Number Feature Mismatches under Ellipsis

Sabrina T. Grimberg
LING 245: Methods in Psycholinguistics
Spring 2017

1 Introduction

Generally speaking, the term *ellipsis* is used to denote the absence of phonological material that is otherwise expected to be present. Although there are many different kinds of ellipses, the following instance of VP-ellipsis in English can serve as an illustration[1]:

(1) *Which films did he refuse to see, and which films$_i$ did he agree to <[$_{VP}$ see t_i]>?*[2].

Several arguments have been posited in support of the claim that ellipsis resolution involves computing silent structures at the ellipsis site. For example, analogously to overt structures, ellipsis sites can be extracted out of. This is shown in (1), where the fact that the ellipsis site contains the origin site of the dependency headed by *which films* can be accommodated if the missing VP is indeed structurally present. Among other arguments for the presence of syntactic structure in ellipsis sites are the possibility of triggering agreement on items outside of the elliptical site, as in (2), and the possibility for quantifiers inside ellipsis sites of taking wide scope, as in (3a) versus (3b) (Merchant 2013).

(2) a. *First, there were bananas available, and then there weren't <bananas available>.*
b. *First, there were going to be bananas available, and then there weren't <going to be bananas available>.*

(3) a. *A doctor examined every patient, and then a nurse did <examine every patient>.* ($\exists\forall$, $\forall\exists$)
b. * *A doctor examined every patient, and then a nurse did it.* ($\exists\forall$, $^{*}\forall\exists$)

A matter that has been observed with respect to ellipsis is that the understood material in ellipsis must be identical or parallel to its antecedent, in some sense of "identity" or "parallelism" to be defined. This identity (or parallelism, or resolution) may be semantic or syntactic, or some mix of the two (Merchant 2016). The need for some sort of identity condition is illustrated by examples such as (4b), a case where morphosyntactic mismatches between the elided material and its antecedent result in ungrammaticality. In contrast, the matched control example (4a) is grammatical, which suggests that the degraded status of (4b) is due to the absence of morphosyntactic isomorphism (Aparicio, Franich, and Xiang 2014).

(4) a. *John was here, and Mary too <was here>* (adapted from Lasnik 1999)
b. * *John was here, and Mary will too <be here>* (adapted from Lasnik 1999)

However, researchers have claimed that this lack of morphosyntactic parallelism is grammatical in many languages (Depiante & Masullo 2001, Merchant 2014, Nunes & Zocca 2005, *inter alia*). Greek is one of such languages, as it allows gender mismatches under ellipsis for some types of nouns: despite having

[1] Examples (1)-(3) are adapted from Merchant (2013: 539-538).
[2] Angled brackets will be used throughout this paper to indicate elided material.

different lexically determined gender features, a masculine noun can serve as the antecedent to a putative feminine form, and vice versa, as shown in (5) (Merchant 2014).

(5) Greek (adapted from Merchant 2014: 15)

a. *O Petros ine* ***kalos jatros****, ala i Maria ine* ***mia kakia****.*
the Petros is **good.M doctor** but the Maria is **a.F bad.F**
'Petros is a good doctor, but Maria is a bad one.'

b. *I Maria ine* ***kali jatros****, ala o Petros ine* ***enas kakos****.*
the Maria is **good.F doctor** but the Petros is **a.M bad.M**
'Maria is a good doctor, but Petros is a bad one.'

In the same vein, Nunes & Zocca (2005, 2009) observe that Brazilian Portuguese tolerates the lack of isomorphism between morphological features on adjectival predicates under VP ellipsis: while (6a) shows a mismatch in gender between the two conjuncts, a mismatch in number occurs in (6b).

(6) Brazilian Portuguese (adapted from Nunes & Zocca 2005: 39)

a. *O João é* ***alto*** *e a Maria também é <****alta****>.*
the Joao is **tall-M.SG** and the Maria also is **tall-F.SG**
'John is tall and Mary is too.'

b. *O João é* ***alto*** *e aqueles meninos também são <****altos****>.*
the Joao is **tall-M.SG** and those boys also are **tall-M.PL**
'John is tall and those boys are too.'

A third language where mismatches seem to be tolerated under ellipsis is Spanish. Although Spanish does not allow VP ellipsis, in instances of stripping (also called "bare argument ellipsis") both gender and number features on the antecedent may differ from those on the elided material. For instance, in (7a), the predicate *entretenido* is singular, but the subject of the second clause, *los conductores*, is plural. Similarly, unlike the masculine subject of the first conjunct in (8a), the subject of the second conjunct (*la conductora*) is feminine. These (a) cases contrast with their non-elliptical (b) counterparts, where the lack of agreement causes a robust ungrammaticality.

(7) Spanish

a. *El* ***programa*** *es entretenid****o*** *y los* ***conductores*** *también <son*
the **show.M.SG** is entertaining.**M.SG** and the **host.M.PL** too are
*entretenid****os****>.*
entertaining.**M.PL**
'The show is entertaining and the hosts, too.'

b. * *Los* ***conductores*** *son* ***entretenido****.*
the **host.M.PL** are **entertaining.M.SG**

(8) Spanish

a. *El* ***programa*** *es entretenid****o*** *y la* ***conductora*** *también <es*
the **show.M.SG** is entertaining.**M.SG** and the **host.F.SG** too is
*entretenid****a****>.*
entertaining.**F.SG**
'The show is entertaining and the hostess, too.'

b. * *La* ***conductora*** *es* ***entretenido***.
the **host.F.SG** is **entertaining.M.SG**

In this context, Aparicio, Franich, and Xiang (henceforth, AFX) (2014) set to explore whether morphological feature matching between the unpronounced material at the ellipsis site and its antecedent is relevant or not for ellipsis computation. Their linking hypothesis is that reading time reflects processing effort. Based on this, they hold as their behavioral prediction that subjects ought to display an increased reading time at the critical region for the mismatched ellipsis conditions, if it is indeed the case that dissimilar morphological features are relevant for ellipsis computation. Contrastingly, they predict that in instances of ellipsis where the features match, no such increasing effect on the reading time should occur.

The current paper builds on AFX by reporting the results of an experiment that partially replicates AFX's study. §2 presents the details of the experiment: §2.1 refers to different aspects of the experimental design, while §2.2 focuses on the results obtained. §3 expands on the results collected and includes some comments about the ways in which my study diverges from AFX's, a matter to which, I suspect, it can be attributed that the original results by AFX were not replicated.

2 Experiment

2.1 Methods

2.1.1 Participants

Nineteen subjects participated in this study, all of whom considered themselves native Spanish speakers. They were paid U$S 0.9 for participating in this task. One subject's data were excluded given that the answers she provided to the linguistic background questions seemed to suggest she was not in fact a Spanish native speaker. This number of participants differs from that of AFX's study, who recruited twenty-eight native Spanish speakers for their Experiment 1.

It might be relevant to point out that the subjects in AFX's study were native speakers of Iberian Spanish, whereas the participants in my replication were all from Central or North America (either themselves born there, their parents born there and/or residents there).

2.1.2 Materials

The materials used in the experiment were a modified version of those in AFX's study. Their original Experiment 1 included two sets of 40 items each: while the first set manipulated the gender feature and kept the number feature constant (the items displayed all singular number), the second set of items manipulated the number feature and kept the gender feature constant (all showing instances of masculine gender). In addition to these 80 items, AFX included 50 filler sentences, all of which were ungrammatical due to subject-verb agreement errors.

Each of the critical items was tested for two conditions. First, "Match", that is, whether the number or gender features on the subject of the second clause matches or mismatches with its antecedent. The second condition is "Subject of the Second Clause" (SSC), which refers to the feature markedness on

the subject of the second clause: the subject of the second clause either has an unmarked feature (i.e., singular for the number set or masculine for the gender set) or a marked one (i.e., plural for the number set or feminine for the gender set). A sample item is presented in (9), from AFX.

(9) Elliptical Examples, Number Set

a. **Match, Unmarked SSC**

El *fugitivo* *es peligroso* *y* *el* *preso* *<es peligroso>*
the.M.SG fugitive.M is dangerous.M.SG and the.M.SG prisoner.M is dangerous.M.SG
también.
too.

b. **Mismatch, Unmarked SSC**

Los *fugitivos* *son peligrosos* *y* *el* *preso* *<es*
the.M.PL fugitives.M are dangerous.M.PL and the.M.SG prisoner.M is
peligroso> ***también.***
dangerous.M.SG **too**.

c. **Mismatch, Marked SSC**

El *fugitivo* *es peligroso* *y* *los* *presos* *<son*
the.M.SG fugitive.M is dangerous.M.SG and the.M.PL prisoners.M are
peligrosos> ***también.***
dangerous.M.PL **too**.

d. **Match, Marked SSC**

Los *fugitivos* *son peligrosos* *y* *los* *presos* *<son*
the.M.PL fugitives.M are dangerous.M.PL and the.M.PL prisoners.M are
peligrosos> ***también.***
dangerous.M.PL **too**.

For every combination of the Match and SSC conditions, AFX included its non-elliptical counterpart, which they considered fillers and whose results they did not analyze. (10) illustrates the non-elliptical sentences corresponding to (9).

(10) Non-elliptical Fillers, Number Set

a. **Match, Unmarked SSC**

El *fugitivo* *es peligroso* *y* *el* *preso* *es peligroso*
the.M.SG fugitive.M is dangerous.M.SG and the.M.SG prisoner.M is dangerous.M.SG
también.
too.

b. **Mismatch, Unmarked SSC**

Los *fugitivos* *son peligrosos* *y* *el* *preso* *es peligroso*
the.M.PL fugitives.M are dangerous.M.PL and the.M.SG prisoner.M is dangerous.M.SG
también.
too.

c. **Mismatch, Marked SSC**

El fugitivo es peligroso y los presos son peligrosos
the.M.SG fugitive.M is dangerous.M.SG and the.M.PL prisoners.M are dangerous.M.PL
también.
too.

d. **Match, Marked SSC**

Los fugitivos son peligrosos y los presos son
the.M.PL fugitives.M are dangerous.M.PL and the.M.PL prisoners.M are
peligrosos también.
dangerous.M.PL too.

Unlike the sample stimuli in (9)-(10), the items in the experiment I conducted included additional material following the critical word *también* 'too'. The reason for this was that spillover effects might be observed after the critical word. Therefore, it was necessary to add material after it in order to test if this was so.

All of the continuations included a connector (either *entonces* 'so', *por eso* or *por lo que* 'that's why', or *así que* 'so that') plus a subject-verb clause that did not include any reference to the previous coordinated clause. The critical items in (11) and the control items in (12) are the modified versions of (9)-(10), respectively, that were included in the experiment. Note that they are identical up until the critical word *también*.

(11) Elliptical Examples, Number Set

a. **Match, Unmarked SSC**

El fugitivo es peligroso y el preso <es peligroso>
the.M.SG fugitive.M is dangerous.M.SG and the.M.SG prisoner.M is dangerous.M.SG
***también**, por eso los policías están preocupados.*
too.

b. **Mismatch, Unmarked SSC**

Los fugitivos son peligrosos y el preso <es
the.M.PL fugitives.M are dangerous.M.PL and the.M.SG prisoner.M is
peligroso> ***también**, por eso los policías están preocupados.*
dangerous.M.SG **too**.

c. **Mismatch, Marked SSC**

El fugitivo es peligroso y los presos <son
the.M.SG fugitive.M is dangerous.M.SG and the.M.PL prisoners.M are
peligrosos> ***también**, por eso los policías están preocupados.*
dangerous.M.PL **too**.

d. **Match, Marked SSC**

Los fugitivos son peligrosos y los presos <son
the.M.PL fugitives.M are dangerous.M.PL and the.M.PL prisoners.M are
peligrosos> ***también**, por eso los policías están preocupados.*
dangerous.M.PL **too**.

(12) Non-elliptical Fillers, Number Set

a. **Match, Unmarked SSC**

El *fugitivo* *es peligroso* *y* *el* *preso* *es peligroso*
the.M.SG fugitive.M is dangerous.M.SG and the.M.SG prisoner.M is dangerous.M.SG
también, por eso los policías están preocupados.
too.

b. **Mismatch, Unmarked SSC**

Los *fugitivos* *son peligrosos* *y* *el* *preso* *es peligroso*
the.M.PL fugitives.M are dangerous.M.PL and the.M.SG prisoner.M is dangerous.M.SG
también, por eso los policías están preocupados.
too.

c. **Mismatch, Marked SSC**

El *fugitivo* *es peligroso* *y* *los* *presos* *son peligrosos*
the.M.SG fugitive.M is dangerous.M.SG and the.M.PL prisoners.M are dangerous.M.PL
también, por eso los policías están preocupados.
too.

d. **Match, Marked SSC**

Los *fugitivos* *son peligrosos* *y* *los* *presos* *son*
the.M.PL fugitives.M are dangerous.M.PL and the.M.PL prisoners.M are
peligrosos *también, por eso los policías están preocupados.*
dangerous.M.PL too.

Another substantial difference with respect to AFX's study is that, as mentioned earlier, I worked only with the number set of their Experiment 1, although I still run both the Match and the SSC conditions.

2.1.3 Procedure

AFX's original experiment consisted of an eye-tracking task: subjects underwent a calibration procedure first, then had a series of practice trials, and later went on to the actual experiment. This involved the participants reading the stimuli on a monitor and performing a yes-no grammaticality judgment about the sentence they had just read, all while their right-eye-movements were being tracked. The instructions were in Spanish.

The procedure for my replication diverged significantly, in that instead of running an eye-tracking experiment, my study was a self-paced reading on Amazon's Mechanical Turk. Subjects were randomly assigned a set of ten sentences. In each set, four items were elliptical combinations of the Match and SSC conditions (as illustrated in (11)), four items were the non-elliptical counterparts which functioned as control items (as in (12)), and the other two sentences were ungrammatical non-elliptical fillers. The order in which these ten sentences of the set appeared was random.

At the beginning of each trial, the sentence appeared on the screen with all non-space characters replaced by a dash. Subjects pressed the space bar to view each consecutive word in the sentence. Durations between space bar presses were recorded. At each press of the space bar, the currently viewed word reverted to dashes as the next word was converted to letters (Fine *et al.* 2013). A yes/no grammaticality judgment followed all experimental and filler sentences, as in AFX's experiment.

Again, like in AFX's study, the participants had some practice sentences first. In my case, these consisted of four sentences without ellipsis –two were grammatical and two ungrammatical due to subject-verb agreement problems. In this initial practice stage, participants received feedback after each item, which

told them whether their response was correct or not. In the latter case, they were asked to pay more attention and, in the former, they were congratulated. It is unclear whether having four practice trials with feedback, out of which two are grammatical and two are not, reproduces AFX's study, as they do not provide information about what this practice stage consisted of.

Similarly, given the lack of information about the experimental design, it is unclear to what extent my study replicated theirs with respect to the procedure. More concretely, AFX do not report, for example, how many items each participant saw, or whether items from different sets were used for each condition or not. Along the same line, the specific instructions participants received were not provided, although I was able to obtain an approximation of such instructions from the first author of the original paper (Helena Aparicio) via a personal communication. The participants in AFX's experiment were asked, after reading each sentence, whether the item was grammatical. Given that the technical meaning of "grammatical" differs greatly with respect to what non-linguists tend to intepret as "grammatical" (for instance, the expressions that are considered "correct" according to some norm, such as that prescribed by the *Real Academia Española* in the case of Spanish), I opted to formulate the instructions for my study in a more detailed and precise way (Schütze 1996, Tremblay 2005). In addition to explaining the mechanics of the self-paced reading task, these specified that following each sentence they would have to answer whether it sounds good or bad and emphasized how this does not amount to what is considered prescriptively correct or incorrect in the language. On a related note, the reason for providing feedback in the practice trials was that subjects could see those items as examples of sentences that "sound good" versus sentences that "sound bad".

2.2 Results

AFX examined the critical region (CR), defined as the first region in which the morphological feature information on the antecedent becomes relevant for the ellipsis condition. While the CR was the word *también* (sentence-final, for them) in the elliptical conditions, the copula plus the adjective following it were considered the CR for the grammatical non-elliptical cases. Their fixation data were analyzed according to four types: First Fixation duration, First Pass duration, Regression Path duration, and Total duration and using mixed effect linear regression models. AFX also employed a mixed effect logistic regression model for the grammaticality judgments. All models included the factors Match (either Match or Mismatch), SSC (either Marked (plural or feminine) or Unmarked (singular or masculine)), and Feature (either Number or Gender), as well as Subject and Item as random effects.

Since the experimental method used in the replication differs, the fixation data that seem the most comparable to that of an online self-paced reading task are those of the Total Time measure. AFX's results from the TT measure indicate a main effect of Match for the elliptical conditions ($\beta = 0.08,\ t = 3.77,\ p < 0.0001$). This points to the fact that sentences with mismatched features incurred in higher processing costs than did the elliptical sentences with matched features, assuming underlyingly that increased reading time reflects processing effort. The same trend seems to be observed for both their number and gender sets, so the interaction between Match and Feature does not seem to be significant, and neither is the interaction between Match and SSC. These results from AFX can be visualized in the following figure.

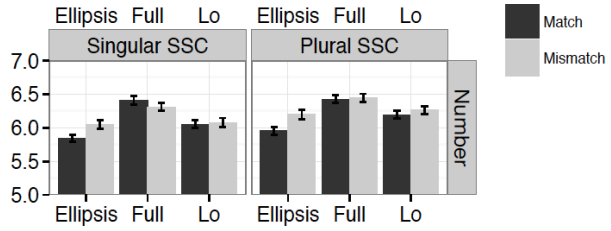

Figure 1: AFX's results for TT, CR

For the experiment I conducted, the factors to consider were Match and SSC (given that my replication covered only the number set of AFX's study, the Feature factor was always Number). Moreover, I considered the factor Type (either critical (that is, the elliptical sentences) or control (that is, the non-elliptical counterparts)). Although AFX included these non-elliptical instances, they considered them fillers and did not analyze them.

The trend observed by AFX with respect to the increased reading time in the mismatched elliptical cases was not replicated by my experiment. The first thing to note is that the interaction between the three factors mentioned (i.e., Match, SSC, and Type) is not relevant, as shown in the following figure.

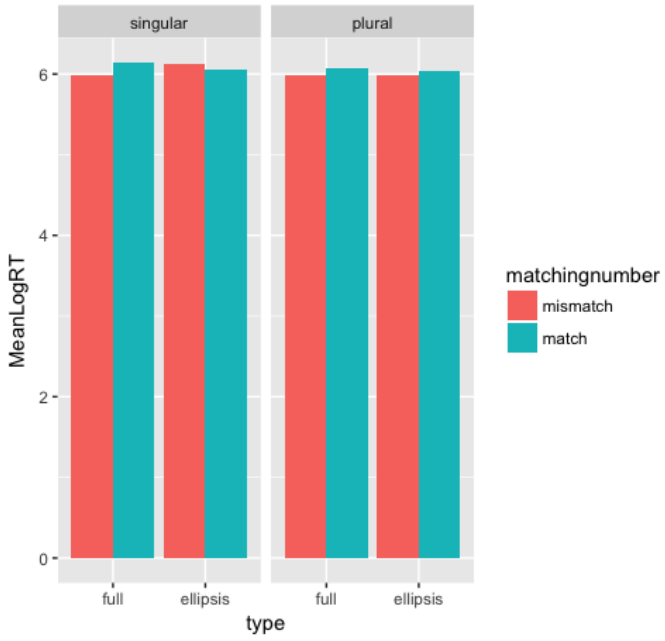

Figure 2: Non-significant interaction between Match, SSC, and Type for CR RTs

The lack of relevance of SSC was also pointed out by AFX, who observe that no significant effects for the interaction between Match and SSC were found. Therefore, I exclude SSC from the following visualizations.

Second, according to my results, the relevance of Match is only marginal. Including now error bars as well, Figure 3 displays the interaction between Match and Type, by the mean reading times, log transformed.

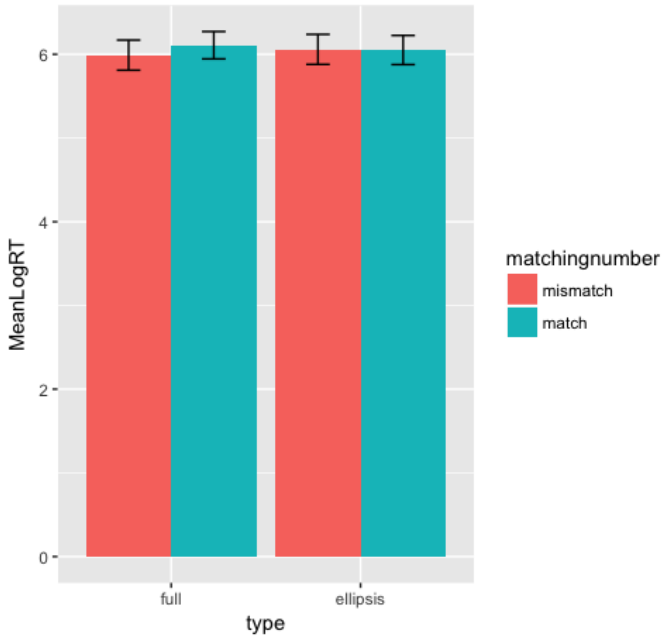

Figure 3: Marginal effect of Match for CR RTs

Given the online nature of this study, one major issue was controlling for the actual native Spanish speaker status of the participants. It could be plausible, in this context, that the reason for the subtle results and very marginal effects is related to the quality of the participants. In order to test if this was so, I filtered out those participants who either stated that their language of preference was English (as opposed to Spanish, both English and Spanish, and English or Spanish depending who they are talking to) or they stated that between their birth and five years old the language spoken in their home was English[3]. These participants were assigned the label "bad subjects", given their doubtful status as native speakers of Spanish. However, as can be seen in Figure 4, the results remain similar to those in Figure 3: there seems to be almost no difference with respect to reading times when comparing matched and mismatched cases, especially in the ellipsis condition.

[3]This information was extracted from a questionnaire that subjects were required to fill out at the end of the experiment. This questionnaire included questions mainly about their linguistic background, such as their preferred language, the language they heard at home between the ages of zero and five, their country of birth, their parents' country of birth, and their country of residence, among others.

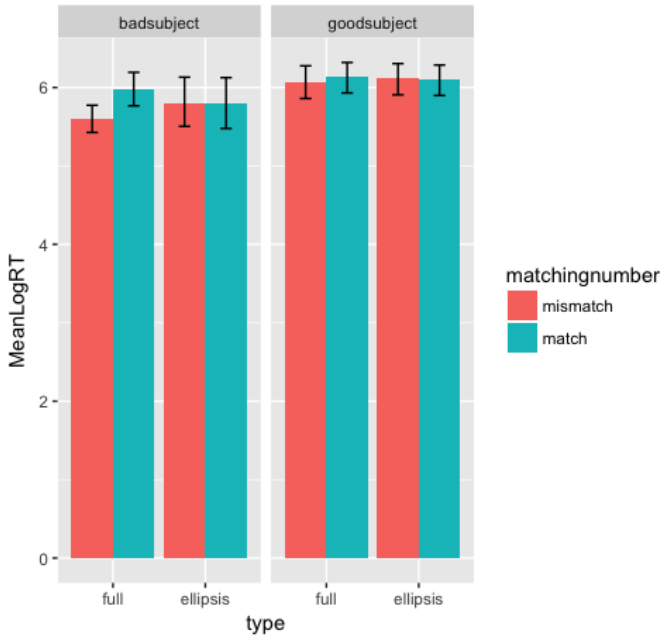

Figure 4: Marginal effect of Match with subject quality for CR RTs

Instead of using this criterion for distinguishing between "good" and "bad" subjects, the extent to which the participants' grammaticality judgments were adequate can be considered the relevant criterion[4]. But, again, the data do not change in any significant way.

All of these models were run considering the mean of the log-transformed reading times for the critical item (i.e., *también*). However, it is possible that some relevant effects are observed in the spillover region that follows. Figure 5 shows, nevertheless, that this does not seem to be the case: the reading times for the item following the critical word (*también*)[5] essentially mirror those for the critical word, if we compare Figure 5 with 4.

[4]This is possible because the grammaticality judgment for all the stimuli are very clear and uncontroversial. AFX ran a mixed effect logistic regression model for the grammaticality judgments and report that participants preferred sentences with matching features in elliptical conditions.

[5]The item following the critical word was always a connector in my stimuli. Four different connectors were used throughout the experiment (each set containing one of these four) and their word length was not uniform (*entonces* is one word long, *por eso* and *así que* are two words long, and *por lo que* is three words long). Hence, I would need to account for these differences in my model. However, for present purposes, I will limit my references to the spillover region to the first item following the critical word.

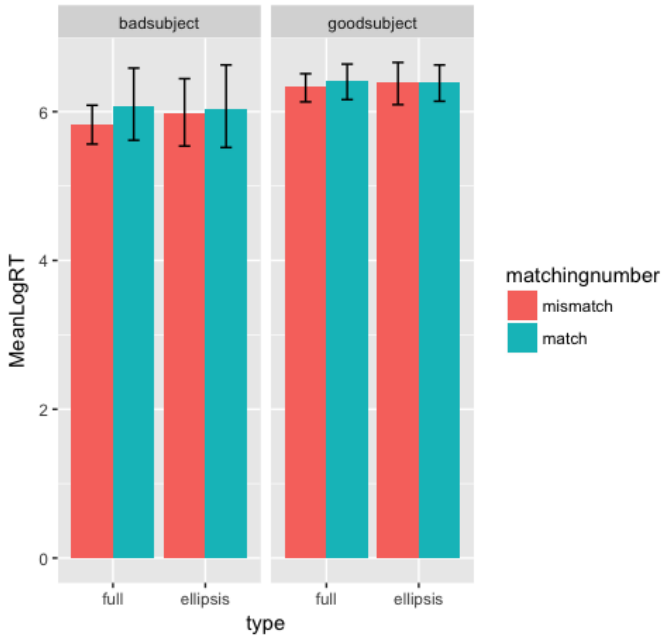

Figure 5: Marginal effect of Match with subject quality for CR + 1 RTs

3 Discussion

In this paper I tested the hypothesis that morphological mismatches in ellipsis, although grammatical in some languages, incur in additional processing cost. This was the motivation for AFX's 2014 study, who conducted two experiments which led them to the conclusion that this is indeed the case, at least in (Iberian) Spanish. Based on their eye-tracking experiments, these researchers observed that the instances of mismatches in ellipsis conditions led to an increase in the subject's reading time at the critical region, when compared to the matched versions of those same elliptical sentences.

Unlike AFX's study, my replication painted a much more puzzling picture, which might cast doubt on the claim that mismatches are relevant for ellipsis computation. However, my suspicion is that this difference may have to do with aspects of my experimental design. One of the main issues is the fact that I collected my data from Mechanical Turk, which, although has the advantage of recruiting large numbers of participants almost instantly, it makes it extremely hard to control for the quality of subjects that participate in the task. In this specific case, there was no straightforward way to allow only Spanish native speakers to participate, as linguistic background is not one of the filters MTurk has available. This is especially important in an experiment like this one, where the alternation being tested is fairly subtle and also is not a feature of English. In order to better control for this, a pre-screening survey could be designed, including text as well as audio, to test the subject's command of the language. Manually supervising the results of the pre-screening and assigning a "Spanish native speaker" qualification to those who pass it would create a pool of suitable participants. Of course, the problem is then to make sure that those who obtain the qualification participate in the actual experiment.

A second aspect of the design that might have affected the results has to do with how many items were presented to each participant and, especially, with whether different items were used for the different conditions. In my experiment, each participant was (randomly) assigned a set of ten sentences, eight of

which were variants of the same item but slightly modified as required by each condition (the combinations of Match, SSC, and Type). This means that it is possible that reading times decrease overall for each participant after the first few sentences had been shown. Unfortunately, AFX do not provide information with respect to how they handled these aspects of their experimental design.

In conclusion, this partial replication did not replicate the original results. However, I do not take this is an indication that AFX's claim is on the wrong track. In my opinion, this is merely an example of the importance of being explicit and clear in experimental papers. If the goal is for experiments to be replicable, it becomes crucial to be as precise as possible in contexts where the results depend on a combination of multiple factors organized in a certain specific way. Otherwise, vagueness and ambiguities when reporting experimental data can –and likely will– lead to dissimilar, perhaps even opposite, results.

4 References

Aparicio, H., K. Franich, & M. Xiang (2014), Morphological Feature Mismatches under Ellipsis: An Eye-tracking Study. *Proceedings of NELS*, 45.

Fine, A., T. F. Jaeger, T. A. Farmer, & T. Qian (2013), Rapid Expectation Adaptation during Syntactic Comprehension. *PLoS Online*, 8, 10.

Depiante, M. & P. Masullo (2001), Genero y Numero en la Elipsis Nominal: Consecuencias para la Hipotesis Lexicalista. Paper presented at *I Encuentro de Gramatica Generativa*, Universidad Nacional del Comahue.

Lasnik, H. (1999), Verbal Morphology: *Syntactic Structures* Meets the Minimalist Program. In H. Lasnik (ed.), *Minimalist Analysis*,

Merchant, J. (2013), Diagnosing Ellipsis. In L. Lai-Shen & N. Corver (eds.), *Diagnosing Syntax*, Oxford: Oxford University Press, 537-542.

Merchant, J. (2014), Gender Mismatches under Nominal Ellipsis. *Lingua*, 151, 9-32.

Merchant, J. (2016), Ellipsis: A Survey of Analytical Approaches. In: J. van Craenenbroeck & T. Temmerman, *A Handbook of Ellipsis*, Oxford: OUP.

Nunes, J. & C. Zocca (2005), Morphological Identity in Ellipsis. In N. Elouazizi *et al.* (eds.), *Leiden Papers in Linguistics*, 2.2, 29-42.

Nunes, J. & C. Zocca (2009), Lack of Morphological Identity and Ellipsis Resolution in Brazilian Portuguese. In J. Nunes (ed)., *Minimalist Essays on Brazilian Portuguese Syntax*, Amsterdam/Philadeplhia: John Benjamins, 29-42.

Schütze, C. (1996) *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistics Methodology*. Chicago: The University of Chicago.

Tremblay, A. (2005) Theoretical and Methodological Perspectives on the Use of Grammaticality Judgments Tasks in Linguistic Theory. *Second Language Studies*, 24, 1, 129-167.