Instituto de Física Gleb Wataghin, UNICAMP

Tarefa 1: Entropia da Linguagem

Letícia S Magalhães 156245

O número de palavras que compõem uma língua é da ordem de centenas de milhares de palavras. A língua inglesa, por exemplo, possui 171.476 palavras em uso e outras 47.156 em desuso, segundo o Dicionário de Oxford [1]. Por outro lado, estudos linguísticos apontam que o vocabulário de falantes nativos possui da ordem de 20.000 palavras [2]. Considerando todo o conjunto de palavras de uma língua, as 1.000 palavras mais frequentes cobrem de 74% a 81% do idioma escrito e 81% a 84% do idioma falado. As seguintes 1.000 palavras mais frequentes cobrem 8% e 5% do idioma escrito e falado, respectivamente [2].

Por isso, para analisar a entropia das linguagens em seu estado atual, buscamos um repositório das 1.000 palavras mais utilizadas [3] numa língua. Os dados são obtidos através das legendas criadas para filmes e disponíveis online [4]. Para a análise, ignoramos acentos e outras variantes de letras (como por exemplo 'ç'). A frequência de cada letra é mostrada na figura 1.

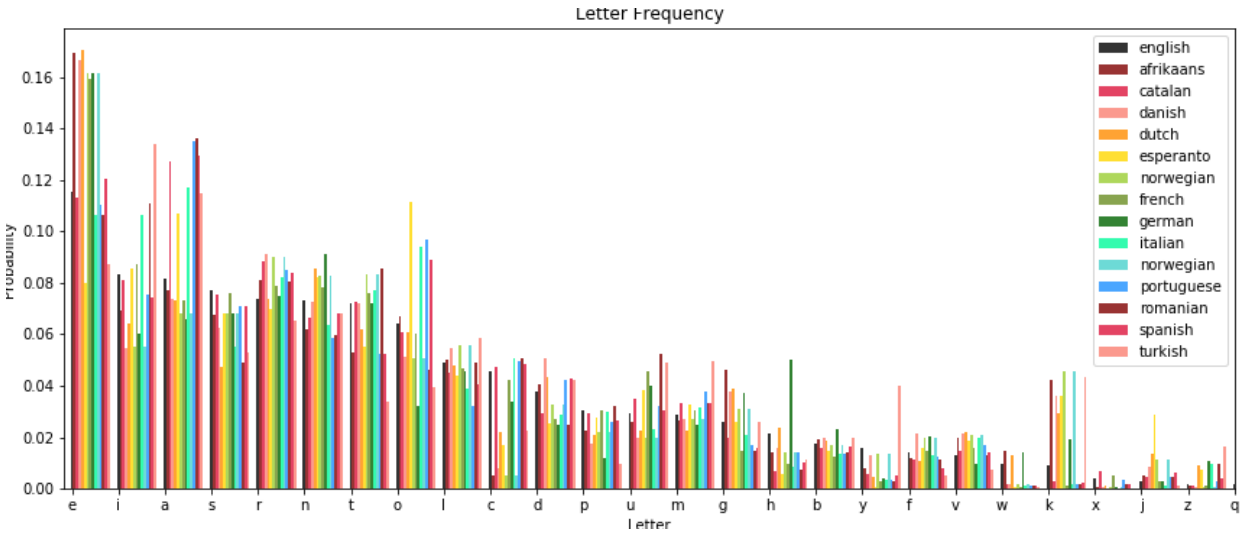

Figura 1: Distribuição de probabilidades para o alfabeto, ordenado com base na frequência do inglês.

Do ponto de vista da Teoria da Informação, podemos investigar a incerteza associada às linguas, como um conceito quantitativo, através da entropia de Shannon. A figura 2 mostra as entropias calculadas.

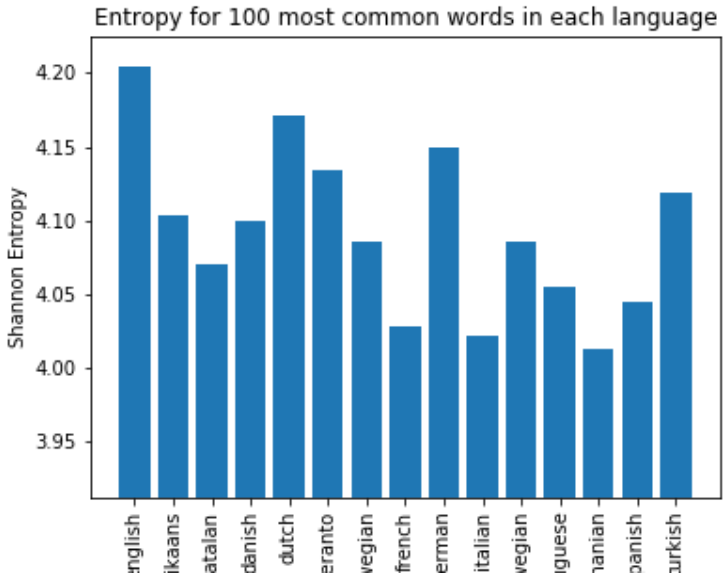

Figura 2: Entropia calculada para as amostras de 1000 palavras mais utilizadas em legendas de filmes para diversas línguas.

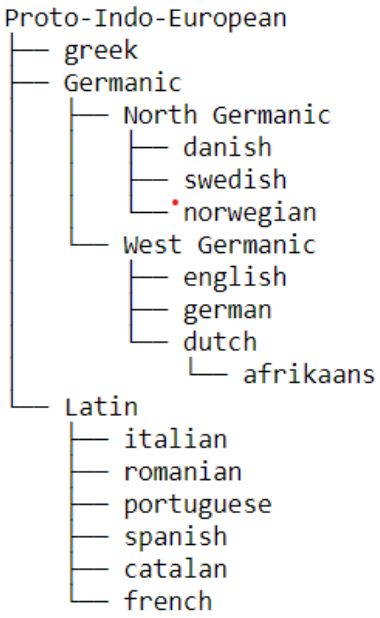

Figura 3: Árvore filogenética linguística para o tronco Proto-Indo-Europeu.

Observamos que, dentre os idiomas investigados, o inglês possui a maior entropia, seguida do holandês e do alemão. A figura 3 mostra a árvore linguística do tronco Proto-Indo-Europeu. Observamos que estas três línguas fazem parte do ramo Oeste-Germânico, juntamente com o afrikaans -

língua derivada do holandês e falada na África do Sul e Namibia [6]. É curioso o fato do afrikaans possuir entropia menor que o holandês, uma vez que se desenvolveu a partir da creolização do holandêns e portanto seria esperado que esta mistura gerasse uma perda das estruturas peculiares - e consequentemente aleatorizando o idioma.

Analisamos a correlação entre a probabilidades dos caracteres do inglês e as línguas do tronco Oeste-Germânico. O resultado é mostrado na figure 4. Os números na legenda indicam o desvio padrão entre as probabilidades em relação ao inglês. De forma simplista, imaginamos que o desvio entre as probabilidades dos caracteres carregue informação sobre o grau de proximidade das línguas. Neste caso, o inglês é quase "equidistante" do alemão e holandês, porém um pouco mais distante do afrikaans.

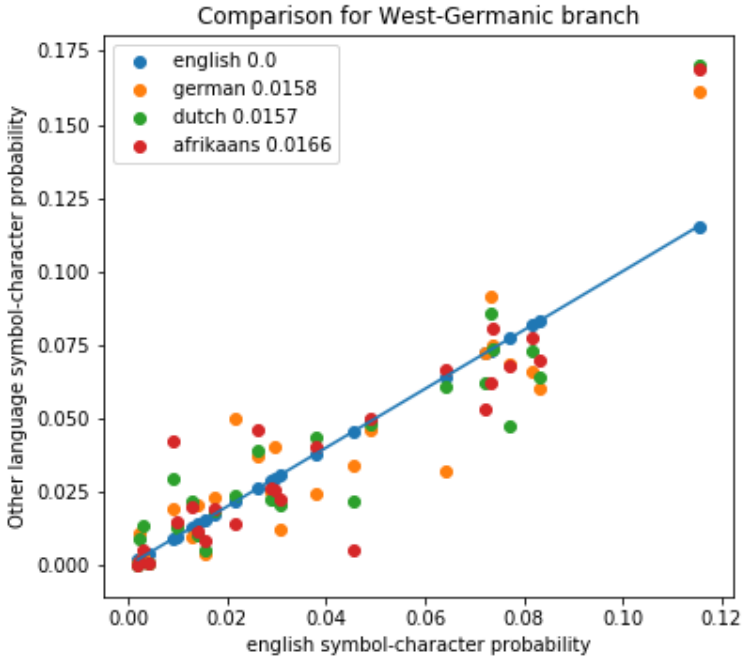

Figura 4: Correlação entre as distribuições de probabilidade de uso dos caracteres para os idiomas do ramo Oeste-Germânico. Na legenda, os números ao lado dos nomes indicam o desvio padrão com relação ao inglês (linha de referência).

A figura 5 mostra a mesma análise para o ramo Latino, desta vez em comparação com o português. Observamos que a língua mais correlacionada é o espanhol, seguida do catalão. Em contrapartida, o francês é o idioma mais distante, segundo o desvio padrão calculado.

Para uma visualização mais abrangente, a figura 6 mostra a correlação entre o português e todas as línguas exploradas na figura 1. Vemos que a língua artificial Esperanto possui o desvio padrão menor que outras linguas latinas como romeno e frances. Além disso, o idioma mais distante, segundo este critério, seria o turco. O turco não faz parte do tronco Proto-Indo-Europeu (apesar de utilizar o alfabeto latino), faz parte da família Turkic [7], originária do leste asiático. O uso

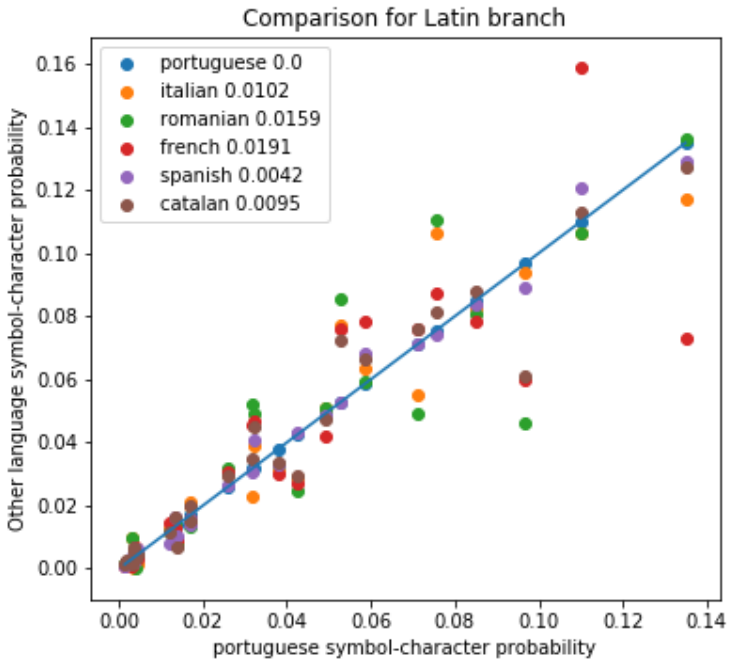

Figura 5: Correlação entre as distribuições de probabilidade de uso dos caracteres para os idiomas do ramo Latino. Na legenda, os números ao lado dos nomes indicam o desvio padrão com relação ao português (linha de referência).

do alfabeto latino veio de uma reforma cultural, introduzida em 1928, que substituiu o alfabeto otomano[8]. Alguns linguistas consideram o tronco Turkic como ramo da família Altaic, que inclui também línguas japonesa e coreana mas não há consenso sobre isto.

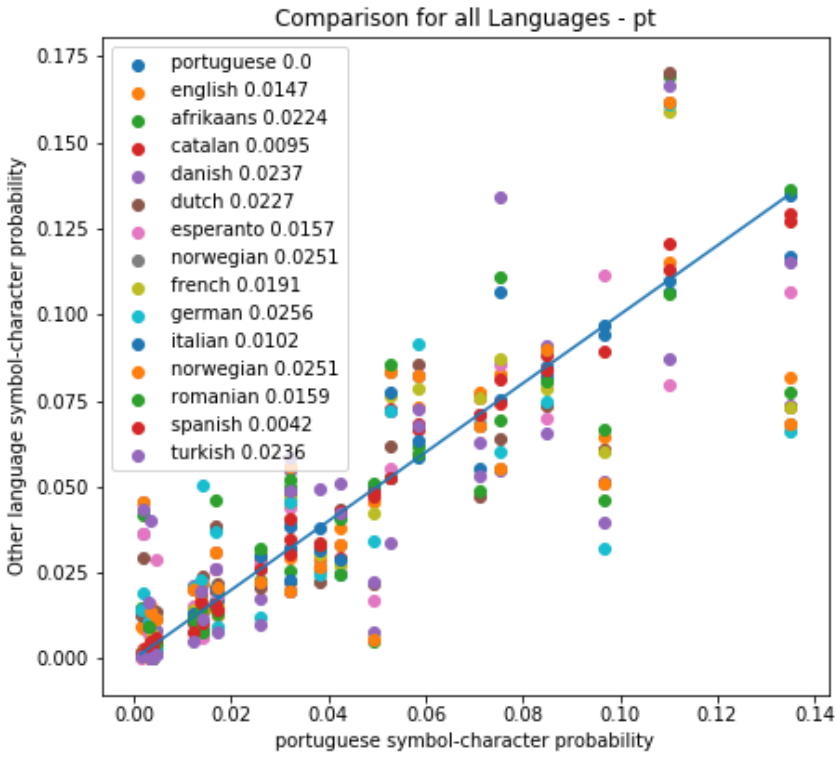

Figura 6: Correlação entre as distribuições de probabilidade de uso dos caracteres para os idiomas explorados na figura 1. Na legenda, os números ao lado dos nomes indicam o desvio padrão com relação ao português (linha de referência).

Referências

[1] Oxford Dictionaries. *How many words are there in the English language?* `https://en.oxforddictionaries.com/explore/how-many-words-are-there-in-the-english-language/`. Acesso em 23 de abril de 2019.

[2] Nation, I., 2006. *How large a vocabulary is needed for reading and listening?*. Canadian modern language review, 63(1), pp.59-82.

[3] *Most common words by language.* `https://github.com/oprogramador/most-common-words-by-language`

[4] Legendas de filmes. *OpenSubtitles*. `http://www.opensubtitles.org/`. Acesso em 22 de abril de 2019.

[5] Wikipedia, *Indo-European languages.* https://en.wikipedia.org/wiki/Indo-European_languages. Acesso em 22 de abril de 2019.

[6] Wikipedia, *Afrikaans.* https://en.wikipedia.org/wiki/Afrikaans. Acesso em 22 de abril de 2019.

[7] Wikipedia, *Turkic Languages.* https://en.wikipedia.org/wiki/Turkic_languages. Acesso em 22 de abril de 2019.

[8] Wikipedia, *Turkish language.* https://en.wikipedia.org/wiki/Turkish_language#Writing_system. Acesso em 22 de abril de 2019.

[9] Wikipedia, *Altaic languages.* https://en.wikipedia.org/wiki/Altaic_languages. Acesso em 22 de abril de 2019.