

Gender prominence across text genres: a corpus study on English, French and Spanish

Yanis da Cunha

University of Graz

yanis.da-cunha@uni-graz.at

A key observation about prominence features (animacy, definiteness, pronominality...) lies in their association with syntactic functions: For example, across languages, animate, definite and pronominal arguments are more likely to be subject, while inanimate, indefinite and nominal arguments tend to be objects [9][12]. Few studies report such an effect for referential gender (*ie.* the linguistic distinction between female and male referents [10]). Looking at syntax examples in linguistics textbooks and papers, several studies found a higher frequency of female objects and male subjects in English [4][11][13] and French [16]. However, this result is limited to a very specific text genre. Adopting an experimental approach, Esaulova (2015) reports a processing preference for female objects and male subjects in French and German, suggesting again that gender associates with syntactic functions [8].

In this context, we conducted a crosslinguistic and cross-genre corpus study on English, French and Spanish. Our first goal is to extend previous results taking into account a larger and more diverse sample, including another language to the current picture (Spanish). We also aim at exploring several explanatory factors of gender prominence. We combined 14 existing corpora: FTB, Contemporary Frantext, FrWac, Democrat, C-Oral-Rom, CFPP, CRFP for French; EWT, Lines, GUM, Partut, PUD for English; and AnCora, GSD, PUD for Spanish. All corpora include a UD annotation [6], except FrWac and Frantext for which we automatically parsed a random sample using Stanza [15]. These corpora show a large diversity of written and spoken text genres: fiction, non fiction, web, newspaper, Wikipedia, interviews, conversations, talks. We added several layers of annotation on top of the dependency annotations: animacy (automatic annotation in French with Flexique [5], manual annotation in Spanish and English), gender (manual annotation in English), semantic roles (manual annotation in French). We extended these annotation to pronouns using pre-existent coreference annotations in Democrat, GUM and AnCora corpora [14]. Our final corpus contains 5 070 654 tokens, of which we extracted only subject/object nouns and 3rd person personal pronouns featuring animacy/gender annotations. To avoid discrepancies between referential and grammatical gender in French and Spanish, we removed coordinations, masculine plurals and lemmas of epicene nouns (*eg.* Fr. *personne* ‘person’, *individu* ‘individual’, Sp. *persona* ‘person’, *gente* ‘people’). This yielded to a sample of 208 809 data points. Data are analyzed with Bayesian Poisson regression models with the *brms* package on R [3].

We report a consistent prominence effect of referential gender (Figure 1): across languages and text genres, female arguments are more often objects and male arguments subjects ($E(\beta_{\text{Male:Subj}}) = 0.34$, $\text{CrI} = [-0.27, 0.85]$, $pd = 0.93$). In grammatical-gender languages, inanimate (pro)nouns do not exhibit any effect ($E(\beta_{\text{Masc:Subj}}) = 0.05$, $\text{CrI} = [-0.75, 0.87]$, $pd = 0.68$). The effect is also seen in English, a language lacking grammatical gender. These two facts support the hypothesis that referential gender (female/male), not grammatical gender (feminine/masculine), is at play in prominence effects. Nouns and pronouns do not show any significant difference ($E(\beta_{\text{Masc:Subj:Pronoun}}) = 0.1$, $\text{CrI} = [-0.38, 0.69]$, $pd = 0.74$). The presence of the gender prominence effect is robust to text genre variation, but its effect size varies. We show that it correlates with the male-over-female ratio: the more a text contains male referents, the more they will be attracted to the subject function ($\rho(\beta_{\text{Male}}, \beta_{\text{Male:Subj}}) = 0.83$, $\text{CrI} = [0.4, 0.96]$, $pd = 0.99$). We also report a difference between female and male speakers: women show a more balanced usage, while men are biased towards male subjects ($E(\beta_{\text{Masc:Subj:Men}}) = 0.2$, $\text{CrI} = [-0.07, 0.46]$, $pd = 0.93$). Finally, we considered semantic roles to disentangle semantic and syntactic prominence. We crucially show that the prominence effect goes beyond semantics, as it is found among patient-like roles (*eg.* passive subjects, $E(\beta_{\text{Masc:Subj}}) = 0.38$, $\text{CrI} = [0.15, 0.59]$, $pd = 0.99$).

Our study supports the hypothesis that gender is a prominence feature [8]. This crucially predicts that some languages can grammaticalize its effect in alternation phenomena [2][9][12]. Gender prominence could also be counted as a gendered linguistic structure [10], alongside masculine resolution in agreement or male-first order in coordinations. We found correlations between gender prominence, female/male frequency and speaker gender. This suggests that gender prominence is driven by both social factors, like in-group empathy and androcentrism, and discourse factors: overmentioned male referents access to the subject function more easily, and overfrequent male subjects make better topics for subsequent sentences [1]. A comparison between gender and other prominence features remains to

be carried out, especially since the gender effect we found seems to be smaller than animacy or definiteness effects.

References:

- [1] Brennan, S. E. (1995). Centering attention in discourse. *Language and Cognitive Processes*, 10(2), 137–167.
- [2] Bresnan, J., Dingare, S., & Manning, C. D. (2001). Soft constraints mirror hard constraints: Voice and person in English and Lummi. In M. Butt & T. Holloway King (Eds.), *Proceedings of the LFG01 Conference* (pp. 13–32). CSLI Publications.
- [3] Bürkner, P.-C. (2017). *brms*: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80, 1–28.
- [4] Cépeda, P., Kotek, H., Pabst, K., & Syrett, K. (2021). Gender bias in linguistics textbooks: Has anything changed since Macaulay & Brice 1997? *Language*, 97(4), 678–702.
- [5] Chlebowski, A., & Bonami, O. (2015). *Annotation sémantique des noms de Flexique* [Report]. Laboratoire de Linguistique Formelle.
- [6] de Marneffe, M.-C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, 47(2), 255–308.
- [7] Esaulova, Y. (2015). *The Prominence of Gender Information in On-Line Language Processing: Cross-Linguistic Evidence of Implicit Gender Hierarchies* [Thèse de Doctorat, Heidelberg University].
- [8] Esaulova, Y., & Von Stockhausen, L. (2015). Cross-linguistic evidence for gender as a prominence feature. *Frontiers in Psychology*, 6.
- [9] Haspelmath, M. (2020). Role-reference associations and the explanation of argument coding splits. *Linguistics*, 59(1), 45.
- [10] Hellinger, M., & Bußmann, H. (2015). The linguistic representation of women and men. In M. Hellinger & H. Motschenbacher (Eds.), *Gender across languages* (Vol. 4, pp. 1–26). John Benjamins Publishing Company.
- [11] Kotek, H., Dockum, R., Babinski, S., & Geissler, C. (2021). Gender bias and stereotypes in linguistic example sentences. *Language*, 97(4), 653–677.
- [12] Levshina, N. (2021). Communicative efficiency and differential case marking: A reverse-engineering approach. *Linguistics Vanguard*, 7(3).
- [13] Macaulay, M., & Brice, C. (1997). Don't touch my projectile: Gender bias and stereotyping in syntactic examples. *Language*, 73(4), 798–825.
- [14] Nedoluzhko, A., Novák, M., Popel, M., Žabokrtský, Z., Zeldes, A., & Zeman, D. (2022). CorefUD 1.0: Coreference meets universal dependencies. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 4859–4872.
- [15] Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. *Preprint arXiv:2003.07082*.
- [16] Richey, C., & Burnett, H. (2020). Jean does the dishes while Marie fixes the car: A qualitative and quantitative study of social gender in French syntax articles. *Journal of French Language Studies*, 30(1), 47–72.

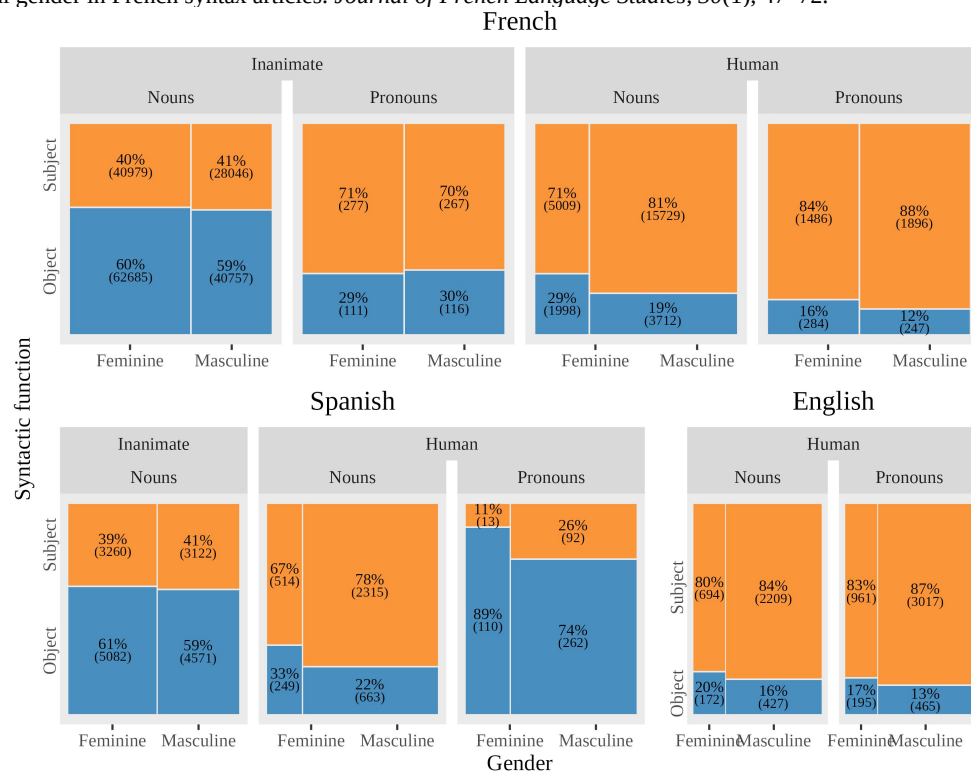


Figure 1: Distribution of arguments given their gender, function, animacy and category across languages. Width indicates the relative frequency of masculine/feminine arguments. Height indicates the frequency of subject/object for each gender.