# The Linguistic Risk Premia of the Top 20 Languages
## A Quantitative Study

Soumadeep Ghosh

Kolkata, India

**Abstract**

In a previous paper, I introduced the concept of a *linguistic risk premium* [1] to quantify the structural disadvantage borne by languages with fewer speakers relative to a more widely spoken language. In this article, we operationalise this concept by computing the linguistic risk premia for the world's 20 most widely spoken languages, using English as the benchmark and employing approximate short-term government-bond yields for both the United States and the United Kingdom as proxies for the risk-free rate. We present the full derivation, tabulated results, and vector-graphic visualisations produced with PGF/TikZ [6], and discuss the implications and limitations of the model.

The paper ends with "The End"

# 1 Introduction

Language is not merely a medium of communication; it is a form of *human capital* whose economic value depends, among other factors, on the size of the community that uses it [4, 5]. A language spoken by a billion people affords its users access to a vastly larger labour market, body of literature, and digital ecosystem than one spoken by a few million.

I formalised this intuition by proposing a **linguistic risk premium**—a single scalar that captures the "discount" applied to a less widely spoken language when compared with a more widely spoken one [1]. The concept is deliberately analogous to the equity risk premium in finance [3]: just as investors demand additional expected return for holding risky assets rather than risk-free bonds, language users implicitly bear additional "cost" when their primary language has fewer speakers.

Drawing on speaker-count data published by Ethnologue [2] and on government-bond yields for the United States [7] and the United Kingdom [8], we compute the linguistic risk premium for the top 20 languages by total number of speakers (first-language + second-language users).

# 2 Theoretical Framework

## 2.1 Definition of the Linguistic Risk Premium

**Definition 1** (Linguistic Risk Premium [1])**.** *Let $L_2$ denote the total number of speakers of a benchmark language and $L_c$ the total number of speakers of a comparison language with $L_c \leq L_2$. The* linguistic risk premium *$p_l$ is the quantity satisfying*

$$\boxed{L_c \;=\; \frac{L_2}{1 + r_f + p_l}} \tag{1}$$

*where $r_f$ is the prevailing risk-free rate.*

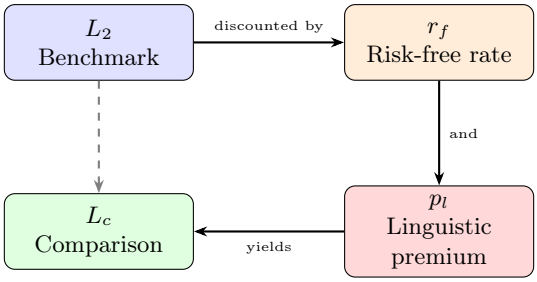


Figure 1: Schematic of the linguistic risk premium model. The benchmark language $L_2$ is discounted by the risk-free rate $r_f$ and the linguistic risk premium $p_l$ to obtain the comparison language $L_c$.

### 2.2 Solving for the Premium

Rearranging Equation (1) for $p_l$:

$$p_l \;=\; \frac{L_2}{L_c} \;-\; 1 \;-\; r_f \tag{2}$$

**Remark 1.** *The premium depends only on the* ratio $L_2/L_c$ *and the risk-free rate. Because typical risk-free rates (≈ 4–5%) are negligible compared with the speaker-count ratio for most language pairs, the premium is overwhelmingly driven by relative community size.*

## 3 Data

### 3.1 Speaker Counts

Total speaker estimates (L1 + L2, in millions) are taken from the Ethnologue 200 listing, 27th edition [2]. We adopt **English** ($L_2$ = 1,456 M) as the benchmark language because it has the largest aggregate speaker population.

### 3.2 Risk-Free Rates

Table 1: Risk-free rate proxies (approx. Feb. 2026).

| Country | Instrument | $r_f$ |
|---|---|---|
| United States [7] | 3-month T-bill | 4.25% |
| United Kingdom [8] | 3-month Gilt | 4.50% |

`The following space was deliberately left blank.`

# 4 Results

Table 2 presents the linguistic risk premium for each of the top 20 languages. Premia are expressed as percentages and computed using Equation (2).

Table 2: Linguistic risk premia for the top 20 languages (English benchmark, $L_2 = 1{,}456\,\mathrm{M}$).

| Rank | Language | Speakers (M) | $L_2/L_c$ | $p_l$ (U.S.) | $p_l$ (U.K.) |
|---|---|---|---|---|---|
| 1 | English (benchmark) | 1 456 | 1.0000 | — | — |
| 2 | Mandarin Chinese | 1 138 | 1.2795 | 23.70% | 23.45% |
| 3 | Hindi | 609 | 2.3908 | 134.83% | 134.58% |
| 4 | Spanish | 559 | 2.6047 | 156.22% | 155.97% |
| 5 | French | 310 | 4.6968 | 365.43% | 365.18% |
| 6 | Modern Standard Arabic | 274 | 5.3139 | 427.14% | 426.89% |
| 7 | Bengali | 273 | 5.3333 | 429.08% | 428.83% |
| 8 | Portuguese | 264 | 5.5152 | 447.27% | 447.02% |
| 9 | Russian | 255 | 5.7098 | 466.73% | 466.48% |
| 10 | Urdu | 232 | 6.2759 | 523.34% | 523.09% |
| 11 | Indonesian | 199 | 7.3166 | 627.41% | 627.16% |
| 12 | Standard German | 134 | 10.8657 | 982.32% | 982.07% |
| 13 | Japanese | 123 | 11.8374 | 1 079.49% | 1 079.24% |
| 14 | Nigerian Pidgin | 121 | 12.0331 | 1 099.06% | 1 098.81% |
| 15 | Marathi | 99 | 14.7071 | 1 366.46% | 1 366.21% |
| 16 | Telugu | 96 | 15.1667 | 1 412.42% | 1 412.17% |
| 17 | Turkish | 90 | 16.1778 | 1 513.53% | 1 513.28% |
| 18 | Tamil | 87 | 16.7356 | 1 569.31% | 1 569.06% |
| 19 | Yue Chinese (Cantonese) | 86 | 16.9302 | 1 588.77% | 1 588.52% |
| 20 | Vietnamese | 85 | 17.1294 | 1 608.69% | 1 608.44% |

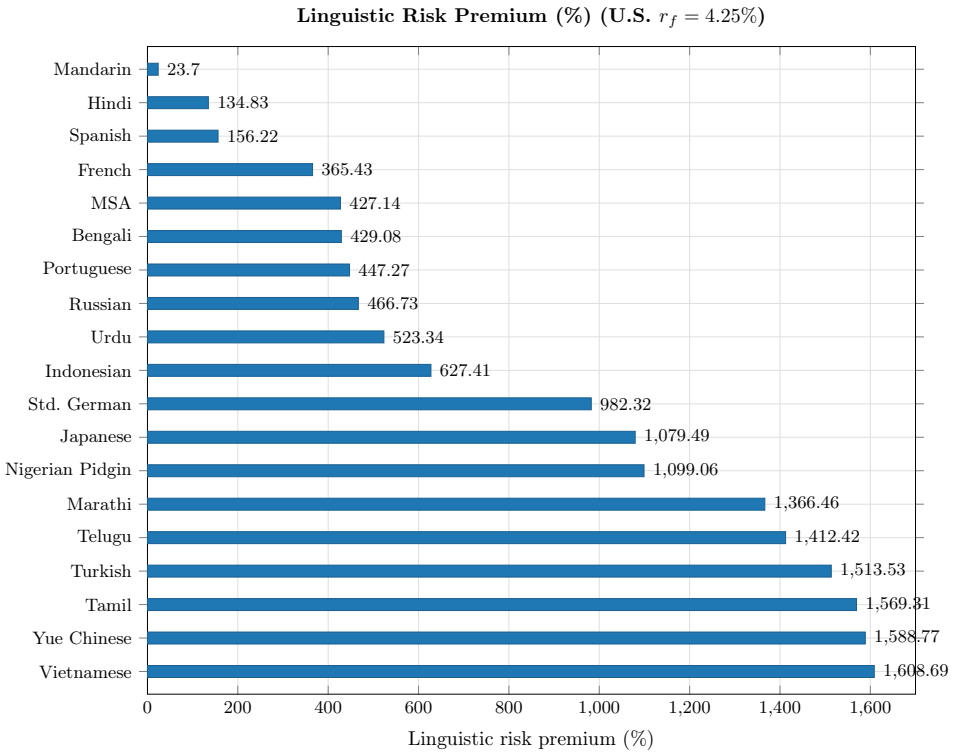


Figure 2: Horizontal bar chart of the linguistic risk premium for the top 20 languages (U.S. risk-free rate). English is the benchmark ($p_l \equiv 0$) and is omitted.

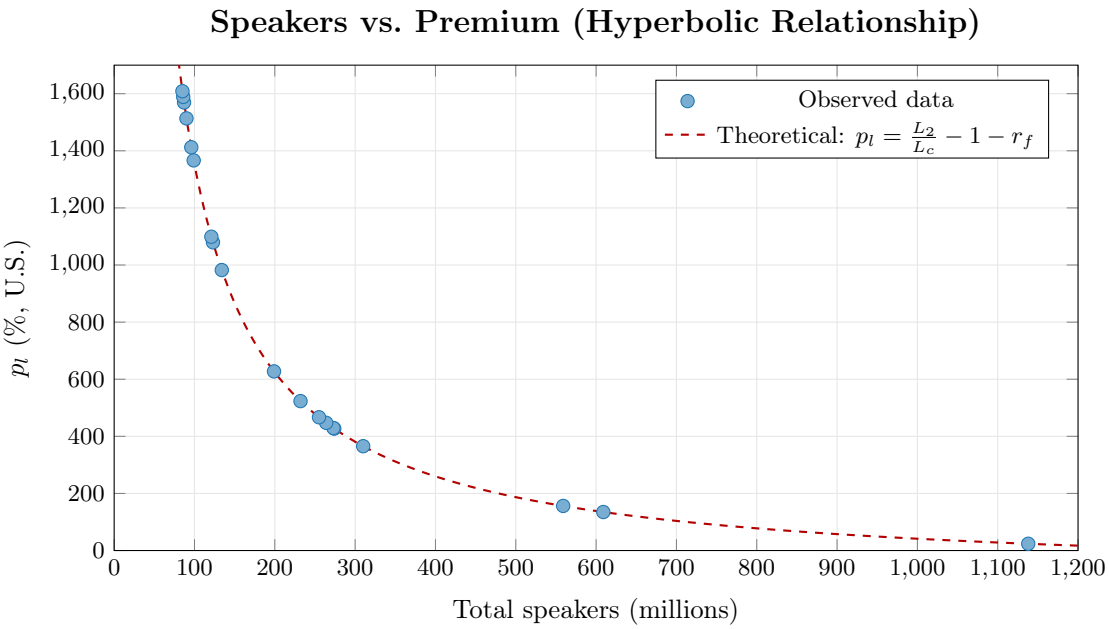


Figure 3: Scatter plot with the theoretical hyperbolic curve $p_l = L_2/L_c - 1 - r_f$ overlaid (dashed). The inverse relationship is clearly visible.

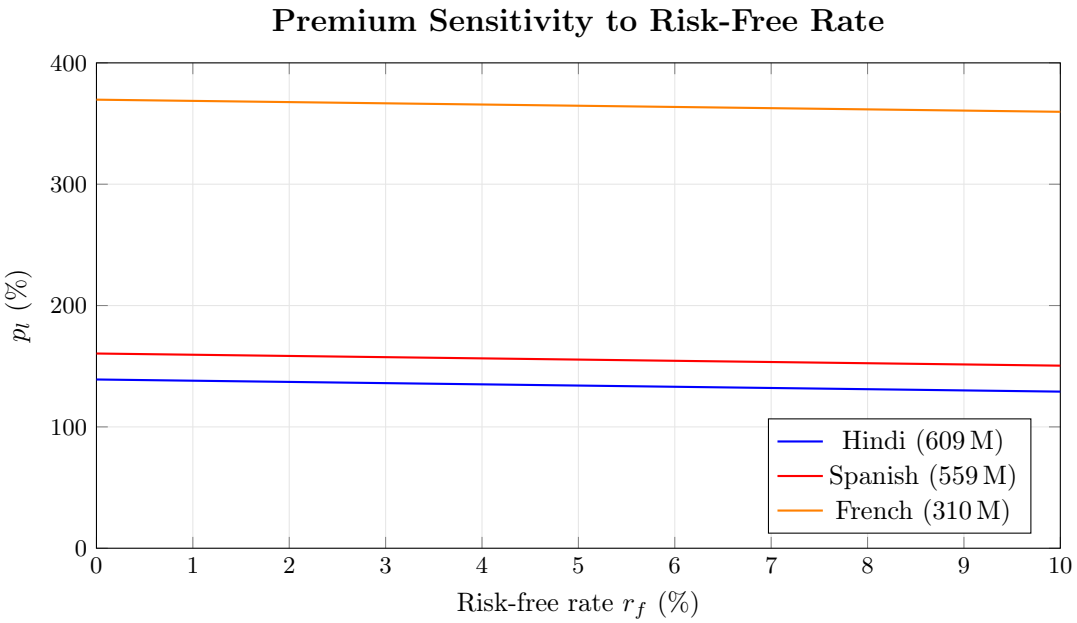


Figure 4: Sensitivity of $p_l$ to the risk-free rate for selected languages. The lines are nearly flat, confirming that $r_f$ has minimal influence on the premium.

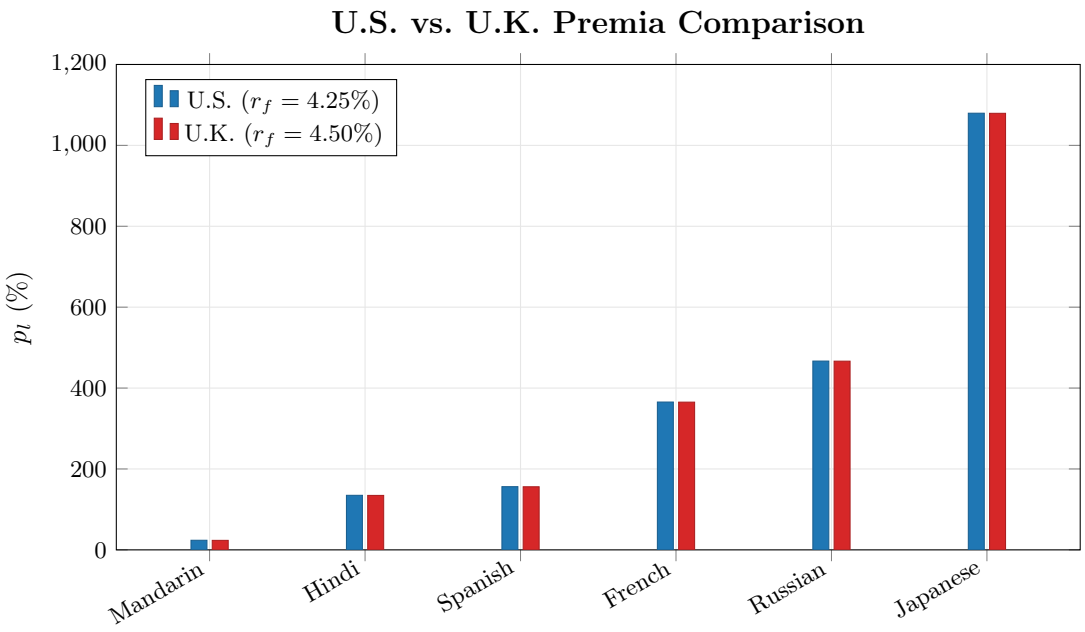


Figure 5: Grouped bar chart comparing U.S. and U.K. linguistic risk premia for six selected languages. The 0.25 percentage-point difference is barely visible at this scale.

# 5 Discussion

## 5.1 Key Observations

**Inverse-proportional growth.** Equation (2) shows that $p_l$ is a *hyperbolic* function of $L_c$. The premium for Mandarin Chinese ($\approx 24\%$) is modest because its speaker base is comparable to English's. By contrast, Vietnamese (85 M speakers) already carries a premium exceeding 1,600%—a figure that would grow to roughly 36,000% for a language with only 4 million speakers [2].

**Negligible influence of $r_f$.** The difference between the U.S. and U.K. premia is exactly $r_f^{\mathrm{UK}} - r_f^{\mathrm{US}} = 0.25\,\mathrm{pp}$ for *every* language (see Figure 5). This confirms that the risk-free rate is a second-order effect; the speaker-count ratio dominates.

**Relation to language economics.** The linguistic risk premium can be viewed as a complement to Selten and Pool's [5] game-theoretic model of language learning and to Grin's [4] broader economics-of-language framework. Where those models examine incentives and returns to multilingualism, the linguistic risk premium distils the *relative structural disadvantage* into a single number.

## 5.2 Limitations

1. **No empirical validation.** The model in [1] is purely definitional; the "premium" has not yet been linked to any observable economic outcome (e.g., wage differentials, trade volumes).
2. **Linear discounting assumption.** The additive structure $1 + r_f + p_l$ implies a linear discount. A logarithmic or power-law specification might yield more interpretable premia for very small languages [3].
3. **Homogeneous speaker counts.** Treating all speakers as equivalent ignores differences in GDP per capita, digital presence, and institutional support [4].
4. **Static snapshot.** Speaker counts change over time; a dynamic model would be more informative.

### 5.3 Potential Extensions

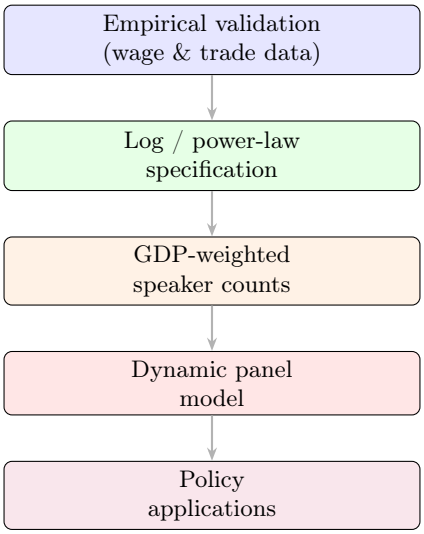


Figure 6: Proposed research roadmap for extending the linguistic risk premium framework.

## 6 Conclusion

We have operationalised the linguistic risk premium for the 20 most widely spoken languages in the world, using English as the benchmark. The resulting premia range from roughly 24% (Mandarin Chinese) to over 1,600% (Vietnamese) and are almost entirely driven by the ratio of speaker populations rather than by the choice of risk-free rate. While the model is elegant in its simplicity, its practical utility hinges on future empirical work that links the premium to measurable economic and social outcomes.

## References

[1] S. Ghosh, "The linguistic risk premium," Kolkata, India, 2024.

[2] Eberhard, David M., Gary F. Simons, and Charles D. Fennig (eds.), *Ethnologue: Languages of the World*, 27th ed., SIL International, Dallas, TX, 2024. `https://www.ethnologue.com/insights/ethnologue200/`

[3] A. Damodaran, *Equity Risk Premiums (ERP): Determinants, Estimation, and Implications – The 2020 Edition*, Stern School of Business, New York University, 2020.

[4] F. Grin, *Language Policy Evaluation and the European Charter for Regional or Minority Languages*, Palgrave Macmillan, 2003.

[5] R. Selten and J. Pool, "The distribution of foreign language skills as a game equilibrium," in R. Selten (ed.), *Game Equilibrium Models IV*, Springer, 1991, pp. 64–87.

[6] T. Tantau, *The TikZ and PGF Packages: Manual for Version 3.1.10*, 2023. `https://ctan.org/pkg/pgf`

[7] U.S. Department of the Treasury, "Daily Treasury Bill Rates," accessed February 2026. `https://home.treasury.gov/`

[8] UK Debt Management Office, "Gilt Market," accessed February 2026. `https://www.dmo.gov.uk/`

## Glossary

**Benchmark language ($L_2$)**
: The language with the greatest number of total speakers against which all other languages are compared. In this study, English (1,456 M speakers) serves as the benchmark.

**Comparison language ($L_c$)**
: Any language whose speaker population is smaller than or equal to that of the benchmark language.

**Linguistic risk premium ($p_l$)**
: A scalar quantity, expressed as a percentage, that measures the structural disadvantage or "discount" of a comparison language relative to the benchmark, after accounting for the risk-free rate [1].

**Risk-free rate ($r_f$)**
: The theoretical rate of return on an investment with zero risk of financial loss, typically proxied by short-term government-bond yields (e.g., 3-month U.S. Treasury bills or U.K. Gilts) [3].

**Ethnologue 200**
: A ranking of the world's 200 most widely spoken languages by total number of speakers (L1 + L2), published by SIL International [2].

**L1 speakers**
: Native or first-language speakers of a given language.

**L2 speakers**
: Second-language or additional-language speakers who have acquired the language after their mother tongue.

**Total speakers**
: The sum of L1 and L2 speakers for a given language.

**Speaker-count ratio ($L_2/L_c$)**
: The ratio of the benchmark language's speaker population to that of the comparison language. This ratio is the dominant determinant of the linguistic risk premium.

**Hyperbolic relationship**
: A mathematical relationship of the form $y = k/x$, characteristic of the premium curve: as speaker count $L_c$ decreases, the premium $p_l$ increases sharply.

**Equity risk premium (ERP)**
: In finance, the excess return that investing in the stock market provides over a risk-free rate; the conceptual analogue from which the linguistic risk premium draws its inspiration [3].

**PGF/TikZ**
: A tandem of LaTeX packages for producing high-quality vector graphics from geometric and algebraic descriptions [6].

# The End