The complex structure of agreement errors: Evidence from distributional analyses of agreement attraction in Arabic[*]

Diogo Almeida[1] & Matthew A. Tucker[1,2]

NYU Abu Dhabi[1] & Oakland University[2]

1. Introduction

Potentially unbounded long-distance dependencies such as subject–verb agreement present a challenge for the language processing system during comprehension, as the formal features of the trigger of the agreement (the subject) need to be successfully resolved at the target site (the verb), but the distance between the two elements is unpredictable in the input. Indeed, that is likely one of the reasons why agreement dependencies can often undergo illusory licensing wherein a *prima facie* ungrammatical string is either accepted or produced by speakers (Bock & Miller 1991):

(1) . . . [T]he warm glow of bygone days **are** collected and safeguarded for all time.[1]

Known as AGREEMENT ATTRACTION, this phenomenon is usually characterized by the presence of a DISTRACTOR DP (*bygone days*, in the example above) which appears in a syntactic position not typically accessible for agreement yet which nevertheless seems to completely or partially license a verb with $\varphi-$features which do not match the correct subject. These errors have been hypothesized to stem from selective failure to construct the correct hierarchical structure for the sentence (Eberhard 1997), or selective failures in working memory retrieval during syntactic processing (Wagers et al. 2009).

Interestingly, despite often being described as by and large a language performance phenomenon, agreement attraction effects in comprehension seem to obey certain structural regularities. One is the GRAMMATICALITY ASYMMETRY: Attraction effects are observed significantly more often when the subject–verb dependency is ungrammatical then when

[*] Thanks to Souaad Al Helou for data collection and help with stimuli creation; to Ali Idrissi for help with stimuli creation as well as with assisting with the design of the original experiment; and to Ellen Lau, Bern Samko, Kevin Schluter, an audience at Yale University, the audiences at NELS 47, and two anonymous NELS abstract reviewers for comments on previous stages of this project.

[1] National Baseball Hall of Fame & Museum. 2014. "This is the Baseball Hall of Fame," television commercial. `https://youtu.be/zCfFe5_kqig`.

it is grammatical (Wagers et al. 2009). The other regularity is the MARKEDNESS ASYMMETRY: Attraction effects are significantly more common when the true subject bears an unmarked $\varphi-$feature and the distractor and verb match for its marked alternant. Thus, attraction is more robust in configurations such as (2a) as opposed to (2b), where it appears very infrequently or not at all (Eberhard 1997).

(2) a. The key to the cabinets **are** rusty from years of disuse.
b. The keys to the cabinet **is** rusty from years of disuse.

The structural regularities that impinge on the processing of subject–verb agreement are important, as any plausible account for them presumably includes some interplay between language–specific structure and language–independent performance models.

1.1 The research question

In order to better characterize these structural asymmetries in agreement processing, here we investigate the process of subject–verb agreement of two different $\varphi-$features in a language that has verbal agreement for both gender and number, Modern Standard Arabic (MSA).

Agreement attraction in comprehension is routinely investigated via self-paced reading studies, where attraction surfaces as a facilitation on mean reading times to ungrammatical verbs relative to ungrammatical controls in non-attraction configurations (Wagers et al. 2009 and references therein). These results are typically analyzed as differences in the means of reaction time (RT) distributions at the critical regions (generally the verb or its spill–over region). This is potentially problematic for a number of reasons. The first is that the mean is a non–robust measure of central tendency when it comes to outliers, and RT distributions are positively skewed with long right tails, which interferes with the identification of potential outliers (Ratcliff 1993). More importantly, RT distributions are known to be potentially differentially affected at their modal and right tail portions, depending on the experimental design or cognitive process of interest. Since both kinds of change will affect the mean of the RT distributions, the specific reason for a difference in means is hard to discern using traditional analysis methods (Balota & Yap 2011). Crucially, recent studies employing more fine–grained distributional analyses of RT in agreement processing suggest that it is the right tail of the RT distribution, rather than its modal portion, that is shifted in attraction configurations (Lago et al. 2015).

Given that agreement attraction effect sizes in self–paced reading studies are relatively small (22 ms, according to a recent meta–analysis in Jäger et al. 2017), it is particularly important to better characterize the nature of attraction effects by means of more fine–grained distributional analyses. Unfortunately, it is extremely challenging within a self-paced reading study to collect enough within-subject observations to support traditional analyses of RT data. Here we take a different approach by presenting two distributional analyses of group–level RT distributions. These analyses are based on two large data sets (each with $N = 330$) of self–paced reading studies of Modern Standard Arabic (MSA), one focusing on subject–verb number agreement and the other on subject–verb gender agreement.

The first analysis involves a traditional approach of fitting ex–Gaussian distributions to our group–level RT data, whereas in the second we present a novel approach to deal with the non–parametric distributional approach based on vincentiles, wherein we summarize the vincentile information by means of PRINCIPAL COMPONENTS ANALYSIS (PCA).

2. Experimental design

Our analyses are based on two experiments with identical methodologies testing gender and number agreement attraction, respectively. Both systematically manipulated attraction contexts in MSA sentences with appropriate distractors. These experiments are a subset of a larger set of studies on agreement attraction in MSA (Tucker et al. 2017), and were selected because of their large sample sizes and matching designs, as well as for evincing attraction for both number and gender subject–verb agreement.

2.1 Participants

Participants were recruited from the student population of the United Arab Emirates University in Al-Ain, UAE. The sample sizes were identical in both experiments ($N = 330$) and all participants were native speakers of Emirati, Palestinian or Sudanese Arabic who self-reported literacy in MSA and no history of language disorders (660 females; mean age 20.65 years).[2] All participants provided written informed consent and were compensated. The experiments were approved by the NYU Abu Dhabi Institutional Review Board and the United Arab Emirates University Ethics Board.

2.2 Materials

The same set of 54 sentences were used in both experiments, constructed with subject relative clauses adjoined to preverbal subjects of the general form *DP_1 — [Comp — V — DP_2 — AdvP] — Target Verb — Continuation*. Subject relative clauses were employed in order to provide a suitable distractor DP in a construction which allowed the appearance of an adjunct between the distractor and target verb (Wagers et al. 2009). DP_1 and DP_2 were animate, human-denoting nouns with masculine-feminine pairs formed with the MSA feminine suffix /-a/. All other stimuli design features were identical to previous work on Arabic agreement attraction (Tucker et al. 2015). An example sentence appears in (3):[3]

[2] The gender asymmetry in our participant population is due to the nature of postsecondary education in the United Arab Emirates: University campuses are segregated by gender and there is a 3:1 female-to-male ratio among students in the country. Data was gathered on the female campus for convenience.

[3] All experimental items and analysis figures are available at the first author's *figshare* located at `https://figshare.com/authors/Diogo_Almeida/386584`.

(3) al-mumarrid alladðii ʕaaladʒ al-mariid bi-ʕanaajat-in
the-nurse(.3MS) who(.3MS) treated the-patient(.3MS) with-care-GEN
ja-drus fii mustaʃfa al-dʒaamaʕa.
3MS-studies at hospital the-university
'The nurse who treated the patient carefully studies at the university hospital.'

The base sentences were manipulated to create eight variants in a $2 \times 2 \times 2$ factorial design in each experiment crossing three factors: (i) SUBJECT-φ, describing the value of the manipulated feature on the subject (MASC/FEM or SG/PL); (ii) MATCH, describing whether DP$_1$ and DP$_2$ had the same value of the manipulated feature; and (iii) GRAMMATICALITY, describing whether DP$_1$ and the target verb had the same value of the manipulated feature. Only the values of SUBJECT-φ differed between the number and gender experiment. All DPs in the number experiment were grammatically feminine in order to avoid confounds from broken/ablauting plural morphology in MSA (Tucker et al. 2015).

2.3 Procedure

Participants took the experiment in groups of up to eight in a computer lab. The experiment employed a self-paced non–cumulative word-by-word moving window paradigm. All items were presented in Courier New Arabic font, 28 point bold face. Following each sentence, participants were asked a simple comprehension question. Participants were instructed to read comfortably but quickly at a speed which would allow them to be accurate on the comprehension questions. The order of trial presentation was randomized by participant.

2.4 Analysis

Trials in which the participant answered the comprehension question incorrectly were removed from subsequent analysis. Because the right tail of the reaction-time distribution was central to our research question, no outlier removal or replacement techniques were employed beyond Winsorization of the extreme 1% of the RTs (Tucker et al. 2015).

2.4.1 Fitting ex-Gaussians

In each experiment, we summarize the 330 subject means for each sentence region in every condition by fitting ex-Gaussian distributions to these values. The ex-Gaussian distribution, a convolution of a Gaussian and an exponential distribution, offers very good fits to RT distributions in general (Heathcote et al. 1991, *i.a.*), and can summarize their morphology using only three parameters: μ, σ (the mean and standard deviation of the Gaussian component, which can represent the modal portion of the RT distribution), and τ (the mean and standard deviation of the exponential component, which represents the degree of skew of the right tail of the RT distribution). Despite providing useful summaries for RT distributions, however, it is unlikely that ex-Gaussian parameters can be directly linked to specific mental processes (Matzke & Wagenmakers 2009). Therefore, ex-Gaussians provide very succinct summaries for the morphology of the RT distributions, at the expense of imposing strong parametric assumptions on the analysis of the data. Here, we estimated the ex-Gaussian parameters using the method of moments as described in Heathcote (1996).

2.4.2 Summarizing vincentiles with PCA

Another way of exploring the morphology of the RT distribution is by binning the collection of RTs into pre-specified quantiles. This technique is called vincentizing, and the quantiles that it generates are sometimes referred to as VINCENTILES (Lago et al. 2015). Here we choose deciles to represent the morphology of the RT distributions. One of the strengths of the vincentile approach is that it provides a view of the similarities and differences between the shape of RT distributions that is free of any parametric assumptions. However, vincentiles provide poor data summaries for statistical inferential purposes, forcing the researcher to tackle a multiple comparisons problem (Lago et al. 2015).

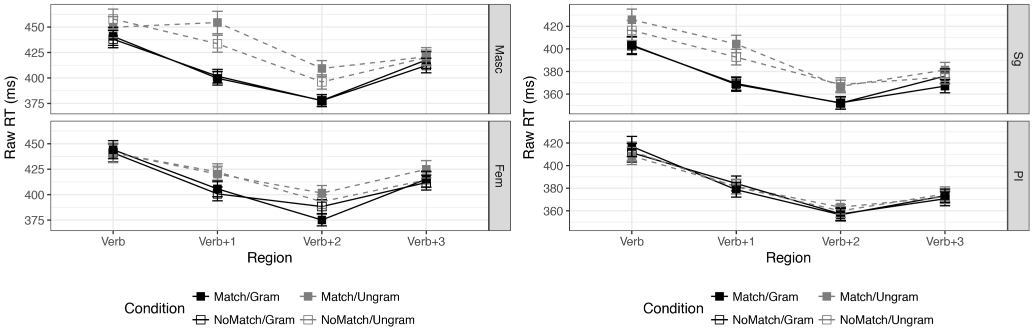
Here, we propose a new approach to the analysis of vincentiles for the study of RT distributions. Since the quantiles of RT distributions are all heavily correlated, these correlations can be exploited by PCA, an exploratory multivariate data analysis method. At an intuitive level, PCA finds the orthogonal dimensions of largest variation in a multivariate dataset (in our case, deciles of RT distributions), and summarizes these dimensions into weighted linear combinations that can be interpreted as synthetic variables, organized in decreasing order of importance. In the case of RT distributions, the expectation is that one or two of these uncovered synthetic variables will correspond to the modal, right tail or both portions of the distributions.

In addition, we take advantage of the PCA framework to construct a decile-shape space based on the RT distribution of every region except the critical ones (*Verb*, *Verb+1*, *Verb+2*). Since the critical regions did not contribute to the construction of this decile-shape space, we can check how well their morphologies can be reconstructed on the basis of the morphologies of the other RT distributions. This can be achieved by projecting the critical region deciles onto the PCA space as supplementary variables, and retrieving their factor scores (*i.e.*, how well correlated they are with each principal component).

3. Results

The figures in (4) show the grand average reading times to the critical verb and three immediately postverbal regions (error bars represent the standard error of the mean computed over subject averages). For reasons of space, we do not provide a formal statistical analysis of the raw reading times in this work (see Tucker et al. (2017) for details), but attraction effects are observed in the gender study at the *Verb+1* and *Verb+2* regions when subjects are masculine, but not when they are feminine. In the number study, we observe comparatively smaller attraction effects at the *Verb* and *Verb+1* regions when the subjects are singular, but not when they are plural. In summary, the attraction effects were smaller and earlier in the number study compared to the gender study. Moreover, we observed a markedness asymmetry effect, as attraction only occurred when the subjects were morphologically unmarked for the feature of interest (masculine for gender and singular for number). We also did not observe any attraction effect in grammatical sentences (consistent with the grammatical asymmetry effect).

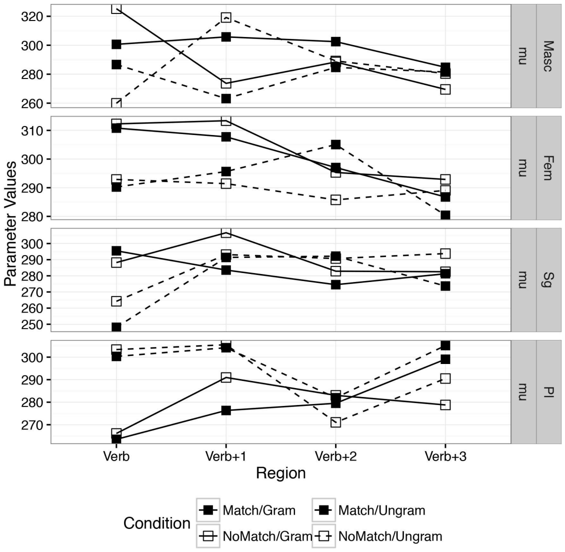
(4) *Grand Average RTs, gender (left) and number (right)*



3.1 Ex–Gaussian results

The results for the fits of μ, which correspond to the modal portion of the RT distribution, are presented in Figure (5). When compared to the raw data in figures in (4), no attraction effect is observed in any condition.

(5) *Fits of* μ *for the critical regions across experiments*

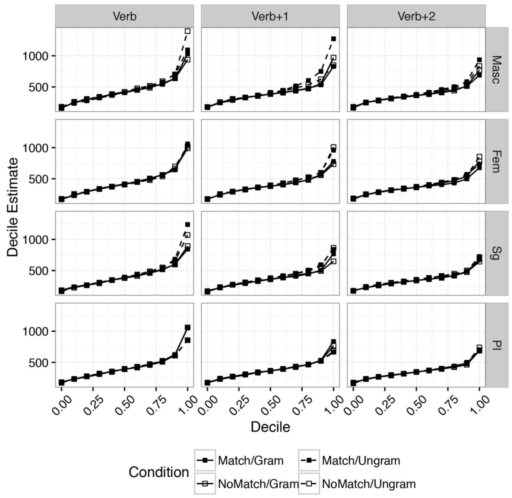


The results for τ, which represents the right tails of the distribution, are presented in Figure (8) for the gender and number studies, respectively, and we observe very similar patterns to the raw data. First, we see clear attraction effects in *Verb+1* region for gender (but only with masculine subjects) and in the *Verb* region for number (but only with singular subjects). Interestingly, the attraction effect on τ seems greatly diminished in the spillover regions in both studies (*Verb+2* for gender and *Verb+1* for number), even though an attraction effect is observed in the raw data at those regions as well.

3.2 Vincentile results

The vincentiles (estimated group–level deciles) for each condition in the verb and post–verbal regions are presented in Figure (6), where it can be seen that the differences between distributions are primarily at the last deciles (*i.e.*, the right tails of the distributions). Therefore, we expect that the first dimension of the PCA analysis on the vincentiles should reflect the high deciles. This prediction is borne out, as can be seen in Figure (7), which illustrates the Cartesian plane represented by the first two principal components (PCs) of the vincentile data.

(6) *Vincentiles of the critical regions for all conditions*



(7) *Circle of correlations from PCA on the combined histograms of the two experiments*

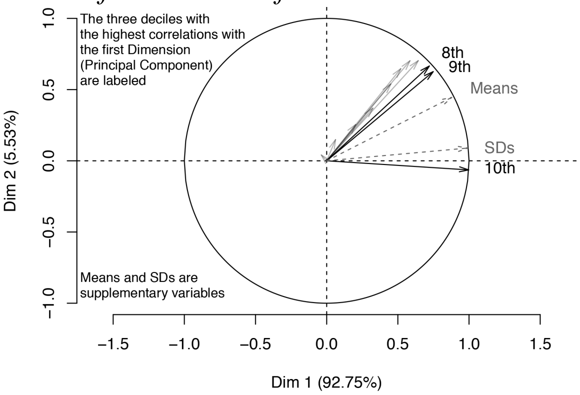
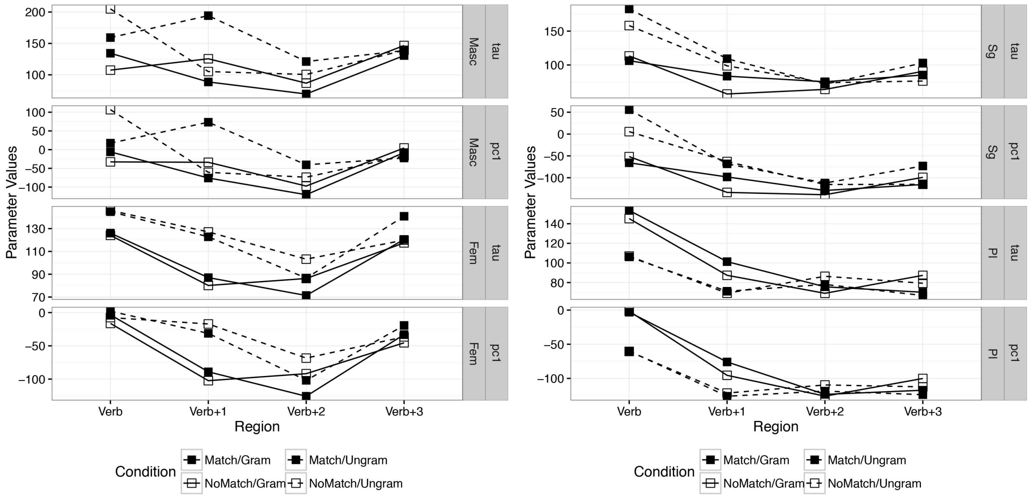


Figure (7) displays the so–called CIRCLE OF CORRELATIONS, which indicates how well our original variables (*i.e.*, our deciles) are represented onto each synthetic variable (*i.e.*, each PC) provided by the PCA. As can be seen in Figure (7), the first PC is represented by the x–axis and the second PC is represented by the y–axis. On the basis of this graph, our conjecture that the largest variance in the shapes of RT distributions occurs at the extreme of the right tail appears to be correct. The first PC accounts for almost 93% of the variance in RT deciles. The 10th decile is also almost perfectly correlated with this PC: this can be seen by the arrow representing the 10th decile lying almost perfectly on top of the x–axis and almost touching the circle, which itself indicates a perfect correlation.

The 8th and 9th deciles seem to be highly correlated with both the first and the second PCs, which strengthens our interpretation that the first PC primarily captures the variance at the right tail of the RT distribution. Finally, the means and standard deviations of the decile distributions are projected as supplementary variables (*i.e.*, they were not used to construct the PCA space), and both are highly correlated with the first PC. This means that, in this dataset, the RT distributions with the heaviest right tails are also the distributions with the larger means and standard deviations. This illustrates the problems of relying solely on the means of the RT distribution, as they can be caused either by a full shift of the distribution, or by a stretching of the right tail (the latter of which seems to be the case in this situation).

Now that we have a clear interpretation for the first PC, we can use each critical region factor score as a synthetic variable representing the weight of the right tail of the distribution, much like τ behaves in the ex-Gaussian analysis. The figures in (8) show the comparison of the fits of τ and the factor scores obtained from the first PC of the PCA vincentile analysis, and the patterns are almost identical. This result is even more striking when we take into consideration that none of the deciles from critical regions were used in the construction of the PCA space. These scores are obtained purely as a projection of new observations onto an existing decile–shape space, and yet the results are virtually identical with the results of τ in the ex–Gaussian analysis.

(8) *Fits of τ and scores of PC1 from the PCA analysis for the gender (left) and number (right) dataset*



4. Discussion

The results of our distributional analyses clearly support the notion that agreement errors are not uniform in their expression. First, the timing of the agreement attraction effects differs by $\varphi-$feature. They occurred immediately at the verb in the number manipulation and in the first spillover region for gender, though both spilled over to their subsequent region with diminished effect size.

The distributional analysis showed that this increase in the mean RT response was distributionally different across conditions as well as regions of interest. Only two conditions showed increases in the right tail of the RT distribution and these correlated with the $\varphi-$feature of the subject: right tail changes were elicited only when the subject had unmarked $\varphi-$features (masculine in the gender experiment, singular in the number experiment). Interestingly, attraction effects were observed only in exactly those conditions, suggesting a novel interpretation of the grammaticality and markedness asymmetries in which the visibility of attraction is conditional upon a right-tail effect being possible in the first place. Finally, an interesting pattern also emerges in the agreement attraction cases: Both lasted for two regions of interest, and in both the first region of the effect displayed changes in the right tail of the RT distribution, but this effect was greatly diminished or absent altogether in the second region, even though a mean RT difference is still observable in those regions.

These distributional analyses show that there are subtle and complicated patterns underlying agreement errors that go beyond simply observing mean differences across conditions. Agreement attraction, for instance, seems to rely on a complex pattern of factors. It depends on the markedness of the $\varphi-$features insofar as marked subjects do not engender an error response to ungrammatical agreement morphology which can be modulated by attraction conditions. Moreover, the error signal at or after the verb has to be a lengthening of the right tail, which is not as pronounced in the presence of a distractor DP, rather than a shift in the modal portion of the distribution. This signal occurs only in the first region where the attraction effect is observed, and any lingering spill-over effects do not seem to be predicated on differences in the right tail of the RT distribution.

In addition to this theoretical conclusion, our results also provide a methodological cautionary tale about data transformation and cleaning: The right tail of reaction-time distributions is clearly important. Outlier handling methods such as trimming, exclusion, or Winsorization alter this right tail and are potentially distortive of results. The same is true of the log transformation of RT data, which is often used to approximate it to a normal distribution. Here, this transformation would have prevented us from observing any of the attraction findings we report. Future research in psycholinguistics must therefore remain attentive to the source of a shift in RT means between two conditions, as this shift could be the result of a shift in either central tendency or the right tail of the distribution, and depending on which situation is being investigated, seemingly simple data preprocessing decisions may impact qualitative results.

While our findings support an analytical decomposition of multiple sources of ungrammaticality and $\varphi-$features, the discussion in this paper has left the formal underpinning of these findings deliberately underspecified. While much previous literature suggests that the markedness asymmetry is a product of attraction configurations *per se*, our results suggest that markedness may interact more strongly with simple agreement errors than with attraction specifically. Thus, future research must focus not only on developing articulated models of memory search during processing, but also articulated models of how grammatical information such as markedness is represented during processing. Only then can one understand how the *a priori* availability of simple agreement errors precludes the availability of agreement attraction.

References

Balota, David A, & Melvin J Yap. 2011. Moving beyond the mean in studies of mental chronometry the power of response time distributional analyses. *Current Directions in Psychological Science* 20:160–166.

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

Eberhard, Kathleen M. 1997. The marked effect of number on subject–verb agreement. *Journal of Memory and Language* 36:147–164.

Heathcote, Andrew. 1996. RTSYS: A DOS application for the analysis of reaction time data. *Behavior Research Methods* 28:427–445.

Heathcote, Andrew, Stephen J. Popiel, & D.J. Mewhort. 1991. Analysis of response time distributions: An example using the Stroop task. *Psychological Bulletin* 109:340–347.

Jäger, Lena A, Felix Engelmann, & Shravan Vasishth. 2017. Similarity-based interference in sentence comprehension: Literature review and bayesian meta-analysis. *Journal of Memory and Language* 94:316–339.

Lago, Sol, Diego E. Shalom, Mariano Sigman, Ellen F. Lau, & Colin Phillips. 2015. Agreement in Spanish comprehension. *Journal of Memory and Language* 82:133–49.

Matzke, Dora, & Eric-Jan Wagenmakers. 2009. Psychological interpretation of the ex-gaussian and shifted wald parameters: A diffusion model analysis. *Psychonomic bulletin & review* 16:798–817.

Ratcliff, Roger. 1993. Methods for dealing with reaction time outliers. *Psychological Bulletin* 114:510–532.

Tucker, Matthew A., Ali Idrissi, & Diogo Almeida. 2015. Representing number in the real-time processing of agreement: Self-paced reading evidence from Arabic. *Frontiers in Psychology* 6.

Tucker, Matthew A., Ali Idrissi, & Diogo Almeida. 2017. Attraction effects for verbal gender and number are similar but not identical: Self-paced reading evidence from Modern Standard Arabic. Under review.

Wagers, Matthew W., Ellen F. Lau, & Colin Phillips. 2009. Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language* 61:206–237.

Diogo Almeida, Matthew Tucker
diogo@nyu.edu, matthewtucker@oakland.edu