the transcriptions. We recorded the number of discordances for each heuristic and gathered them as disfluencies.

Table 11 presents the proportion of disfluencies observed within each heuristic. It is important to note that the disfluencies are not mutually exclusive, meaning some trials included multiple errors. To detect this, instead of directly discarding trials, we encoded the presence of a disfluency as a binary variable for each trial.

Table 11
Counts of disfluencies by condition. Conditions are shortened into three character abbreviations. P = plural, S = singular, A = unaccusative, E = unergative. The first P/S indicates the head number, the second P/S indicates the attractor number.

name	PPA	PPE	PSA	PSE	SPA	SPE	SSA	SSE	rowCount
Wrong_structure	101 (13%)	89 (11%)	89 (11%)	104 (13%)	95 (12%)	102 (13%)	105 (13%)	96 (12%)	781 (19.8%)
No_aux	24 (13%)	21 (12%)	28 (15%)	23 (13%)	27 (15%)	19 (10%)	14 (8%)	25 (14%)	181 (4.59%)
No_the	2 (5%)	4 (11%)	8 (21%)	3 (8%)	2 (5%)	7 (18%)	9 (24%)	3 (8%)	38 (0.96%)
Wrong_num	43 (18%)	32 (13%)	50 (20%)	39 (16%)	32 (13%)	43 (18%)	3 (1%)	3 (1%)	245 (6.21%)
Verb_deviate	77 (12%)	81 (13%)	91 (15%)	96 (15%)	82 (13%)	73 (12%)	64 (10%)	58 (9%)	622 (15.77%)
Head_deviate	58 (11%)	58 (11%)	74 (14%)	82 (15%)	66 (12%)	72 (13%)	72 (13%)	57 (11%)	539 (13.66%)

Some participants, despite receiving excessive instructions, did not use the intended sentence structure. Instead, they first said the subject head and the verb, and then added the modifier, as in *The doctor is boiling by the pirate* (`wrong_structure`). We removed these trials from the analysis.

We also removed participants who did not use a determiner or used an incorrect determiner (`no_the`), or failed to use an auxiliary verb (`no_aux`) in their responses. Additionally, participants who did not use the correct verb in their responses (`verb_deviate`) or the correct subject head (`head_deviate`) were excluded.

Finally, we excluded participants who used the incorrect number-marked subjects. For example, even though the picture indicated a single doctor, some participants used the plural form of the verb (`wrong_num`).

In our analyses for attraction, we excluded only the trials marked as `wrong_structure`, `no_the`, `no_aux`, and `wrong_num`. We also excluded trials that did not have a clear auxiliary verb, in other words, those that were not fully pronounced. The total percentage of trials excluded from the attraction error analysis was 28.26%.

For our timing analyses, we excluded every disfluency, similar to the previous experiment. The total percentage of trials excluded from the timing analyses based on these criteria was 39.11%.

6.3 Analysis

The decisions related to the analyses were the same as in Experiment 1. We conducted two error-related analyses: disfluency and agreement attraction. Additionally, we conducted two timing analyses: the time to the first word and the time to utter the preverb and the auxiliary verb.

We also conducted an analysis that combines errors and timing: pause likelihood analyses.

6.4 Results

6.4.1 Disfluencies

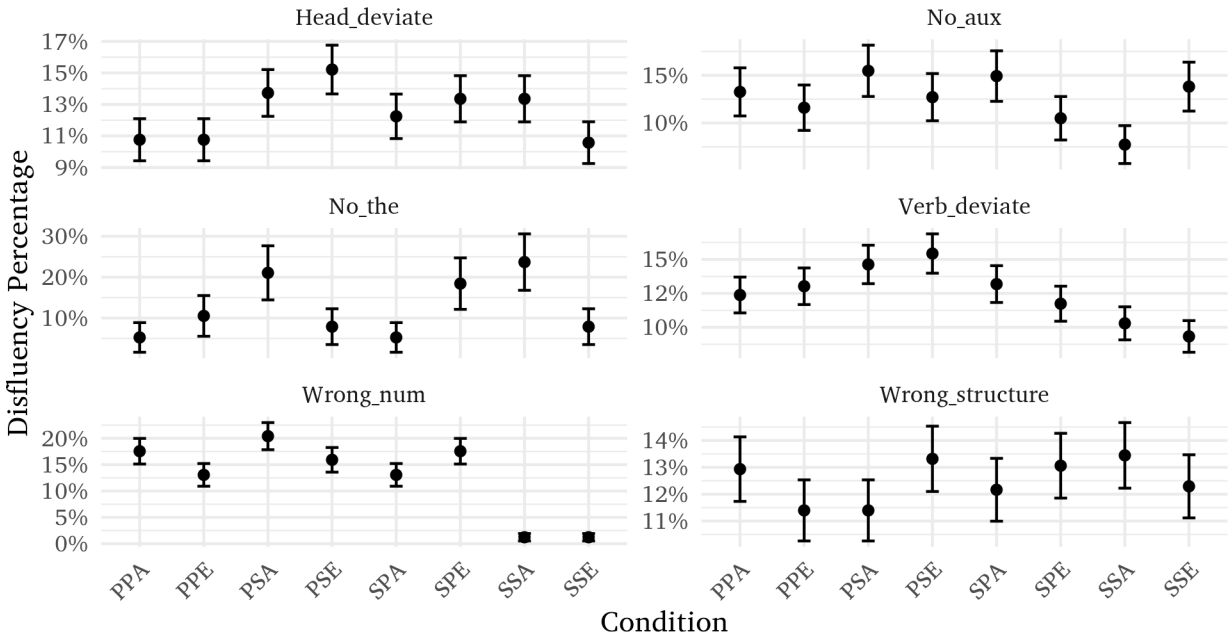
To simplify the presentation, we used three-letter abbreviations. The first letter indicates the head number (P = plural, S = singular), and the second letter indicates the attractor number (P = plural, S = singular). The final letter indicates the verb type (A = unaccusative, E = unergative).

Figure 45 illustrates the mean percentage of each disfluency across these conditions. One of the conditions that consistently showed increased disfluency was the plural head singular attractor condition. We also observed that in certain disfluencies, such as `wrong_num`, `head_deviate`, or `verb_deviate`, singular heads induced fewer overall errors.

We fitted a Bayesian probit regression model to examine predictors of disfluency error likelihood in Experiment 2. The predictors included verb type (unaccusative vs. unergative), the number of the head noun (plural vs. singular), and the number of the attractor noun (plural vs. singular), along with their interactions. The model included by-subject random slopes for all fixed effects and a random intercept for head due to convergence related issues.

The posterior distributions for our model, along with the probability of direction is provided in Figure 46. Among the main effects, the number of the head noun significantly increased the likelihood of disfluency errors ($\hat{\beta} = 0.21; CI = [0.04; 0.38];$

Figure 45
Mean Percentages with Standard Error for disfluencies in Exp2 for each condition.



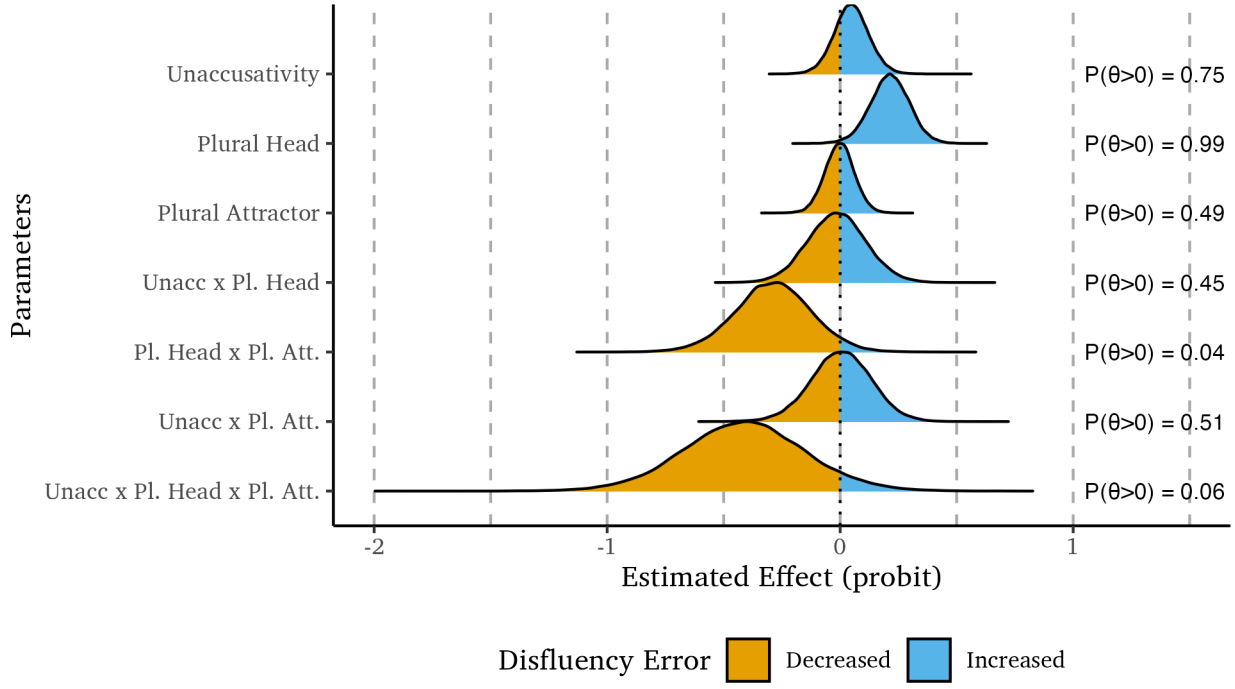
$P(\beta > 0) = .992$), suggesting that participants made more disfluencies with plural heads and had increased production difficulty. The effects of verb type ($\hat{\beta} = 0.05$; $CI = [-0.10; 0.20]$; $P(\beta > 0) = .75$) and attractor number ($\hat{\beta} = -0.00$; $CI = [-0.13; 0.12]$; $P(\beta > 0) = .49$) were not conclusive, showing no evidence of consistent influence on disfluency.

Most of the two-way interactions did not show compelling effects: the verb type $\times$ head number interaction ($\hat{\beta} = -0.01$; $CI = [-0.26; 0.23]$; $P(\beta > 0) = .45$), and the verb type $\times$ attractor number interaction ($\hat{\beta} = 0.00$; $CI = [-0.26; 0.26]$; $P(\beta > 0) = .51$). Both of these interactions yielded posterior distributions centered near zero, with low probability of meaningful effects. However, the interaction between attractor number and the head number showed strong evidence for an effect on the disfluencies ($\hat{\beta} = -0.29$; $CI = [-0.62; 0.04]$; $P(\beta < 0) = .96$). We also found a strong evidence for a small negative effect of the three-way verb type $\times$ attractor number $\times$ head number ($\hat{\beta} = -0.41$; $CI = [-0.94; 0.11]$; $P(\beta < 0) = .94$), suggesting a modulation of disfluency likelihood by the combined presence of these structural features.

These results suggest that **plural head nouns** reliably increase disfluency rates. However, participant had an easier time when both heads were plural. The combined effects of matching number were even more amplified when the verb was unaccusative.

Figure 46

Posterior distribution and the degree of belief for the probit regression coefficients for the model of disfluency errors in Experiment 2.



6.4.2 Agreement Attraction

One of our research questions was whether the effect of a plural attractor would change according to the verb type. Given that previous work by Momma and Ferreira (2019) has shown that verb type determines the timing of verb planning, we hypothesized that if the planning of agreement is dependent on the verb planning, late-planned verbs might be more prone to attraction errors than early-planned verbs due to the relative timing of the distractor and verb planning. However, our data did not show a clear effect of verb type on the agreement attraction errors.

Figure 47 presents the average proportions of agreement attraction errors by experimental condition. The first thing we notice is that our changes in the experiment

were successful. We now observe attraction errors as large as those reported in Veenstra et al. (2014) and even larger than in our previous experiment and Nozari and Omaki (2022).

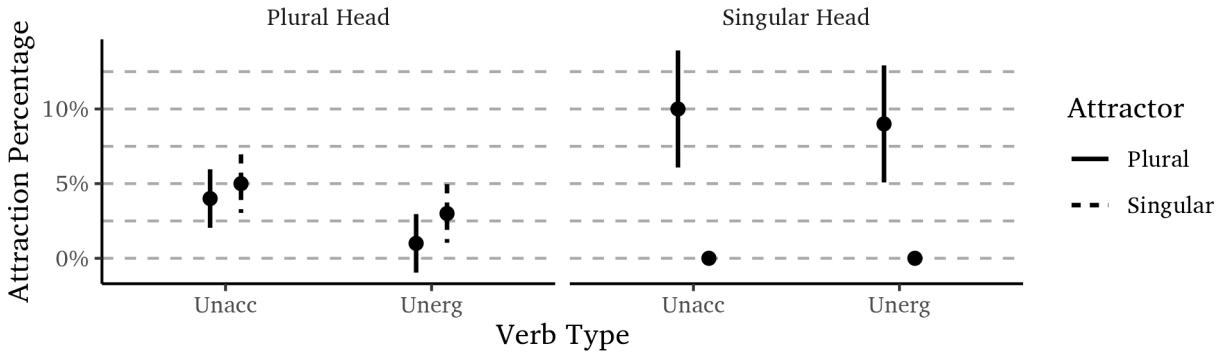
Within the singular heads, we see that participants made more agreement errors overall when the attractor was plural (M = 0.1, SE = 0) than when it was singular (M = 0, SE = 0). This effect was observed in both unaccusatives ($\Delta_{PL\text{-}SG}$M = 0.10 [0.06, 0.14]) and unergatives ($\Delta_{PL\text{-}SG}$M = 0.09 [0.05, 0.13]) in a symmetric manner.

However, within plural heads, the picture was slightly different. First of all, although there is a suggestion of an overall difference based on the attractor number—i.e., participants made more errors with singular attractors compared to plural attractors—the magnitude of this difference is quite small. Moreover, when the verb type manipulation is introduced, this difference becomes less prevalent.

The difference based on attractor number was very small, and possible values for this difference were observable with either sign for both unaccusatives ($\Delta_{PL\text{-}SG}$M = -0.01 [-0.04, 0.02]) and unergatives ($\Delta_{PL\text{-}SG}$M = -0.02 [-0.05, 0.01]).

Figure 47

*The average percentages of agreement errors according to the experimental conditions (excluding semantic relatedness) in our Exp2. Error bars signal Morey-Cousineau 95% confidence intervals (1.96*Standard Error) corrected for by-subject variance and not by-item variance.*



We fitted a nested Bayesian probit regression models to investigate agreement attraction errors in Experiment 2. Our research question was whether attraction effects

are modulated by verb type, as this would provide evidence for early planning of agreement. To this end, we are not interested how the head number interacts with the other predictors. Our descriptive results also suggested that the locus of the differences would be with singular heads. To isolate the effects of attractor number and verb type on agreement attraction, we fitted our model using only items with singular heads. Predictors were the same as the disfluency model, except for the head number. The model included random slopes for all predictors by subject and by head.

Figure 48 shows the posterior distributions and our degree of belief for a given effect in a model with only singular head conditions. Our model shows that verb type, i.e. unaccusativity, did not have a main effect on attraction errors ($\hat{\beta} = 0.02$; $CI = [-1.69; 1.72]$; $P(\beta > 0) = .51$), suggesting that participants' overall error was comparable in both verb types. We were able to find a strong evidence for an increased overall attraction error with plural attractors (mismatch conditions) ($\hat{\beta} = 3.45$; $CI = [1.87; 5.61]$; $P(\beta > 0) > .999$).

For us the important posterior is the interaction between the unaccusativity and the presence of the plural attractor. We did not find any evidence for this interaction ($\hat{\beta} = 0.06$; $CI = [-3.30; 3.49]$; $P(\beta > 0) = .51$), verifying our results from the descriptive statistics.

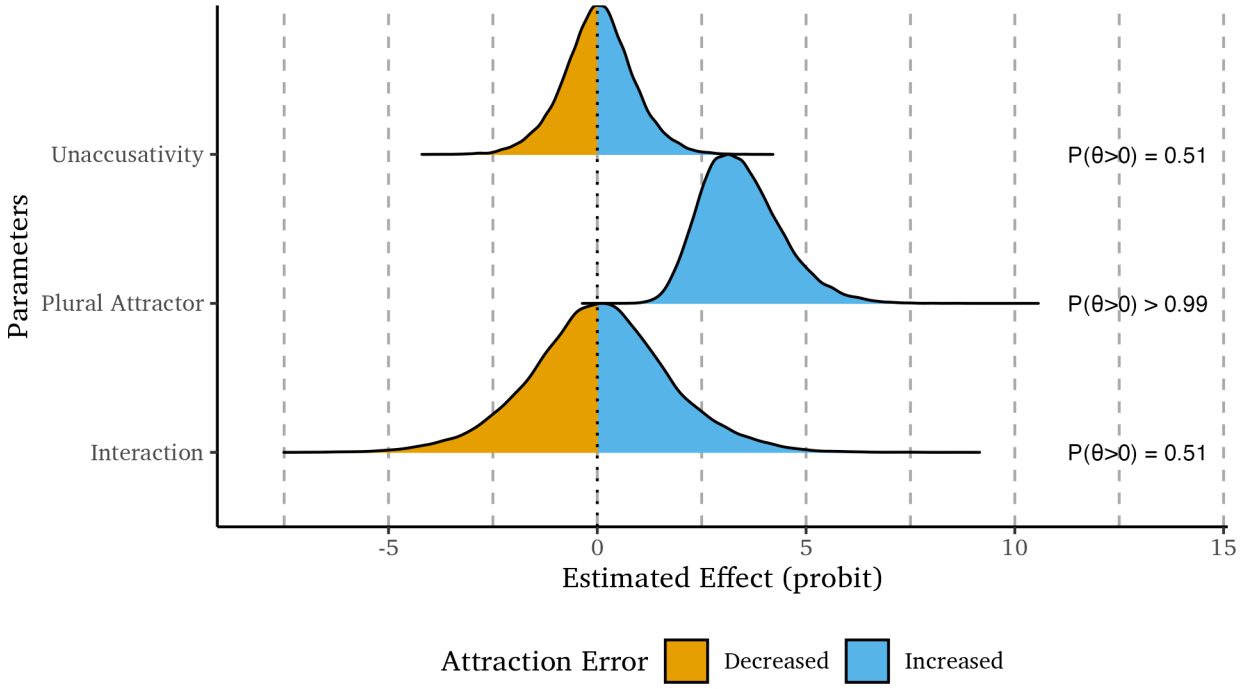
Together, these results suggest that **unaccusativity did not modulate agreement errors**, particularly in the presence of plural attractors. This supports the view that agreement is planned independent of the verb planning.

We also report results from a pooled model that includes all items and all predictors in a single regression. While this model is less informative than the condition-specific models discussed above, it provides a useful **overview of the general effects** of the predictors across the full dataset.

In this model, we replaced standard head number with a match/mismatch variable, allowing us to more directly assess whether attraction errors are more likely

Figure 48

Posterior distribution and the degree of belief for the probit regression coefficients with for the model of attraction errors in Experiment 2 with singular heads only.



when the attractor mismatches the head in number. The fixed effects included verb type (unaccusative vs. unergative), attractor number (plural vs. singular), mismatch (mismatch vs. match), and their two- and three-way interactions. The model included by-subject and by-head random slopesfor all predictors.

Figure 49 shows the posterior distributions and our degree of belief for a given effect in a model with all conditions. Main effects revealed that both plural attractors ($\hat{\beta} = 1.81$; $CI = [0.83; 3.18]$; $P(\beta > 0) > .999$) and mismatching number between the head and attractor ($\hat{\beta} = 2.09$; $CI = [1.09; 3.44]$; $P(\beta > 0) > .999$) substantially increased the likelihood of an attraction error. These results align with standard agreement attraction effects.

However, none of the interactions involving **verb type** showed compelling evidence. The verb type $\times$ attractor number interaction was small and uncertain ($\hat{\beta} = 0.39$; $CI = [-1.85; 2.71]$; $P(\beta > 0) = .65$), and neither the verb type $\times$ mismatch

($\hat{\beta} = -0.31$; $CI = [-2.58; 2.02]$; $P(\beta > 0) = .38$) nor the three-way interaction with mismatch and attractor number ($\hat{\beta} = -0.93$; $CI = [-5.51; 3.55]$; $P(\beta > 0) = .33$) showed strong modulation of attraction effects by unaccusativity.

A notable interaction emerged between attractor number and mismatch ($\hat{\beta} = -2.52$; $CI = [-5.26; -0.44]$; $P(\beta > 0) = .007$), indicating that the overall effect of plural attractor were not amplified when it mismatches with the subject head. This is surprising given our descriptive statistics, but one possiblity is that the model overpredicts the main effect of the plural attractor and counterbalances it with a negative interaction.

In summary, these findings reinforce that mismatching attractors are key drivers of agreement attraction. We were able to replicate the previous findings in attraction. Our data also showed the patterns dubbed as plural markedness effects, i.e. the attraction seems to be harder to induce with plural heads. However, there is no strong evidence that unaccusativity modulates attraction effects, suggesting that the planning of the agreement does not follow the timeline of the planning of the verbs.

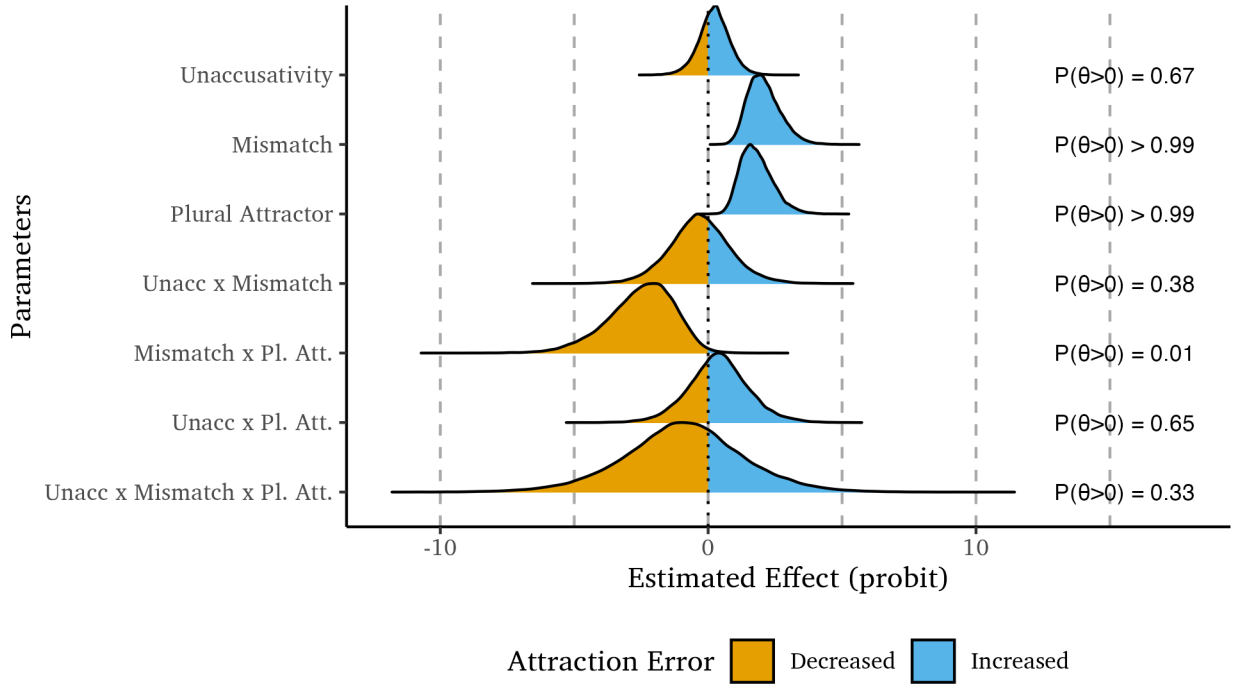
6.4.3 Pause Likelihood

Kandel and Phillips (2022) recently showed that the increased pause likelihood immediately before the auxiliary verb can be a good indicator of the agreement computation. Following the difference between the planning of the verb, we expected that the pause likelihood would differ between verb types if the agreement and the verb are planned separately. However, our data did not show a clear effect of verb type on the pause likelihood. Both unaccusative and unergative verbs showed similar pause likelihoods when the subject was singular. Yet, we observed a difference in pause likelihood when the subject was plural.

Kandel and Phillips (2022) found an overall 11% pause likelihood within their data. In our previous experiment, we found an overall 29% pause likelihood, which we attributed to the overall uncertainty about the verb, given the sheer number of verbs

Figure 49

Posterior distribution and the degree of belief for the probit regression coefficients with for the model of attraction errors in Experiment 2.



that had to be remembered. In this experiment, the overall pause likelihood was 20%, which is lower than our previous experiments but still higher than the 11% found by Kandel and Phillips (2022). The fact that the number of verbs was reduced in this experiment might have contributed to the lower pause likelihood.

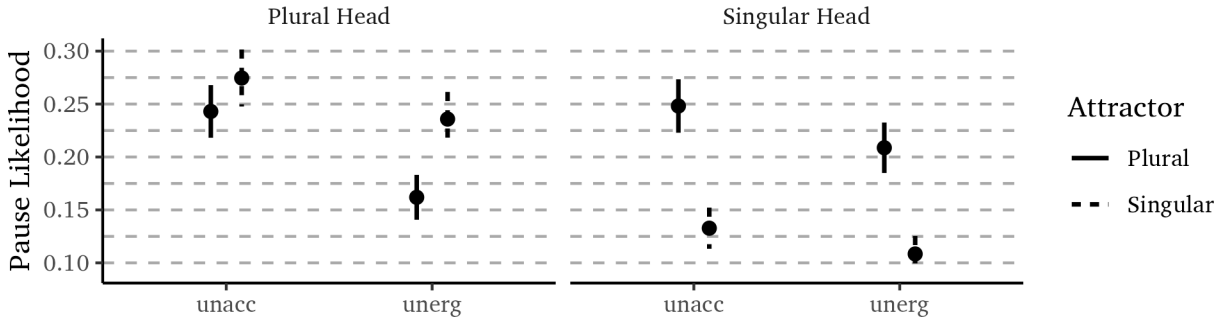
Figure 47 presents the average proportions of pause likelihoods by experimental condition. Within singular head conditions, we notice that the pause likelihood was higher when the attractor was plural compared to when it was singular. This effect was observed in both unaccusatives ($\Delta_{PL\text{-}SG}$M = 0.12 [0.05, 0.18]) and unergatives ($\Delta_{PL\text{-}SG}$M = 0.10 [0.04, 0.16]). This points towards an increased difficulty in computing the agreement as a function of the attractor number.

However, the same cannot be said for the items with a plural head. Even though we did not find any effect of attractor number in plural heads, we found that in unergatives, the number of the attractor mattered for pause likelihoods. Unergatives in

the PS condition had a higher pause likelihood (M = 0.24, SE = 0.03) compared to the PP condition (M = 0.16, SE = 0.02). However, this difference was not observed in unaccusative conditions with plural heads ($\Delta_{PL\text{-}SG}$M = -0.03 [-0.10, 0.04]), as indicated by a confidence interval for differences including 0.

Figure 50

*The average likelihood of pause between the second noun and the auxilliary according to the experimental conditions in our Exp2. Error bars signal Morey-Cousineau 95% confidence intervals (1.96*Standard Error) corrected for by-subject variance and not by-item variance.*



We analyzed pause likelihood by fitting three Bayesian models to our data in Experiment 2. We will first present two nested models, in which the models were only fit to subset of the data, singular head conditions only and plural head conditions only. We will lastly present the model fitted to the pooled data, i.e. all conditions in Experiment 2. The models included fixed effects for verb type (unaccusative vs. unergative), attractor number (plural vs. singular), and their interaction, along with by-subject and by-head random slopes. For the pooled model, we also included the predictor mismatch (mismatch vs. match) and all two- and three-way interactions.

Figure 51 shows the posterior distributions and our degree of belief for a given effect in a model with only singular head conditions. The model revealed a significant positive effect of plural attractors on pause likelihood ($\hat{\beta} = 0.53$; $CI = [0.10; 0.94]$; $P(\beta > 0) = .99$), suggesting that encountering a plural noun in the attractor position increases the likelihood of a pause by a participant before an auxiliary verb. The effect of verb type was positive but the evidence for it is weaker ($\hat{\beta} = 0.18$; $CI = [-0.14; 0.50]$; $P(\beta > 0) = .89$), indicating a possible—but not strong—tendency for unaccusative verbs

to elicit pauses more often overall.

To answer our research question, we have to look at the interaction interaction between verb type and attractor number. The posterior of the interaction spans over 0 and the the effect being negative or positive is at chance level ($\hat{\beta} = 0.04$; $CI = [-0.56; 0.65]$; $P(\beta > 0) = .55$), suggesting no clear evidence that the effect of attractor number differs as a function of verb type.

Figure 51

Posterior distribution and the degree of belief for the probit regression coefficients with for the model of pause likelihood in Experiment 2 with singular heads only.

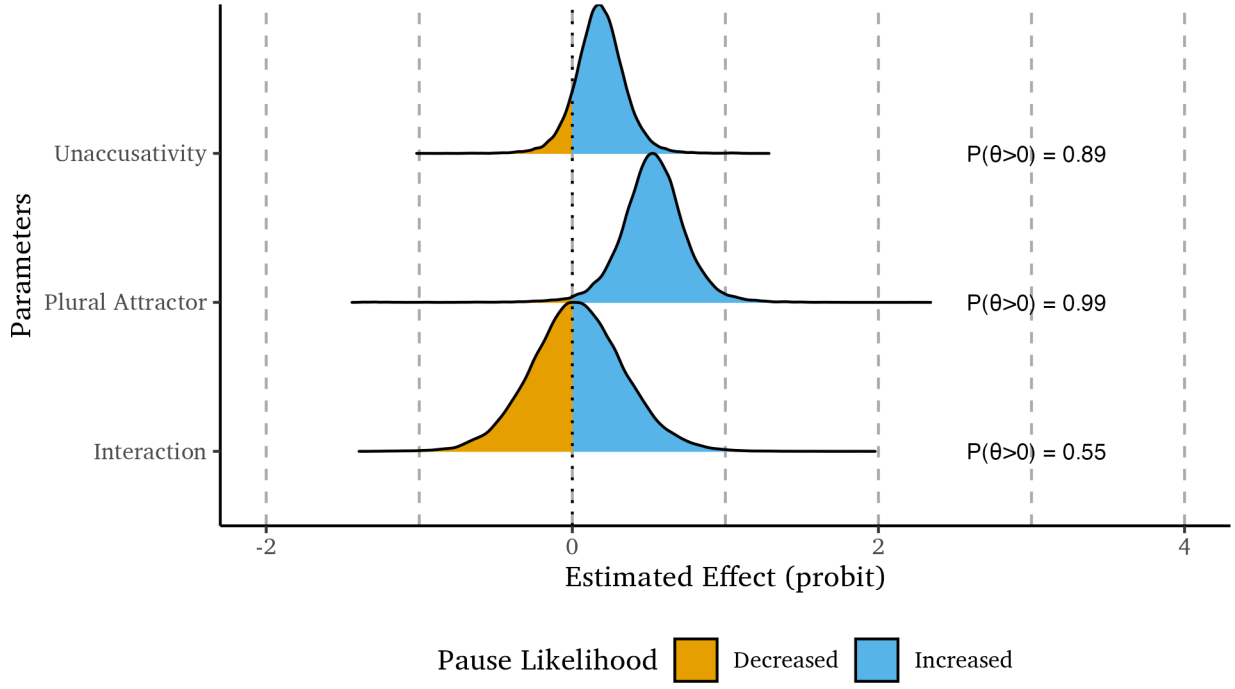


Figure 52 shows the posterior distributions and our degree of belief for a given effect in a model with only plural head conditions. Our model found a strong evidence for a negative main effect of attractor number within plural head conditions ($\hat{\beta} = -0.21$; $CI = [-0.48; 0.05]$; $P(\beta < 0) = .95$), suggesting when the attractor and head matches with number, people produced less pauses as predicted by previous findings. In contrast, verb type, unaccusativity, showed a strong positive influence on the pause likelihood ($\hat{\beta} = 0.25$; $CI = [-0.08; 0.57]$; $P(\beta > 0) = .95$), indicating that the upcoming

unaccusative verbs leads participants to be slow down in their speech. However, we were not able to find any evidence for the interaction between verb type and attractor number ($\hat{\beta} = 0.17$; $CI = [-0.28; 0.61]$; $P(\beta > 0) = .78$) when the subject head is plural.

Figure 52

Posterior distribution and the degree of belief for the probit regression coefficients with for the model of pause likelihood in Experiment 2 with plural heads only.

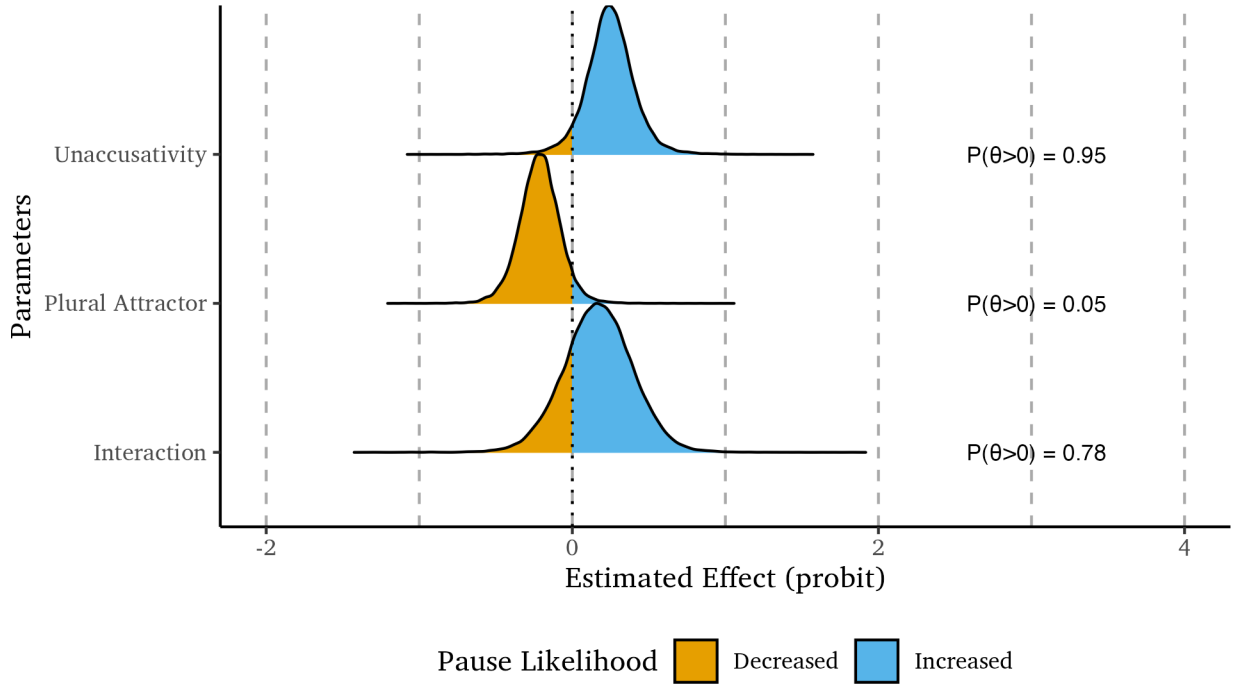


Figure 53 shows the posterior distributions and our degree of belief for a given effect in a model with all conditions. The present posteriors are comparable to the posteriors of the agreement attraction errors. Main effects revealed that both plural attractors ($\hat{\beta} = 0.15$; $CI = [-0.07; 0.36]$; $P(\beta > 0) = .92$) and mismatching number between the head and attractor ($\hat{\beta} = 0.38$; $CI = [0.11; 0.67]$; $P(\beta > 0) = .992$) substantially increased the likelihood of an attraction error. These results align with standard agreement attraction effects.

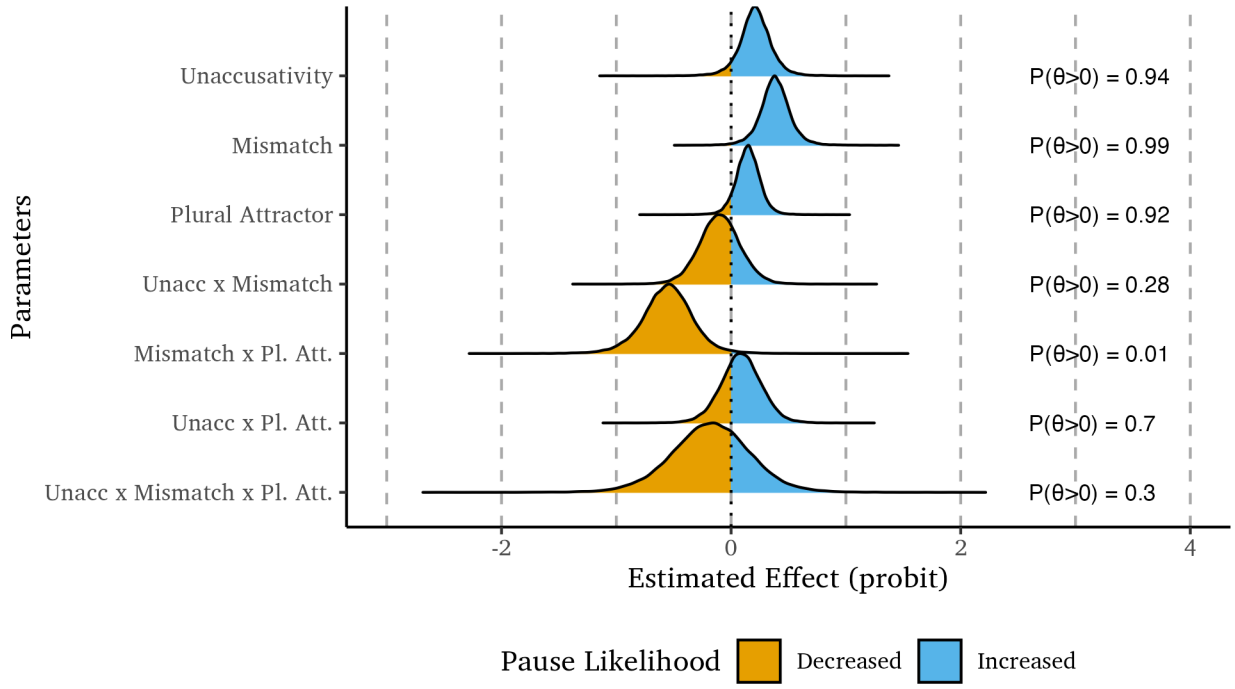
Even though we found a strong evidence for a main effect of unaccusativity ($\hat{\beta} = 0.21$; $CI = [-0.09; 0.50]$; $P(\beta > 0) = .94$), none of the interactions involving **verb type** showed compelling evidence. The verb type $\times$ attractor number interaction was

small and uncertain ($\hat{\beta} = 0.09$; $CI = [-0.27; 0.44]$; $P(\beta > 0) = .70$), and neither the verb type × mismatch ($\hat{\beta} = -0.10$; $CI = [-0.46; 0.27]$; $P(\beta > 0) = .28$) nor the three-way interaction with mismatch and attractor number ($\hat{\beta} = -0.18$; $CI = [-0.88; 0.51]$; $P(\beta > 0) = .30$) showed strong modulation of attraction effects by unaccusativity.

Similar to the posteriors of the attraction model, the interaction between attractor number and mismatch had strong evidence for a negative effect ($\hat{\beta} = -0.55$; $CI = [-1.02; -0.10]$; $P(\beta > 0) = .01$), indicating that the effect of mismatch and the plural attractor were not fully additive.

Figure 53

Posterior distribution and the degree of belief for the probit regression coefficients with for the model of pause likelihood in Experiment 2.



6.4.4 Onset Latency

Our question related to the agreement planning depends on early verb planning. Even though we did not use a semantic interference task in this experiment, previous studies allow us to make a prediction about the onset latency. Previous studies on morphological production in determiner phrases, as covered in Section 3.1, have shown

that the presence of mismatching features (gender or number) might affect the onset latency and interfere with the planning. In Section 3.1, we discussed the experiments conducted by Schriefers (1993). In his experiments, he showed that participants slowed down when the superimposed word and the picture had mismatching features. Even though the locus of this effect is still under discussion, i.e., whether it is at the lexeme (morpho-phonological) or lemma (morpho-syntactic) level, it is possible that the relevant morpho-syntactic computation can be done earlier, resulting in a facilitation effect when the features match.

Moreover, the name agreement can also offer another prediction. Griffin (2001) shows that the onset latency of the verb is affected by the number of potential names for a picture. Even though there is no replication of this study with verbs, we will assume that this effect may be generalizable to verbs as well. We will discuss possible future directions related to this idea in the General Discussion section.

Additionally, previous studies on morphological production in determiner phrases, as covered in Section 3.1, have shown that the presence of mismatching features (gender or number) might affect onset latency and interfere with the planning process. In Section 3.1, we discussed the experiments conducted by Schriefers (1993), who found that participants slowed down when the superimposed word and the picture had mismatching features.

Even though the locus of this effect is still under discussion—whether it arises at the lexeme (morpho-phonological) or lemma (morpho-syntactic) level—it is possible that the relevant morpho-syntactic computation can be done earlier, potentially resulting in a facilitation effect when the features match.

Figure 54 presents the average onset latencies by experimental condition. We observe that onset latencies were overall higher ($M = 1268.07$, $SE = 9.2$) in this experiment compared to our Experiment 1 ($M = 1037.91$, $SE = 6.09$).

Our descriptive plots suggest two main findings. First, overall onset latencies

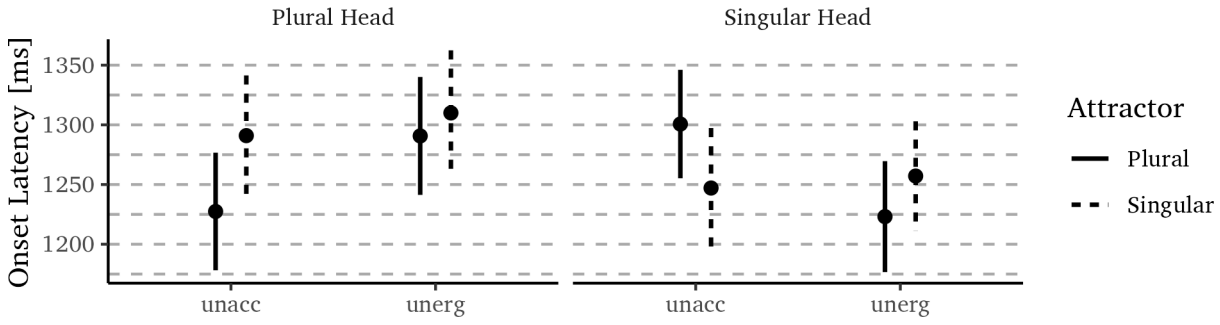
were higher when the head was plural (M = 1279.48, SE = 12.84) compared to when it was singular (M = 1257.6, SE = 12). Second, in the unaccusative verb conditions, we observed an effect of attractor number across each condition.

With singular heads, unaccusative sentences with a plural attractor—i.e., the condition where the head and the attractor number mismatched—led to longer onset latencies (M = 1300.67, SE = 23.17) compared to the condition where the head and attractor number matched (M = 1247.07, SE = 25.65). This effect was not present in unergative sentences with a singular head ($\Delta_{PL\text{-}SG}$M = -33.99 [-99.28, 31.29]).

With plural heads, we see a similar picture. In conditions where the verb is unaccusative and the number of the head and the attractor does not match, participants' onset latency changed. While the overall latency was high in plural head conditions, when the unaccusative verb had a plural attractor (i.e., matching features), participants took less time to start uttering the sentence (M = 1227.43, SE = 25.13) compared to when the unaccusative verb had a singular attractor (M = 1290.96, SE = 25.73). This effect was not present in unergative sentences ($\Delta_{PL\text{-}SG}$M = -19.27 [-91.29, 52.76]).

Figure 54

*The average onset times according to the experimental conditions in our Exp2. Error bars signal Morey-Cousineau 95% confidence intervals (1.96*Standard Error) corrected for by-subject variance and not by-item variance.*



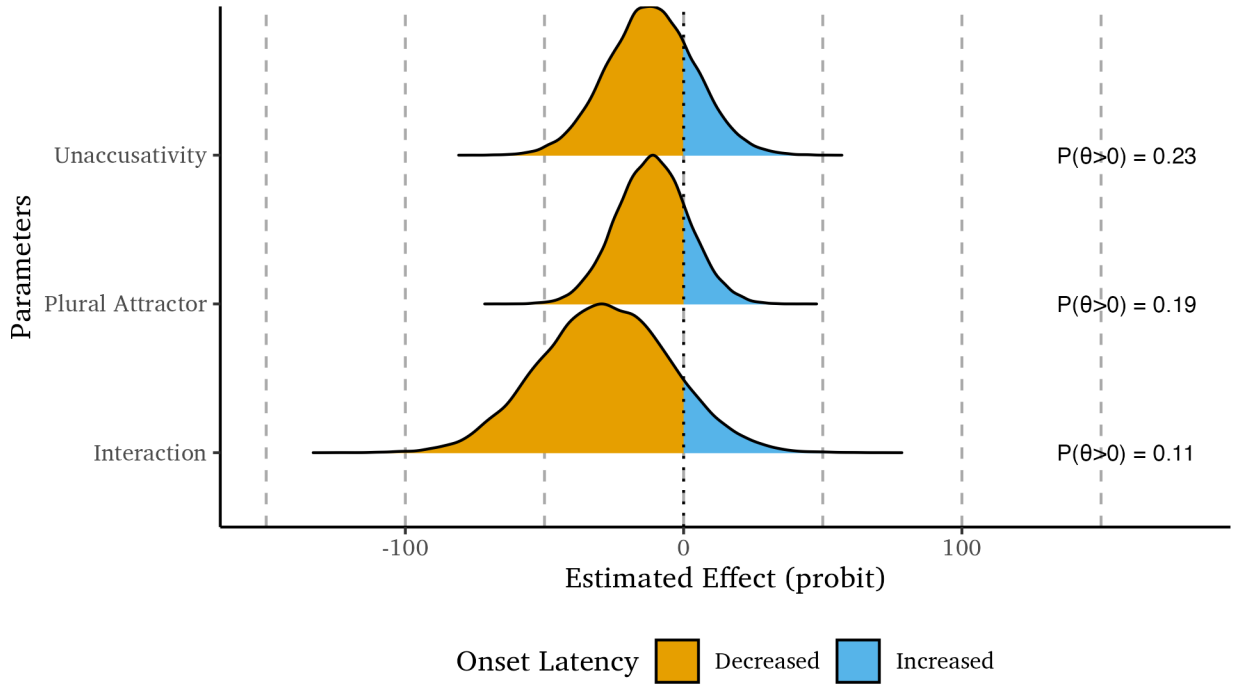
We analyzed onset latencies by fitting two nested Bayesian models in Experiment 2, assuming an ex-Gaussian mixture distribution. The models included fixed effects for verb type (unaccusative vs. unergative), attractor number (plural vs. singular), and their interaction, as well as by-subject and by-head random slopes. In the current analysis,

we restricted the dataset to plural head conditions, in order to isolate the effects of attractor and verb structure without the influence of head number.

As shown in Figure 55, the posterior distributions provide our degree of belief that an effect exists in a particular direction. Our model did not find substantial evidence for the negative effect of unaccusativity ($\hat{\beta} = -11.65$; $CI = [-42.39; 19.32]$; $P(\beta < 0) = .77$) or the plural attractor (match) ($\hat{\beta} = -11.29$; $CI = [-36.51; 14.17]$; $P(\beta < 0) = .81$). Meaning that overall onset latencies were not significantly different between unaccusative and unergative verbs, nor between plural and singular attractors when the head noun is plural. However, the interaction between verb type and attractor number ($\hat{\beta} = -28.26$; $CI = [-73.51; 17.33]$; $P(\beta < 0) = .89$) showed a negative estimate with posterior mass leaning below zero. Although this effect did not reach a high certainty threshold, the directionality is consistent with a trend toward greater planning cost in unaccusative structures and when an attractor noun and the head noun matches in number.

Figure 55

Posterior distribution and the degree of belief for the exGaussian regression coefficients for the model of onset latencies in Experiment 2 with plural heads only.



To complement our analysis of onset latencies in plural-head conditions, we also fitted a Bayesian ex-Gaussian mixture model to singular-head items only in Experiment 2, posteriors of which is shown in Figure 56. All three effects—verb type ($\hat{\beta} = 8.75$; $CI = [-22.24; 39.28]$; $P(\beta > 0) = .71$), plural attractor (mismatch) ($\hat{\beta} = 14.62$; $CI = [-15.76; 44.97]$; $P(\beta > 0) = .82$), and their interaction ($\hat{\beta} = 26.02$; $CI = [-22.02; 73.83]$; $P(\beta > 0) = .86$)—trended positively, suggesting that these features may be associated with longer onset latencies and this latency is amplified within unaccusatives. However, the credible intervals were wide and included zero in all cases, and posterior probabilities remained below conventional thresholds for even moderate evidence.

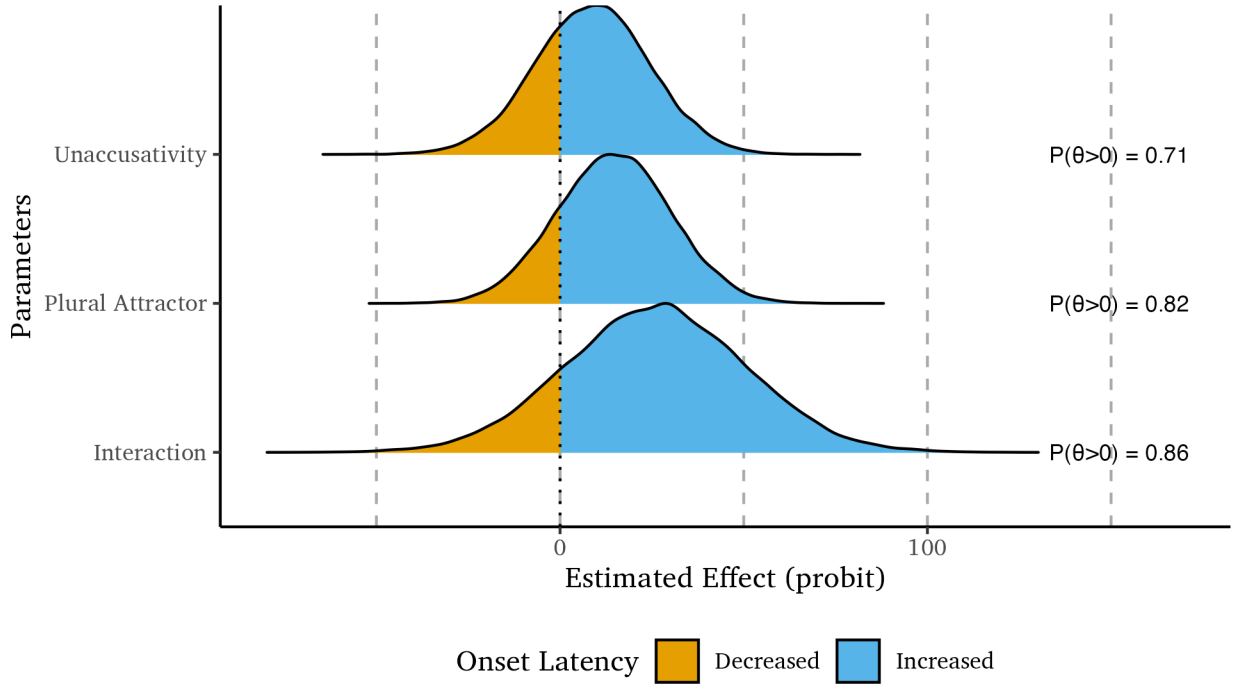
These findings provide tentative evidence that there is a relationship between the number match between the head and the attractor, only in unaccusative scenes. This is unexpected given that the previous measures like agreement attraction errors and pause likelihood did not show a different picture between unaccusative and unergative verbs. We will discuss the implications of this finding in the General Discussion section.

6.4.5 Preverbal Time

Similar to onset latencies, we did not have strong a priori predictions for preverbal times, as our experiment did not include direct manipulations targeting verb planning. However, we consider potential effects related to agreement computation, which we discuss in more detail in the discussion section—particularly in relation to name agreement. As far as we are aware, this is the first set of experiments to report preverbal production times in the context of agreement attraction. Previous production studies have primarily focused on onset latencies, especially in preamble completion paradigms. While picture-description tasks have been used in related work, they have not typically reported preverbal timing measures. Given the structure of our design, we expect preverbal times to broadly align with the patterns observed in pause likelihood, reflecting the timing of agreement computation across conditions.

Figure 56

Posterior distribution and the degree of belief for the exGaussian regression coefficients for the model of onset latencies in Experiment 2 with singular heads only.



In our calculation, preverbal time was defined as the interval between the onset of the second noun phrase and the onset of the main verb, signaled by the square-bracketed region in sentences like *the doctor(s) by the [pirate(s) (is/are)]jumping*. To eliminate potential confounds between conditions, we subtracted the duration of the plural morpheme on the second noun from the overall measure. Additionally, we verified that the pronunciation of *is* versus *are* did not systematically differ in duration, ensuring that variation in preverbal time was not driven by phonetic differences between the auxiliary forms.

Figure 57 presents the average preverbal production times across experimental conditions. We observed that the average preverbal production time in this experiment (M = 674.89, SE = 4.64) was lower than that in Experiment 1 (). This reduction may be attributed to the absence of a semantic interference task in the current design, which likely decreased processing demands during sentence planning. Additionally, the

smaller set of verbs used in this experiment may have contributed to the shorter preverbal times, as participants had fewer lexical alternatives to retrieve and manage during production.

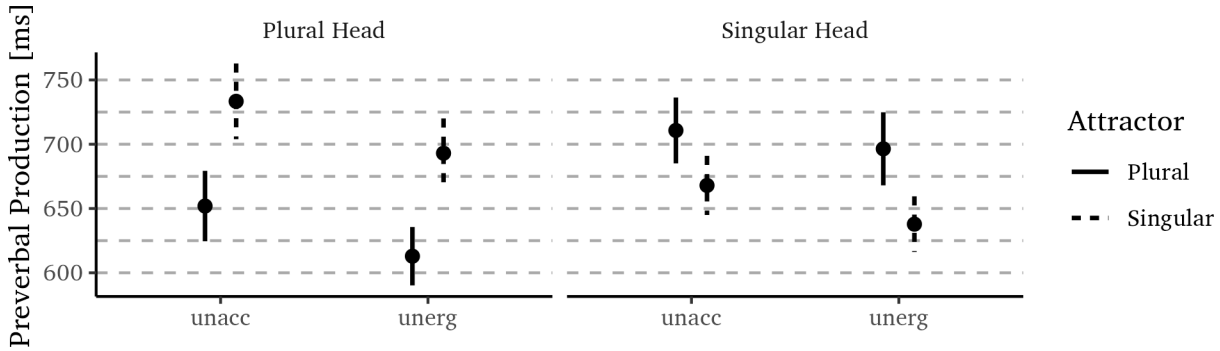
In addition to the overall reduction in preverbal production times for this experiment, we observed an effect of number mismatch between the head noun and the attractor. Specifically, within plural-headed noun phrases, preverbal production times were higher when the attractor was singular (M = 713.67, SE = 10.22), compared to when it was plural (M = 631.48, SE = 9), indicating increased processing time in number mismatch conditions. This effect appears to be independent of verb type: both unaccusative ($\Delta_{PL\text{-}SG}$M = -81.45 [-121.66, -41.25]) and unergative ($\Delta_{PL\text{-}SG}$M = -80.11 [-115.42, -44.80]) verbs exhibited comparable differences in preverbal time as a function of attractor number. These findings suggest that preverbal timing is sensitive to agreement-related mismatch effects, but not strongly modulated by the structural differences between unaccusative and unergative verbs.

In contrast to the clear pattern observed within plural-headed noun phrases, the effect of attractor number was less pronounced in singular-headed conditions. The difference in preverbal production time based on attractor number was smaller, and the observed differences were not consistently signed across verb types. For both unaccusative ($\Delta_{PL\text{-}SG}$M = 42.74 [8.32, 77.17]) and unergative ($\Delta_{PL\text{-}SG}$M = 58.60 [22.88, 94.31]) verbs, the difference in preverbal time between singular and plural attractors under singular heads could plausibly vary in either direction. Nonetheless, a general trend was still detectable: preverbal production time was higher when the singular head was followed by a plural attractor (M = 703.88, SE = 9.72), compared to when the attractor was also singular (M = 652.32, SE = 8.04). This suggests that number mismatch may continue to influence production timing under singular heads, though the effect is weaker and more variable than in plural-headed constructions.

We analyzed preverbal time—the time between utterance onset and the start of

Figure 57

*The average preverbal production times according to the experimental conditions in our Exp2. Error bars signal Morey-Cousineau 95% confidence intervals (1.96*Standard Error) corrected for by-subject variance and not by-item variance.*



the verb—using a Bayesian ex-Gaussian regression model. The model included fixed effects of verb type (unaccusative vs. unergative), attractor number (plural vs. singular), and subject-verb number match (mismatch vs. match), as well as their interactions. We also included by-subject random slopes for all predictors and by-head random slopes for main effects. Our contrasts were centered such that unaccusative = 0.5, plural = 0.5, and mismatch = 0.5.

As shown in Figure 58, we observed a strong positive effect of mismatch, where preverbal times were longer in conditions where the nouns have mismatching number ($\hat{\beta} = 28.97$; $CI = [8.57; 46.70]$; $P(\beta > 0) = .996$). This effect was amplified when the attractor was plural, as evidenced by a strong interaction between the attractor number and mismatch ($\hat{\beta} = 27.00$; $CI = [5.89; 47.94]$; $P(\beta > 0) = .994$), suggesting that plural attractors in mismatching configurations (singular head) lead to increased preverbal production times overall

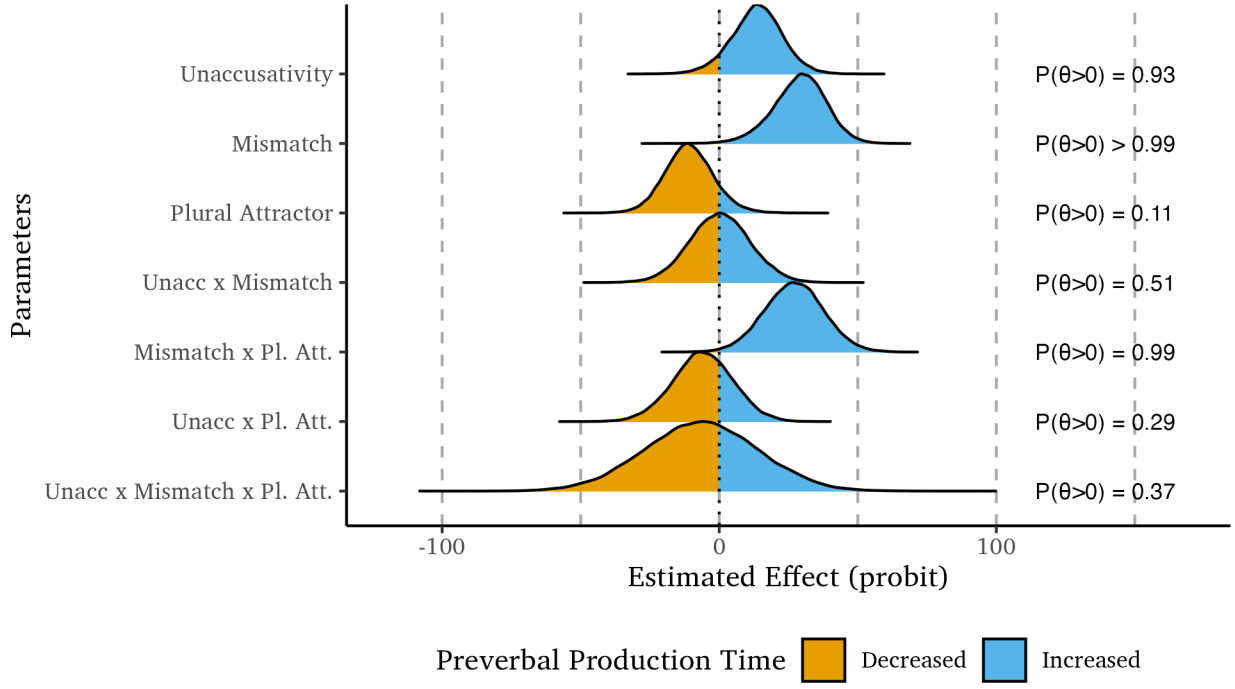
There was also evidence for a main effect of verb type, with unaccusative verbs associated with longer preverbal times ($\hat{\beta} = 13.57$; $CI = [-5.51; 31.50]$; $P(\beta > 0) = .93$), suggesting that unaccusativity overall increases the time required to initiate the verb. The main effect of attractor number ($\hat{\beta} = -10.83$; $CI = [-27.79; 7.43]$; $P(\beta > 0) = .11$) showed a weaker evidence in the same direction.

Our main research question was whether any of these effects were specifically modulated by the verb type. We see that interactions involving verb type—including verb type × match ($\hat{\beta} = 0.28$; $CI = [-21.00; 21.85]$; $P(\beta > 0) = .51$), verb type × attractor number ($\hat{\beta} = -5.87$; $CI = [-27.36; 15.23]$; $P(\beta > 0) = .29$), and the three-way interaction ($\hat{\beta} = -6.50$; $CI = [-46.72; 33.31]$; $P(\beta > 0) = .37$)—did not yield reliable evidence for modulation.

Together, these results suggest that preverbal planning is sensitive to subject-verb agreement configurations. However, we find little evidence that these effects are modulated by verb type, implying that unaccusativity does not reliably influence preverbal timing in this context.

Figure 58

Posterior distribution and the degree of belief for the exGaussian regression coefficients for the model of preverbal production time in Experiment 2.



6.5 Discussion

The primary goal of Experiment 2 was to obtain clearer and more robust agreement attraction effects, particularly to better observe how such effects interact

with timing-related measures and differences between unaccusative and unergative verbs. To achieve this, we made two key changes to the experimental design: (i) we manipulated the availability of the attractor noun as a potential agreement controller within the sentence context and (ii) we manipulated our scenes such that there were a communicative need for a modifier. In addition to these targeted changes, we also adjusted broader aspects of the experimental setup to improve task clarity and reduce working memory demands. Specifically, we reduced the number of verbs and nouns used across trials, aiming to simplify the production task and enhance the reliability of the measured effects.

With these changes, we were able to elicit levels of agreement attraction that are comparable to those reported in other picture description studies. However, unlike previous experiments, the attraction effects in our data were limited to singular head conditions—a pattern commonly referred to as plural markedness, where attraction errors are more likely when the subject head is singular and the attractor is plural. Notably, this pattern did not extend to the timing-based measure of agreement attraction, as pause likelihood exhibited clearer effects in both singular and plural head conditions.

We also found evidence that attractor number influenced pause likelihood in both singular and plural head conditions—even in cases where no attraction errors were observed. This suggests that pause likelihood may be more sensitive to number mismatch than overt attraction errors, potentially reflecting a computation in which the number features are integrated or conflict between possible features is detected, even if not ultimately realized as an error in production.

Turning to our central research question, we found no evidence that unaccusativity modulates agreement attraction. The pattern of attraction errors and pause likelihood was comparable in both unaccusative and unergative sentence contexts, suggesting that verb type did not meaningfully influence the likelihood or

timing of agreement attraction. The only suggestive evidence of a non-parallel effect of attractor number across verb types came from the analysis of onset latencies. Specifically, participants showed signs of facilitation when the attractor and head matched in number in plural head conditions, which appeared overall harder to initiate, and displayed inhibitory effects of mismatching number in singular head conditions. Importantly, this interaction was observed only in unaccusative sentences but our models only showed weak evidence for this interaction. While intriguing, this finding must be interpreted cautiously, and does not offer strong support for the hypothesis that unaccusativity shapes the time course or resolution of agreement attraction or answer our question in a way that suggest early agreement planning. Overall, our results provide ample evidence that agreement is not planned at the point of verb retrieval, but it is planned rather late at the point of utterance. However, measures like onset latency may offer important insights into the incremental processes underlying agreement planning.

While our study was not designed to explicitly contrast competing models of attraction, one notable finding is that the status of the attractor as a possible subject within the experiment influenced the magnitude of attraction errors between experiments. This pattern is difficult to explain using purely representational accounts, which do not predict sensitivity to the discourse or features like subject-like-ness of the attractor. In contrast, this effect aligns well with cue-based retrieval models, in which shared features between attractors and subjects (e.g., case, structural position, subjecthood cues) increase the likelihood of retrieval interference.

In the upcoming sections, we will discuss our findings more in depth.

6.5.1 Association with subjecthood modulates attraction

Our results in Experiment 2 diverged from those of Experiment 1 in that we were able to elicit attraction effects that are comparable in magnitude and pattern to those reported in previous production studies. We attribute this improvement to two

important design changes. First, we modified our materials so that all attractor nouns were plausible subjects within the experiment. This manipulation likely increased their salience as potential agreement controllers, thus enhancing their ability to interfere with agreement processing.

Recent work on the comprehension of agreement has highlighted the importance of language-specific subjecthood cues in shaping agreement processing (Bhatia & Dillon, 2022; Dillon & Keshev, n.d.; Lago et al., 2019; Slioussar, 2018; Türk & Logačev, 2024). Our results are consistent with these findings, suggesting that both experimental design and language-specific properties can modulate the extent to which attractors interfere with agreement. Specifically, shared features between attractors and subjects—such as structural position or case marking—appear to increase the likelihood of retrieval interference during agreement computation.

For example, Bhatia and Dillon (2022) leveraged aspectual properties of Hindi to manipulate which noun within a sentence was a plausible agreement controller. In Hindi, either the subject or the object can control agreement, depending on case marking. When the verb appears in the perfective aspect, the subject is marked with overt case and controlling the agreement shifts to the object. Their findings showed that only elements matching the grammatical role of the agreement controller in a given context—objects when the object controlled agreement, and subjects when the subject controlled agreement—served as attractors. Importantly, objects did not induce attraction when another object was not a licit agreement controller within the sentence, despite being potential controllers in other contexts.

In addition to the within-sentence manipulation of subjecthood, Türk (2022) introduced a between-trials manipulation of subjecthood in a comprehension tasks. In his study, participants were presented reduced relative clauses that can serve as a nominal modifiers. These relative-clause based modifiers due to their nominal nature can also marked with plural markers. Türk (2022) first showed that these plural marked

reduced-relative-clause modifiers do not induce attraction. More importantly, they showed that when participants were presented both attraction-inducing and non-attraction-inducing modifiers, overall attraction errors for attraction-inducing modifiers were significantly lowered.

Our results in Experiment 1 and 2 align with the interpretation of these findings from the comprehension literature that both experimental constraints and language-specific cues influence whether another noun can induce attraction.

Since our primary research question centered on the timing of agreement computation, we did not initially set out to distinguish between cue-based retrieval theories and representational accounts of agreement. This was partly because neither theoretical framework makes strong or explicit predictions about when in time agreement should be computed. However, our findings regarding the influence of an attractor's association with subjecthood appear to align more naturally with the assumptions of cue-based retrieval theories.

In cue-based retrieval accounts, agreement is computed through a memory retrieval process guided by feature-based cues shared between the subject and potential agreement controllers (Lewis & Vasishth, 2005). A straightforward prediction of this framework is that as the overlap of subject-related cues between the attractor and the subject increases, the likelihood of retrieval interference—and thus attraction errors—should also increase. In contrast, representational accounts do not predict sensitivity to subjecthood without additional assumptions. In these models, agreement attraction is determined primarily by three factors: the notional number of the subject head, the grammatical number of the attractor noun(s), and the structural distance between the attractor and the root of the sentence (Eberhard et al., 2005; Hammerly et al., 2019).

We think that the observed influence of subjecthood cues provides indirect support for cue-based retrieval mechanisms, suggesting that feature overlap between the

subject and attractor plays a key role in agreement computation.

It is important to interpret the theoretical implications of our findings on subjecthood and attraction with caution. While our results appear to parallel recent findings in the comprehension literature showing that association with subjecthood modulates attraction (Bhatia & Dillon, 2022; Dillon & Keshev, n.d.), these conclusions have not yet been systematically tested in production experiments. Moreover, although cue-based retrieval accounts more naturally predict that attraction is modulated by shared subject-like features between the attractor and the subject, it is possible for representational accounts to accommodate such findings under additional assumptions.

A key theoretical distinction between the two frameworks concerns what goes wrong during encoding. In representational accounts, the number feature of the subject head is misencoded, leading to agreement errors. In contrast, cue-based retrieval theories assume that all elements are correctly encoded, but the retrieval process erroneously selects the wrong element as the agreement controller (see Yadav et al., 2023 for nuances and additional models). For representational models to predict effects of subjecthood, one would have to assume not only that the number feature of the subject was misrepresented, but also that participants were uncertain about which noun was the subject head. This is a strong assumption that requires further empirical validation. Indeed, comprehension studies suggest that misidentification of the subject head is relatively rare, and that attraction effects are primarily driven by misencoding of number, not misinterpretations (Schlueter et al., 2019).

In light of these considerations, while our findings are suggestive of a role for subjecthood in production-based attraction, further targeted research is needed to determine whether this influence is best explained by retrieval-based processes, representational accounts with added complexity, or a combination of both (Yadav et al., 2023).

6.5.2 Restrictive modifiers modulate attraction

A second factor we believe contributed to the increased attraction effects observed in Experiment 2 is the communicative relevance of the modifier attractor nouns. In our first experiment, we argued that the attractor nouns were not central to the communicative task—that is, they did not help distinguish between the visible events participants were asked to describe. In the discussion of Experiment 1, we outlined both syntactic and experimental reasons why the communicative status of the attractor should influence its likelihood of interfering with agreement.

Syntactically, we showed that the non-restrictive modifiers differ from restrictive modifiers in their structural characteristics. Using binding as a diagnostic, we demonstrated that non-restrictive modifiers may attach at a different level or with an unaccessible inner structure (Lasnik & Uriagereka, 2022), resulting in modified syntactic distance between the attractor and the subject head as a function of restrictivity of the attractor. Both representational and cue-based retrieval accounts of agreement attraction predict that syntactic distance and nature modulates the strength of attraction (Avetisyan et al., 2020; Dillon et al., 2013; Eberhard et al., 2005). Additionally, recent comprehension research shows that non-restrictive modifiers induce attenuated attraction effects compared to restrictive ones (Kim & Xiang, 2024), and similar distinctions have been observed across other dependency types (Dillon et al., 2014, 2017).

Our results from Experiment 2 align with these findings. By embedding attractors in a task which made them more likely to be part of the at-issue content of the utterance, we observed stronger attraction effects. This supports the view that the syntactic and at-issue-related status of the attractor plays a critical role in determining its potential to interfere with agreement computation.

6.5.3 Plural Markedness Effect

Early studies on agreement attraction consistently found that attraction errors are more common when the attractor is plural, but not when it is singular—a phenomenon known as the plural markedness effect. This asymmetry has been well-documented in both production and comprehension studies (Bock et al., 2004; Bock & Miller, 1991; Eberhard, 1999; Eberhard et al., 2005; Vigliocco et al., 1995; Wagers et al., 2009). Recent work with picture description tasks, on the other hand, has shown that this asymmetry is not always present (Kandel & Phillips, 2022; Veenstra et al., 2014). Our results diverge from these recent studies using picture description tasks in that we did not observe a clear attraction effect in singular attractor conditions when the subject head was plural.

Interestingly, picture description studies—which arguably provide a more naturalistic production context—have at times failed to replicate this strong asymmetry, showing greater attraction from singular attractors in plural-head conditions than previously expected. One key difference between our study and prior picture-description-based attraction experiments is the use of lexical verbs in our materials, whereas previous studies typically used auxiliary verbs (e.g., to be) or a single nonce verb throughout. Notably, Kandel et al. (2022), in their meta-analysis, found that production experiments using the verb *to be* were more likely to exhibit attraction effects in singular attractor conditions and when the head was plural.

Taken together, these findings suggest that our results are consistent with the broader trend observed in picture description studies, where attraction from singular attractors is weaker or absent—particularly when lexical verbs are used. Our data also align with corpus findings: attraction errors are more frequently observed in conditions with plural attractors and singular heads, while singular attractors in plural-head contexts rarely induce errors (Bock & Miller, 1991; Pfau, 2009). Thus, while our findings differ in surface detail from some previous picture description studies, they

appear to follow a systematic pattern across tasks, materials, and spontaneous speech.

6.5.4 Pause Likelihood as a time-signal of attraction

Another key finding from our experiment was the effect of attractor number on pause likelihood. This timing-based measure was proposed by Kandel and Phillips (2022) as a potential signal of agreement attraction, reflecting the point in time at which agreement computation occurs. It refers to the presence of non-zero pauses—defined in our study as those longer than 50 milliseconds—between the second noun and the element that carries agreement, namely the auxiliary verb.

In Experiment 2, we found that pause likelihood was sensitive to number mismatch, even in conditions where no attraction errors were observed. For example, while attraction effects were only present when the subject head was singular, the effect of attractor number on pause likelihood was observed in both singular and plural head conditions. That said, the effect was more robust in singular head conditions and less stable in plural head conditions. The variability in the plural head context appeared to stem specifically from unaccusative verb constructions.

Despite some indications that attractor number may behave differently in unaccusative sentences, the overall pattern of pause likelihood was comparable across unaccusative and unergative structures. This suggests that the timing of agreement computation is not modulated by verb type.

Recall that our hypothesis predicted that if agreement is planned early—alongside the verb—we should observe differences in agreement-related measures between unaccusative and unergative verbs. However, we found no evidence in support of this hypothesis. Instead, our findings point to the conclusion that agreement is likely planned relatively late, closer to the point of articulation, and independently from the planning of the verb.

6.5.5 Severed morpho-syntax and morpho-phonology

One interesting result we observed was an interaction between verb type and number match in onset latencies. Sentences with plural subject heads, such as *The doctors by the pirate(s) are jumping*, exhibited longer initiation times compared to sentences with singular subject heads, like *The doctor by the pirate(s) is jumping*. Within the plural subject head condition, the number of the attractor did not affect onset latencies for unergative scenes involving verbs like *jumping* or *running*. However, in unaccusative sentences, the attractor number did influence initiation time: sentences with plural attractors—matching the subject in number—were initiated faster than those with singular attractors.

A version of this pattern extended to singular subject head conditions as well. Participants showed increased onset latencies when the subject and attractor mismatched in number, but again, this effect was only present in unaccusative sentences. These results suggest that number match between the subject and the attractor affects sentence initiation timing, but primarily within unaccusative structures.

The asymmetry is striking given that we did not observe verb-type-specific effects in our other attraction output-related measures. Our output measures—agreement errors (i.e., production of the wrong auxiliary) and pause likelihood—showed symmetric effects of attractor number across both unaccusative and unergative verbs. Thus, the underlying reason for this interaction observed here seems unique to onset latency.

This finding is also surprising in light of Momma and Ferreira's (2019) results regarding semantic interference from the second noun phrase. They showed that while unaccusative verbs were planned early along with the subject, the second noun (e.g., in the modifier phrase) was not planned at the same time. When participants were shown semantically related distractors for the second noun, there was no delay in onset latency, leading them to conclude that the second NP was not part of the initial planning scope.

Our observation that attractor number affected onset latency in unaccusative

sentences—but not in unergatives—is not easily reconcilable with this conclusion. If participants in our study were excluding the second NP from initial planning, as argued by Momma and Ferreira (2019), we would not expect number mismatch at the attractor to influence onset timing in unaccusatives. The fact that it does suggests that either the second NP was sometimes included in early planning, or that number mismatch has an early, perhaps due to conceptual number, effect that influences initiation time without full syntactic integration.

There are two possible interpretations of this finding. The first has recent support in the literature. In a recent paper, Roeser et al. (2024) argue that the syntactic scope of advance planning, particularly for conjoined noun phrases, is not fixed. They compared multiple statistical models to account for previous findings on early noun planning and found that models assuming two distributions with the same location parameter but differing tail parameters fit the data better. This approach challenges the view that observed mean differences necessarily reflect categorical differences in planning scope. Instead, they propose that differences in means may be driven by a subset of participants in particular conditions, rather than reflecting uniform early planning across trials.

One interpretation of their findings is that the inclusion of the second noun in the initial planning window may vary depending on experimental conditions or participant populations. Applied to our case, it is possible that unaccusative sentences encouraged more frequent or prolonged early planning, occasionally extending to the second noun—especially given prior findings of early verb and subject planning in unaccusatives (Momma & Ferreira, 2019). This could account for the early sensitivity to attractor number we observed in unaccusative conditions.

However, our results offer a challenge to this interpretation. We did not observe generally prolonged onset latencies for unaccusatives relative to unergatives. In fact, utterance onset times were similar across verb types. If unaccusative trials systematically involved the early planning of both the subject and the second noun, we

would expect to see longer initiation times compared to unergatives. The lack of this difference suggests that, while attractor number influenced onset latency in unaccusatives, this cannot be straightforwardly attributed to extended early planning of the second noun.

Let us now consider a second interpretation. Our initial assumption was that unaccusative verbs are planned early alongside the subject head, and that at that point in time, speakers have either no or only limited access to the second noun phrase, which is not yet fully planned. However, we may entertain the possibility that some elements related to the second NP are nonetheless available during early stages of planning—even if the second NP has not yet been lexically or syntactically encoded.

For instance, while participants may not have constructed a lemma representation for the second NP—something that appears necessary for triggering semantic interference—the conceptual features associated with the second NP may already be accessible. One such feature could be the multiplicity of the attractor noun. This idea is not new: Meyer and Bock (1999) proposed that conceptual features like number may influence the production of pronouns. More recently, this hypothesis has been tested empirically by Kandel et al. (2025)[13].

Under this view, we can assume that the conceptual multiplicity feature of the attractor may interfere with the planning of the unaccusative verb, even if the second NP is not fully planned. However, this interference appears to be strong enough to modulate onset latencies, but too subtle to manifest in our output-related measures. In both the proportion of attraction errors and the pause likelihood data, participant behavior was not modulated by verb type. This apparent mismatch between early timing measures and downstream production outcomes can be reconciled by assuming a two-step model of agreement computation.

---

[13] It is important to note that this interpretation diverges from the view of Roelofs (1992), who argued that the conceptual network comprises only content-word-related nodes and excludes syntactic-semantic features such as multiplicity or temporality.

In this model, the first step involves the encoding of the number feature—or diacritic—onto the verb, which occurs early, during verb retrieval. In unaccusative sentences, this diacritic assignment is hypothesized to be eager and initiated alongside verb access. The second step, however, involves the retrieval and production of the correct auxiliary verb form, such as *is* or *are*, which happens later, closer to articulation. The agreement attraction effects we observe in output measures, then, may not stem from errors in the early diacritic assignment process. Instead, they likely reflect competition or interference during the later retrieval of the auxiliary. In this framework, the early multiplicity-related interference seen in onset timing reflects the conceptual encoding of number, while the attraction errors and pause likelihood reflect a later stage of lexical retrieval. This view of attraction also aligns with how Schiller and Caramazza (2003) and Schriefers et al. (2002) thought about gender incongruency effects, namely that there might be a component that is based on the output of the lemma and that component might be responsible for the gender incongruency effects.

This understanding of agreement as a two-step process also helps illuminate the results from our first experiment. Recall that in Experiment 1, when the subject was singular and the attractor was plural—a mismatch condition—we observed effects of semantic relatedness late in production, particularly in the preverbal region. However, when the subject and the attractor shared the same number (i.e., no mismatch), we observed differences in onset latency for unaccusative sentences as a function of semantic relatedness.

If we adopt the view that agreement involves two components—early morpho-syntactic assignment and later lexical retrieval—we can more easily account for these findings from Experiment 1. In the mismatched conditions, the early assignment of number diacritics to the verb may have introduced processing demands that postponed or dampened the effects of lemma-level semantic interference, pushing them later in the production timeline. Conversely, in matched-number conditions, where

agreement encoding was likely more straightforward, semantic interference effects could surface earlier and be observed at the level of sentence initiation. This framework allows for a unified explanation of both timing and error-based measures across our two experiments.

6.5.6 Detecting Advance Planning Without Semantic Interference

In our first experiment, we included a semantic interference manipulation to test the advance planning of unaccusative verbs. This manipulation yielded weak but suggestive evidence in favor of early planning. Notably, our independent replication of Momma and Ferreira (2019) confirmed that unaccusative verbs are planned early, aligning with a growing body of work showing advance planning effects for certain verb types and syntactic dependencies. Based on this body of evidence, we believe there are strong a priori reasons to assume that unaccusative verbs are planned early in production. Consequently, we did not include a direct test of this hypothesis in our second experiment.

Nonetheless, it remains important to explore additional ways of verifying this assumption, even in the absence of an explicit manipulation. In our second experiment, onset latencies showed sensitivity to verb type, suggesting that some aspects of early planning for unaccusatives persisted. Crucially, this sensitivity was not uniform across all conditions—it emerged only in certain configurations (number mismatch), which suggests that early planning is not due to a general characteristics of the unaccusative pictures

A promising direction for further work is to assess verb name agreement using codability measures (Griffin, 2001; Konopka & Kuchinsky, 2015; Kuchinsky & Bock, 2010). For instance, Griffin (2001) examined the effect of codability—defined as the number of acceptable lexical alternatives for naming an object—on naming latencies. She found that lower codability (i.e., more naming alternatives) led to longer production times. Importantly, she linked codability to semantic interference: whereas

semantic interference tasks introduce competition experimentally through superimposed or auditory distractors, codability reflects naturally occurring lexical competition triggered by simply viewing the image.

Unfortunately, our current study lacks codability norms for the verbs we used. One way to address this in future research would be to conduct a free-production experiment with no prior habituation, in which participants are asked to describe the depicted actions without prompts. From this data, we could calculate codability scores for each verb and examine whether codability differentially modulates onset latencies for unaccusative and unergative verbs. This approach would allow us to test for early verb planning effects without relying on artificial semantic interference, offering a more naturalistic window into the planning dynamics of verb retrieval.

As a preliminary test, we calculated a proxy for codability using the erroneous utterances from Experiment 2. Codability was quantified using Shannon entropy, computed from the proportion of correct and incorrect verb responses for each item, following the standard formulation (Shannon, 1948):

$$H = -\sum_{i=1}^{n} p_i \log_2 p_i$$

This approach parallels prior work in production studies where entropy is used to estimate codability from naming variability (Snodgrass & Vanderwart, 1980). Verbs with more varied responses across participants yield higher entropy values, suggesting greater lexical uncertainty or competition.

We fit a preliminary model to our onset latency data using this entropy-based codability measure as a predictor, including an interaction term with verb type. While the model revealed strong evidence for a main positive effect of codability ($\hat{\beta} = 31.44$; $CI = [-3.98; 67.06]$; $P(\beta > 0) = .96$), we did not find strong evidence for its interaction with verb type ($\hat{\beta} = 28.52$; $CI = [-39.93; 97.46]$; $P(\beta > 0) = .79$). However, in more complex models, we observed a weak effect of a three-way interaction between

codability, verb type, and attractor number ($\hat{\beta} = 98.19$; $CI = [-86.92; 282.38]$; $P(\beta > 0) = .85$).

Despite these intriguing trends, we urge caution in interpreting these results. First, our codability estimates are derived solely from erroneous utterances in Experiment 2, which may not reflect participants' true uncertainty about the verb labels. Second, we do not know the underlying reasons for participants' use of alternative verbs—whether they reflect lexical uncertainty, misperception, or task-related confusion. For these reasons, we do not take this analysis as strong evidence, but rather as an exploratory step toward developing more principled codability-based measures of verb planning in future work.

6.6 Conclusion

Our central research question was whether attraction-related effects are modulated by verb type. We hypothesized that if unaccusative verbs are retrieved early in production, then the number diacritic could also be assigned early—potentially at the point of verb retrieval. This would suggest that number marking on unaccusative and unergative verbs occurs at different points in the planning timeline. Specifically, we expected that in unaccusative conditions, number would be assigned when only the subject head is planned, whereas in unergative conditions, number assignment would occur after both noun phrases are entertained. This difference in the timing of NP activation was predicted to lead to differences in pause likelihood and attraction error rates across verb types.

In Experiment 1, although we observed symmetrical effects of attractor number in both unaccusative and unergative conditions for attraction error rates and pause likelihoods, the overall rate of attraction errors was too low to confidently interpret timing-related effects. In the current experiment, we successfully elicited attraction errors at a magnitude comparable to previous studies in the literature, allowing us to better evaluate our predictions.

With respect to our core hypothesis, we found no evidence that attraction effects were modulated by verb type. Both unaccusative and unergative verbs exhibited similar attraction error rates and pause likelihoods. This symmetry supports an interpretation in which attraction arises late in production—during the retrieval of the auxiliary—rather than during early syntactic encoding.

However, we did observe a difference in onset latencies between unaccusative and unergative verbs as a function of attractor number. This modulation was present only in unaccusative conditions. Combined with the absence of verb-type effects in the attraction-related measures, we interpret this pattern as evidence that number diacritics are assigned early—at the point of verb retrieval—in unaccusative structures. In contrast, the retrieval of the auxiliary, which is where attraction errors are likely to arise, occurs later in production. These findings point toward a two-stage model of agreement in which morpho-syntactic marking and auxiliary retrieval are temporally decoupled.

7 General Discussion

7.1 Main Question

The central question of this study is: when do speakers plan verbal agreement during sentence production? To investigate this question behaviorally, we turned to the phenomenon of agreement attraction. This phenomenon is robustly observed in production: speakers systematically produce verbs or auxiliary verbs with erroneous number agreement when another noun phrase with mismatching number appears near the subject head Kandel & Phillips (2022). For example, when describing a scene with a single red bird and multiple green snakes, participants systematically produce errors like **The bird next to the green snakes are red* (Nozari & Omaki, 2022).

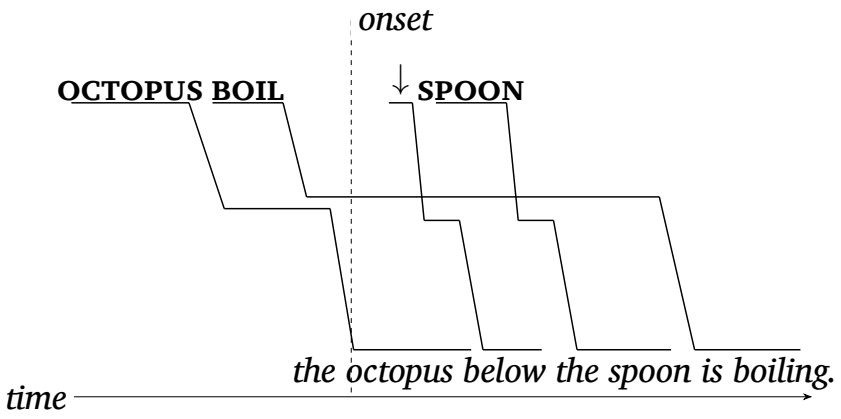
However, a standard picture description task is insufficient to address our central question. To systematically manipulate sentence planning processes, we exploited known differences in planning between verb types. Previous work by Momma and Ferreira (2019) and Momma et al. (2016) has shown that speakers can deviate from strictly incremental sentence planning routines. Across multiple experiments, constructions, and languages, Momma and colleagues have demonstrated that syntactic structure can license or even require advance planning of certain elements.

One particularly influential finding concerns the planning of intransitive verbs. Momma and Ferreira (2019), using a series of extended Picture–Word Interference (ePWI) experiments, found that unaccusative verbs are affected by semantically related distractors even before the sentence is initiated. This interference suggests that unaccusative verbs are planned early, alongside the subject noun phrase. This key finding is visualized in Figure 59 and forms the theoretical basis for our experimental design.

Building on these findings, we aimed to test whether the planning of agreement is tied to the planning of the verb itself. Specifically, we asked whether the timing of verb planning also includes the planning of agreement features, such as number

Figure 59

Planning procedure of the sentence with an unaccusative verb, such as 'The octopus under the spoons is boiling,' following Momma and Ferreira's (2019) findings.

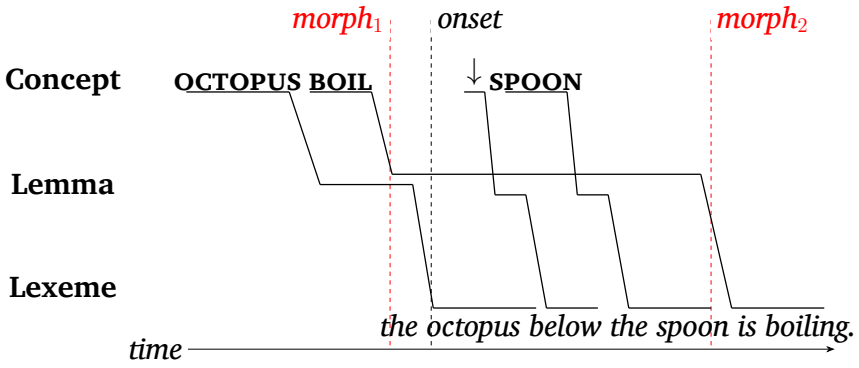


marking. One possibility is that agreement is planned at the same time as the verb—an early process shown as *morph*$_1$ in Figure 60. In this scenario, the number diacritic would be assigned during verb retrieval, before any verbal morphology is produced.

Alternatively, agreement might be planned later, at the moment the agreement-bearing element—such as an auxiliary verb—is produced. This later planning process is illustrated as *morph*$_2$ in Figure 60. The distinction between these two possibilities—eager versus lazy agreement—is central to our investigation and motivates the design of our experiments.

Figure 60

Possible times for the morphing process given an unaccusative event.

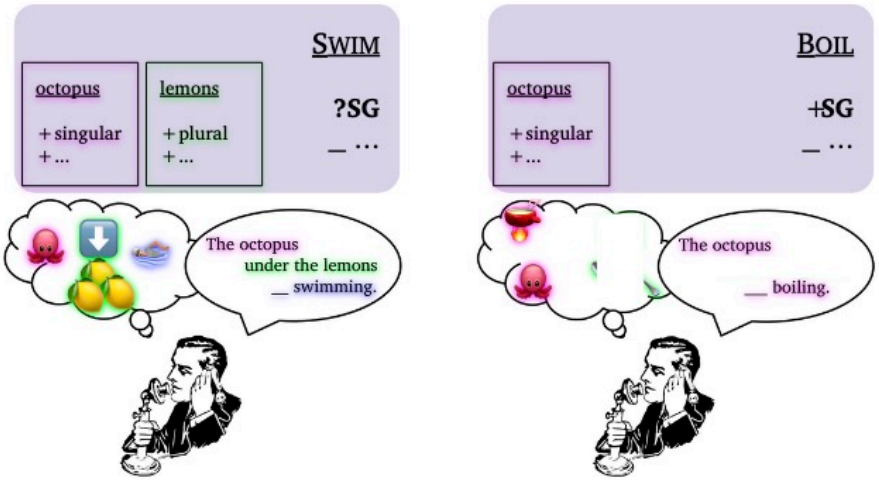


The early planning of agreement would align with previous theoretical accounts

of morphological encoding, particularly the model proposed by Schiller and Caramazza (2003). In their framework, morphological information is accessed directly and without competition, suggesting that features like number can be encoded early in production. If this account holds, and agreement planning occurs early, we would expect to see agreement attraction effects in unergative verbs—where the verb is planned later—but not in unaccusative verbs, where the verb is planned early.

This prediction follows from the assumption that, during the planning of an unaccusative verb, only the subject head is available in the speaker's mental representation. The modifier noun, which could serve as a potential attractor, is not yet active in the planning process. This configuration is schematized in Figure 61, where early verb and agreement planning occur in a syntactically minimal scope. As a result, no agreement attraction effect should be observed in unaccusative conditions if agreement is planned eagerly with the verb. On the other hand, both nouns should be available at the point of planning of the verb, thus the uncertainty for the number feature of the verb *swim* should increase, signalled by a question mark in the figure.

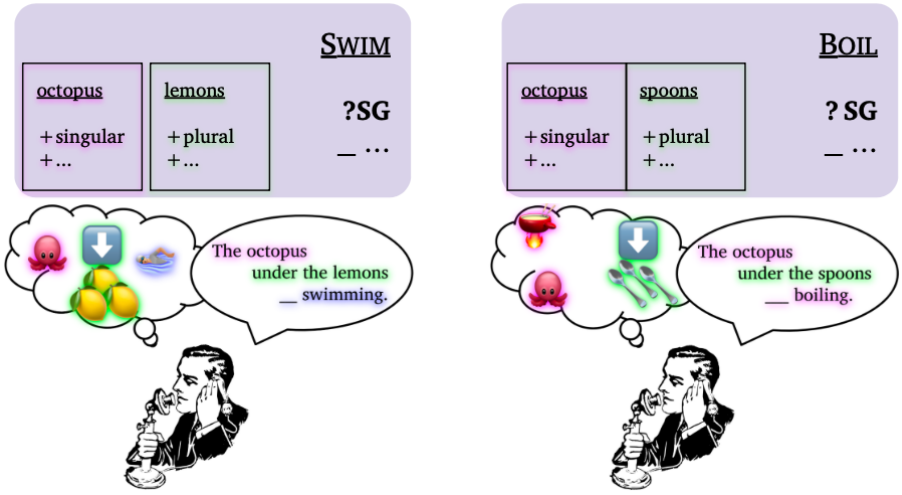
Figure 61
Hypothesized access to nominal information in early agreement planning. Since the agreement is planned early, Only the auxiliaries of unergatives would be planned independent of the verb planning.



On the other hand, if agreement planning is driven by the immediate needs of the utterance and follows a just-in-time planning strategy, we would expect to observe agreement attraction effects in both unaccusative and unergative conditions. Under this view, agreement is not planned alongside the verb, but rather at the point when the agreement-marked element—such as an auxiliary—is about to be produced. Since this occurs after both noun phrases have already been planned in either type of intransitive sentence, both the head and the attractor noun are active in the production system. This broader activation window increases the potential for interference and would make attraction effects possible regardless of verb type.

Figure 62

Hypothesized access to nominal information in late agreement planning. Since the agreement is planned late, both nouns would be planned independent of the verb planning.



7.2 Summary of Experimental Findings

In our experiments, we asked whether the type of verb (unaccusative vs. unergative) influences agreement attraction—that is, the erroneous production of agreement-carrying auxiliary verbs. Across both experiments, we found that participants' attraction error rates were not modulated by verb type. That is, whether the sentence contained an unaccusative or unergative verb did not affect the likelihood of producing an incorrect auxiliary. We observed a similar pattern in pause likelihood,

another measure linked to agreement computation, which also remained unaffected by verb type.

However, our verb-planning-related timing measures told a different story. In Experiment 1, which incorporated a semantic interference task, we found that number mismatch between the subject and the modifier noun modulated where semantic interference effects emerged. In conditions where both the subject head and the modifier noun were singular, we observed early semantic interference effects, consistent with the findings of Momma and Ferreira (2019). This early effect was restricted to unaccusative sentences. In contrast, when the subject was singular and the modifier was plural, the semantic interference effect appeared later in production, specifically in the preverbal region—again, only for unaccusative verbs.

In Experiment 2, a simpler sentence production task without semantic interference, we found that number mismatch between the subject and the modifier noun influenced onset latencies. When the subject head was singular, sentences with plural modifiers took longer to initiate than those with singular modifiers. Conversely, when the subject head was plural, sentences with plural modifiers were initiated more quickly than those with singular modifiers. As in Experiment 1, this interaction between number mismatch and initiation timing was present only in unaccusative sentences.

These findings suggest a dissociation between timing measures and overt agreement errors, and point to verb type affecting early planning processes—especially in the context of unaccusative verbs—even when error-based measures remain unaffected.

7.3 Synthesis

Much of our pre-experimental reasoning about agreement focused on the idea that the timing of morphosyntactic feature selection would drive attraction effects. For instance, we assumed that if agreement-relevant features—such as number—are assigned early at the lemma level or on chunks, before the second noun is planned, then

the morpho-phonological form of the auxiliary should reflect only that initial selection.

However, we found effects both in early and late measures. While the early effects were selective as a function of verb type, late ones were not. This suggest that, even if there is an early morphosyntactic assignment, it alone is not sufficient to explain the observed attraction effects. This indicates that other processes—possibly during later stages of production—play a key role in shaping agreement behavior.

The model we propose accounts for three key empirical patterns observed across our experiments:

- **Unaccusative-specific number match effect on the onset**: We observed that number matching between the subject and modifier noun influenced onset latencies, but only in unaccusative sentences.
- **Parallel attraction and timing effects**: Agreement attraction errors and pause likelihood showed aligned patterns, suggesting that both reflect the same underlying computation process.
- **Feature-based modulation of attraction**: Attraction was modulated by the degree of feature overlap between the subject head and its modifier, particularly when the modifier was a plausible agreement controller.

We adopt a cue-based retrieval model of agreement, in which the auxiliary is retrieved based on the features of the subject head and the predictive cues available at the point of verb retrieval (Badecker & Lewis, 2007; Lewis & Vasishth, 2005). In this framework:

- The locus of agreement attraction is not the representation or encoding of agreement-related features.
- Rather, attraction arises from erroneous retrieval—specifically, retrieving an incorrect agreement controller due to partial cue overlap.

This account aligns with our findings. Although we observed some evidence of early planning of agreement-relevant features—particularly in unaccusative conditions—attraction effects were not modulated by verb type. This suggests that while morpho-syntactic information that is marked on lemmas or in chunks may be available early, the agreement computation that leads to attraction errors likely occurs later, during the retrieval of the auxiliary.

In our model, at the point of picture presentation, participants plan only the chunks that are immediately necessary for initiating the utterance. For unaccusative verbs, following Momma and Ferreira (2019), participants plan both the subject head and the verb before the sentence onset. As part of this early planning, diacritics such as category, aspect, and tense are also specified for the verb. Crucially, we argue that number diacritics are also planned at this stage. However, this number assignment is not automatic, as proposed by Schiller and Caramazza (2003). Instead, the other possible number information within the visual information may affect the early assignment (Schriefers et al., 2002). Even though the second noun (e.g., in a modifier PP) is not fully planned at this stage, its number is readily visible and can influence the planned features of the verb. For instance, in a target sentence like: *The doctor by the wizards is boiling*, participants may construct a propositional representation in which the number of the second noun is encoded, even before its full lexical identity is retrieved. This is plausible given the repetitive nature of the task: participants are exposed to similar sentence structures and constrained to produce sentences in a uniform format. Moreover, the visual similarity across scenes makes extraction of number information relatively effortless.

$$\left[\begin{array}{l|l|l} \textit{doctor} & & \\ Cat: & DP & DomCat: \quad TP \\ \textsc{nom}: & + & \textsc{def}: \quad + \\ Num: & \textsc{sg} & \end{array}\right] \quad \left[\begin{array}{l|l|l} x & & \\ Cat: & DP & DomCat: \quad PP \\ \textsc{nom}: & + & \textsc{def}: \quad + \\ Num: & \textsc{pl} & \end{array}\right] \quad \left[\begin{array}{l|l|l} \textit{boil} & & \\ Cat: & T & Dom \\ Tense: & \textsc{prs} & Aspe \\ Num: & \textsc{sg} & Pers \end{array}\right.$$

One of the key insights that gave rise to cue-based retrieval—and contributed to its prominence—is its prediction about cue similarity and timing. Lewis and Vasishth (2005) and Smith and Vasishth (2020) explains how feature mismatch between the target and the attractor influences both the timing of syntactic dependency formation and the likelihood of interference.

This prediction has two direct consequences for our results. First, in both experiments, utterance onset times were modulated by whether the subject and modifier nouns matched or mismatched in number. Second, we only observed attraction effects when the attractor was similar to the subject in its potential to serve as an agreement controller—a point we return to in the discussion of how attraction arises in this model.

In the case of unergatives, since the verb is planned later, we do not see any effects of shared features at sentence onset. While it is not entirely clear why number mismatch does not yield a timing effect in these structures, this absence is consistent with previous findings. Notably, Momma and Ferreira (2019) also reported difficulty in detecting the time of semantic interference effects for unergative verbs. We believe that our failure to observe a localized timing effect from number mismatch between the attractor and the head in unergatives follows from the same underlying principles.

The question, then, is why we observe attraction effects in unaccusatives at all, if the agreement diacritic is set early. We argue that this pattern is fully compatible with cue-based retrieval accounts of agreement attraction. Except for hybrid models such as those proposed by Yadav et al. (2023), most cue-based approaches assume that

agreement errors do not arise from faulty representations, but from retrieval failures. These models typically posit that sentence elements are correctly encoded. With these encoding, speakers posit or predict the form of the agreement later on (akin to Wagers et al., 2009). At the time of retrieval of the agreement bearing verb, the wrong noun might be selected as the agreement controller, which would occasionally force participants to change their predicted agreement form.

This misretrieval is driven by feature overlap and noise in the system. When another noun—such as a plural attractor—shares key features with the subject head, the likelihood of erroneously retrieving it as the controller increases, raising the probability of producing a mismatching auxiliary. Our finding that the magnitude of attraction varies as a function of shared features between the subject and the attractor, such as subjecthood or discourse relevance, supports this view. As the overlap in features increases, so does the risk of misretrieval—and thus the strength of attraction effects.

This explanation also aligns with findings from pause likelihood measures, present both in our experiments and in Kandel and Phillips (2022). Pause likelihood reflects the time required to retrieve the controller noun and select the appropriate verb form. When the attractor and head noun share more agreement-relevant features, retrieval becomes more difficult, increasing pause likelihood. This follows naturally from the fan effect discussed in Chapter 4: increased similarity among potential retrieval candidates leads to greater interference and processing cost. A surprising fact is that this pause likelihood as a function of distractor number is still seem to be visible in the experiment where there is very small magnitutde of attraction effects.

One natural question is why speakers would need to retrieve the agreement controller at all if the number diacritic has already been set earlier in the sentence. To address this, we turn to the discussion by Eberhard et al. (2005). They explain that the grammatical number of the verb is not solely determined by the morphological features of the subject head. Instead, they argue that lexical, notional, and distributivity related

information-possibly introduced by other nouns in the sentence—can all influence the final agreement outcome.

This account implies that English speakers may redundantly re-evaluate or confirm the agreement controller, or other relevant cues, at the point of auxiliary production. This leads to two predictions. First, in a language where verb agreement is strictly determined by the immediate morphosyntactic features of the subject head, we would expect no attraction effects in unaccusatives—assuming the number diacritic is set early and no additional retrieval or verification is performed. Second, if we replicate this experiment with additional manipulations targeting features like distributivity—whether introduced by the attractor, the verb, or the visual scene—we should see increased attraction effects or pause likelihood at the preverbal region. Crucially, these effects should not be observable at utterance onset.

7.4 Conclusion

In this study, we used agreement attraction as a diagnostic tool to investigate the timing of morphosyntactic planning during sentence production. By contrasting unaccusative and unergative verbs, we aimed to test whether agreement is planned early—alongside the verb—or later, closer to articulation. Across two experiments, we found no verb-type differences in attraction error rates or pause likelihood, suggesting that attraction effects are not tightly coupled to when the verb itself is planned.

However, onset latencies did reveal signs of early number encoding for unaccusative verbs, consistent with prior work suggesting that these verbs are often planned alongside their subjects. Taken together, these findings support a two-stage view of agreement computation: number diacritics may be assigned early, but the use of these diacritics—particularly in retrieving the correct auxiliary form—occurs later and appears to be the locus of attraction effects.

Finally, we found that attraction was modulated by structural and discourse-level properties of the attractor, including its subjecthood status and whether it appeared in a

restrictive or non-restrictive modifier. These findings align more naturally with cue-based retrieval models, where attraction arises not from representational encoding errors but from interference during retrieval.

Abbreviations

ACC	Accusative	NOM	Nominative
C	Common gender	PL	Plural
DET	Determiner	PROG	Progressive
F	Feminine	PRS	Present
M	Masculine	SG	Singular
N	Neuter		

References

Allaire, J., Xie, Y., Dervieux, C., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., Chang, W., & Iannone, R. (2024). *rmarkdown: Dynamic documents for r*. https://github.com/rstudio/rmarkdown

Anderson, J. R. (1996). ACT: A simple theory of complex cognition. *American Psychologist*, *51*(4), 355.

Avetisyan, S., Lago, S., & Vasishth, S. (2020). Does case marking affect agreement attraction in comprehension? *Journal of Memory and Language*, *112*, 104087.

Badecker, W., & Kuminiak, F. (2007). Morphology, agreement and working memory retrieval in sentence production: Evidence from gender and case in Slovak. *Journal of Memory and Language*, *56*(1), 65–85. https://doi.org/10.1016/j.jml.2006.08.004

Badecker, W., & Lewis, R. L. (2007). *Agreement and sentence formulation: The role of working memory (retrievals) in language production*.

Badecker, W., Miozzo, M., & Zanuttini, R. (1995). The two-stage model of lexical retrieval: Evidence from a case of anomia with selective preservation of grammatical gender. *Cognition*, *57*(2), 193–216.

Bhatia, S., & Dillon, B. (2022). Processing agreement in hindi: When agreement feeds attraction. *Journal of Memory and Language*, *125*, 104322.

Bock, K. (1986). Syntactic persistence in language production. *Cognitive Psychology*, *18*(3), 355–387.

Bock, K., & Cutting, J. C. (1992). Regulating mental energy: Performance units in language production. *Journal of Memory and Language*, *31*(1), 99–127. https://doi.org/10.1016/0749-596X(92)90007-K

Bock, K., & Eberhard, K. M. (1993). Meaning, sound and syntax in English number agreement. *Language and Cognitive Processes*, *8*(1), 57–99. https://doi.org/10.1080/01690969308406949

Bock, K., Eberhard, K. M., & Cutting, J. C. (2004). Producing number agreement: How pronouns equal verbs. *Journal of Memory and Language*, *51*(2), 251–278.

Bock, K., & Ferreira, V. (2014). Syntactically speaking. *The Oxford Handbook of Language Production*, 21–46.

Bock, K., & Miller, C. A. (1991). Broken agreement. *Cognitive Psychology*, *23*(1), 45–93.

Bürki, A., Elbuy, S., Madec, S., & Vasishth, S. (2020). What did we learn from forty years of research on semantic interference? A bayesian meta-analysis. *Journal of Memory and Language*, *114*, 104125.

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

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

Bürkner, P.-C. (2018a). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*(1), 395–411. https://doi.org/10.32614/RJ-2018-017

Bürkner, P.-C. (2018b). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*(1), 395–411. https://doi.org/10.32614/RJ-2018-017

Bürkner, P.-C. (2021). Bayesian item response modeling in R with brms and Stan.

Journal of Statistical Software, *100*(5), 1–54. https://doi.org/10.18637/jss.v100.i05
Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*, 1–32.
Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2024). *shiny: Web application framework for r*. https://CRAN.R-project.org/package=shiny
Chkhetiani, L., Vanzo, A., Khare, Y., Peyash, T., Sklyar, I., Liang, M., Botros, R., Bousbib, R., Etefy, A., Ghahremani, P., Oexle, G., Trueba, J. L., Ferreira, W. P., Gotthold, B., Bahadoori, S., Chiang, M., Mitov, A., Fakhan, E., Yoshioka, T., … Martin, S. R. (n.d.). *Universal-2: Advancing multilingual speech understanding at scale*. AssemblyAI. Retrieved October 27, 2024, from https://www.assemblyai.com/research/universal-2
Cousineau, D. (2005). 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. https://doi.org/10.20982/tqmp.01.1.p042
Damian, M. F., & Martin, R. C. (1999). Semantic and phonological codes interact in single word production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(2), 345.
Dillon, B., Clifton Jr, C., & Frazier, L. (2014). Pushed aside: Parentheticals, memory and processing. *Language, Cognition and Neuroscience*, *29*(4), 483–498.
Dillon, B., Clifton Jr, C., Sloggett, S., & Frazier, L. (2017). Appositives and their aftermath: Interference depends on at-issue vs. Not-at-issue status. *Journal of Memory and Language*, *96*, 93–109.
Dillon, B., & Keshev, M. (n.d.). *Syntactic dependency formation in sentence processing: A comparative perspective*.
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.
Drenhaus, H., Saddy, D., & Frisch, S. (2005). Processing negative polarity items: When negation comes through the backdoor. *Linguistic Evidence: Empirical, Theoretical, and Computational Perspectives*, 145–165.
Drummond, A. (2013). Ibex farm.
Eberhard, K. M. (1999). The accessibility of conceptual number to the processes of subject–verb agreement in English. *Journal of Memory and Language*, *41*(4), 560–578. https://doi.org/10.1006/jmla.1999.2662
Eberhard, K. M., Cutting, J. C., & Bock, K. (2005). Making syntax of sense: Number agreement in sentence production. *Psychological Review*, *112*(3), 531.
Ferreira, V., Morgan, A., & Slevc, L. R. (2007). *Grammatical encoding*. The Oxford Handbook of Psycholinguistics/Oxford University Press.
Franck, J., Lassi, G., Frauenfelder, U. H., & Rizzi, L. (2006). Agreement and movement: A syntactic analysis of attraction. *Cognition*, *101*(1), 173–216.
Frank, R. (2004). *Phrase structure composition and syntactic dependencies* (Vol. 38). Mit Press.
Gabry, J., Češnovar, R., Johnson, A., & Bronder, S. (2024). *cmdstanr: R interface to*

"CmdStan". https://github.com/stan-dev/cmdstanr

Ghodsi, M., Liu, X., Apfel, J., Cabrera, R., & Weinstein, E. (2020). Rnn-transducer with stateless prediction network. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7049–7053. https://doi.org/10.1109/ICASSP40776.2020.9054419

Gillespie, M., & Pealmutter, N. J. (2011). Effects of semantic integration and advance planning on grammatical encoding in sentence production. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *33*.

Griffin, Z. M. (2001). Gaze durations during speech reflect word selection and phonological encoding. *Cognition*, *82*(1), B1–B14. https://doi.org/https://doi.org/10.1016/S0010-0277(01)00138-X

Griffin, Z. M., & Ferreira, V. (2006). Properties of spoken language production. In *Handbook of psycholinguistics* (pp. 21–59). Elsevier.

Gussow, A. E., & MacDonald, M. C. (2023). Utterance planning under message uncertainty: Evidence from a novel picture-naming paradigm. *Cognitive, Affective, & Behavioral Neuroscience*, *23*(3), 957–972.

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.

Hope, R. M. (2022). *Rmisc: Ryan miscellaneous*. https://CRAN.R-project.org/package=Rmisc

Humphreys, K. R., & Bock, K. (2005). Notional number agreement in english. *Psychonomic Bulletin & Review*, *12*, 689–695.

Hwang, H., & Kaiser, E. (2014). The role of the verb in grammatical function assignment in english and korean. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(5), 1363.

Jäger, L. A., Engelmann, F., & Vasishth, S. (2015). Retrieval interference in reflexive processing: Experimental evidence from mandarin, and computational modeling. *Frontiers in Psychology*, *6*, 617.

Jescheniak, J. D., & Schriefers, H. (2001). Priming effects from phonologically related distractors in picture–word interference. *The Quarterly Journal of Experimental Psychology A*, *54*(2), 371–382. https://doi.org/10.1080/02724980042000273

Jescheniak, J. D., Schriefers, H., & Lemhöfer, K. (2014). Selection of freestanding and bound gender-marking morphemes in speech production: A review. *Language, Cognition and Neuroscience*, *29*(6), 684–694. https://doi.org/10.1080/01690965.2012.654645

Joshi, A. K., Levy, L. S., & Takahashi, M. (1975). Tree adjunct grammars. *Journal of Computer and System Sciences*, *10*(1), 136–163.

Kandel, M., Pañeda, C., Bahmanian, N., Bruera, M. M., Phillips, C., & Lago, S. (2025). Number and grammatical gender attraction in spanish pronouns: Evidence for a syntactic route to their features. *Journal of Cognition*, *8*(1), 10.

Kandel, M., & Phillips, C. (2022). Number attraction in verb and anaphor production. *Journal of Memory and Language*, *127*, 104370.

Kandel, M., Wyatt, C. R., & Phillips, C. (2022). Agreement attraction error and timing profiles in continuous speech. *Glossa Psycholinguistics*, *1*(1).

Kempen, G., & Hoenkamp, E. (1987). An incremental procedural grammar for sentence formulation. *Cognitive Science*, *11*(2), 201–258.
Kempen, G., & Huijbers, P. (1983). The lexicalization process in sentence production and naming: Indirect election of words. *Cognition*, *14*(2), 185–209.
Kim, S. J., & Xiang, M. (2024). Incremental discourse-update constrains number agreement attraction effect. *Cognitive Science*, *48*(9), e13497.
Konopka, A. E., & Kuchinsky, S. E. (2015). How message similarity shapes the timecourse of sentence formulation. *Journal of Memory and Language*, *84*, 1–23.
Kuchinsky, S., & Bock, K. (2010). *Paper presented at the 23rd meeting of the CUNY human sentence processing conference*.
Kush, D., & Phillips, C. (2014). Local anaphor licensing in an SOV language: Implications for retrieval strategies. *Frontiers in Psychology*, *5*, 1252.
Lago, S., Gračanin–Yuksek, M., Şafak, D. F., Demir, O., Kırkıcı, B., & Felser, C. (2019). Straight from the horse's mouth: Agreement attraction effects with Turkish possessors [Journal Article]. *Linguistic Approaches to Bilingualism*, *9*(3), 398–426. https://doi.org/10.1075/lab.17019.lag
Lago, S., Shalom, D. E., Sigman, M., Lau, E. F., & Phillips, C. (2015). Agreement attraction in Spanish comprehension. *Journal of Memory and Language*, *82*, 133–149. https://doi.org/10.1016/j.jml.2015.02.002
Lasnik, H., & Uriagereka, J. (2022). *Structure: Concepts, consequences, interactions*. MIT Press.
Levelt, W. J. (1989). *Speaking: From intention to articulation*. MIT press.
Levelt, W. J., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*(1), 1–38.
Lewis, R. L., & Vasishth, S. (2005). An Activation–Based Model of Sentence Processing as Skilled Memory Retrieval. *Cognitive Science*, *29*(3), 375–419. https://doi.org/10.1207/s15516709cog0000/_25
Logačev, P., & Vasishth, S. (2011). Case matching and conflicting bindings interference. In *Case, word order and prominence: Interacting cues in language production and comprehension* (pp. 187–216). Springer.
Makowski, D., Ben-Shachar, M. S., & Lüdecke, D. (2019). bayestestR: Describing effects and their uncertainty, existence and significance within the bayesian framework. *Journal of Open Source Software*, *4*(40), 1541. https://doi.org/10.21105/joss.01541
McAuliffe, M., & Sonderegger, M. (2023). *English (US) MFA G2P model v3.0.0*.
McAuliffe, M., & Sonderegger, M. (2024a). *English MFA acoustic model v3.1.0*.
McAuliffe, M., & Sonderegger, M. (2024b). *English MFA dictionary v3.1.0*.
Meyer, A. S. (1996). Lexical access in phrase and sentence production: Results from picture–word interference experiments. *Journal of Memory and Language*, *35*(4), 477–496.
Meyer, A. S., & Bock, K. (1999). Representations and processes in the production of pronouns: Some perspectives from dutch. *Journal of Memory and Language*, *41*(2), 281–301.
Meyer, A. S., & Schriefers, H. (1991). Phonological facilitation in picture-word interference experiments: Effects of stimulus onset asynchrony and types of interfering stimuli. *Journal of Experimental Psychology: Learning, Memory, and*

Cognition, *17*(6), 1146.
Miozzo, M., & Caramazza, A. (1997a). On knowing the auxiliary of a verb that cannot be named: Evidence for the independence of grammatical and phonological aspects of lexical knowledge. *Journal of Cognitive Neuroscience*, *9*(1), 160–166.
Miozzo, M., & Caramazza, A. (1997b). Retrieval of lexical–syntactic features in tip-of-the tongue states. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(6), 1410.
Momma, S., & Ferreira, V. (2019). Beyond linear order: The role of argument structure in speaking. *Cognitive Psychology*, *114*, 101228.
Momma, S., Slevc, L. R., & Phillips, C. (2016). The timing of verb selection in japanese sentence production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(5), 813.
Momma, S., & Yoshida, M. (2023). Planning multiple dependencies in sentence production. *Language, Cognition and Neuroscience*, *38*(9), 1183–1213.
Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Reason*, *4*(2), 61–64. https://doi.org/10.20982/tqmp.04.2.p061
Nozari, N., & Omaki, A. (2022). *An investigation of the dependency of subject-verb agreement on inhibitory control processes in sentence production.*
Parker, D., & Phillips, C. (2017). Reflexive attraction in comprehension is selective. *Journal of Memory and Language*, *94*, 272–290.
Pedersen, T. L., Ooms, J., & Govett, D. (2025). *systemfonts: System native font finding.* https://CRAN.R-project.org/package=systemfonts
Perlmutter, D. M. (1978). Impersonal passives and the unaccusative hypothesis. *Annual Meeting of the Berkeley Linguistics Society*, 157–190.
Perlmutter, D. M., & Postal, P. (1984). The inadequacy of some monostratal theories of passive. *Studies in Relational Grammar*, *2*, 3–37.
Pfau, R. (2009). *Grammar as processor*. John Benjamins Publishing Company. http://digital.casalini.it/9789027289636
Pickering, M. J., & Branigan, H. P. (1998). The representation of verbs: Evidence from syntactic priming in language production. *Journal of Memory and Language*, *39*(4), 633–651.
R Core Team. (2024a). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. https://www.R-project.org/
R Core Team. (2024b). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. https://www.R-project.org/
Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. *International Conference on Machine Learning*, 8748–8763.
Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108. https://doi.org/10.1037/0033-295X.85.2.59
Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, *9*(5), 347–356.
Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences*, *20*(4), 260–281.

Ristic, B., Molinaro, N., & Mancini, S. (2016). Agreement attraction in Serbian: Decomposing markedness [Journal Article]. *The Mental Lexicon*, *11*(2), 242–276. https://doi.org/10.1075/ml.11.2.04ris

Roelofs, A. (1992). A spreading-activation theory of lemma retrieval in speaking. *Cognition*, *42*(1-3), 107–142.

Roelofs, A. (2001). Set size and repetition matter: Comment on caramazza and costa (2000). *Cognition*, *80*(3), 283–290.

Roeser, J., Torrance, M., Andrews, M., & Baguley, T. (2024). *No default syntactic scope for advance planning in sentence production: Evidence from finite mixture models*. OSF.

Rosinski, R. R., Golinkoff, R. M., & Kukish, K. S. (1975). Automatic semantic processing in a picture-word interference task. *Child Development*, 247–253.

Sartori, G., Job, R., & Coltheart, M. (1992). The organization of object knowledge: Evidence from neuropsychology. In D. E. Meyer & S. Kornblum (Eds.), *Attention and performance, XN*. Lawrence Erlbaum.

Sauppe, S. (2017). Word order and voice influence the timing of verb planning in german sentence production. *Frontiers in Psychology*, *8*, 1648.

Schiller, N. O., & Caramazza, A. (2003). Grammatical feature selection in noun phrase production: Evidence from German and Dutch. *Journal of Memory and Language*, *48*(1), 169–194. https://doi.org/10.1016/S0749-596X(02)00508-9

Schlueter, Z., Parker, D., & Lau, E. (2019). Error-driven retrieval in agreement attraction rarely leads to misinterpretation. *Frontiers in Psychology*, *10*, 1002.

Schnur, T. T. (2011). Phonological planning during sentence production: Beyond the verb. *Frontiers in Psychology*, *2*, 319.

Schriefers, H. (1993). Syntactic processes in the production of noun phrases. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(4), 841–850. https://doi.org/10.1037/0278-7393.19.4.841

Schriefers, H., Jescheniak, J. D., & Hantsch, A. (2002). Determiner selection in noun phrase production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(5), 941–950. https://doi.org/10.1037/0278-7393.28.5.941

Schriefers, H., Meyer, A. S., & Levelt, W. J. (1990). Exploring the time course of lexical access in language production: Picture-word interference studies. *Journal of Memory and Language*, *29*(1), 86–102.

Schriefers, H., Teruel, E., & Meinshausen, R.-M. (1998). Producing simple sentences: Results from picture–word interference experiments. *Journal of Memory and Language*, *39*(4), 609–632.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Slioussar, N. (2018). Forms and features: The role of syncretism in number agreement attraction. *Journal of Memory and Language*, *101*, 51–63.

Slowikowski, K. (2024). *ggrepel: Automatically position non-overlapping text labels with "ggplot2"*. https://CRAN.R-project.org/package=ggrepel

Smith, G., & Vasishth, S. (2020). A Principled Approach to Feature Selection in Models of Sentence Processing. *Cognitive Science*, *44*(12), e12918. https://doi.org/10.1111/cogs.12918

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms

for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(2), 174.
Solomon, E. S., & Pearlmutter, N. J. (2004). Semantic integration and syntactic planning in language production. *Cognitive Psychology*, *49*(1), 1–46.
Stan Development Team. (n.d.). *RStan: The R interface to Stan*. https://mc-stan.org/
Stan Development Team. (2024). *Stan modeling language users guide and reference manual, version 2.36*. https://mc-stan.org
Szekely, A., Jacobsen, T., D'Amico, S., Devescovi, A., Andonova, E., Herron, D., Lu, C. C., Pechmann, T., Pléh, C., Wicha, N., et al. (2004). A new on-line resource for psycholinguistic studies. *Journal of Memory and Language*, *51*(2), 247–250.
Tucker, M. A., Idrissi, A., & Almeida, D. (2015). Representing number in the real–time processing of agreement: Self-paced reading evidence from Arabic. *Frontiers in Psychology*, *6*, 347. https://doi.org/10.3389/fpsyg.2015.00347
Türk, U. (2022). *Agreement attraction in turkish*. Bogaziçi University.
Türk, U., & Logačev, P. (2024). Agreement attraction in turkish: The case of genitive attractors. *Language, Cognition and Neuroscience*, *39*(4), 448–454.
Van Buuren, S. (2018). *Flexible imputation of missing data, second edition* (p. 414). Chapman & Hall/CRC Press.
Van Dyke, J. A., & McElree, B. (2011). Cue-dependent interference in comprehension. *Journal of Memory and Language*, *65*(3), 247–263.
Vasishth, S., Brüssow, S., Lewis, R. L., & Drenhaus, H. (2008). Processing polarity: How the ungrammatical intrudes on the grammatical. *Cognitive Science*, *32*(4), 685–712.
Vasishth, S., & Engelmann, F. (2021). *Sentence comprehension as a cognitive process: A computational approach*. Cambridge University Press.
Veenstra, A., Acheson, D. J., Bock, K., & Meyer, A. S. (2014). Effects of semantic integration on subject–verb agreement: Evidence from dutch. *Language, Cognition and Neuroscience*, *29*(3), 355–380.
Vigliocco, G., Antonini, T., & Garrett, M. F. (1997). Grammatical gender is on the tip of italian tongues. *Psychological Science*, *8*(4), 314–317.
Vigliocco, G., Butterworth, B., & Semenza, C. (1995). Constructing subject-verb agreement in speech: The role of semantic and morphological factors. *Journal of Memory and Language*, *34*(2), 186–215. https://doi.org/10.1006/jmla.1995.1009
Vigliocco, G., & Nicol, J. (1998). Separating hierarchical relations and word order in language production: Is proximity concord syntactic or linear? *Cognition*, *68*(1), B13–B29.
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pubmed&cmd=Retrieve&dopt=AbstractPlus&list_uids=24202153947149733 14related:gvTqIRRVliEJ
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.
Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686
Wickham, H., Pedersen, T. L., & Seidel, D. (2023). *scales: Scale functions for visualization*.

https://CRAN.R-project.org/package=scales
Wilke, C. O. (2024). *cowplot: Streamlined plot theme and plot annotations for "ggplot2"*. https://CRAN.R-project.org/package=cowplot
Williams, A. (2015). *Arguments in syntax and semantics*. Cambridge University Press.
Xiang, M., Dillon, B., & Phillips, C. (2009). Illusory licensing effects across dependency types: ERP evidence. *Brain and Language*, *108*(1), 40–55.
Xiang, M., Grove, J., & Merchant, J. (2019). Structural priming in production through "silence": An investigation of verb phrase ellipsis and null complement anaphora. *Glossa: A Journal of General Linguistics*, *4*(1).
Xiao, N. (2024). *ggsci: Scientific journal and sci-fi themed color palettes for "ggplot2"*. https://CRAN.R-project.org/package=ggsci
Xie, Y. (2014). knitr: A comprehensive tool for reproducible research in R. In V. Stodden, F. Leisch, & R. D. Peng (Eds.), *Implementing reproducible computational research*. Chapman; Hall/CRC.
Xie, Y. (2015). *Dynamic documents with R and knitr* (2nd ed.). Chapman; Hall/CRC. https://yihui.org/knitr/
Xie, Y. (2024). *knitr: A general-purpose package for dynamic report generation in r*. https://yihui.org/knitr/
Xie, Y., Allaire, J. J., & Grolemund, G. (2018). *R markdown: The definitive guide*. Chapman; Hall/CRC. https://bookdown.org/yihui/rmarkdown
Xie, Y., Dervieux, C., & Riederer, E. (2020). *R markdown cookbook*. Chapman; Hall/CRC. https://bookdown.org/yihui/rmarkdown-cookbook
Yadav, H., Smith, G., Reich, S., & Vasishth, S. (2023). Number feature distortion modulates cue-based retrieval in reading. *Journal of Memory and Language*, *129*, 104400.
Zehr, J., & Schwarz, F. (2018). *PennController for internet based experiments (IBEX)*. https://doi.org/10.17605/OSF.IO/MD832
Zhao, J., Gao, R., & Brennan, J. R. (2024). Decoding the neural dynamics of headed syntactic structure building. *bioRxiv*. https://doi.org/10.1101/2024.11.07.622560
Zhu, H. (2024). *kableExtra: Construct complex table with "kable" and pipe syntax*. https://CRAN.R-project.org/package=kableExtra
Allaire, J., Xie, Y., Dervieux, C., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., Chang, W., & Iannone, R. (2024). *rmarkdown: Dynamic documents for r*. https://github.com/rstudio/rmarkdown
Anderson, J. R. (1996). ACT: A simple theory of complex cognition. *American Psychologist*, *51*(4), 355.
Avetisyan, S., Lago, S., & Vasishth, S. (2020). Does case marking affect agreement attraction in comprehension? *Journal of Memory and Language*, *112*, 104087.
Badecker, W., & Kuminiak, F. (2007). Morphology, agreement and working memory retrieval in sentence production: Evidence from gender and case in Slovak. *Journal of Memory and Language*, *56*(1), 65–85. https://doi.org/10.1016/j.jml.2006.08.004
Badecker, W., & Lewis, R. L. (2007). *Agreement and sentence formulation: The role of working memory (retrievals) in language production*.
Badecker, W., Miozzo, M., & Zanuttini, R. (1995). The two-stage model of lexical retrieval: Evidence from a case of anomia with selective preservation of grammatical

gender. *Cognition*, *57*(2), 193–216.
Bhatia, S., & Dillon, B. (2022). Processing agreement in hindi: When agreement feeds attraction. *Journal of Memory and Language*, *125*, 104322.
Bock, K. (1986). Syntactic persistence in language production. *Cognitive Psychology*, *18*(3), 355–387.
Bock, K., & Cutting, J. C. (1992). Regulating mental energy: Performance units in language production. *Journal of Memory and Language*, *31*(1), 99–127. https://doi.org/10.1016/0749-596X(92)90007-K
Bock, K., & Eberhard, K. M. (1993). Meaning, sound and syntax in English number agreement. *Language and Cognitive Processes*, *8*(1), 57–99. https://doi.org/10.1080/01690969308406949
Bock, K., Eberhard, K. M., & Cutting, J. C. (2004). Producing number agreement: How pronouns equal verbs. *Journal of Memory and Language*, *51*(2), 251–278.
Bock, K., & Ferreira, V. (2014). Syntactically speaking. *The Oxford Handbook of Language Production*, 21–46.
Bock, K., & Miller, C. A. (1991). Broken agreement. *Cognitive Psychology*, *23*(1), 45–93.
Bürki, A., Elbuy, S., Madec, S., & Vasishth, S. (2020). What did we learn from forty years of research on semantic interference? A bayesian meta-analysis. *Journal of Memory and Language*, *114*, 104125.
Bürkner, P.-C. (2017a). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01
Bürkner, P.-C. (2017b). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01
Bürkner, P.-C. (2018a). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*(1), 395–411. https://doi.org/10.32614/RJ-2018-017
Bürkner, P.-C. (2018b). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*(1), 395–411. https://doi.org/10.32614/RJ-2018-017
Bürkner, P.-C. (2021). Bayesian item response modeling in R with brms and Stan. *Journal of Statistical Software*, *100*(5), 1–54. https://doi.org/10.18637/jss.v100.i05
Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*, 1–32.
Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2024). *shiny: Web application framework for r*. https://CRAN.R-project.org/package=shiny
Chkhetiani, L., Vanzo, A., Khare, Y., Peyash, T., Sklyar, I., Liang, M., Botros, R., Bousbib, R., Etefy, A., Ghahremani, P., Oexle, G., Trueba, J. L., Ferreira, W. P., Gotthold, B., Bahadoori, S., Chiang, M., Mitov, A., Fakhan, E., Yoshioka, T., … Martin, S. R. (n.d.). *Universal-2: Advancing multilingual speech understanding at scale*. AssemblyAI. Retrieved October 27, 2024, from https://www.assemblyai.com/research/universal-2
Cousineau, D. (2005). 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. https://doi.org/10.20982/tqmp.01.1.p042
Damian, M. F., & Martin, R. C. (1999). Semantic and phonological codes interact in

single word production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(2), 345.
Dillon, B., Clifton Jr, C., & Frazier, L. (2014). Pushed aside: Parentheticals, memory and processing. *Language, Cognition and Neuroscience*, *29*(4), 483–498.
Dillon, B., Clifton Jr, C., Sloggett, S., & Frazier, L. (2017). Appositives and their aftermath: Interference depends on at-issue vs. Not-at-issue status. *Journal of Memory and Language*, *96*, 93–109.
Dillon, B., & Keshev, M. (n.d.). *Syntactic dependency formation in sentence processing: A comparative perspective*.
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.
Drenhaus, H., Saddy, D., & Frisch, S. (2005). Processing negative polarity items: When negation comes through the backdoor. *Linguistic Evidence: Empirical, Theoretical, and Computational Perspectives*, 145–165.
Drummond, A. (2013). Ibex farm.
Eberhard, K. M. (1999). The accessibility of conceptual number to the processes of subject–verb agreement in English. *Journal of Memory and Language*, *41*(4), 560–578. https://doi.org/10.1006/jmla.1999.2662
Eberhard, K. M., Cutting, J. C., & Bock, K. (2005). Making syntax of sense: Number agreement in sentence production. *Psychological Review*, *112*(3), 531.
Ferreira, V., Morgan, A., & Slevc, L. R. (2007). *Grammatical encoding*. The Oxford Handbook of Psycholinguistics/Oxford University Press.
Franck, J., Lassi, G., Frauenfelder, U. H., & Rizzi, L. (2006). Agreement and movement: A syntactic analysis of attraction. *Cognition*, *101*(1), 173–216.
Frank, R. (2004). *Phrase structure composition and syntactic dependencies* (Vol. 38). Mit Press.
Gabry, J., Češnovar, R., Johnson, A., & Bronder, S. (2024). *cmdstanr: R interface to "CmdStan"*. https://github.com/stan-dev/cmdstanr
Ghodsi, M., Liu, X., Apfel, J., Cabrera, R., & Weinstein, E. (2020). Rnn-transducer with stateless prediction network. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7049–7053. https://doi.org/10.1109/ICASSP40776.2020.9054419
Gillespie, M., & Pealmutter, N. J. (2011). Effects of semantic integration and advance planning on grammatical encoding in sentence production. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *33*.
Griffin, Z. M. (2001). Gaze durations during speech reflect word selection and phonological encoding. *Cognition*, *82*(1), B1–B14. https://doi.org/https://doi.org/10.1016/S0010-0277(01)00138-X
Griffin, Z. M., & Ferreira, V. (2006). Properties of spoken language production. In *Handbook of psycholinguistics* (pp. 21–59). Elsevier.
Gussow, A. E., & MacDonald, M. C. (2023). Utterance planning under message uncertainty: Evidence from a novel picture-naming paradigm. *Cognitive, Affective, & Behavioral Neuroscience*, *23*(3), 957–972.
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.
Hope, R. M. (2022). *Rmisc: Ryan miscellaneous*. https://CRAN.R-project.org/package=Rmisc
Humphreys, K. R., & Bock, K. (2005). Notional number agreement in english. *Psychonomic Bulletin & Review*, *12*, 689–695.
Hwang, H., & Kaiser, E. (2014). The role of the verb in grammatical function assignment in english and korean. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(5), 1363.
Jäger, L. A., Engelmann, F., & Vasishth, S. (2015). Retrieval interference in reflexive processing: Experimental evidence from mandarin, and computational modeling. *Frontiers in Psychology*, *6*, 617.
Jescheniak, J. D., & Schriefers, H. (2001). Priming effects from phonologically related distractors in picture–word interference. *The Quarterly Journal of Experimental Psychology A*, *54*(2), 371–382. https://doi.org/10.1080/02724980042000273
Jescheniak, J. D., Schriefers, H., & Lemhöfer, K. (2014). Selection of freestanding and bound gender-marking morphemes in speech production: A review. *Language, Cognition and Neuroscience*, *29*(6), 684–694. https://doi.org/10.1080/01690965.2012.654645
Joshi, A. K., Levy, L. S., & Takahashi, M. (1975). Tree adjunct grammars. *Journal of Computer and System Sciences*, *10*(1), 136–163.
Kandel, M., Pañeda, C., Bahmanian, N., Bruera, M. M., Phillips, C., & Lago, S. (2025). Number and grammatical gender attraction in spanish pronouns: Evidence for a syntactic route to their features. *Journal of Cognition*, *8*(1), 10.
Kandel, M., & Phillips, C. (2022). Number attraction in verb and anaphor production. *Journal of Memory and Language*, *127*, 104370.
Kandel, M., Wyatt, C. R., & Phillips, C. (2022). Agreement attraction error and timing profiles in continuous speech. *Glossa Psycholinguistics*, *1*(1).
Kempen, G., & Hoenkamp, E. (1987). An incremental procedural grammar for sentence formulation. *Cognitive Science*, *11*(2), 201–258.
Kempen, G., & Huijbers, P. (1983). The lexicalization process in sentence production and naming: Indirect election of words. *Cognition*, *14*(2), 185–209.
Kim, S. J., & Xiang, M. (2024). Incremental discourse-update constrains number agreement attraction effect. *Cognitive Science*, *48*(9), e13497.
Konopka, A. E., & Kuchinsky, S. E. (2015). How message similarity shapes the timecourse of sentence formulation. *Journal of Memory and Language*, *84*, 1–23.
Kuchinsky, S., & Bock, K. (2010). *Paper presented at the 23rd meeting of the CUNY human sentence processing conference*.
Kush, D., & Phillips, C. (2014). Local anaphor licensing in an SOV language: Implications for retrieval strategies. *Frontiers in Psychology*, *5*, 1252.
Lago, S., Gračanin–Yuksek, M., Şafak, D. F., Demir, O., Kırkıcı, B., & Felser, C. (2019). Straight from the horse's mouth: Agreement attraction effects with Turkish possessors [Journal Article]. *Linguistic Approaches to Bilingualism*, *9*(3), 398–426. https://doi.org/10.1075/lab.17019.lag
Lago, S., Shalom, D. E., Sigman, M., Lau, E. F., & Phillips, C. (2015). Agreement

attraction in Spanish comprehension. *Journal of Memory and Language*, *82*, 133–149. https://doi.org/10.1016/j.jml.2015.02.002

Lasnik, H., & Uriagereka, J. (2022). *Structure: Concepts, consequences, interactions*. MIT Press.

Levelt, W. J. (1989). *Speaking: From intention to articulation*. MIT press.

Levelt, W. J., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*(1), 1–38.

Lewis, R. L., & Vasishth, S. (2005). An Activation–Based Model of Sentence Processing as Skilled Memory Retrieval. *Cognitive Science*, *29*(3), 375–419. https://doi.org/10.1207/s15516709cog0000/_25

Logačev, P., & Vasishth, S. (2011). Case matching and conflicting bindings interference. In *Case, word order and prominence: Interacting cues in language production and comprehension* (pp. 187–216). Springer.

Makowski, D., Ben-Shachar, M. S., & Lüdecke, D. (2019). bayestestR: Describing effects and their uncertainty, existence and significance within the bayesian framework. *Journal of Open Source Software*, *4*(40), 1541. https://doi.org/10.21105/joss.01541

McAuliffe, M., & Sonderegger, M. (2023). *English (US) MFA G2P model v3.0.0*.

McAuliffe, M., & Sonderegger, M. (2024a). *English MFA acoustic model v3.1.0*.

McAuliffe, M., & Sonderegger, M. (2024b). *English MFA dictionary v3.1.0*.

Meyer, A. S. (1996). Lexical access in phrase and sentence production: Results from picture–word interference experiments. *Journal of Memory and Language*, *35*(4), 477–496.

Meyer, A. S., & Bock, K. (1999). Representations and processes in the production of pronouns: Some perspectives from dutch. *Journal of Memory and Language*, *41*(2), 281–301.

Meyer, A. S., & Schriefers, H. (1991). Phonological facilitation in picture-word interference experiments: Effects of stimulus onset asynchrony and types of interfering stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(6), 1146.

Miozzo, M., & Caramazza, A. (1997a). On knowing the auxiliary of a verb that cannot be named: Evidence for the independence of grammatical and phonological aspects of lexical knowledge. *Journal of Cognitive Neuroscience*, *9*(1), 160–166.

Miozzo, M., & Caramazza, A. (1997b). Retrieval of lexical–syntactic features in tip-of-the tongue states. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(6), 1410.

Momma, S., & Ferreira, V. (2019). Beyond linear order: The role of argument structure in speaking. *Cognitive Psychology*, *114*, 101228.

Momma, S., Slevc, L. R., & Phillips, C. (2016). The timing of verb selection in japanese sentence production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(5), 813.

Momma, S., & Yoshida, M. (2023). Planning multiple dependencies in sentence production. *Language, Cognition and Neuroscience*, *38*(9), 1183–1213.

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Reason*, *4*(2), 61–64. https://doi.org/10.20982/tqmp.04.2.p061

Nozari, N., & Omaki, A. (2022). *An investigation of the dependency of subject-verb*

agreement on inhibitory control processes in sentence production.
Parker, D., & Phillips, C. (2017). Reflexive attraction in comprehension is selective. *Journal of Memory and Language*, *94*, 272–290.
Pedersen, T. L., Ooms, J., & Govett, D. (2025). *systemfonts: System native font finding.* https://CRAN.R-project.org/package=systemfonts
Perlmutter, D. M. (1978). Impersonal passives and the unaccusative hypothesis. *Annual Meeting of the Berkeley Linguistics Society*, 157–190.
Perlmutter, D. M., & Postal, P. (1984). The inadequacy of some monostratal theories of passive. *Studies in Relational Grammar*, *2*, 3–37.
Pfau, R. (2009). *Grammar as processor.* John Benjamins Publishing Company. http://digital.casalini.it/9789027289636
Pickering, M. J., & Branigan, H. P. (1998). The representation of verbs: Evidence from syntactic priming in language production. *Journal of Memory and Language*, *39*(4), 633–651.
R Core Team. (2024a). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. https://www.R-project.org/
R Core Team. (2024b). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. https://www.R-project.org/
Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. *International Conference on Machine Learning*, 8748–8763.
Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108. https://doi.org/10.1037/0033-295X.85.2.59
Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, *9*(5), 347–356.
Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences*, *20*(4), 260–281.
Ristic, B., Molinaro, N., & Mancini, S. (2016). Agreement attraction in Serbian: Decomposing markedness [Journal Article]. *The Mental Lexicon*, *11*(2), 242–276. https://doi.org/10.1075/ml.11.2.04ris
Roelofs, A. (1992). A spreading-activation theory of lemma retrieval in speaking. *Cognition*, *42*(1-3), 107–142.
Roelofs, A. (2001). Set size and repetition matter: Comment on caramazza and costa (2000). *Cognition*, *80*(3), 283–290.
Roeser, J., Torrance, M., Andrews, M., & Baguley, T. (2024). *No default syntactic scope for advance planning in sentence production: Evidence from finite mixture models*. OSF.
Rosinski, R. R., Golinkoff, R. M., & Kukish, K. S. (1975). Automatic semantic processing in a picture-word interference task. *Child Development*, 247–253.
Sartori, G., Job, R., & Coltheart, M. (1992). The organization of object knowledge: Evidence from neuropsychology. In D. E. Meyer & S. Kornblum (Eds.), *Attention and performance, XN*. Lawrence Erlbaum.
Sauppe, S. (2017). Word order and voice influence the timing of verb planning in german sentence production. *Frontiers in Psychology*, *8*, 1648.
Schiller, N. O., & Caramazza, A. (2003). Grammatical feature selection in noun phrase

production: Evidence from German and Dutch. *Journal of Memory and Language*, *48*(1), 169–194. https://doi.org/10.1016/S0749-596X(02)00508-9

Schlueter, Z., Parker, D., & Lau, E. (2019). Error-driven retrieval in agreement attraction rarely leads to misinterpretation. *Frontiers in Psychology*, *10*, 1002.

Schnur, T. T. (2011). Phonological planning during sentence production: Beyond the verb. *Frontiers in Psychology*, *2*, 319.

Schriefers, H. (1993). Syntactic processes in the production of noun phrases. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(4), 841–850. https://doi.org/10.1037/0278-7393.19.4.841

Schriefers, H., Jescheniak, J. D., & Hantsch, A. (2002). Determiner selection in noun phrase production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(5), 941–950. https://doi.org/10.1037/0278-7393.28.5.941

Schriefers, H., Meyer, A. S., & Levelt, W. J. (1990). Exploring the time course of lexical access in language production: Picture-word interference studies. *Journal of Memory and Language*, *29*(1), 86–102.

Schriefers, H., Teruel, E., & Meinshausen, R.-M. (1998). Producing simple sentences: Results from picture–word interference experiments. *Journal of Memory and Language*, *39*(4), 609–632.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Slioussar, N. (2018). Forms and features: The role of syncretism in number agreement attraction. *Journal of Memory and Language*, *101*, 51–63.

Slowikowski, K. (2024). *ggrepel: Automatically position non-overlapping text labels with "ggplot2"*. https://CRAN.R-project.org/package=ggrepel

Smith, G., & Vasishth, S. (2020). A Principled Approach to Feature Selection in Models of Sentence Processing. *Cognitive Science*, *44*(12), e12918. https://doi.org/10.1111/cogs.12918

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(2), 174.

Solomon, E. S., & Pearlmutter, N. J. (2004). Semantic integration and syntactic planning in language production. *Cognitive Psychology*, *49*(1), 1–46.

Stan Development Team. (n.d.). *RStan: The R interface to Stan*. https://mc-stan.org/

Stan Development Team. (2024). *Stan modeling language users guide and reference manual, version 2.36*. https://mc-stan.org

Szekely, A., Jacobsen, T., D'Amico, S., Devescovi, A., Andonova, E., Herron, D., Lu, C. C., Pechmann, T., Pléh, C., Wicha, N., et al. (2004). A new on-line resource for psycholinguistic studies. *Journal of Memory and Language*, *51*(2), 247–250.

Tucker, M. A., Idrissi, A., & Almeida, D. (2015). Representing number in the real–time processing of agreement: Self-paced reading evidence from Arabic. *Frontiers in Psychology*, *6*, 347. https://doi.org/10.3389/fpsyg.2015.00347

Türk, U. (2022). *Agreement attraction in turkish*. Bogaziçi University.

Türk, U., & Logačev, P. (2024). Agreement attraction in turkish: The case of genitive attractors. *Language, Cognition and Neuroscience*, *39*(4), 448–454.

Van Buuren, S. (2018). *Flexible imputation of missing data, second edition* (p. 414).

Chapman & Hall/CRC Press.
Van Dyke, J. A., & McElree, B. (2011). Cue-dependent interference in comprehension. *Journal of Memory and Language*, *65*(3), 247–263.
Vasishth, S., Brüssow, S., Lewis, R. L., & Drenhaus, H. (2008). Processing polarity: How the ungrammatical intrudes on the grammatical. *Cognitive Science*, *32*(4), 685–712.
Vasishth, S., & Engelmann, F. (2021). *Sentence comprehension as a cognitive process: A computational approach*. Cambridge University Press.
Veenstra, A., Acheson, D. J., Bock, K., & Meyer, A. S. (2014). Effects of semantic integration on subject–verb agreement: Evidence from dutch. *Language, Cognition and Neuroscience*, *29*(3), 355–380.
Vigliocco, G., Antonini, T., & Garrett, M. F. (1997). Grammatical gender is on the tip of italian tongues. *Psychological Science*, *8*(4), 314–317.
Vigliocco, G., Butterworth, B., & Semenza, C. (1995). Constructing subject-verb agreement in speech: The role of semantic and morphological factors. *Journal of Memory and Language*, *34*(2), 186–215. https://doi.org/10.1006/jmla.1995.1009
Vigliocco, G., & Nicol, J. (1998). Separating hierarchical relations and word order in language production: Is proximity concord syntactic or linear? *Cognition*, *68*(1), B13–B29.
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pubmed&cmd=Retrieve&dopt=AbstractPlus&list_uids=2420215394714973314related:gvTqIRRVliEJ
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.
Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686
Wickham, H., Pedersen, T. L., & Seidel, D. (2023). *scales: Scale functions for visualization*. https://CRAN.R-project.org/package=scales
Wilke, C. O. (2024). *cowplot: Streamlined plot theme and plot annotations for "ggplot2"*. https://CRAN.R-project.org/package=cowplot
Williams, A. (2015). *Arguments in syntax and semantics*. Cambridge University Press.
Xiang, M., Dillon, B., & Phillips, C. (2009). Illusory licensing effects across dependency types: ERP evidence. *Brain and Language*, *108*(1), 40–55.
Xiang, M., Grove, J., & Merchant, J. (2019). Structural priming in production through "silence": An investigation of verb phrase ellipsis and null complement anaphora. *Glossa: A Journal of General Linguistics*, *4*(1).
Xiao, N. (2024). *ggsci: Scientific journal and sci-fi themed color palettes for "ggplot2"*. https://CRAN.R-project.org/package=ggsci
Xie, Y. (2014). knitr: A comprehensive tool for reproducible research in R. In V. Stodden, F. Leisch, & R. D. Peng (Eds.), *Implementing reproducible computational research*. Chapman; Hall/CRC.
Xie, Y. (2015). *Dynamic documents with R and knitr* (2nd ed.). Chapman; Hall/CRC. https://yihui.org/knitr/
Xie, Y. (2024). *knitr: A general-purpose package for dynamic report generation in r*.

https://yihui.org/knitr/
Xie, Y., Allaire, J. J., & Grolemund, G. (2018). *R markdown: The definitive guide*. Chapman; Hall/CRC. https://bookdown.org/yihui/rmarkdown
Xie, Y., Dervieux, C., & Riederer, E. (2020). *R markdown cookbook*. Chapman; Hall/CRC. https://bookdown.org/yihui/rmarkdown-cookbook
Yadav, H., Smith, G., Reich, S., & Vasishth, S. (2023). Number feature distortion modulates cue-based retrieval in reading. *Journal of Memory and Language*, *129*, 104400.
Zehr, J., & Schwarz, F. (2018). *PennController for internet based experiments (IBEX)*. https://doi.org/10.17605/OSF.IO/MD832
Zhao, J., Gao, R., & Brennan, J. R. (2024). Decoding the neural dynamics of headed syntactic structure building. *bioRxiv*. https://doi.org/10.1101/2024.11.07.622560
Zhu, H. (2024). *kableExtra: Construct complex table with "kable" and pipe syntax*. https://CRAN.R-project.org/package=kableExtra

Appendices

Appendix A Materials for Experiment 1

Type	Verb Type	Target Sentence	Related Dist.	Unrelated Dist.
control	Unerg	The babies below the waffle(s) are hiding	find	consider
control	Unacc	The babies below the trophy(s) are crying	yell	write
control	Unerg	The girls below the brush(es) are standing	squat	squash
control	Unacc	The girls below the globe(s) are smiling	cry	launch
control	Unerg	The men above the skateboard(s) are laughing	wink	swear
control	Unacc	The men above the strawberry(ies) are sweating	wash	purify
control	Unerg	The professors below the umbrella(s) are teaching	show	stalk
control	Unacc	The professors below the watch(es) are listening	audit	orbit
control	Unerg	The soldiers above the candle(s) are shooting	strike	find
control	Unacc	The soldiers above the hat(s) are saluting	respect	restart
control	Unerg	The women above the broom(s) are surfing	swim	stink
control	Unacc	The women above the fan(s) are sunbathing	sleep	sell
experimental	Unacc	the ballerina above the axe(s) is shrinking	grow	float
experimental	Unerg	the ballerina above the harp(s) is running	swim	sneeze
experimental	Unacc	the boy below the desk(s) is floating	drown	shrink
experimental	Unerg	the boy below the lighthouse(s) is yawning	sleep	bark
experimental	Unacc	the chef above the dresser(s) is drowning	float	bounce
experimental	Unerg	the chef above the windmill(s) is yelling	bark	sleep
experimental	Unacc	the clown below the violin(s) is growing	shrink	shake
experimental	Unerg	the clown below the violin(s) is walking	crawling	cough
experimental	Unacc	the cowboy above the sword(s) is falling	sink	boil
experimental	Unerg	the cowboy above the piano(s) is winking	smile	crawl
experimental	Unacc	the dog above the hammer(s) is spinning	trip	sink
experimental	Unerg	the dog above the apple(s) is barking	yell	yawn
experimental	Unacc	the monkey above the carrot(s) is tripping	spin	melt
experimental	Unerg	the monkey above the knife(s) is sleeping	yawn	yell
experimental	Unacc	the octopus below the spoon(s) is boiling	melt	fall
experimental	Unerg	the octopus below the lemon(s) is swimming	run	smile
experimental	Unacc	the penguin below the drill(s) is bouncing	shake	drown
experimental	Unerg	the penguin below the tomato(es) is sneezing	cough	run
experimental	Unacc	the pirate below the guitar(s) is sinking	fall	spin
experimental	Unerg	the pirate below the gun(s) is coughing	sneeze	walk
experimental	Unacc	the rabbit above the church(es) is shaking	bounce	grow
experimental	Unerg	the rabbit above the chair(s) is smiling	wink	swim
experimental	Unacc	the snail below the castle(s) is melting	boil	trip
experimental	Unerg	the snail below the desk(s) is crawling	walk	wink

Table 1

Materials for Experiment 1

Appendix B Materials for Experiment 2

The items for Experiment 2 were generated dynamically through randomization. The set of entities used in the experiment consisted of the following six characters: **pirate**, **clown**, **doctor**, **wizard**, **chef**, and **cowboy**.

Two types of intransitive verbs were used to represent different syntactic structures:

- **Unaccusative verbs**: *fall*, *shrink*, *boil*, *bounce*, *burn*, *float*, *freeze*, *grow*, *shake*, *drown*, *slip*, *melt*
- **Unergative verbs**: *run*, *laugh*, *surf*, *cough*, *smile*, *sing*, *shout*, *jump*, *fly*, *cry*, *dance*, *climb*

Participants were divided into two counterbalanced lists to control for lexical variability.

- In **List 1**, participants saw the **first six unaccusatives** and **first six unergatives** paired with the entities **pirate**, **clown**, and **doctor**. The **remaining six unaccusatives** and **remaining six unergatives** were paired with **wizard**, **chef**, and **cowboy**. - In **List 2**, this pairing was reversed: the **first six unaccusatives** and **first six unergatives** were shown with **wizard**, **chef**, and **cowboy**, while the **remaining six verbs** of each type were shown with **pirate**, **clown**, and **doctor**.

On each trial, the two “standing” entities were randomly selected from the remaining five entities not assigned as the agent for that trial.