CESAR AUGUSTO ALVES CAMILLO

O DNA DAS LÍNGUAS - USANDO ALGORTIMOS DE SEMELHANÇA GENÉTICA PARA IDENTIFICAR SIMILARIDADE ENTRE LINGUAGENS

Trabalho apresentado como requisito para a obtenção do título de Bacharel em Ciência da Computação, Setor de Exatas da Universidade Federal do Paraná.

Orientador: Prof Fabiano Silva, DINF
Coorientadora: Profª Adelaide Hercília Pescatori Silva, DELLIN

CURITIBA

2023

RESUMO

Neste trabalho é realizada a análise do uso de algoritmos de análise filogenética para verificar a similaridade entre línguas através do cálculo do alinhamento entre várias palavras para cada par de linguagens, com base na ideia que, de forma similar a variação genética de uma espécie dando origem a outras, existe um rastro que pode ser detectado pela distância de edição entre palavras com o mesmo significado nas diversas línguas. Para tal, usamos cerca de 1200 palavras retiradas de um dicionário online de português e tiramos as traduções das mesmas para inglês, espanhol, italiano, romeno, francês e galego através do *Google Translator*, gerando um arquivo csv com as mesmas. Com esse arquivo em mãos, calculamos a distância de edição entre todas as palavras para cada par de línguas utilizadas para então gerar uma matriz das distâncias normalizadas para cada par de linguagens utilizadas no trabalho. Finalmente, construímos duas árvores filogenéticas, usando os algoritmos NJ e UPGMA, para então fazermos uma análise crítica dos resultados obtidos para verificar a validade do uso de algoritmos da filogenética para uso na área linguística. Os resultados do trabalho mostram que, mesmo com problemas na escolha do modelo de cálculo de distância e nas traduções utilizadas para tais cálculos, algoritmos simples da área de análise filogenética são capazes de reproduzir resultados muito próximos de pesquisas mais elaboradas encima do assunto, provando a usabilidade dos mesmos na linguística comparativa.

Palavras-chaves: Línguas, distância de edição, árvores filogenéticas, genômica computacional, linguística comparativa

ABSTRACT

This work performs an analysis about the use of phylogenetic algorithms to verify similarity between languages through the use of edit distance between words of every pair of the languages used in the study based on the idea that, similarly to the tracking of genetic drift through different but related species, the passage of time leaves a similar expression on the written word. We use about 1200 words taken from an online portuguese dictionary and their translations in english, spanish, italian, romanian, french and galician through the use of *Google Translator*, creating a csv file with them. With that in hands, we calculate the edit distance, pairwise for each word with the same meaning for each pair of languages used. Finally, we construct two phylogenetic trees, using the algorithms NJ and UPGMA, ending with a critical analysis of the results obtained to verify the use of phylogenetics algorithms in linguistics. The results of this works proves that, even with the issues introduced with a faulty model for edit distance calculations and the low quality translations used in those, the use of simple phylogenetic analysis algorithms can reproduce results similar to those of more elaborate research in the subject, proving their usability in comparative linguistics.

Key-words: Languages, edit distance, phylogenetic trees, computational genomics, comparative linguistics

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E DE SIGLAS

CSV *Comma-separated values*

NJ *Neighbor-joining*

PDF *Portable Document Format*

UPGMA *Unweighted Pair Group Method with Arithmetic Averages*

normal Normalizada

LISTA DE SÍMBOLOS

Θ	Score de alinhamento
Δ	Distância normalizada
Λ	Total de palavras

SUMÁRIO

1 INTRODUÇÃO

A introdução apresenta as motivações, propostas de projeto, desafios enfrentados e o que se espera desse trabalho. Em seu término, apresenta a disposição dos capítulos do mesmo.

1.1 MOTIVAÇÃO

Alguns dos maiores mistérios na área da linguística histórica e comparativa são como as linguagens se relacionam entre si e como elas evoluíram ao passar do tempo (RAMA et al., 2015), com vários métodos existindo na literatura para realizar a reconstrução de famílias linguísticas e suas protolínguas, mas apenas a partir de meados de 1990 que essas áreas começaram a utilizar de recursos da computação para seus fins (RAMA et al., 2015; MCMAHON; MCMAHON, 2003), utilizando conceitos como distância de edição para auxiliar nesta análise.

Um conceito equivalente, vindo da área de genômica computacional, é a sequência de alinhamento. De fato, é matematicamente provável que este conceito é equivalente a distância de edição utilizada na linguística (SELLERS, 1974), e numa visão longe de ambas as áreas, alguns dos outros conceitos utilizados são similares o suficiente para criar a pergunta de se algoritmos de uma podem ser utilizados em outra.

1.2 PROPOSTA

Este trabalho se propõe a verificar a viabilidade do uso de algoritmos e conceitos de genômica computacional, como alinhamento global e inferência de árvores filogenéticas, no contexto da filologia de forma a verificar a ideia de que o vocabulário de uma língua pode ser tratado como parte de seu DNA para identificar relações entre linguagens diferentes. Para verificar a corretude desta hipótese, pegamos um conjunto de palavras em português e, com ajuda do *Google Translate*, traduzimos as mesmas para outras 6 línguas: francês, romeno, galego, italiano, espanhol e inglês. Com um arquivo mantendo as relações entre todas as nossas palavras em mãos, calculamos o alinhamento global para cada par de palavras em duas línguas diferentes para então utilizar os algoritmos de inferência de árvores filogenéticas para podermos analizar os resultados e verificar a viabilidade do uso destas ferramentas no contexto linguístico.

1.3 DESAFIOS

Os desafios encontrados para a realização do trabalho foram os seguintes:

- Encontrar trabalhos que utilizassem de técnicas similares para análise dos resultados com a literatura da área.
- Encontrar uma base de dados com uma palavra sendo traduzida para várias línguas diferentes.
- Criar uma possível métrica para ser utilizada para cálculo e geração dos mapas filogenéticos.
- Realizar uma análise crítica dos resultados obtidos com os experimentos.

1.4 CONTRIBUIÇÃO

A contribuição deste trabalho consiste na análise dos resultados obtidos ao utilizar algoritmos de genômica computacional para verificar a usabilidade dos mesmos no estudo da evolução das linguagens e como elas se relacionam.

1.5 ORGANIZAÇÃO

O trabalho consiste de cinco capítulos.

Neste capítulo são definidas as motivações e a proposta para o estudo, os desafios enfrentados durante o percurso do mesmo e sua contribuição para a comunidade. Também é apresentada a disposição das informações contidas no trabalho como um todo.

O capítulo 2, Fundamentação, apresenta conceitos e definições necessárias para compreender a obra, assim como trabalhos relacionados. Seu objetivo é dar um embasamento teórico mínimo para o entendimento dos métodos e da análise crítica deste estudo.

O capítulo 3, Materiais e Métodos, descreve quais foram os procedimentos utilizados para obtenção de dados e geração dos mapas filogenéticos, mostrando os trechos de código e discutindo sobre os mesmos.

O capítulo 4, Análise Experimental, faz uma análise manual e crítica dos resultados obtidos e faz uma comparação com o que existe de literatura em áreas similares ao tema.

O capítulo 5, Conclusão, apresenta o fechamento deste trabalho e dá sugestões de possíveis trabalhos futuros relacionados.

2 FUNDAMENTAÇÃO

Neste capítulo serão introduzidos os conceitos e definições fundamentais para estabelecer a proposta estudada, focando na geração dos mapas filogenéticos para as linguagens trabalhadas. Logo depois, alguns trabalhos relacionados serão apresentados, passando brevemente por suas propostas e conclusões.

2.1 CONCEITOS

2.1.1 Genômica Computacional

Um campo da bioinformática, a genômica computacional consiste no uso de análise estática e computação científica para decifrar sequências de genes e dados relacionados, como sequências de DNA e RNA com o objetivo de descobrir informações além da biologia (KOONIN, 2001). É sabido que a comparação entre *strings* genéticas é a forma mais fácil e garantida de identificar partes do genoma, além de possibilitar a predição de interações e funções das mesmas.

Isso se deve ao fato de que, através do tempo e sem intervenção externa, as tripas de nucleotídeos vão acumular mudanças e divergir sem ser possível reconhecer qualquer ancestralidade entre os genes de seres distintos. Assim, se é possível a detecção de similaridade entre genomas, e a literatura nos diz que tal similaridade é extensiva mesmo entre espécies muito distintas, indica que existe uma informação vital na sequência, que pode ser então analisada e compreendida.

2.1.2 Alinhamento de Sequência & Distância de Edição

Na bioinformática, alinhamento de sequência é uma forma de arranjar sequências de DNA, RNA ou proteína de forma a encontrar similaridades que sejam consequência de relações entre um conjunto diferente *strings* genéticas.(GOLLERY, 2005) As técnicas de alinhamento são divididas entre alinhamentos globais e locais: os primeiros realizam sua análise por todo o comprimento de um par de sequências, enquanto que os locais se concentram em subsequências na disposição das tripas utilizadas. Independente do método utilizado, ele retorna um valor numérico que indica a proximidade entre duas *strings*.

Ela é equivalente a chamada distância de edição(SELLERS, 1974), que em linguística computacional é uma métrica para indicar o quão diferentes duas *strings* são entre elas, minimizando a quantidade de operações necessárias para transformar uma em outra (NAVARRO, 2001).

Em ambos os conceitos, algoritmos diferentes comportam pelo menos uma entre as seguintes operações, com cada uma recebendo um score diferente dependendo de como é usado e dando outro score caso os caracteres sejam iguais:

- Inserção: inserir um caractere novo para eliminar um espaço em branco.
- Apagamento: deletar um caractere para criar um espaço em branco.
- Substituição: trocar um caractere por outro.

Como exemplo, tomemos o seguinte par de palavras no português e inglês, emergir-*emerge* e apliquemos um alinhamento de sequência dando os seguintes valores: dois caracteres iguais recebem o valor 1, e qualquer uma das três operações realizadas tem valor 0.

- Ambas as palavras começam com e, portanto, adicionemos 1 ao valor.
- A segunda letra de ambas as palavras é m, elevando o valor total para 2.
- As três letras seguintes de ambas as palavras, em**erg**ir e *em**erg**e* também são iguais, assim deixando nosso placar em 5.
- A próxima letra de ambas as palavras, emerg**i**r e *emerg**e*** difere, e como nossas operações não alteram nosso valor, ele continua em 5.
- Finalmente, existe a abertura de um espaço em branco devido ao fato de emergir ter uma letra a mais que *emerge*, resultando num placar final de 5.

2.1.3 Árvores Filogenéticas

Árvores filogenéticas são representações gráficas para representar inferências de relações entre espécies, uma ferramenta importante em campos diversos como biologia molecular, ecologia e fisiologia. (KHALAFVAND, 2015) Existem diversas formas de criar tais árvores, porém, as utilizadas neste trabalho são as chamadas árvores de distância de pares com raiz, indicando que elas são criadas através de matrizes de distância entre diversas sequências analizadas e que elas possuem uma direção e um nodo raiz. (LEMEY; SALEMI; VANDAMME, 2009)

Em ambos os algoritmos utilizados, NJ (TREES, 1987) e UPGMA (RR, 1958), as imagens geradas tem eixo vertical indicando os grupamentos mais próximos inferidos pelo algoritmo utilizado e o eixo horizontal indicando a evolução do espécime indicado em relação ao nodo pai.

2.2 TRABALHOS RELACIONADOS

2.2.1 Uma abordagem etimológica para semelhança ortográfica entre idiomas e sua aplicação em Romeno

Uma abordagem computacional para determinar a similaridade ortográfica entre idiomas foi proposta por(CIOBANU; DINU, 2014a), dando como exemplo o uso do romeno comparado com outras 20 línguas. O trabalho também leva em conta cognatos e étimos, investigando não apenas o número de palavras relacionadas mas também suas várias formas.

Os resultados do estudo suportam as teorias vigentes na literatura, incluindo a teoria de Darwin (DARWIN, 1859) que dita que a genealogia das línguas segue a genealogia dos povos que as falam, fator importante para validar a proposta deste estudo.

2.2.2 Detecção automática de cognatos usando alinhamento ortográfico

(CIOBANU; DINU, 2014b) propõem um método para identificação automática de cognatos quando há uma falta de informação etimológica necessária para estudos mais profundos, extraindo características ortográficas que possam servir para identificar as mudanças mais prováveis entre pares de línguas através da distância de edição para cognatos conhecidos e o uso de aprendizado de máquina, com o trabalho focando na relação do romeno com as outras línguas românicas.

Os autores concluem que, apesar da simplicidade das análises utilizadas, o método proposto é superior a métodos que utilizam métricas ortográficas com características individuais apenas, e propõe mudanças que podem melhorar o mesmo.

3 MATERIAIS E MÉTODOS

Neste capítulo são apresentados os instrumentos e métodos utilizados nos experimentos, os resultados destes descritos no capítulo 4. O capítulo começa descrevendo a base de dados utilizada e sua expansão, as ferramentas utilizadas para processar os dados e encerra com a explicação do cálculo utilizado para a distância entre os pares de palavras e línguas.

3.1 DESCRIÇÃO DOS EXPERIMENTOS

Os experimentos se dividem em três partes, sendo elas:

1. A geração dos arquivos csv com $\triangle$ para todos os pares de palavras para cada par de línguas
2. A geração da matriz de distâncias, através do cálculo da distância $\triangle$ entre pares de línguas
3. A inferência das árvores filogenéticas através da matriz de distância

Cada uma dessas partes foi executada onze vezes; uma vez com a base completa obtida de (PORTUGUÊS, s.d.) e dez vezes através de um *split* aleatório feito na base completa para validação maior dos resultados.

3.2 BASE DE DADOS

Como base de dados, foram utilizadas duas listas diferentes de palavras. A primeira consiste num pdf da Academia Brasileira de Letras que indica os verbos, adjetivos e substantivos mais frequentes em português (LETRAS, s.d.), utilizada apenas devido a seu pequeno tamanho, 109 palavras, para construir os *scripts* necessários para o experimento sem necessidade de demora na execução dos algoritmos. A segunda base de dados provém do dicionário online de português Dicio, contendo segundo o site, as palavras mais utilizadas de acordo com o número de letras da mesma (de 3 a 15 letras), contendo 1193 palavras após retirada de substantivos compostos na tentativa de simplificar a tradução (PORTUGUÊS, s.d.).

É de se notar que, devido ao método utilizado para cálculo do alinhamento global entre cada par de palavras não aceitar diacríticos, foi necessária a retirada dos mesmos de todas as palavras utilizadas como dados neste trabalho.

3.2.1 Expansão da Base de Dados

Apesar de termos um dataset razoável em português, houve uma dificuldade em encontrar de forma fácil as traduções para as línguas que gostaríamos de comparar neste trabalho. Para fim de conseguir os dados necessários, portanto, foi necessário realizar um trabalho de expansão através do uso da biblioteca *Translate Shell*, um tradutor de linha de comando que dá acesso fácil a diferentes ferramentas de tradução. (YAO, s.d.) No caso deste trabalho, utilizamos o *Google Translator* para, uma a uma, traduzir as palavras do português para as outras línguas utilizadas no estudo: francês, romeno, galego, italiano, espanhol e francês, e então adicionar a combinação de todas elas num arquivo csv através do uso de um *script* em python.

Uma tradução que depende de um tradutor *online* não se compara a uma base de dados feita por especialistas, o que irá impactar nos resultados obtidos. Porém, para efeitos deste estudo, que contempla apenas a viabilidade inicial do uso de algoritmos da genômica computacional na área linguística, foi considerado uma alternativa aceitável.

Além disso, também foi feita uma divisão de forma aleatória, chamada de *split*, de 10 partes na base de dados após a tradução das línguas para um experimento comparativo, de forma a verificar a viabilidade do algoritmo utilizado em bases menores e a robustez dos resultados obtidos no geral.

3.3 FERRAMENTAS

Os experimentos foram realizados com as ferramentas descritas a seguir devido a facilidade de uso e documentação extensiva. Nesta seção será apresentada a biblioteca *Biopython* e os dois módulos da mesma utilizados para a execução do trabalho.

3.3.1 *Biopython*

O projeto *Biopython* é uma iniciativa de software livre desenvolvida por uma associação internacional de desenvolvedores para disponibilizar ferramentas de acesso gratuito de biologia molecular computacional. (COCK et al., 2009) Entre seus diversos módulos, dois foram utilizados nos experimentos descritos no capítulo 4: o módulo *Pairwise2*, que disponibiliza métodos de alinhamento de sequência pareada, e o módulo *Phylo*, que provê suporte para trabalhos de com árvores filogenéticas.

3.3.1.1 *Pairwise2*

Este módulo prove métodos para utilização dos dois tipos de alinhamento de sequência: alinhamentos globais e locais. Por conta das diferenças entre ambas as técnicas, e considerando que nossas maiores *strings* contêm poucos caracteres e não

diferem muito de tamanho, foi decidido pela utilização do alinhamento global como técnica.

Para o cálculo do alinhamento, o método implementado pelo módulo recebeu os seguintes parâmetros:

- Caso dois caracteres ocupem a mesma posição, dê um score de +1 para o alinhamento.
- Caso dois caracteres sejam diferentes numa posição, dê um score de -0,5 para o alinhamento.
- Caso ocorra a abertura de um novo espaço em branco, dê um score de -0,2 para o alinhamento.
- Caso ocorra a expansão de um espaço em branco existente, dê um score de -0,1 para o alinhamento.

É importante notar que não houve uma análise para identificar os melhores valores neste trabalho, já que, conforme (CIOBANU; DINU, 2014b) indica uma variação para detecção de cognatos, há uma necessidade de gerar matrizes de substituição específicas para cada par de línguas de forma a otimizar os resultados, o que foge do escopo aqui apresentado.

Para exemplificar o cálculo do alinhamento, vamos resgatar o exemplo dado no capítulo 2.1.2, emergir-*emerge*.

- É fácil identificar que existe um alinhamento entre as cinco primeiras letras, **emerg**ir e ***emerg****e*, portanto, adicionamos 5 no nosso score de alinhamento.
- A próxima letra de ambas as palavras, emerg**i**r e *emerg**e*** difere, portanto, subs-traímos 0,5 do nosso score anterior, totalizando 4,5 neste ponto.
- Finalmente, existe a abertura de um espaço em branco devido ao fato de emergir ter uma letra a mais que *emerge*, então subtraímos mais 0,1, totalizando um score de 4,4.

3.3.1.2 *Phylo*

Este módulo prove métodos para lidar com árvores filogenéticas, desde ler e plotar uma nova árvore até criar uma através de uma matriz de distâncias. Para este trabalho, foram utilizados os métodos do submódulo *DistanceTreeConstructor*, já que estamos trabalhando com distâncias entre strings, para gerar as árvores NJ e UPGMA.

Neste trabalho, devido a natureza dos dados utilizados, geramos manualmente a matriz de distâncias através do cálculo descrito na seção 3.3 para então chamar os construtores das árvores utilizadas.

3.4 CÁLCULO DA DISTÂNCIA ENTRE LÍNGUAS

Conforme descrito na seção 3.2.1.1, o módulo *Pairwise2* nos dá o score de alinhamento Θ entre um par de *strings*. A partir então do score de alinhamento, a proposta é normalizar e depois tirar a inversa para conseguirmos a distância normalizada de edição Δ, e, após isso, tirar a média de todas as distâncias entre cada par de línguas. Portanto, para cada par de palavras (P1, P2), podemos obter Δ através da seguinte fórmula:

$$\underset{normal}{\Theta(P1,P2)} = \frac{\Theta(P1,P2) + (\frac{menor(P1,P2)}{2})}{3 * (\frac{menor(P1,P2)}{2})}$$

$$\Delta(P1,P2) = 1 - \underset{normal}{\Theta(P1,P2)}$$

Tendo em mãos todos os Δ(P1,P2) para um par de línguas (L1,L2) contendo um total de Λ pares de palavras, podemos calcular a distância normalizada Δ entre L1 e L2 sendo:

$$\Delta(L1,L2) = \frac{\sum_{n=1}^{\Lambda} \Delta(P1,P2)}{\Lambda}$$

3.5 CONSIDERAÇÕES FINAIS

Os dados, código e resultados gerados por este trabalho estão disponíveis no GitHub do autor via link: `https://github.com/CesarCamillo/TCC/tree/master`, na seguinte disposição:

- O diretório resultados1200 contém os arquivos .csv para cada par de línguas, as imagens das árvores geradas e um arquivo de texto contendo a matriz de distância para todo o banco de dados utilizado
- O diretório resultadossplit contém os subdiretórios para as 10 divisões da base de dados completa, contendo todos os arquivos usados e gerados nos experimentos com os *splits*

- O script tocsv.py pega um arquivo de texto contendo uma palavra por linha em português e cria um arquivo .csv contendo cada palavra e as traduções geradas pelo *Google Translate*
- O script tensplit.py pega o arquivo .csv gerado pelo script anterior e faz a divisão em dez para os experimentos realizados
- O script alinhamento.py faz, para cada par de línguas, o cálculo da distância entre cada palavra utilizada
- O script treemaker.py realiza a geração das matrizes de distância e das árvores filogenéticas
- Os arquivos maisfreq1200.csv contém todas as palavras e suas traduções utilizadas nos experimentos

4 ANÁLISE EXPERIMENTAL

Neste capítulo são descritos os resultados dos dois experimentos realizados com a geração das distâncias entre palavras e línguas e com a inferência das árvores filogenéticas, realizando uma comparação das distâncias do romeno para outras línguas com os resultados obtidos por (CIOBANU; DINU, 2014a). Por fim, as árvores de ambos os experimentos são comparadas entre si para identificar padrões e discrepâncias de forma a analisar a validade do trabalho para além de possíveis problemas com a base ou os cálculos.

4.1 ANÁLISE DAS MATRIZES DE DISTÂNCIA

Para todos os efeitos, as distâncias apresentadas na tabela abaixo estão truncadas se comparado com o valor obtido no cálculo do experimento na terceira casa decimal por uma questão de espaçamento da tabela:

4.1.1 Base Completa

A matriz triangular inferior de distância de edição normalizada da base completa como um todo pode ser encontrada abaixo na tabela 1:

Línguas	Português	Inglês	Espanhol	Francês	Italiano	Galego	Romeno
Português	0						
Inglês	0.327	0					
Espanhol	0.195	0.314	0				
Francês	0.308	0.285	0.287	0			
Italiano	0.267	0.333	0.242	0.289	0		
Galego	0.189	0.324	0.131	0.292	0.243	0	
Romeno	0.189	0.339	0.314	0.315	0.304	0.316	0

Tabela 1 – Matriz das Distâncias de Edição Normalizadas feitas com a Base Completa

Para efeitos de comparação, segue uma parte da tabela de resultados de (CIOBANU; DINU, 2014a) mostrando o valor de similaridade calculado pelos pesquisadores com seu método com relação ao romeno, demonstrado abaixo na tabela 2. É importante frisar que, apesar da diferença dos nomes dos resultados entre os dois estudos, devido a equivalência dos algoritmos utilizados(SELLERS, 1974), o resultado final de ambos indica a mesma coisa - quão próximas duas línguas são.

Analisando os cálculos de ambos os trabalhos, podemos facilmente notar a similaridade entre os valores entre os dois trabalhos no que diz respeito ao romeno: português, italiano e francês se encontram dentro da gama de resultados obtidos por

Língua/% para cada corpus	Parliament	Eminescu	Chronicles	RVR
Português	24.0	21.0	18.4	18.9
Inglês	14.2	10.1	6.2	10.4
Espanhol	27.0	23.7	17.2	21.0
Francês	48.8	38.2	23.1	33.3
Italiano	34.5	31.2	21.4	28.0

Tabela 2 – Similaridade entre línguas inferida em (CIOBANU; DINU, 2014a) para o romeno, em porcentagem, desconsiderando diacríticos e contando cognatos e étimos

(CIOBANU; DINU, 2014a), enquanto espanhol se encontra relativamente próximo. A maior discrepância, o inglês, pode estar relacionada tanto as palavras usadas quanto as traduções feitas, sendo necessário uma investigação que foge ao escopo deste trabalho para avaliar melhor este caso específico.

4.1.2 Base após *Split*

A tabela 3 abaixo traz a distância entre línguas de todas as divisões criadas.

Línguas	Português	Inglês	Espanhol	Francês	Italiano	Galego	Romeno
Português 1	0.000						
Inglês 1	0.303	0.000					
Espanhol 1	0.182	0.304	0.000				
Francês 1	0.309	0.284	0.275	0.000			
Italiano 1	0.275	0.335	0.231	0.289	0.000		
Galego 1	0.182	0.310	0.124	0.276	0.252	0.000	
Romeno 1	0.182	0.371	0.306	0.333	0.311	0.317	0.000
Português 2	0.000						
Inglês 2	0.324	0.000					
Espanhol 2	0.199	0.309	0.000				
Francês 2	0.302	0.277	0.278	0.000			
Italiano 2	0.263	0.332	0.249	0.287	0.000		
Galego 2	0.196	0.351	0.138	0.306	0.256	0.000	
Romeno 2	0.196	0.313	0.299	0.274	0.279	0.307	0.000
Português 3	0.000						
Inglês 3	0.388	0.000					
Espanhol 3	0.205	0.358	0.000				
Francês 3	0.332	0.328	0.312	0.000			
Italiano 3	0.289	0.377	0.249	0.278	0.000		
Galego 3	0.190	0.349	0.135	0.309	0.253	0.000	
Romeno 3	0.190	0.390	0.354	0.334	0.328	0.347	0.000
Português 4	0.000						

Inglês 4	0.319	0.000					
Espanhol 4	0.202	0.295	0.000				
Francês 4	0.330	0.267	0.288	0.000			
Italiano 4	0.292	0.306	0.246	0.304	0.000		
Galego 4	0.194	0.297	0.119	0.298	0.230	0.000	
Romeno 4	0.194	0.345	0.320	0.336	0.327	0.326	0.000
Português 5	0.000						
Inglês 5	0.311	0.000					
Espanhol 5	0.203	0.299	0.000				
Francês 5	0.302	0.273	0.287	0.000			
Italiano 5	0.274	0.335	0.258	0.315	0.000		
Galego 5	0.204	0.309	0.143	0.291	0.253	0.000	
Romeno 5	0.204	0.327	0.333	0.322	0.324	0.321	0.000
Português 6	0.000						
Inglês 6	0.304	0.000					
Espanhol 6	0.184	0.303	0.000				
Francês 6	0.295	0.272	0.287	0.000			
Italiano 6	0.229	0.311	0.250	0.268	0.000		
Galego 6	0.167	0.289	0.133	0.276	0.233	0.000	
Romeno 6	0.167	0.334	0.324	0.304	0.305	0.314	0.000
Português 7	0.000						
Inglês 7	0.361	0.000					
Espanhol 7	0.221	0.347	0.000				
Francês 7	0.320	0.307	0.308	0.000			
Italiano 7	0.266	0.359	0.253	0.313	0.000		
Galego 7	0.250	0.366	0.173	0.309	0.264	0.000	
Romeno 7	0.250	0.367	0.315	0.344	0.315	0.342	0.000
Português 8	0.000						
Inglês 8	0.321	0.000					
Espanhol 8	0.182	0.311	0.000				
Francês 8	0.301	0.280	0.287	0.000			
Italiano 8	0.259	0.323	0.233	0.304	0.000		
Galego 8	0.189	0.339	0.121	0.307	0.239	0.000	
Romeno 8	0.189	0.340	0.306	0.322	0.291	0.308	0.000
Português 9	0.000						
Inglês 9	0.315	0.000					
Espanhol 9	0.197	0.298	0.000				

Francês 9	0.313	0.274	0.291	0.000			
Italiano 9	0.254	0.312	0.221	0.277	0.000		
Galego 9	0.171	0.304	0.117	0.283	0.223	0.000	
Romeno 9	0.171	0.302	0.298	0.312	0.294	0.304	0.000
Português 10	0.000						
Inglês 10	0.325	0.000					
Espanhol 10	0.178	0.318	0.000				
Francês 10	0.282	0.288	0.265	0.000			
Italiano 10	0.273	0.346	0.236	0.257	0.000		
Galego 10	0.154	0.330	0.112	0.268	0.233	0.000	
Romeno 10	0.154	0.310	0.287	0.275	0.268	0.282	0.000

Tabela 3 – Matriz das distâncias de todos os splits

Da mesma forma que a tabela 1, a tabela 3 também realiza um truncamento na terceira casa decimal. Uma simples vista sobre dos valores da tabela mostra que, apesar de haver uma variação considerável, os valores estão consistentes - nenhum dos *splits* destoa muito dos outros ou dos valores obtidos, o que indica que o cálculo utilizado é consistente o suficiente.

Abaixo, as tabelas 4 e 5 trazem a média e da soma dos quadrados do desvio padrão entre toda a amostragem dos dados pela tabela 3 por par de línguas.

Línguas	Português	Inglês	Espanhol	Francês	Italiano	Galego	Romeno
Português	0						
Inglês	0.3271	0					
Espanhol	0.1953	0.3142	0				
Francês	0.3086	0.285	0.2878	0			
Italiano	0.2674	0.3336	0.2426	0.2892	0		
Galego	0.1897	0.3244	0.1315	0.2923	0.2436	0	
Romeno	0.1897	0.3399	0.3142	0.3156	0.3042	0.3168	0

Tabela 4 – Média dos valores dos *splits*

Línguas	Português	Inglês	Espanhol	Francês	Italiano	Galego	Romeno
Português	0						
Inglês	0.0064949	0					
Espanhol	0.0016561	0.0041376	0				
Francês	0.0021924	0.00319	0.0017856	0			
Italiano	0.0029504	0.0045404	0.0012304	0.0034356	0		
Galego	0.0060981	0.0061524	0.0028245	0.0022041	0.0016724	0	
Romeno	0.0060981	0.0074929	0.0034356	0.0054724	0.0038656	0.0032056	0

Tabela 5 – Soma dos quadrados do desvio das distâncias normalizadas dos *splits*

Como é possível ver, os resultados da média se aproximam muito dos obtidos durante o experimento com todos os dados, e a soma dos quadrados é baixa. Isso mostra a robustez do algoritmo utilizado, já que funciona mesmo com amostras muito pequenas e não varia muito entre elas.

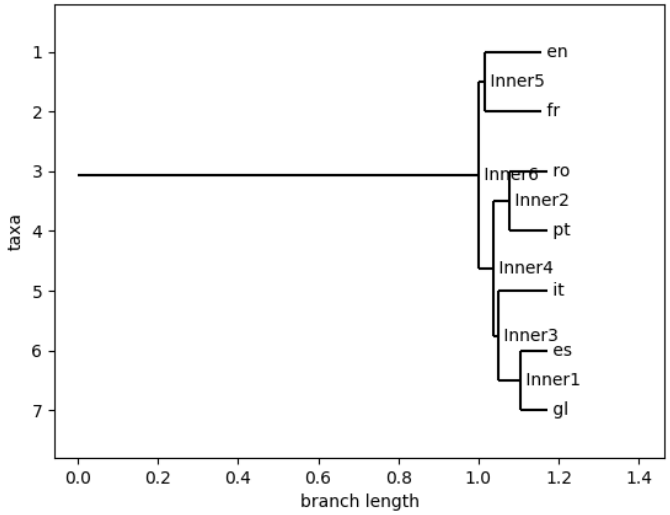
4.2 ANÁLISE DAS ÁRVORES DE INFERÊNCIA

No apêndice A, estão concentradas as imagens das árvores de inferência geradas por todas as matrizes de distância elaboradas nas tabelas 1 e 3. Para exemplificação dos resultados, algumas delas serão exibidas aqui.

4.2.1 UPGMA

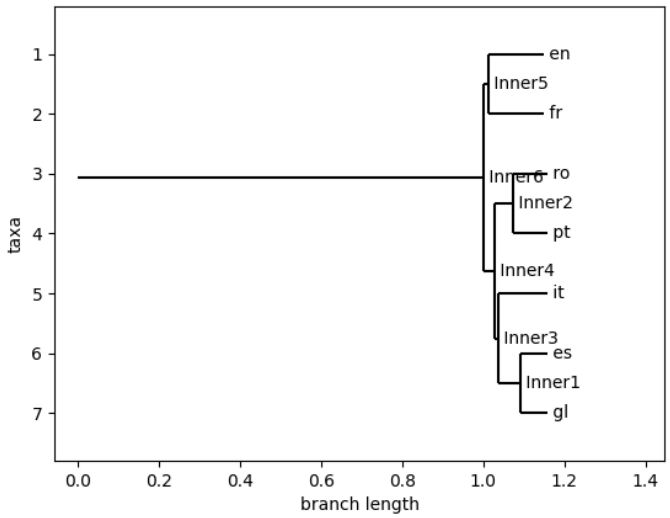
As figuras 1 a 4, logo abaixo, comparam 4 árvores filogenéticas geradas através do algoritmo UPGMA:

Figura 1 – UPGMA BASE COMPLETA



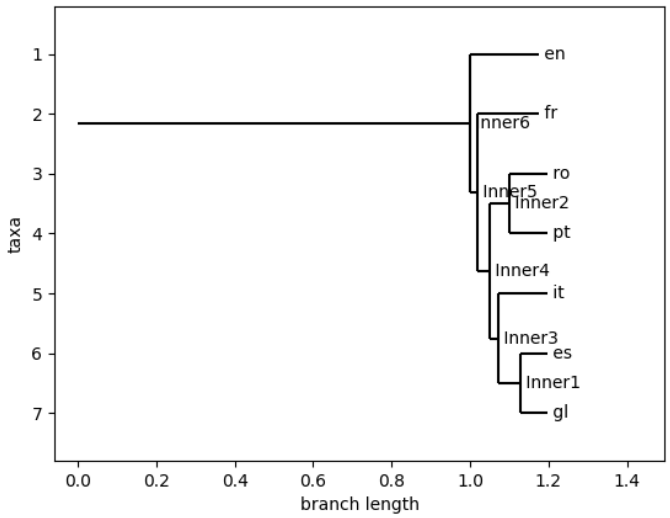
FONTE: o Autor (2023)

Figura 2 – UPGMA BASE *SPLIT* 6



FONTE: o Autor (2023)

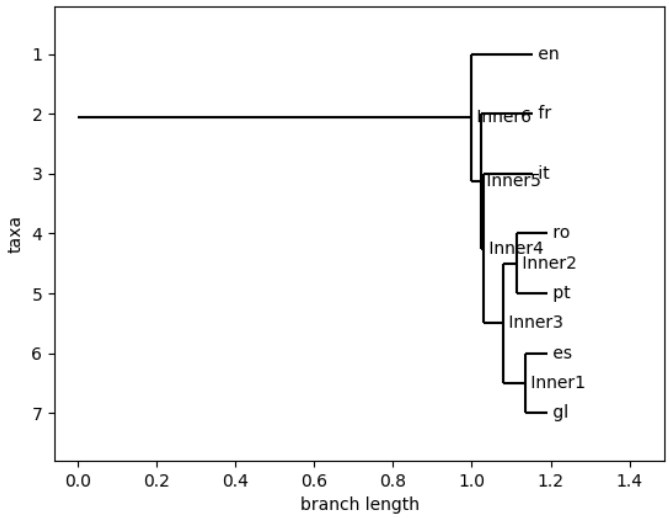
Figura 3 – UPGMA BASE *SPLIT* 3



FONTE: o Autor (2023)

Considerando os valores levemente diferentes entre as distâncias em cada experimento realizado, as árvores correspondem bem ao que era esperado - seus formatos são bastante similares, e a maioria dos seus agrupamentos se repetem, mas a posição deles é levemente diferente em cada um dos mesmos, o que condiz com os resultados demonstrados nos capítulos 4.1.1 e 4.1.2

Figura 4 – UPGMA BASE *SPLIT* 10

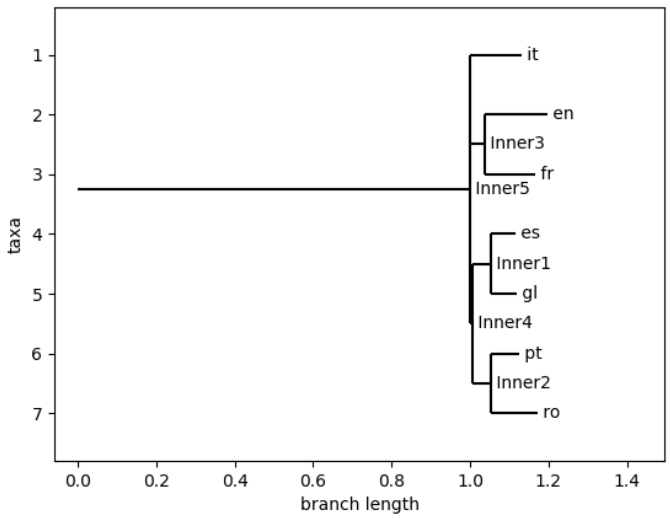


FONTE: o Autor (2023)

4.2.2 NJ

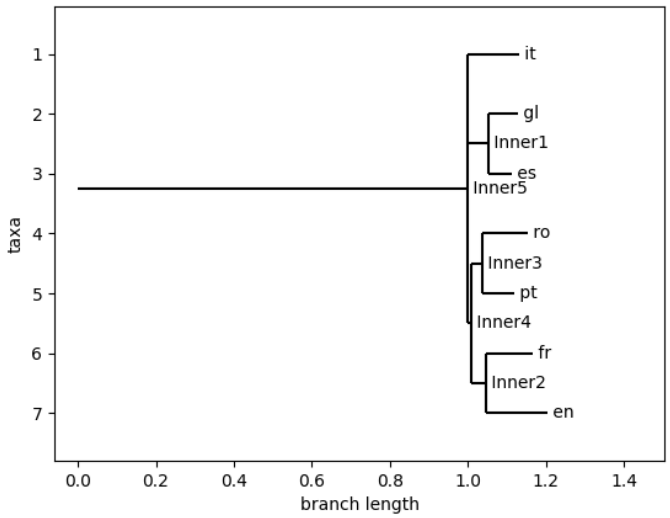
As figuras 5 a 8 abaixo comparam 4 árvores filogenéticas geradas através do algoritmo NJ:

Figura 5 – NJ BASE COMPLETA



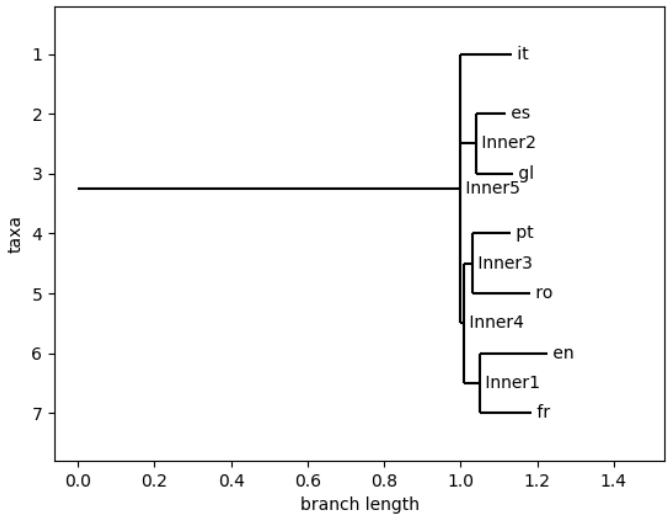
FONTE: o Autor (2023)

Figura 6 – NJ BASE *SPLIT* 2



FONTE: o Autor (2023)

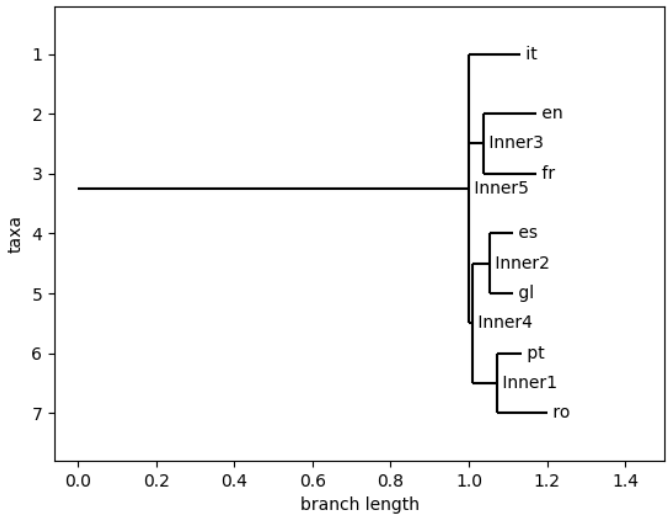
Figura 7 – NJ BASE *SPLIT* 7



FONTE: o Autor (2023)

Da mesma forma que nas árvores filogenéticas provenientes do algoritmo UPGMA, podemos notar grande similaridade entre todos os dendrogramas apresentados aqui. Uma diferença interessante entre ambas é que, no algoritmo NJ, o grupo mais externo costuma ser o italiano ao invés do inglês, este último sendo agrupado

Figura 8 – NJ BASE *SPLIT* 4



FONTE: o Autor (2023)

diretamente com o francês, diferença tal que pode ser atribuída ao algoritmo e base de dados não confiável utilizados no estudo.

5 CONCLUSÃO

Este estudo iniciou a investigação de algortimos da área da gênomica computacional num ambiente em que tais capacidades parecem ter sido pouco exploradas. Apesar de todos os desafios encontrados durante a realização dos experimentos, os dados incluídos e a similaridade entre as árvores filogenéticas indicam que o uso dos algoritmos de genômica computacional na área de etimologia é viável.

Futuros refinamentos sem fugir de uma análise apenas ortográfica podem se basear nos seguintes fatores:

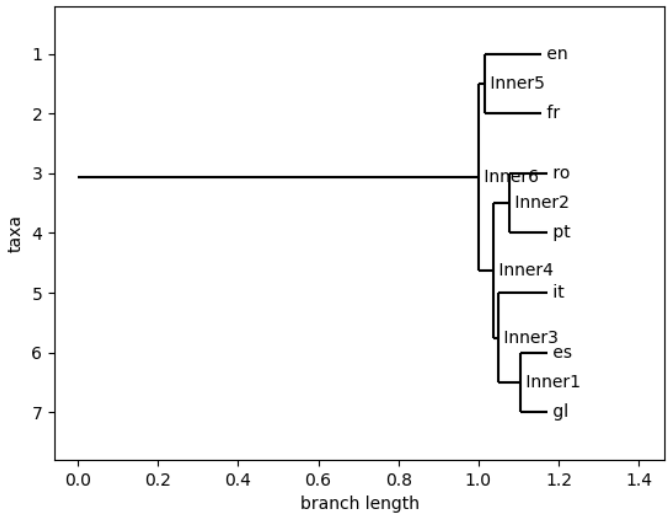
- Obtenção de um maior banco de dados pré expansão tradutória, garantindo assim um vocabulário maior, o que permite a extração de mais informações.
- Uso de dicionários mais robustos ao invés de recorrer ao *Google Translate*, melhorando qualitativamente os dados a serem analisados.
- Uso de um grupo de cognatos e étimos conhecidos entre as línguas a serem analisadas para a criação de uma matriz de substituição de forma a melhorar o cálculo da distância, de forma similar ao realizado em (CIOBANU; DINU, 2014b).

Já fora do âmbito discutido neste estudo, as seguintes modificações podem ser realizadas para ampliar as informações obtidas durante a experimentação:

- Uso de dados semânticos, sintáticos e/ou fonéticos na construção da base de dados.
- Geração de árvores filogenéticas através de algoritmos que não se utilizem de matrizes de distância como forma de expandir o grau de confiabilidade na solução de acordo com a hipótese de que mudanças ortográficas podem ser equiparadas a mudanças no genoma.

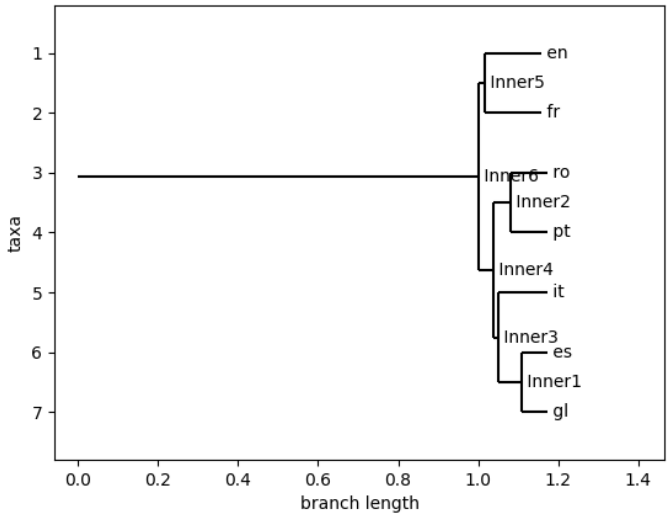
A MAPAS FILOGENÉTICOS GERADOS

Figura 9 – UPGMA BASE COMPLETA



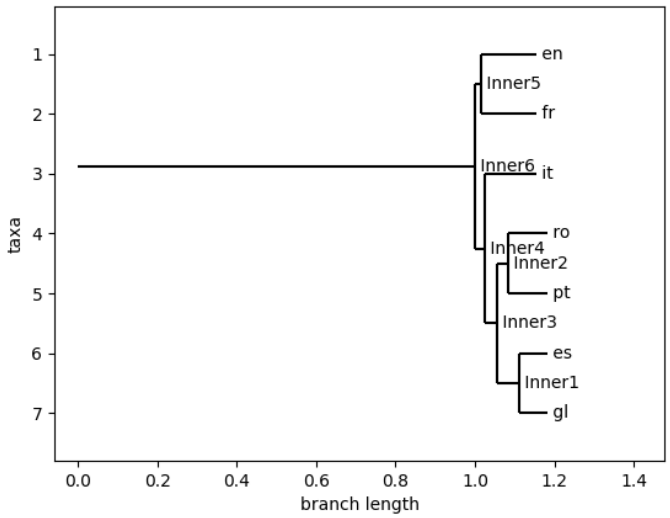
FONTE: o Autor (2023)

Figura 10 – UPGMA BASE *SPLIT* 1



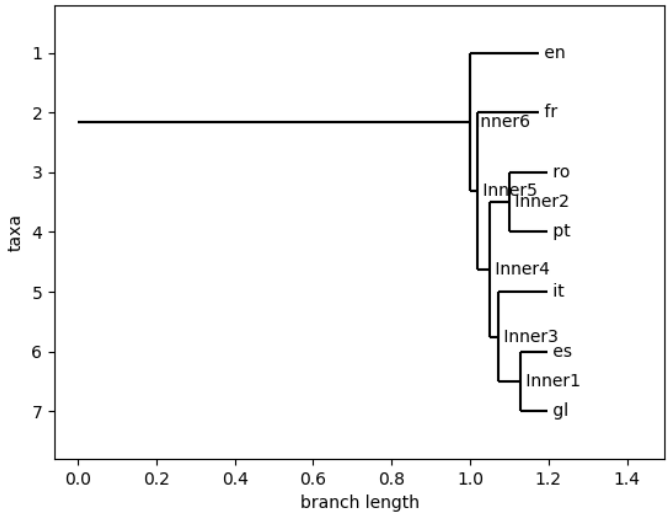
FONTE: o Autor (2023)

Figura 11 – UPGMA BASE *SPLIT* 2



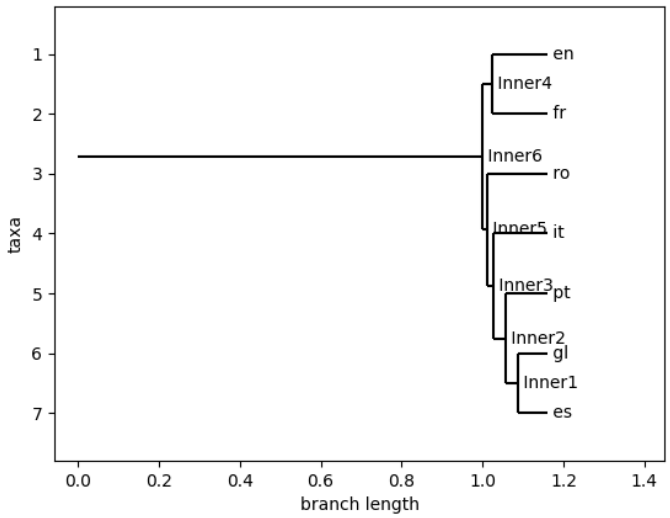
FONTE: o Autor (2023)

Figura 12 – UPGMA BASE *SPLIT* 3



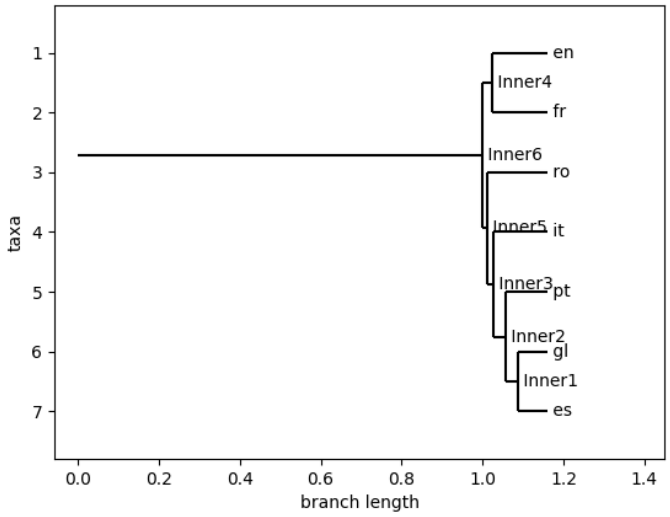
FONTE: o Autor (2023)

Figura 13 – UPGMA BASE *SPLIT* 4



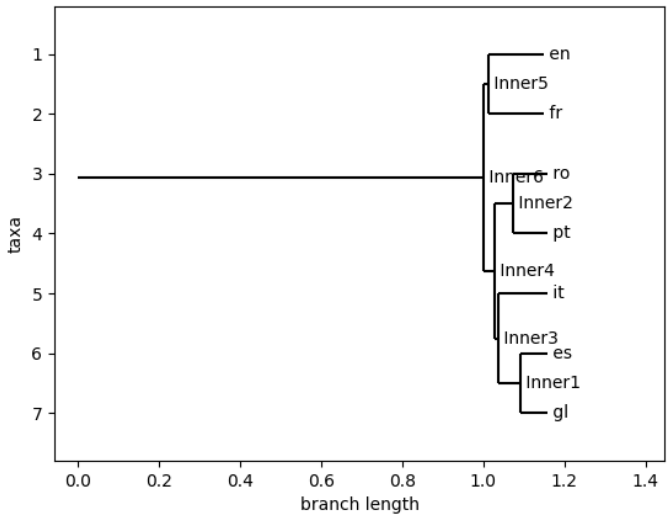
FONTE: o Autor (2023)

Figura 14 – UPGMA BASE *SPLIT* 5



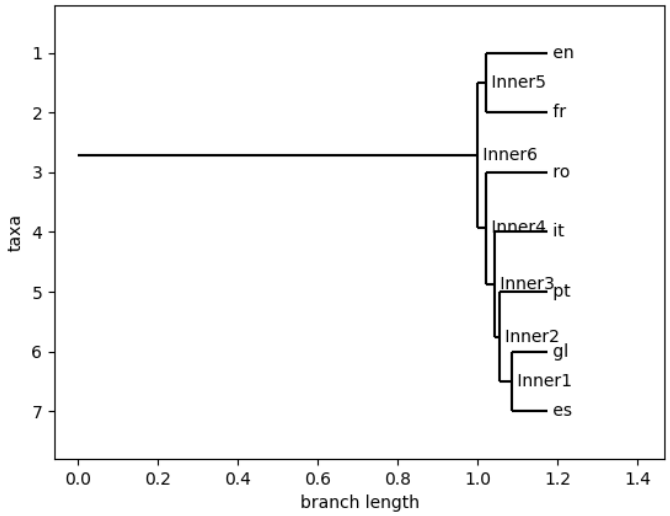
FONTE: o Autor (2023)

Figura 15 – UPGMA BASE *SPLIT* 6



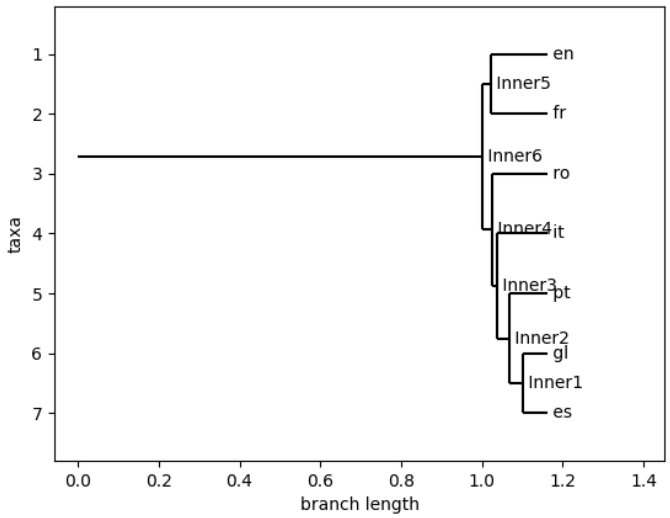
FONTE: o Autor (2023)

Figura 16 – UPGMA BASE *SPLIT* 7



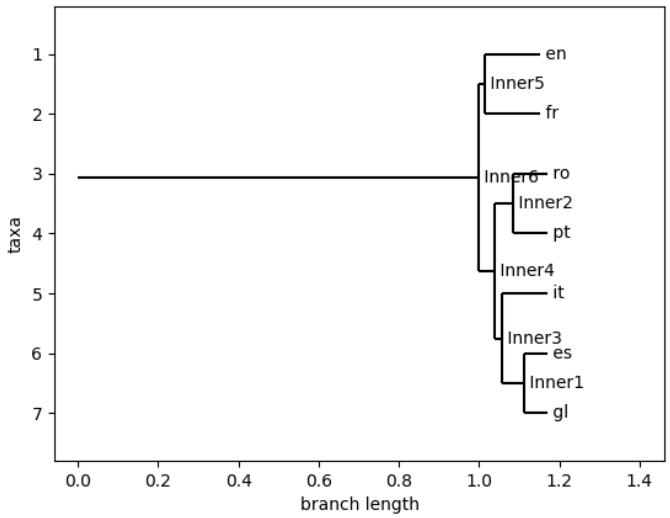
FONTE: o Autor (2023)

Figura 17 – UPGMA BASE *SPLIT* 8



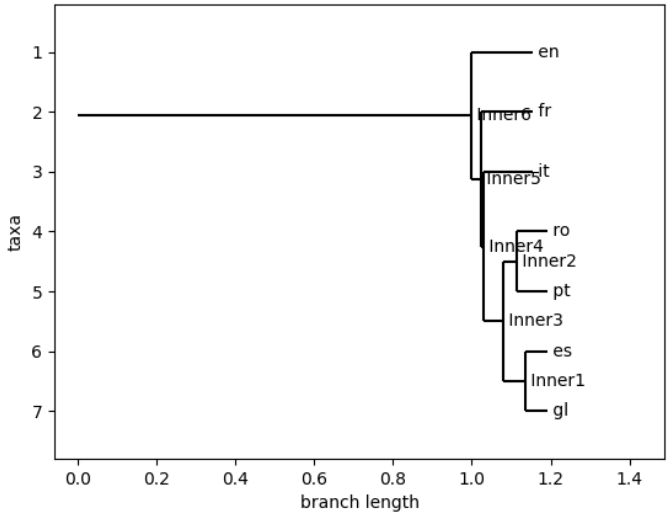
FONTE: o Autor (2023)

Figura 18 – UPGMA BASE *SPLIT* 9



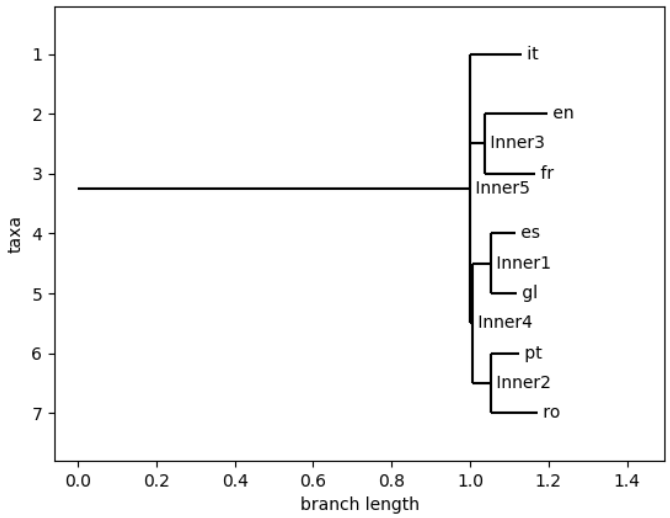
FONTE: o Autor (2023)

Figura 19 – UPGMA BASE *SPLIT* 10



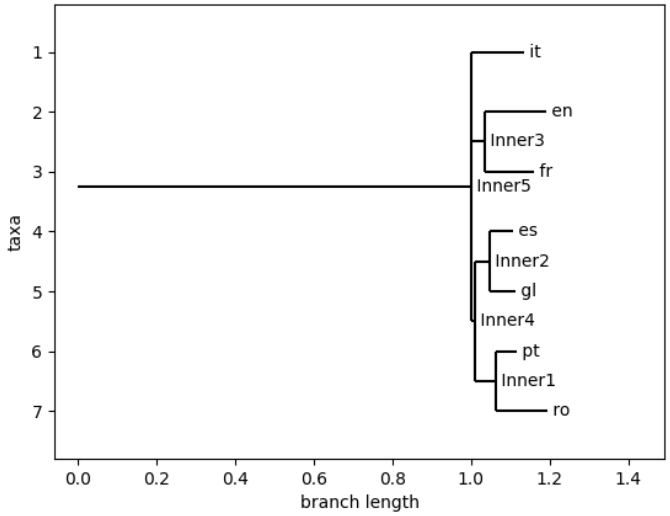
FONTE: o Autor (2023)

Figura 20 – NJ BASE COMPLETA



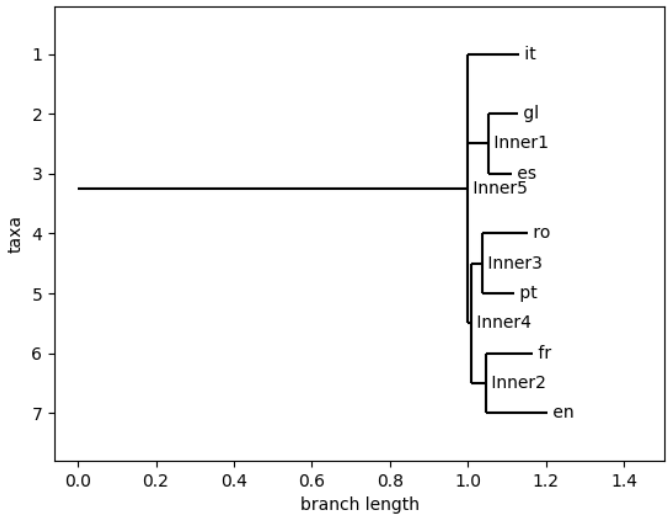
FONTE: o Autor (2023)

Figura 21 – NJ BASE *SPLIT* 1



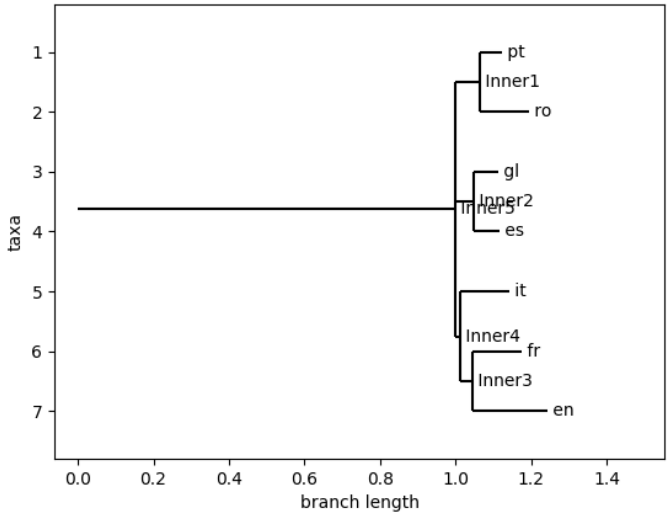
FONTE: o Autor (2023)

Figura 22 – NJ BASE *SPLIT* 2



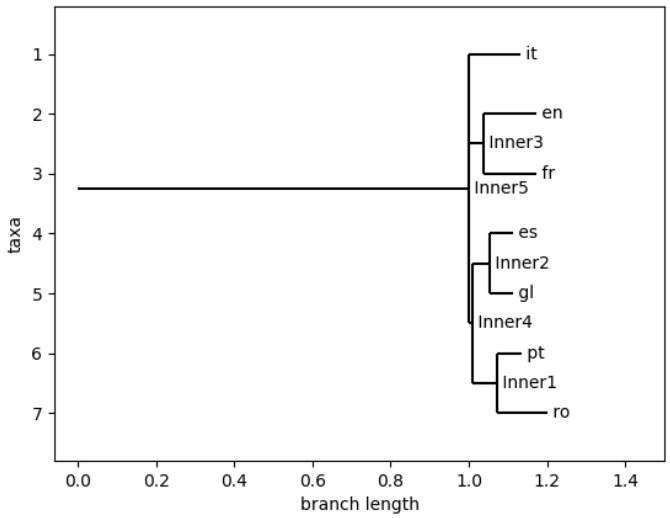
FONTE: o Autor (2023)

Figura 23 – NJ BASE *SPLIT* 3



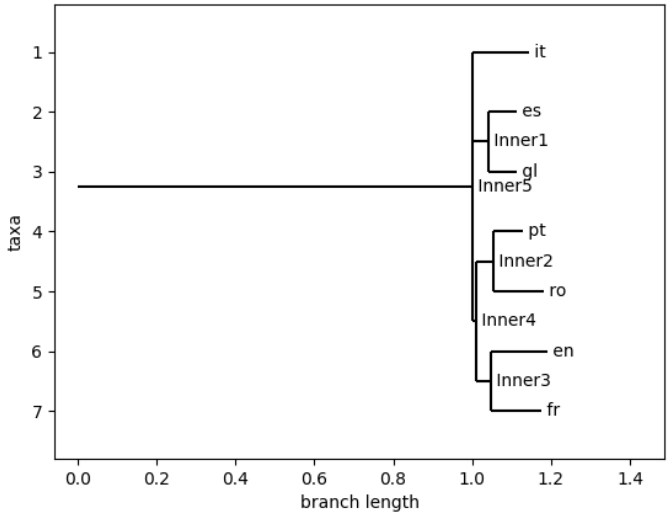
FONTE: o Autor (2023)

Figura 24 – NJ BASE *SPLIT* 4



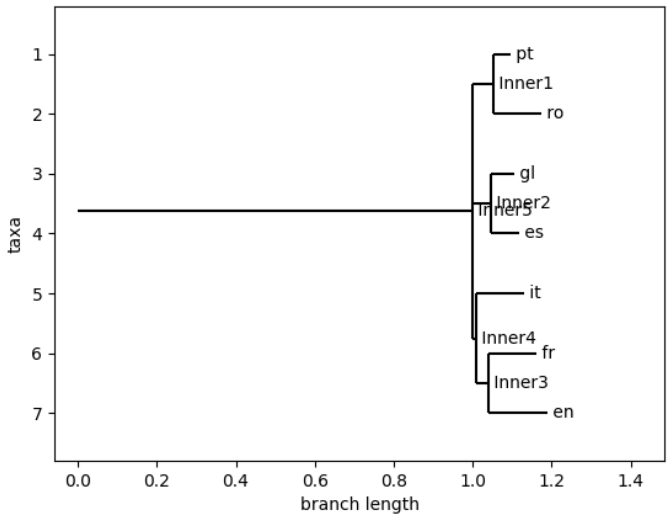
FONTE: o Autor (2023)

Figura 25 – NJ BASE *SPLIT* 5



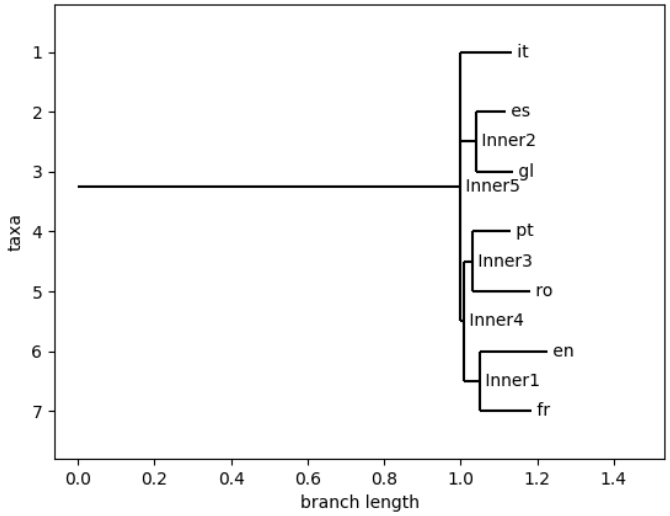
FONTE: o Autor (2023)

Figura 26 – NJ BASE *SPLIT* 6



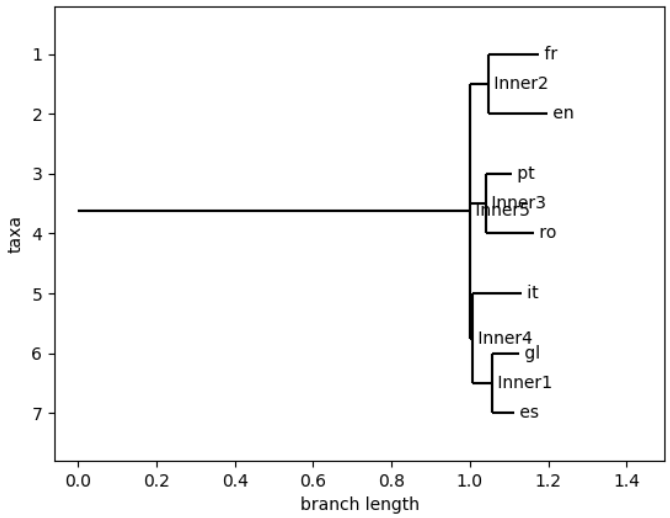
FONTE: o Autor (2023)

Figura 27 – NJ BASE *SPLIT* 7



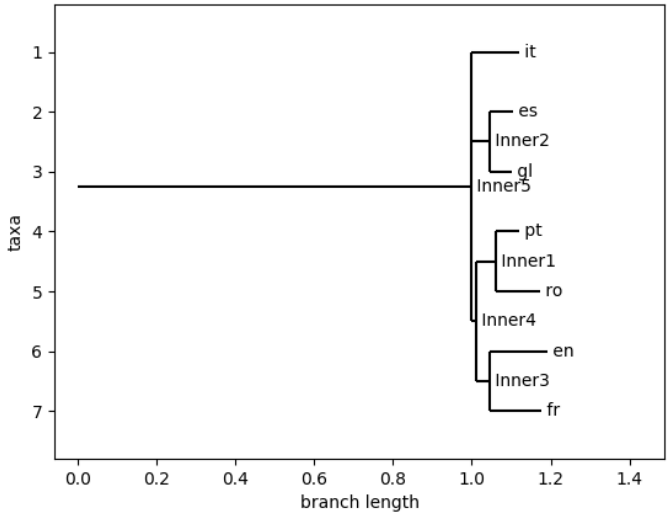
FONTE: o Autor (2023)

Figura 28 – NJ BASE *SPLIT* 8



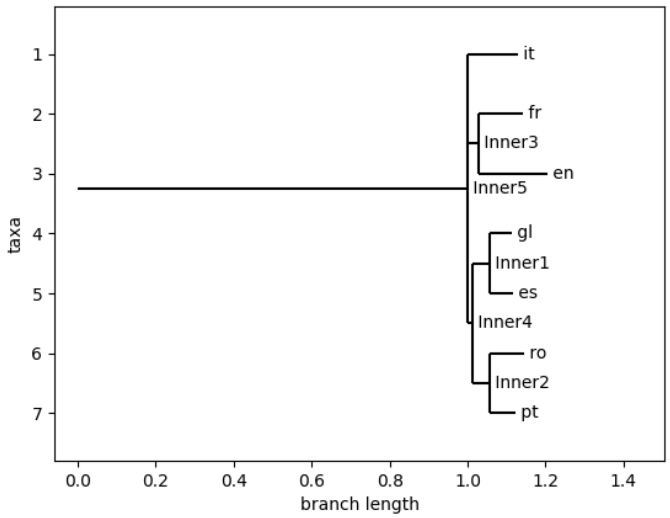
FONTE: o Autor (2023)

Figura 29 – NJ BASE *SPLIT* 9



FONTE: o Autor (2023)

Figura 30 – NJ BASE *SPLIT* 10



FONTE: o Autor (2023)

BIBLIOGRAFIA

CIOBANU, Alina Maria; DINU, Liviu P. An etymological approach to cross-language orthographic similarity. Application on Romanian. In: PROCEEDINGS of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). [S.l.: s.n.], 2014. P. 1047–1058. Citado 4 vezes nas páginas 14, 20, 21.

_____. Automatic detection of cognates using orthographic alignment. In: PROCEEDINGS of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). [S.l.: s.n.], 2014. P. 99–105. Citado 3 vezes nas páginas 14, 17, 29.

COCK, Peter JA et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. **Bioinformatics**, Oxford University Press, v. 25, n. 11, p. 1422–1423, 2009. Citado 1 vez na página 16.

DARWIN, Charles. On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. **London: Murray**, 1859. Citado 1 vez na página 14.

GOLLERY, Martin. Bioinformatics: Sequence and Genome Analysis, 2nd ed. David W. Mount. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press, 2004, 692 pp., $75.00, paperback. ISBN 0-87969-712-1$.. **Clinical Chemistry**, v. 51, n. 11, p. 2219–2219, nov. 2005. ISSN 0009-9147. DOI: 10.1373/clinchem.2005.053850. eprint: https://academic.oup.com/clinchem/article-pdf/51/11/2219/32684429/clinchem2219.pdf. Disponível em: <https://doi.org/10.1373/clinchem.2005.053850>. Citado 1 vez na página 12.

KHALAFVAND, Tyler Seyed Amin. **Finding Structure in the Phylogeny Search Space**. 2015. Tese (Doutorado). Citado 1 vez na página 13.

KOONIN, Eugene V. Computational genomics. **Current Biology**, Elsevier, v. 11, n. 5, r155–r158, 2001. Citado 1 vez na página 12.

LEMEY, Philippe; SALEMI, Marco; VANDAMME, Anne-Mieke. **The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing**. [S.l.]: Cambridge University Press, 2009. Citado 1 vez na página 13.

LETRAS, Academia Brasileira de. [S.l.: s.n.]. Acessado 17 de Setembro de 2023. https://www.academia.org.br/sites/default/files/relacao_das_palavras_mais_frequentes.pdf. Citado 1 vez na página 15.

MCMAHON, April; MCMAHON, Robert. Finding families: Quantitative methods in language classification. **Transactions of the Philological Society**, Wiley Online Library, v. 101, n. 1, p. 7–55, 2003. Citado 1 vez na página 10.

NAVARRO, Gonzalo. A guided tour to approximate string matching. **ACM computing surveys (CSUR)**, ACM New York, NY, USA, v. 33, n. 1, p. 31–88, 2001. Citado 1 vez na página 12.

PORTUGUÊS, Dicio - Dicionário Online de. [S.l.: s.n.]. Acessado 17 de Setembro de 2023. https://www.dicio.com.br/lista-de-palavras/. Citado 2 vez na página 15.

RAMA, Taraka et al. **Comparative evaluation of string similarity measures for automatic language classification.** [S.l.: s.n.], 2015. Citado 2 vez na página 10.

RR, SOKAL. A statiscal method for evaluating systematic relationships. **Univ Kans sci bull**, v. 38, p. 1409–1438, 1958. Citado 1 vez na página 13.

SELLERS, Peter H. On the theory and computation of evolutionary distances. **SIAM Journal on Applied Mathematics**, SIAM, v. 26, n. 4, p. 787–793, 1974. Citado 3 vezes nas páginas 10, 12, 20.

TREES, Reconstructing Phylogenetic. The Neighbor-joining Method: A New Method for. **Mol. Biol. Evol**, v. 4, n. 4, p. 406–425, 1987. Citado 1 vez na página 13.

YAO, Mort. **Translate Shell**. [S.l.: s.n.]. Disponível em: <https://www.soimort.org/translate-shell/>. Citado 1 vez na página 16.