Fine-grained gender typicality systematically modulates anaphora resolution: Evidence from eye movements

Vishal Arvindam (New York University) & Brian Dillon (UMass Amherst), arvindam@nyu.edu

Introduction: Prior work[1-3] has shown that comprehenders are sensitive to gender typicality, or the likelihood that a role noun (e.g., *nurse*) is associated with male or female referent, during anaphora resolution. Specifically, comprehenders incur a processing cost when reading anaphors that mismatch the gender typicality of their antecedents (e.g., reading *himself* following *nurse*). Surprisal theory leads us to expect that comprehenders have precise estimates of a word's probability in context, even at the low end of the probability range[4]. This view can explain sensitivity to gender mismatch on anaphora as unexpected wordforms given referential context, but the model makes the stronger prediction: that the mismatch effect should be fully continuous across the range of gender typicalities. Previous work[5] that manipulated a fully gradient sample of typicalities with gender marked (e.g., *him*, *her*) and neutral pronouns (i.e., singular *they*) did not find any reliable effect of typicality. The current study extends this work by using reflexives, to disambiguate the referent, and eye-tracking, for greater temporal resolution. In doing so, we test the strong prediction of a surprisal based view.

Methods: 42 college-aged adults (M_{age}= 19) estimated the extent to which a role noun consisted of women and men (i.e., the ratio) on an 11-point rating scale ranging from 0% women and 100% men on one end to 100% women and 0% men on the other. From these norms, 30 frequency matched nouns were selected that ranged in typicality from being male (0.0) to female (1.0) in increments of 0.05 (Figure 1). Two nouns were chosen for each point on this continuum. In addition, 15 definitional nouns (e.g., *boy*) were included at either end of the continuum for a total of 45 items. We monitored participants' eye-movements (N = 48, M_{age}= 21) as they read sentences similar to "**The nurse** embarrassed /**himself/**$_{Critical}$ /**while/**$_{Spillover}$ ..", which manipulated the gender typicality of the antecedent (0.0_{female} - 1.0_{female}) and the reflexive pronoun (*himself/herself/themselves*). If readers are sensitive to graded gender typicality, we expected reading times at the reflexives to covary with gender typicality. Log-transformed reading times were analyzed using a linear mixed effects (LME) model with *gender typicality* as a fixed-effects factor and *participant* and *item* as random-effects. For *himself* and *herself, gender typicality* was used a linear predictor, but for *themselves, gender typicality* was also used as a quadratic predictor since we expected *themselves* to be harder with more strongly biased antecedents in either direction. The effect of *gender typicality* is presented for first fixation, first pass, and go-past times at both the critical and spillover regions. The results from the LME effects model are presented in Table 1, and the data are visualized in Figure 2.

Results: At first fixation, first pass, and go-past measures for *himself* and *herself*, gender typicality significantly modulated reading times. We also saw a sharp increase in reading time for mismatched definitional nouns, starting at the first fixation for *himself*, and in go-past times for *herself*. For *themselves,* there was a reading time slow down across all levels of typicality. However, in go-past at the spillover, *gender typicality,* when used as a quadratic predictor, significantly modulated reading time on *themselves*. The data reveal that comprehenders are immediately sensitive to fine-grained gender typicality during anaphora resolution. The sharp increase in reading times for definitional nouns suggests an additional penalty for mismatching in definitional gender above and beyond stereotypical gender mismatch. In the case of *themselves,* although readers incur an early cost, they overcame it when the antecedent had no gender bias, suggesting *themselves* was recognized as a gender neutral-pronoun for singular antecedents in online processing, albeit at a delay.

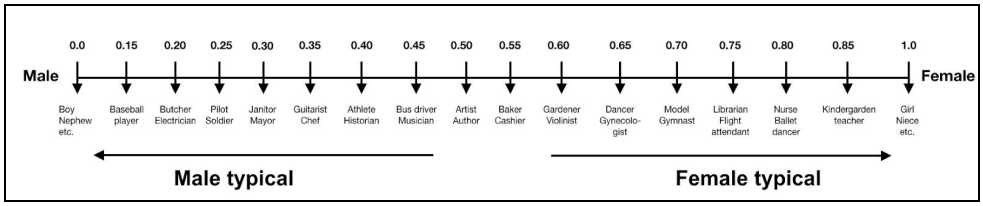

Figure 1: Sample of nouns ranging in gender typicality from 0.0 (male) to 1.0 (female) in increments of 0.05. Nouns beyond 0.85 and less than 0.15 were not included as they were rarely estimated by participants.

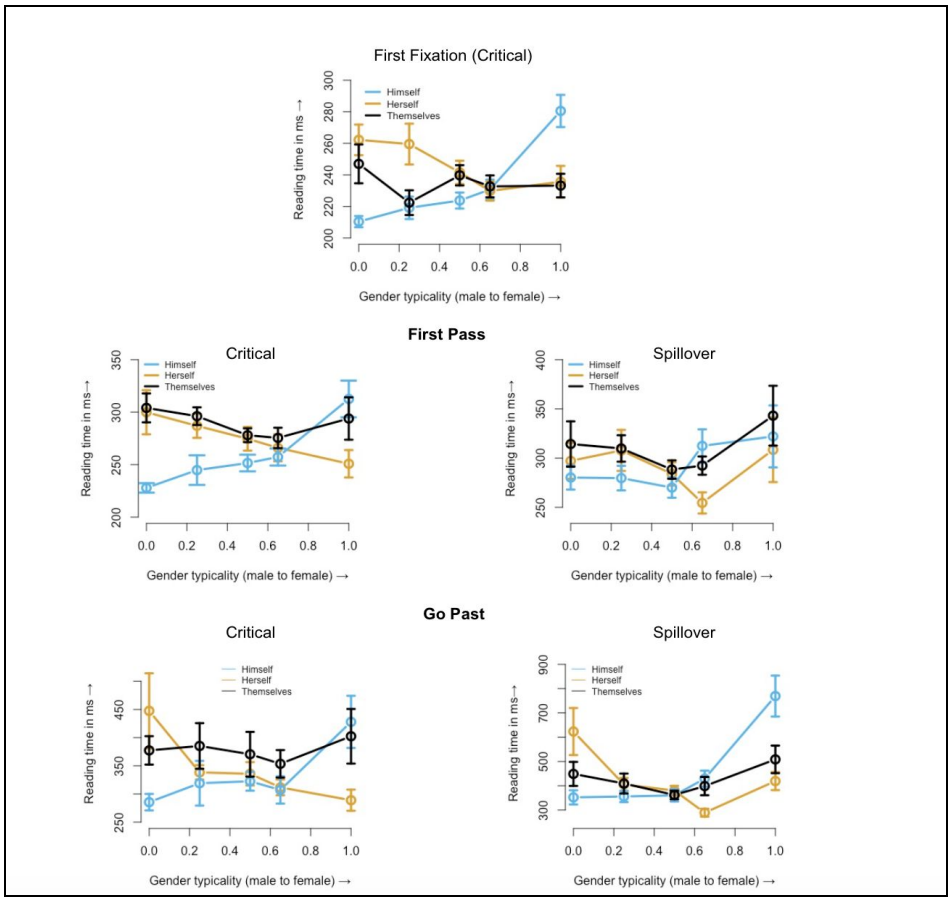

Figure 2: Mean first fixation, first pass, and go-past at the critical and spillover regions grouped into bins by bias. '0' indicates nouns that are definitionally male (e.g. boy) and '1' indicates nouns that are definitionally male (e.g. girl) with nouns that increase in typicality from male to female in between. The error bars are the standard error of the mean by items within each bin.

Reading Measure	Region	Condition			
		Himself	*Herself*	*Themselves (linear)*	*Themselves (quadratic)*
First Fixation	*Critical Region*	$\beta = 0.17(\pm 0.04)$, t = 4.31***	β = -0.11 ($\pm$0.04), t = -2.90**	β =0.01 ($\pm$0.04), t = 0.37	$\beta = 0.1(\pm 0.11)$, t = 0.85
First Pass	*Critical Region*	β = 0.20 ($\pm$ 0.05), t = 4.38***	β = -0.14 ($\pm$ 0.05), t = -2.6*	β= -0.01 ($\pm$ 0.05), t = -0.2	β= 0.23 ($\pm$ 0.16), t = 1.42
	Spillover	β= 0.20 ($\pm$0.05), t = 4.38***	β= -0.14 ($\pm$ 0.05), t = -2.6*	β= 0.03 ($\pm$ 0.06), t = 0.49	β= 0.35 ($\pm$ 0.19), t = 1.88.
Go Past	*Critical Region*	β=0.25 ($\pm$0.09), t = 2.9**	β= -0.21 ($\pm$0.07), t = -2.8**	β= 0.04 ($\pm$0.09), t = 0.44	β= 0.26($\pm$0.28), t = 0.92
	Spillover	β=0.25 ($\pm$0.09), t = 2.9**	β= -0.21($\pm$0.07), t = -2.8**	β= 0.07($\pm$0.1), t = 0.7	β= 0.87 ($\pm$0.27), t = 3.26**
Total Time	*Critical Region*	β= 0.44 ($\pm$0.08), t = 5.2***	β= -0.36 ($\pm$0.09), t = -4.00***	β= -0.01 ($\pm$0.09), t = -0.17	β= 0.5 ($\pm$0.23), t = 2.16*
	Spillover	β=0.44 ($\pm$0.08), t = 5.2***	β= -0.22 ($\pm$0.11),t = -2.05*	β= 0.02 ($\pm$0.09), t = 0.17	β=0.57 ($\pm$ 0.26), t = 2.15*

Table 1: Summary of LME coefficient estimates, standard errors (in parentheses) and associated t-values for first fixation, first pass, and go-past reading times at the critical region and spillover. Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 ' '

References: [1] Kreiner, H., et al. 2008. *JML*, 58 (2), 239-261. [2] Duffy, S.A, & Keir, J.A, 2004. *Cognition, 32(4),* 551-559. [3] Kennison, S.M., & Trofe, J. L. *J Psycholinguist Res.* [4] Smith, N. J., & Levy, R, 2013. *Cognition*, *128*(3), 302-319. [5] Boyce, V., et al., 2018. *Poster at CUNY 2018*, Davis CA