

# coling2024\_loess\_curves

Amber Converse

2024-09-11

```
## # A tibble: 5 x 5
##   lang          mean    sd median   IQR
##   <ord>        <dbl> <dbl> <dbl> <dbl>
## 1 en_general_rarity    0.304 0.161  0.273 0.167
## 2 es_en_DEEP_general_rarity 0.289 0.160  0.258 0.156
## 3 es_en_DEEPL_general_rarity 0.293 0.161  0.269 0.161
## 4 es_en_GOOGLE_general_rarity 0.292 0.159  0.261 0.156
## 5 es_en_TRANSFORMERS_general_rarity 0.294 0.161  0.267 0.165

## # A tibble: 5 x 5
##   lang          mean    sd median   IQR
##   <ord>        <dbl> <dbl> <dbl> <dbl>
## 1 en_general_rarity    0.304 0.161  0.273 0.167
## 2 ar_en_DEEP_general_rarity 0.296 0.163  0.264 0.171
## 3 ar_en_DEEPL_general_rarity 0.300 0.164  0.273 0.164
## 4 ar_en_GOOGLE_general_rarity 0.299 0.163  0.267 0.168
## 5 ar_en_TRANSFORMERS_general_rarity 0.292 0.160  0.267 0.161

## # A tibble: 5 x 5
##   lang          mean    sd median   IQR
##   <ord>        <dbl> <dbl> <dbl> <dbl>
## 1 en_genre_rarity    0.0806 0.113 0.0462 0.111
## 2 es_en_DEEP_genre_rarity 0.0709 0.109 0.0357 0.0952
## 3 es_en_DEEPL_genre_rarity 0.0726 0.108 0.0377 0.1
## 4 es_en_GOOGLE_genre_rarity 0.0737 0.110 0.0392 0.1
## 5 es_en_TRANSFORMERS_genre_rarity 0.0700 0.109 0.0345 0.0909

## # A tibble: 5 x 5
##   lang          mean    sd median   IQR
##   <ord>        <dbl> <dbl> <dbl> <dbl>
## 1 en_genre_rarity    0.0806 0.113 0.0462 0.111
## 2 ar_en_DEEP_genre_rarity 0.0685 0.107 0.0345 0.0909
## 3 ar_en_DEEPL_genre_rarity 0.0764 0.119 0.04  0.1
## 4 ar_en_GOOGLE_genre_rarity 0.0700 0.108 0.0357 0.0952
## 5 ar_en_TRANSFORMERS_genre_rarity 0.0655 0.106 0.0294 0.0909

##
## Pairwise comparisons using Wilcoxon rank sum test with continuity correction
##
## data:  general_es_df$rarity and general_es_df$lang
##
##               en_general_rarity es_en_DEEP_general_rarity
## es_en_DEEP_general_rarity < 2e-16 -
## es_en_DEEPL_general_rarity 1.4e-09 0.0096
## es_en_GOOGLE_general_rarity 8.5e-13 0.1246
```

```

## es_en_TRANSFORMERS_general_rarity 1.2e-09          0.0111
##                                     es_en_DEEPL_general_rarity
## es_en_DEEP_general_rarity          -
## es_en_DEEPL_general_rarity         -
## es_en_GOOGLE_general_rarity        0.3216
## es_en_TRANSFORMERS_general_rarity  0.9326
##                                     es_en_GOOGLE_general_rarity
## es_en_DEEP_general_rarity          -
## es_en_DEEPL_general_rarity         -
## es_en_GOOGLE_general_rarity        -
## es_en_TRANSFORMERS_general_rarity  0.3372
##
## P value adjustment method: BH

##
## Pairwise comparisons using Wilcoxon rank sum test with continuity correction
##
## data:  general_ar_df$rarity and general_ar_df$lang
##
##                                     en_general_rarity ar_en_DEEP_general_rarity
## ar_en_DEEP_general_rarity          1.4e-06          -
## ar_en_DEEPL_general_rarity          0.02199          0.01092
## ar_en_GOOGLE_general_rarity         0.00047          0.18930
## ar_en_TRANSFORMERS_general_rarity  5.3e-10          0.20094
##                                     ar_en_DEEPL_general_rarity
## ar_en_DEEP_general_rarity          -
## ar_en_DEEPL_general_rarity         -
## ar_en_GOOGLE_general_rarity        0.18930
## ar_en_TRANSFORMERS_general_rarity  0.00012
##                                     ar_en_GOOGLE_general_rarity
## ar_en_DEEP_general_rarity          -
## ar_en_DEEPL_general_rarity         -
## ar_en_GOOGLE_general_rarity        -
## ar_en_TRANSFORMERS_general_rarity  0.01174
##
## P value adjustment method: BH

##
## Pairwise comparisons using Wilcoxon rank sum test with continuity correction
##
## data:  genre_es_df$rarity and genre_es_df$lang
##
##                                     en_genre_rarity es_en_DEEP_genre_rarity
## es_en_DEEP_genre_rarity             < 2e-16          -
## es_en_DEEPL_genre_rarity            9.1e-13          0.1260
## es_en_GOOGLE_genre_rarity           5.3e-10          0.0115
## es_en_TRANSFORMERS_genre_rarity     < 2e-16          0.1331
##                                     es_en_DEEPL_genre_rarity
## es_en_DEEP_genre_rarity             -
## es_en_DEEPL_genre_rarity            -
## es_en_GOOGLE_genre_rarity           0.3198
## es_en_TRANSFORMERS_genre_rarity     0.0025
##                                     es_en_GOOGLE_genre_rarity
## es_en_DEEP_genre_rarity             -
## es_en_DEEPL_genre_rarity            -

```

```

## es_en_GOOGLE_genre_rarity      -
## es_en_TRANSFORMERS_genre_rarity 5.6e-05
##
## P value adjustment method: BH

##
## Pairwise comparisons using Wilcoxon rank sum test with continuity correction
##
## data:  genre_ar_df$rarity and genre_ar_df$lang
##
##               en_genre_rarity ar_en_DEEP_genre_rarity
## ar_en_DEEP_genre_rarity      < 2e-16      -
## ar_en_DEEPL_genre_rarity      3.7e-09      2.2e-06
## ar_en_GOOGLE_genre_rarity      < 2e-16      0.0922
## ar_en_TRANSFORMERS_genre_rarity < 2e-16      2.4e-05
##               ar_en_DEEPL_genre_rarity
## ar_en_DEEP_genre_rarity      -
## ar_en_DEEPL_genre_rarity      -
## ar_en_GOOGLE_genre_rarity      0.0017
## ar_en_TRANSFORMERS_genre_rarity < 2e-16
##               ar_en_GOOGLE_genre_rarity
## ar_en_DEEP_genre_rarity      -
## ar_en_DEEPL_genre_rarity      -
## ar_en_GOOGLE_genre_rarity      -
## ar_en_TRANSFORMERS_genre_rarity 4.9e-09
##
## P value adjustment method: BH

Summary of Significance Wilcox Pairwise Test on Rarity:\ General: es->en Deep < en (p < 0.0001) es->en
DeepL < en (p < 0.0001) es->en GT < en (p < 0.0001) es->en OPUS < en (p < 0.00001)

ar->en Deep < en (p < 0.0001) ar->en DeepL < en (p < 0.03) ar->en GT < en (p < 0.001) ar->en OPUS <
en (p < 0.0001)

Genre: es->en Deep < en (p < 0.0001) es->en DeepL < en (p < 0.0001) es->en GT < en (p < 0.0001)
es->en OPUS < en (p < 0.00001)

ar->en Deep < en (p < 0.0001) ar->en DeepL < en (p < 0.0001) ar->en GT < en (p < 0.0001) ar->en
OPUS < en (p < 0.0001)

## # A tibble: 5 x 5
##   lang          mean    sd median   IQR
##   <ord>         <dbl> <dbl> <dbl> <dbl>
## 1 es_general_rarity    0.316 0.144  0.288 0.133
## 2 en_es_DEEP_general_rarity 0.323 0.147  0.3   0.139
## 3 en_es_DEEPL_general_rarity 0.325 0.149  0.303 0.138
## 4 en_es_GOOGLE_general_rarity 0.324 0.146  0.3   0.137
## 5 en_es_TRANSFORMERS_general_rarity 0.316 0.146  0.289 0.133

## # A tibble: 5 x 5
##   lang          mean    sd median   IQR
##   <ord>         <dbl> <dbl> <dbl> <dbl>
## 1 es_genre_rarity    0.0974 0.117 0.0698 0.108
## 2 en_es_DEEP_genre_rarity 0.0966 0.118 0.0667 0.111
## 3 en_es_DEEPL_genre_rarity 0.100 0.120 0.0704 0.113
## 4 en_es_GOOGLE_genre_rarity 0.0977 0.118 0.0682 0.113
## 5 en_es_TRANSFORMERS_genre_rarity 0.0910 0.116 0.0606 0.122

```

```

##
## Pairwise comparisons using Wilcoxon rank sum test with continuity correction
##
## data:  general_df$rarity and general_df$lang
##
##               es_general_rarity en_es_DEEP_general_rarity
## en_es_DEEP_general_rarity      1.3e-06                -
## en_es_DEEPL_general_rarity      1.6e-08                0.48
## en_es_GOOGLE_general_rarity      1.6e-08                0.46
## en_es_TRANSFORMERS_general_rarity 0.97                1.9e-06
##               en_es_DEEPL_general_rarity
## en_es_DEEP_general_rarity      -
## en_es_DEEPL_general_rarity      -
## en_es_GOOGLE_general_rarity      0.97
## en_es_TRANSFORMERS_general_rarity 2.2e-08
##               en_es_GOOGLE_general_rarity
## en_es_DEEP_general_rarity      -
## en_es_DEEPL_general_rarity      -
## en_es_GOOGLE_general_rarity      -
## en_es_TRANSFORMERS_general_rarity 1.6e-08
##
## P value adjustment method: BH
##
## Pairwise comparisons using Wilcoxon rank sum test with continuity correction
##
## data:  genre_df$rarity and genre_df$lang
##
##               es_genre_rarity en_es_DEEP_genre_rarity
## en_es_DEEP_genre_rarity      0.067                -
## en_es_DEEPL_genre_rarity      0.582                0.021
## en_es_GOOGLE_genre_rarity      0.274                0.474
## en_es_TRANSFORMERS_genre_rarity 4.1e-12            1.5e-06
##               en_es_DEEPL_genre_rarity
## en_es_DEEP_genre_rarity      -
## en_es_DEEPL_genre_rarity      -
## en_es_GOOGLE_genre_rarity      0.114
## en_es_TRANSFORMERS_genre_rarity 4.9e-13
##               en_es_GOOGLE_genre_rarity
## en_es_DEEP_genre_rarity      -
## en_es_DEEPL_genre_rarity      -
## en_es_GOOGLE_genre_rarity      -
## en_es_TRANSFORMERS_genre_rarity 2.6e-08
##
## P value adjustment method: BH
##
General: en->es Deep > es (p < 0.0001) en->es DeepL > es (p < 0.0001) en->es GT > es (p < 0.0001)
en->es OPUS > es NOT SIGNIFICANT (p>0.9!!!)

Genre: en->es Deep > es NOT SIGNIFICANT en->es DeepL > es NOT SIGNIFICANT en->es GT > es
NOT SIGNIFICANT en->es OPUS < es (p < 0.0001)

## # A tibble: 5 x 5
##   lang          mean    sd median   IQR
##   <ord>        <dbl> <dbl> <dbl> <dbl>
## 1 ar_general_rarity 0.594 0.155 0.585 0.184

```

```

## 2 en_ar_DEEP_general_rarity      0.579 0.159 0.571 0.188
## 3 en_ar_DEEPL_general_rarity      0.575 0.161 0.571 0.196
## 4 en_ar_GOOGLE_general_rarity     0.579 0.158 0.571 0.185
## 5 en_ar_TRANSFORMERS_general_rarity 0.586 0.159 0.581 0.188

## # A tibble: 5 x 5
##   lang      mean    sd median  IQR
##   <ord>    <dbl> <dbl> <dbl> <dbl>
## 1 ar_genre_rarity      0.249 0.166 0.222 0.196
## 2 en_ar_DEEP_genre_rarity 0.221 0.164 0.2 0.189
## 3 en_ar_DEEPL_genre_rarity 0.223 0.164 0.2 0.189
## 4 en_ar_GOOGLE_genre_rarity 0.220 0.163 0.194 0.186
## 5 en_ar_TRANSFORMERS_genre_rarity 0.224 0.164 0.2 0.193

##
## Pairwise comparisons using Wilcoxon rank sum test with continuity correction
##
## data:  general_df$rarity and general_df$lang
##
##               ar_general_rarity en_ar_DEEP_general_rarity
## en_ar_DEEP_general_rarity      2.6e-12 -
## en_ar_DEEPL_general_rarity      3.6e-16 0.2035
## en_ar_GOOGLE_general_rarity      2.6e-12 0.9568
## en_ar_TRANSFORMERS_general_rarity 0.0016 0.0002
##               en_ar_DEEPL_general_rarity
## en_ar_DEEP_general_rarity      -
## en_ar_DEEPL_general_rarity      -
## en_ar_GOOGLE_general_rarity      0.2035
## en_ar_TRANSFORMERS_general_rarity 6.7e-07
##               en_ar_GOOGLE_general_rarity
## en_ar_DEEP_general_rarity      -
## en_ar_DEEPL_general_rarity      -
## en_ar_GOOGLE_general_rarity      -
## en_ar_TRANSFORMERS_general_rarity 0.0002
##
## P value adjustment method: BH

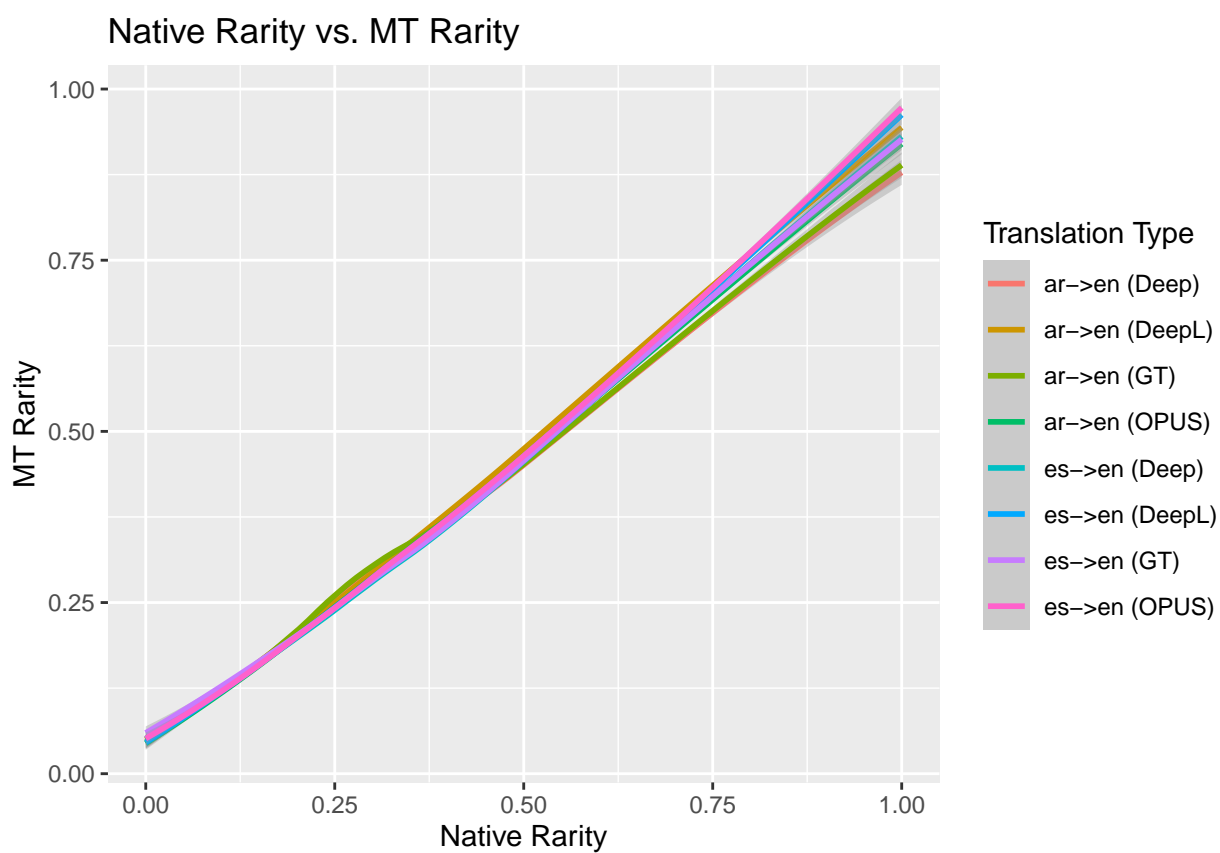
##
## Pairwise comparisons using Wilcoxon rank sum test with continuity correction
##
## data:  genre_df$rarity and genre_df$lang
##
##               ar_genre_rarity en_ar_DEEP_genre_rarity
## en_ar_DEEP_genre_rarity      <2e-16 -
## en_ar_DEEPL_genre_rarity      <2e-16 0.52
## en_ar_GOOGLE_genre_rarity      <2e-16 0.55
## en_ar_TRANSFORMERS_genre_rarity <2e-16 0.30
##               en_ar_DEEPL_genre_rarity
## en_ar_DEEP_genre_rarity      -
## en_ar_DEEPL_genre_rarity      -
## en_ar_GOOGLE_genre_rarity      0.23
## en_ar_TRANSFORMERS_genre_rarity 0.67
##               en_ar_GOOGLE_genre_rarity
## en_ar_DEEP_genre_rarity      -
## en_ar_DEEPL_genre_rarity      -

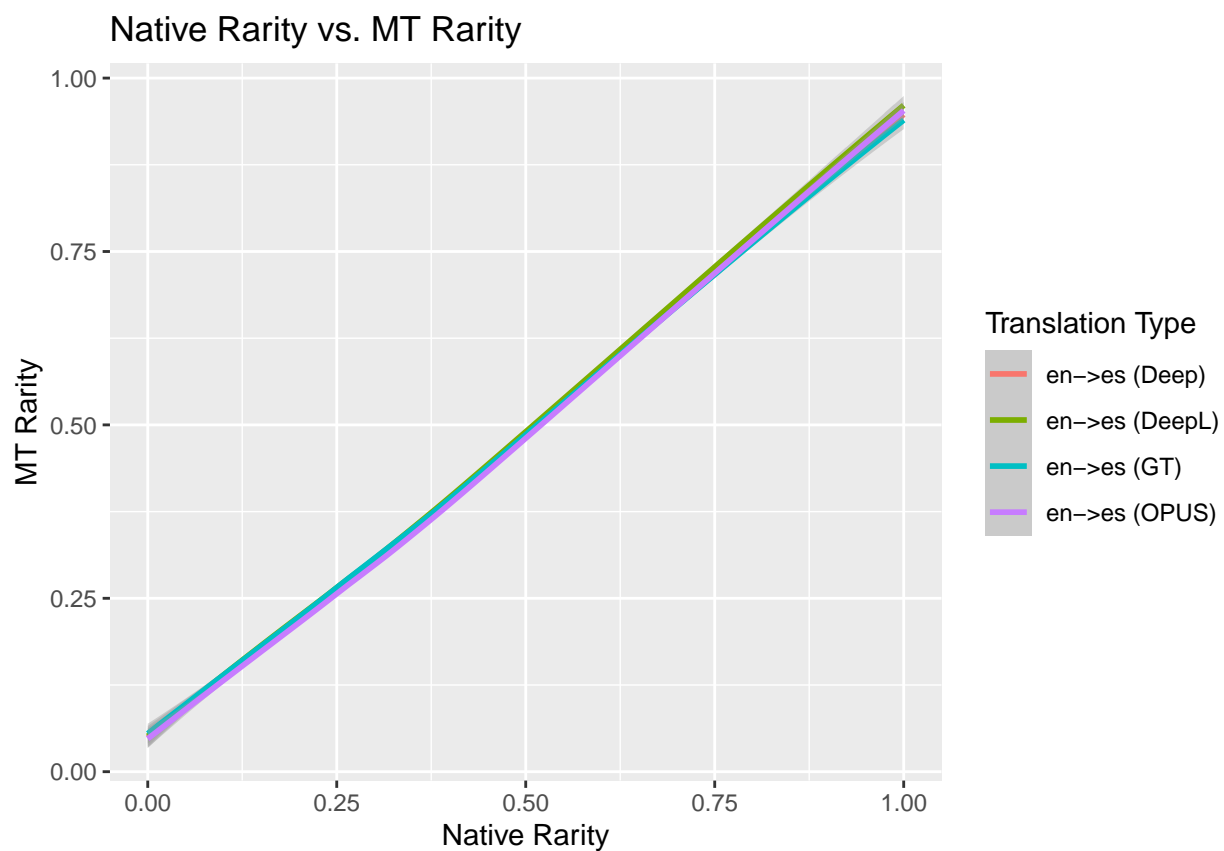
```

```
## en_ar_GOOGLE_genre_rarity -
## en_ar_TRANSFORMERS_genre_rarity 0.11
##
## P value adjustment method: BH
```

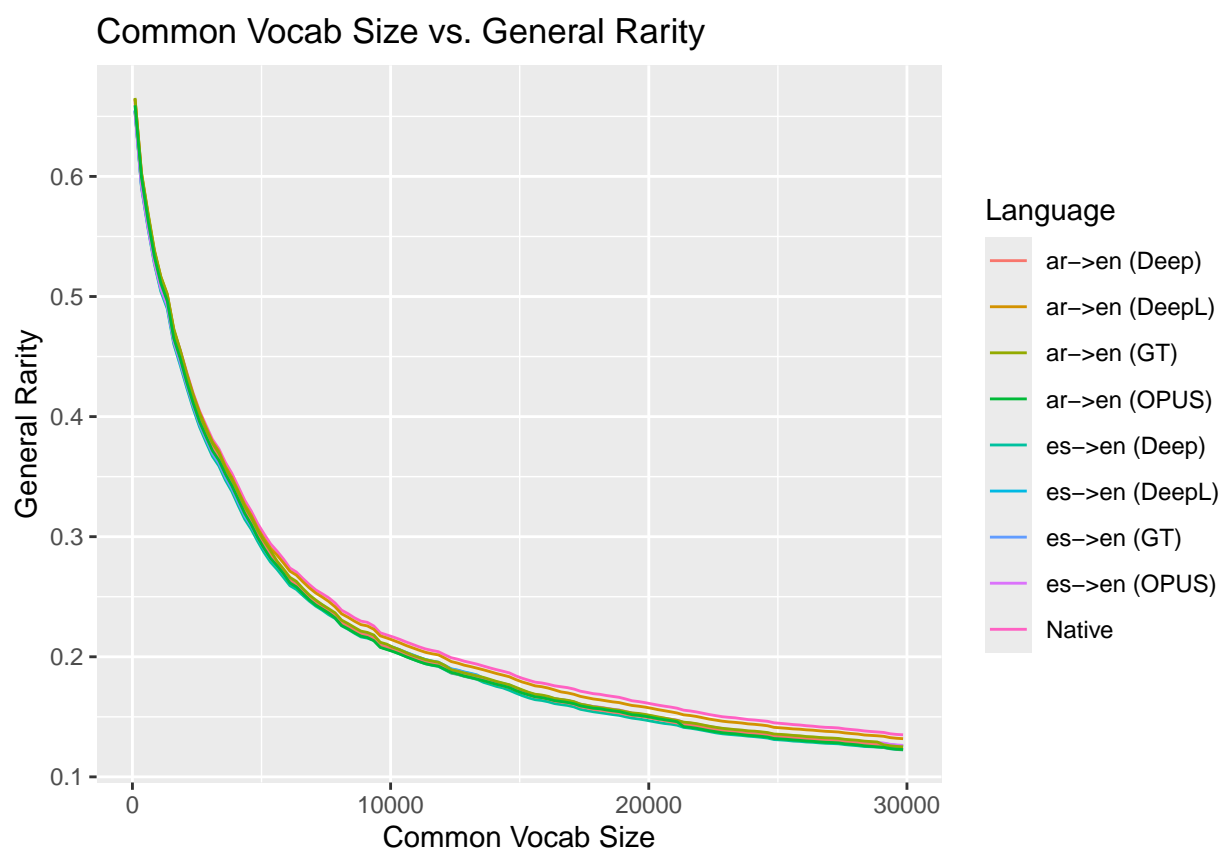
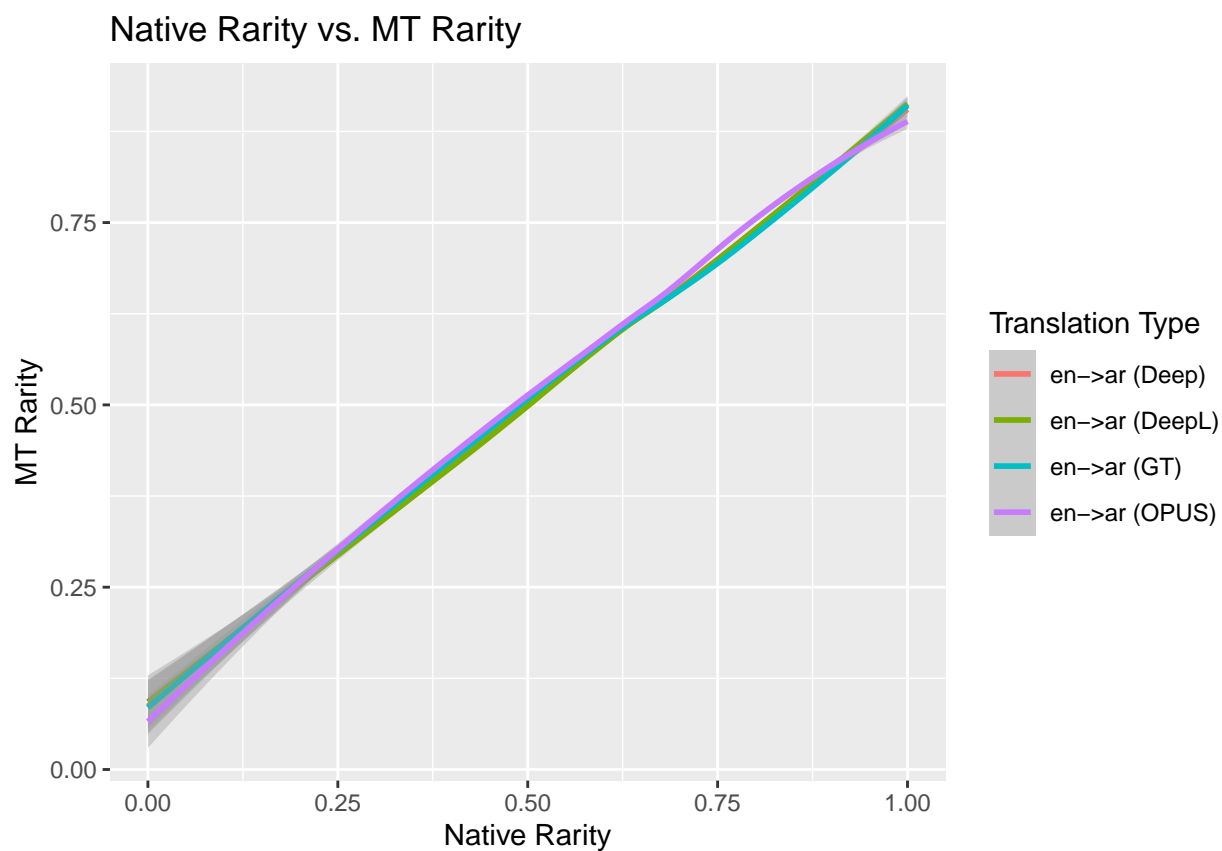
General: en->ar Deep < ar ( $p < 0.0001$ ) en->ar DeepL < ar ( $p < 0.0001$ ) en->ar GT < ar ( $p < 0.0001$ )  
 en->ar OPUS < ar ( $p < 0.002$ )

Genre: en->ar Deep < ar ( $p < 0.0001$ ) en->ar DeepL < ar ( $p < 0.0001$ ) en->ar GT < ar ( $p < 0.0001$ ) en->ar  
 OPUS < ar ( $p < 0.0001$ )



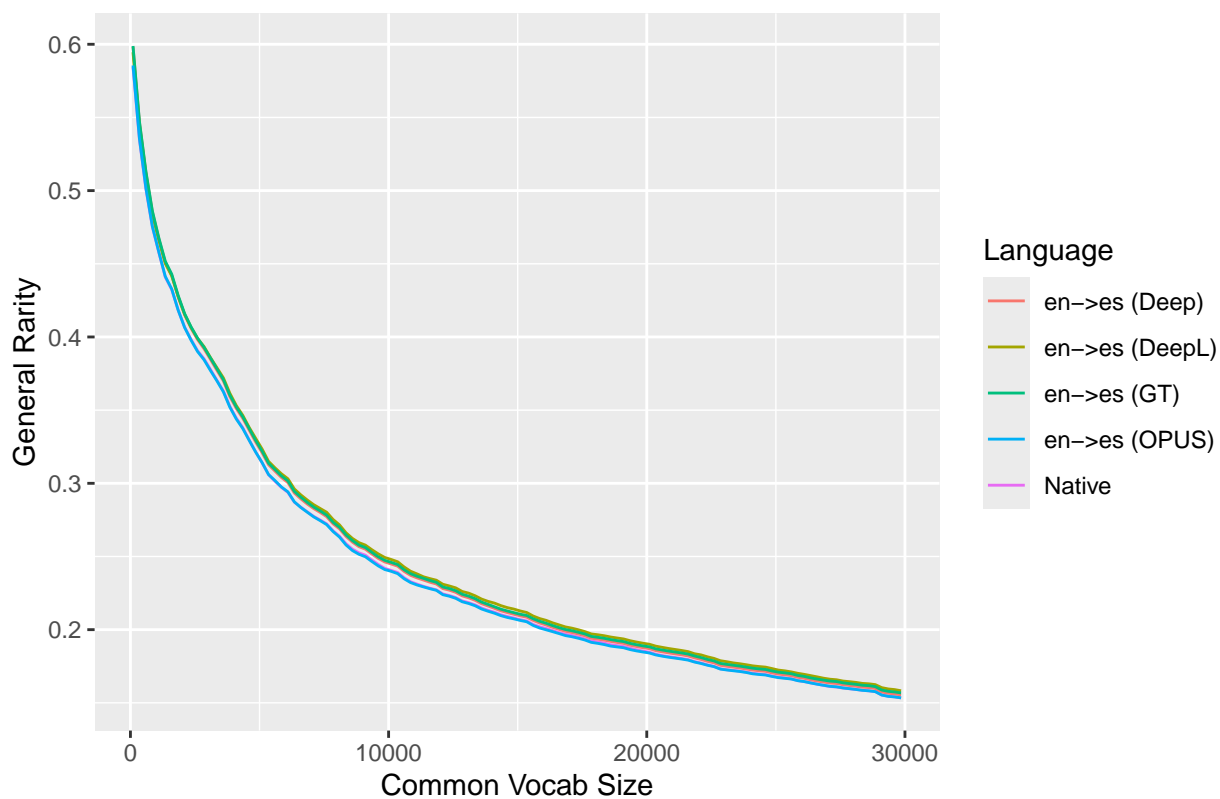


```
## Warning: Removed 6 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```





Common Vocab Size vs. General Rarity



Common Vocab Size vs. General Rarity

