Quantifying syntax similarity with a polynomial representation of dependency trees (supplementary material)

Pengyu Liu[1,2], Tinghao Feng[3], Rui Liu[4,5*]

[1] Department of Microbiology and Molecular Genetics, University of California, Davis
[2] Department of Mathematics, Simon Fraser University
[2] Department of Computer Science, Appalachian State University
[4] Department of Chinese Language and Literature, Beijing Normal University, Zhuhai
[5] Center for Linguistic Sciences, Beijing Normal University, Zhuhai
[*] Corresponding author's email: liu_rui@bnu.edu.cn

1 Supplementary figures and tables

We display the sentence in the ENG dataset with minimum and maximum polynomial distances between its English and French translations in Supplementary Figure 1 and 2. There are more than one sentences in the ENG dataset whose English and French translations have the same dependency tree, so we only display one sentence with the minimum polynomial distance between its English and French translations. We display the sentence in the ENG dataset with minimum and maximum polynomial distances between its English and Spanish translations in Supplementary Figure 3 and 4. There are more than one sentences in the ENG dataset whose English and Spanish translations have the same dependency tree, so we only display one sentence with the minimum polynomial distance between its English and Spanish translations.

We display the language distance matrices of the GER, FRE, ITA and SPA datasets in Supplementary Figure 5-8. We summarize the pairwise language distances in the 5 datasets by listing the three nearest languages and the three farthest languages to each of the 20 languages in the 5 datasets. The summary for Germanic languages is listed in Supplementary Table 1; the summary for Romance languages is listed in Supplementary Table 2; the summary for Balto-Slavic and Hindi is listed in Supplementary Table 3; the summary for non-Indo-European languages is listed in Supplementary Table 4. We display the visualizations of the language distance matrices of the GER, FRE, ITA and SPA datasets in Supplementary Figure 9-12.

The ENG dataset has 750 sentences, so there are 280,875 pairwise sentence distances between the translations of all 750 sentences in a language. We consider the 750 translations in a language as a corpus of the languages. We display the distributions of the 280,875 pairwise sentence distances and the diameters of the 20 corpora in Supplementary Figure 13. The GER dataset has 100 sentences, so there are 4,950 pairwise sentence distances in each corpus. The distributions of the 4,950 pairwise sentence

distances and the diameters of the 20 corpora are displayed in Supplementary Figure 14. Each of the FRE, ITA and SPA datasets has 50 sentences, so there are 1,225 pairwise sentence distances in each corpus of each dataset. We display the distributions and diameters of the pairwise sentence distances of the 3 datasets in Supplementary Figure 14-17 respectively.

2 Supplementary results

We provide more details about the syntax similarity of languages in groups based on the genealogical classification of Glottolog 4.6.

Supplementary Table 1: The nearest and farthest languages to Germanic languages in language distance. The 5 rows for each languages correspond to results in 5 datasets. Language are represented by their ISO 639-2/B codes.

Closeness	1st	2nd	3rd	18th	19th	20th
English						
ENG	4.28 (swe)	5.06 (ger)	5.08 (por)	7.90 (chi)	8.72 (tha)	11.63 (jpn)
GER	3.83 (por)	4.07 (swe)	4.15 (ita)	7.53 (ice)	7.79 (tha)	9.23 (jpn)
FRE	4.64 (spa)	4.66 (ita)	4.78 (por)	7.86 (ice)	8.92 (tha)	10.72 (jpn)
ITA	4.43 (por)	4.81 (swe)	4.94 (spa)	8.26 (chi)	9.16 (tha)	11.13 (jpn)
SPA	4.56 (ita)	4.70 (por)	4.75 (swe)	8.82 (chi)	9.15 (tha)	12.08 (jpn)
German						
ENG	5.06 (eng)	5.61 (swe)	5.91 (por)	8.50 (hin)	9.23 (tha)	12.33 (jpn)
GER	4.53 (eng)	4.74 (swe)	5.01 (por)	7.57 (ice)	7.91 (tha)	9.80 (jpn)
FRE	5.09 (eng)	5.41 (swe)	5.71 (ita)	8.51 (chi)	9.10 (tha)	11.63 (jpn)
ITA	5.35 (eng)	5.59 (por)	6.18 (spa)	9.08 (chi)	9.58 (tha)	12.06 (jpn)
SPA	5.27 (eng)	5.42 (ita)	5.67 (por)	9.06 (chi)	9.51 (tha)	12.51 (jpn)
Swedish						
ENG	4.28 (eng)	5.19 (ind)	5.61 (ger)	7.93 (chi)	8.44 (tha)	12.26 (jpn)
GER	4.07 (eng)	4.32 (cze)	4.59 (fin)	7.15 (chi)	7.24 (tha)	9.75 (jpn)
FRE	4.75 (ind)	4.84 (eng)	5.05 (cze)	7.82 (chi)	8.54 (tha)	11.32 (jpn)
ITA	4.81 (eng)	5.79 (cze)	5.88 (ind)	8.47 (chi)	8.91 (tha)	11.98 (jpn)
SPA	4.75 (eng)	5.58 (cze)	5.61 (ind)	8.54 (chi)	8.85 (tha)	12.44 (jpn)
Icelandic						
ENG	6.95 (swe)	7.10 (fin)	7.24 (ind)	9.37 (hin)	9.44 (tha)	12.40 (jpn)
GER	6.42 (cze)	6.52 (fin)	6.58 (rus)	8.54 (tha)	8.57 (fre)	10.18 (jpn)
FRE	7.12 (rus)	7.31 (pol)	7.32 (swe)	9.26 (chi)	9.71 (tha)	11.84 (jpn)
ITA	7.10 (fin)	7.12 (rus)	7.14 (cze)	9.35 (hin)	9.78 (tha)	12.43 (jpn)
SPA	7.55 (swe)	7.61 (rus)	7.63 (fin)	9.91 (fre)	10.08 (tha)	12.78 (jpn)

Romance languages (French, Italian, Portuguese and Spanish) are close to each other in pairwise language distance. This can be visualized in both the MDS plot and the UPGMA dendrogram; see Figure 7. The mean pairwise language distance in the ENG dataset is 7.73; see Figure 8. We use mean pairwise language distances as references for syntax similarity between languages: Languages with smaller pairwise language distances are considered similar in syntax, and languages with larger pairwise language distances are considered distinct in syntax. The pairwise language distances between Romance languages are from 4.35 to 5.80, all smaller than the mean value 7.73. The nearest languages to Italian are Portuguese and Spanish; the nearest languages to French are Portuguese and Italian; the nearest languages to Portuguese are Spanish and Italian; and the nearest languages to Spanish are Portuguese and Italian.

Supplementary Table 2: The nearest and farthest languages to Romance languages in language distance. The 5 rows for each languages correspond to results in 5 datasets. Language are represented by their ISO 639-2/B codes.

Closeness	1st	2nd	3rd	18th	19th	20th
Italian						
ENG	4.61 (por)	5.18 (spa)	5.49 (fre)	9.33 (chi)	9.67 (tha)	11.72 (jpn)
GER	3.44 (spa)	3.53 (por)	4.15 (eng)	8.14 (tha)	8.18 (chi)	9.49 (jpn)
FRE	4.39 (por)	4.48 (spa)	4.66 (eng)	8.66 (ice)	9.15 (tha)	11.17 (jpn)
ITA	4.63 (por)	5.16 (spa)	5.23 (eng)	9.03 (chi)	9.71 (tha)	11.58 (jpn)
SPA	4.13 (por)	4.34 (fre)	4.56 (eng)	9.70 (chi)	9.88 (tha)	11.83 (jpn)
French						
ENG	5.16 (por)	5.49 (ita)	5.80 (spa)	9.44 (chi)	9.83 (tha)	11.92 (jpn)
GER	4.46 (ita)	4.51 (por)	4.64 (spa)	8.57 (ice)	8.75 (tha)	9.65 (jpn)
FRE	5.13 (por)	5.21 (ita)	5.46 (spa)	8.97 (chi)	9.36 (tha)	11.37 (jpn)
ITA	4.74 (por)	5.00 (eng)	5.06 (spa)	9.51 (chi)	9.92 (tha)	11.66 (jpn)
SPA	4.34 (ita)	4.90 (por)	5.24 (eng)	9.92 (tha)	9.93 (chi)	11.85 (jpn)
Portuguese						
ENG	4.35 (spa)	4.61 (ita)	5.08 (eng)	9.08 (chi)	9.42 (tha)	11.69 (jpn)
GER	3.24 (spa)	3.53 (ita)	3.83 (eng)	8.05 (ice)	8.27 (tha)	9.25 (jpn)
FRE	3.60 (spa)	4.39 (ita)	4.78 (eng)	8.56 (ice)	9.27 (tha)	10.65 (jpn)
ITA	3.99 (spa)	4.43 (eng)	4.63 (ita)	8.65 (chi)	9.22 (tha)	10.83 (jpn)
SPA	4.13 (ita)	4.16 (spa)	4.70 (eng)	9.35 (chi)	9.57 (tha)	11.74 (jpn)
Spanish						
ENG	4.35 (por)	5.18 (ita)	5.72 (eng)	9.12 (chi)	9.37 (tha)	11.55 (jpn)
GER	3.24 (por)	3.44 (ita)	4.16 (eng)	8.28 (tha)	8.37 (ice)	9.35 (jpn)
FRE	3.60 (por)	4.48 (ita)	4.64 (eng)	8.62 (ice)	9.17 (tha)	10.85 (jpn)
ITA	3.99 (por)	4.94 (eng)	5.06 (fre)	9.02 (chi)	9.42 (tha)	11.10 (jpn)
SPA	4.16 (por)	4.72 (ita)	5.46 (eng)	9.71 (chi)	9.99 (tha)	11.62 (jpn)

Actually, Portuguese and Spanish are among the pairs of languages with the smallest pairwise language distance based on available PUD treebanks; see Figure 8. The farthest languages to Romance languages are Japanese, Thai, Chinese and Icelandic. Based on the polynomial distance, the syntax difference between Romance languages and Icelandic is larger than the syntax difference between Chinese and English. These are consistent in the 5 datasets; see Supplementary Figure 5-12 and Supplementary Table 2.

Balto-Slavic languages (Czech, Polish and Russian) are close to each other in pairwise language distance. The MDS plot and the UPGMA dendrogram in Figure 7 show that the three languages are clustered and surrounded by other languages including Arabic, Finnish, Indonesian and Swedish. The pairwise language distances between Balto-Slavic languages are from 4.60 to 5.10, all smaller than the mean value 7.73. The two nearest languages to each Balto-Slavic language are always the other two Balto-Slavic languages. The nearest language to Czech is Polish and the nearest to Polish is Czech, while the nearest to Russian is Polish. Based on available PUD treebanks, Czech and Polish are among the pairs of languages with the smallest pairwise language distance, and Balto-Slavic languages are also among the languages with the smallest average language distances, suggesting that their syntax is on average least different to all other available languages in PUD treebanks; see Figure 8. The farthest languages to Balto-Slavic languages include Japanese, Thai, Chinese and Hindi. Note that Hindi is also an Indo-European language.

Supplementary Table 3: The nearest and farthest languages to Balto-Slavic and Indo-Iranian languages in language distance. The 5 rows for each languages correspond to results in 5 datasets. Language are represented by their ISO 639-2/B codes.

Closeness	1st	2nd	3rd	18th	19th	20th
Czech						
ENG	4.60 (pol)	5.10 (rus)	5.49 (fin)	8.41 (hin)	8.92 (tha)	11.96 (jpn)
GER	3.46 (pol)	4.15 (rus)	4.32 (fin)	6.85 (chi)	7.38 (tha)	9.29 (jpn)
FRE	4.56 (pol)	4.82 (rus)	5.05 (swe)	8.13 (chi)	8.70 (tha)	11.08 (jpn)
ITA	4.21 (pol)	4.65 (rus)	5.25 (fin)	8.26 (hin)	9.22 (tha)	11.75 (jpn)
SPA	4.37 (pol)	5.11 (rus)	5.44 (fin)	8.40 (chi)	9.46 (tha)	12.25 (jpn)
Polish						
ENG	4.60 (cze)	4.79 (rus)	5.58 (fin)	8.21 (hin)	8.80 (tha)	11.75 (jpn)
GER	3.46 (cze)	3.89 (rus)	4.55 (ara)	6.93 (chi)	7.46 (tha)	9.16 (jpn)
FRE	4.56 (rus)	4.56 (cze)	5.41 (ind)	8.00 (chi)	8.77 (tha)	11.00 (jpn)
ITA	4.21 (cze)	4.39 (rus)	5.53 (ara)	8.25 (chi)	9.15 (tha)	11.40 (jpn)
SPA	4.37 (cze)	4.95 (rus)	5.67 (fin)	8.56 (chi)	9.47 (tha)	12.11 (jpn)
Russian						
ENG	4.79 (pol)	5.10 (cze)	5.65 (ind)	8.19 (hin)	8.89 (tha)	11.74 (jpn)
GER	3.89 (pol)	4.15 (cze)	4.81 (ara)	7.29 (chi)	7.53 (tha)	9.29 (jpn)
FRE	4.14 (pol)	4.82 (cze)	5.14 (ind)	7.72 (chi)	8.72 (tha)	10.70 (jpn)
ITA	4.39 (pol)	4.65 (cze)	5.54 (ara)	8.07 (hin)	9.10 (tha)	11.50 (jpn)
SPA	4.95 (pol)	5.11 (cze)	5.53 (fin)	8.23 (chi)	9.45 (tha)	11.71 (jpn)
Hindi						
ENG	7.41 (eng)	7.84 (swe)	7.98 (ara)	9.37 (ice)	9.42 (tha)	10.38 (jpn)
GER	6.27 (eng)	6.46 (ara)	6.48 (por)	8.26 (chi)	8.28 (ice)	8.33 (jpn)
FRE	6.94 (eng)	7.10 (rus)	7.39 (spa)	9.07 (ice)	9.29 (tha)	9.77 (jpn)
ITA	7.74 (eng)	7.80 (por)	7.91 (ara)	9.42 (chi)	9.67 (tha)	9.83 (jpn)
SPA	7.40 (eng)	7.45 (ara)	7.76 (ind)	9.59 (chi)	9.60 (kor)	10.95 (jpn)

We observe that the nearest language to Hindi is English with language distance 7.41, slightly smaller than the mean value 7.73 but larger than the language distance between Arabic and English, suggesting the syntax difference between English and Hindi is larger than between Arabic and English. The farthest languages to Hindi include Japanese, Thai, Chinese and Icelandic. These are consistent in the 5 datasets; see Supplementary Figure 5-12 and Supplementary Table 3.

For Germanic languages (English, German, Swedish and Icelandic), the pairwise language distances between English, German and Swedish are from 4.28 to 5.61, smaller than the mean value 7.73, but the pairwise language distances from Icelandic to English and German are 7.63 and 8.15, close to or larger than the mean value. The pairwise language distance between Icelandic and German is larger than the distance between Chinese and English in all 5 datasets; see Figure 6 and Supplementary Figure 5-12. The nearest language to German is English in all 5 datasets, and the nearest languages to Swedish are English, Indonesian and Czech; see Supplementary Table 1. The nearest language to English is Swedish in the ENG dataset, which is also the smallest pairwise language distance for the ENG dataset; see Figure 8. However, the nearest languages to English are inconsistent in the 5 datasets, and other nearest languages include Italian, Portuguese, Spanish and German; see Supplementary Table 1. According to Figure 8, English and Swedish are among the languages with the smallest average language distances based on

Supplementary Table 4: The nearest and farthest languages to non-Indo-European languages in language distance.

The 5 rows for each languages correspond to results in 5 datasets. Language are represented by their ISO 639-2/B codes.

Closeness	1st	2nd	3rd	18th	19th	20th
Arabic						
ENG	6.02 (rus)	6.07 (pol)	6.20 (ind)	8.44 (chi)	8.93 (tha)	11.19 (jpn)
GER	4.55 (pol)	4.64 (cze)	4.81 (rus)	7.26 (chi)	7.76 (tha)	7.85 (jpn)
FRE	5.33 (rus)	5.75 (pol)	6.08 (ind)	8.51 (chi)	9.05 (tha)	10.59 (jpn)
ITA	5.33 (pol)	5.53 (rus)	6.03 (ind)	8.31 (chi)	8.99 (tha)	11.00 (jpn)
SPA	6.12 (pol)	6.33 (rus)	6.53 (ind)	8.94 (chi)	9.56 (tha)	10.79 (jpn)
Chinese						
ENG	7.21 (kor)	7.69 (ind)	7.83 (fin)	9.33 (ita)	9.44 (fre)	11.36 (jpn)
GER	6.25 (kor)	6.85 (cze)	6.90 (ind)	8.28 (ice)	8.50 (fre)	9.09 (jpn)
FRE	7.50 (kor)	7.62 (eng)	7.67 (ind)	8.97 (fre)	9.26 (ice)	10.09 (jpn)
ITA	7.40 (kor)	7.97 (rus)	7.98 (ind)	9.51 (fre)	9.74 (tha)	11.31 (jpn)
SPA	8.05 (kor)	8.14 (ind)	8.23 (rus)	9.93 (fre)	10.01 (tha)	11.67 (jpn)
Finnish						
ENG	5.49 (cze)	5.58 (pol)	5.59 (ind)	8.85 (hin)	8.89 (tha)	12.65 (jpn)
GER	4.32 (cze)	4.52 (tur)	4.59 (swe)	7.55 (tha)	7.80 (fre)	10.08 (jpn)
FRE	5.42 (ind)	5.51 (rus)	5.68 (tur)	8.27 (chi)	9.13 (tha)	12.22 (jpn)
ITA	5.25 (cze)	5.67 (rus)	5.75 (tur)	9.02 (hin)	9.58 (tha)	12.53 (jpn)
SPA	5.44 (cze)	5.53 (rus)	5.67 (pol)	8.93 (hin)	9.60 (tha)	12.96 (jpn)
Indonesian						
ENG	5.19 (swe)	5.24 (eng)	5.59 (fin)	8.19 (hin)	8.39 (tha)	11.77 (jpn)
GER	4.73 (swe)	4.78 (cze)	4.83 (ara)	6.98 (hin)	7.34 (tha)	8.96 (jpn)
FRE	4.75 (swe)	5.09 (eng)	5.14 (rus)	7.69 (hin)	8.52 (tha)	11.10 (jpn)
ITA	5.74 (rus)	5.83 (eng)	5.85 (pol)	8.15 (hin)	8.81 (tha)	11.31 (jpn)
SPA	5.36 (eng)	5.61 (swe)	5.84 (fin)	8.14 (chi)	8.75 (tha)	11.74 (jpn)
Japanese						
ENG	10.38 (hin)	11.19 (ara)	11.36 (chi)	12.33 (ger)	12.40 (ice)	12.65 (fin)
GER	8.33 (hin)	8.85 (ara)	8.91 (kor)	9.80 (ger)	10.08 (fin)	10.18 (ice)
FRE	9.77 (hin)	10.09 (chi)	10.59 (ara)	11.63 (ger)	11.84 (ice)	12.22 (fin)
ITA	9.83 (hin)	10.83 (por)	11.00 (ara)	12.06 (ger)	12.43 (ice)	12.53 (fin)
SPA	10.79 (ara)	10.95 (hin)	11.62 (spa)	12.51 (ger)	12.78 (ice)	12.96 (fin)
Korean						
ENG	6.13 (fin)	6.25 (tur)	6.75 (ind)	8.96 (hin)	9.11 (fre)	11.52 (jpn)
GER	5.21 (fin)	5.49 (tur)	5.92 (ind)	7.72 (hin)	8.10 (fre)	8.91 (jpn)
FRE	5.95 (fin)	5.98 (tur)	6.37 (swe)	8.35 (tha)	8.52 (hin)	10.85 (jpn)
ITA	6.26 (fin)	6.30 (tur)	6.87 (rus)	9.06 (fre)	9.16 (hin)	11.61 (jpn)
SPA	6.50 (fin)	6.65 (tur)	6.95 (cze)	9.60 (hin)	9.77 (fre)	11.95 (jpn)
Thai						
ENG	8.39 (ind)	8.44 (swe)	8.45 (kor)	9.67 (ita)	9.83 (fre)	12.07 (jpn)
GER	7.24 (swe)	7.34 (ind)	7.38 (cze)	8.54 (ice)	8.75 (fre)	9.65 (jpn)
FRE	8.35 (kor)	8.52 (ind)	8.54 (swe)	9.36 (fre)	9.71 (ice)	10.97 (jpn)
ITA	8.81 (ind)	8.91 (swe)	8.96 (kor)	9.78 (ice)	9.92 (fre)	11.87 (jpn)
SPA	8.75 (ind)	8.85 (swe)	9.15 (eng)	10.01 (chi)	10.08 (ice)	12.08 (jpn)
Turkish						
ENG	5.64 (fin)	6.03 (ind)	6.13 (pol)	8.59 (hin)	8.82 (tha)	11.90 (jpn)
GER	4.52 (fin)	4.82 (cze)	4.95 (pol)	7.44 (fre)	7.60 (tha)	9.33 (jpn)
FRE	5.68 (fin)	5.84 (pol)	5.96 (cze)	8.30 (chi)	8.84 (tha)	11.07 (jpn)
ITA	5.75 (fin)	6.07 (rus)	6.21 (pol)	8.75 (hin)	9.50 (tha)	11.99 (jpn)
SPA	6.21 (fin)	6.31 (rus)	6.33 (pol)	9.02 (fre)	9.50 (tha)	12.24 (jpn)

currently available PUD treebanks. Icelandic also has inconsistent nearest languages in the 5 datasets, and the nearest languages include Swedish, Finnish and Balto-Slavic languages. In the ENG dataset, the two nearest languages to Icelandic are Swedish and Finnish with distance 6.95 and 7.10, which are close to the mean value 7.73. The farthest languages to Germanic languages include Japanese, Thai, Chinese and Hindi. For English and German, Icelandic is among the three farthest languages, and for Icelandic,

French is among the three farthest languages; see Supplementary Table 1.

Japanese is consistently the language with the largest average language distance in the 5 datasets based on the currently available PUD treebanks; see Figure 8. This is also observable in the visualizations of language distance matrices displayed in Figure 7 and Supplementary Figure 9-12, suggesting that the syntax of Japanese is distinct from other 19 languages in this study. The three largest pairwise language distances are consistently between Japanese and Finnish, Japanese and Icelandic and Japanese and German; see Figure 8. Actually, Finnish, Icelandic and German are the farthest languages to Japanese; see Supplementary Table 4. Among the other 19 languages, the nearest language to Japanese is Hindi, with pairwise language distance 10.38, which is larger than all pairwise language distances between other 19 languages. Other languages near Japanese in language distance include Arabic and Chinese, though the distances suggest rather distinct syntax between the languages.

Thai is consistently the language with the second largest average language distance in the 5 datasets based on the currently available PUD treebanks; see Figure 8. The nearest language to Thai is Indonesian, with a pairwise language distance 8.39 in the ENG dataset, which is larger than the mean value 7.73. This suggests that the syntax difference between Thai and Indonesian is as large as the difference between Chinese and German. The other languages near Thai include Swedish and Korean based on the 5 datasets, though the pairwise language distances between the languages are all larger than the mean values of the datasets. The farthest language to Thai is Japanese, and the other far languages include French and Icelandic; see Supplementary Table 4.

Chinese is consistently the language with the third largest average language distance in the 5 datasets based on the currently available PUD treebanks; see Figure 8. The nearest language to Chinese is Korean in all 5 datasets, with a pairwise language distance 7.21 in the ENG dataset, which is close to the mean value 7.73. This suggests that the syntax difference between Chinese and Korean is as large as between English and Korean. The other language near Chinese is Indonesian, with a pairwise language distance 7.69 in the ENG dataset, which is slightly smaller than the pairwise language distance 7.90 between Chinese and English. The farthest language to Chinese is also Japanese, and the second farthest language is French; see Supplementary Table 4.

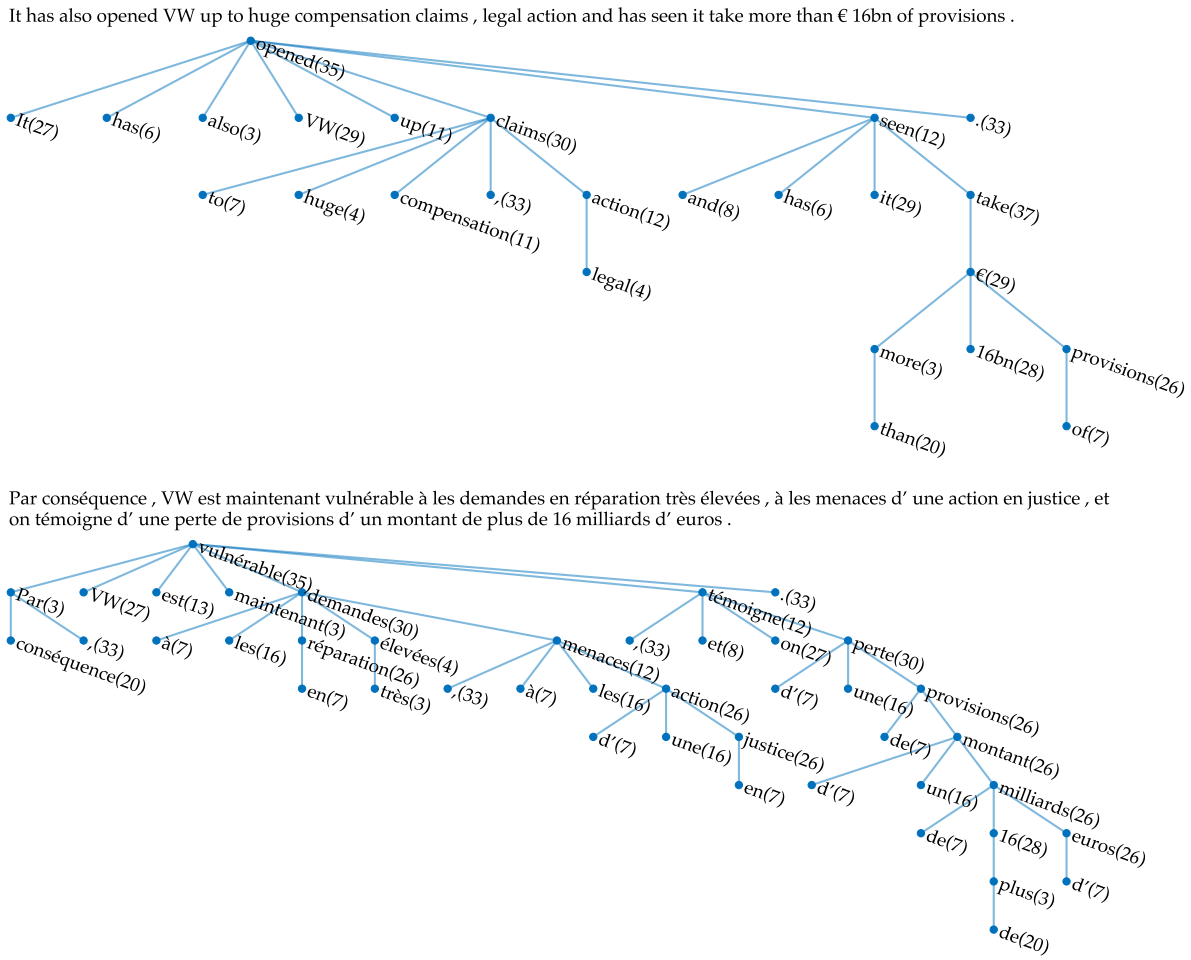
The nearest languages to Indonesian in the 5 datasets of the PUD treebanks are Swedish and English, respectively with pairwise language distances 5.19 and 5.24 in the ENG dataset. This suggests that the syntax difference between Indonesian and Swedish or English is smaller than the difference between English and Czech; see Supplementary Table 4, Figure 6 and Supplementary Figure 5-8. The farthest languages to Indonesian are Japanese, Thai and Hindi.

The nearest languages to Arabic in the 5 datasets of the PUD treebanks are Balto-Slavic languages,

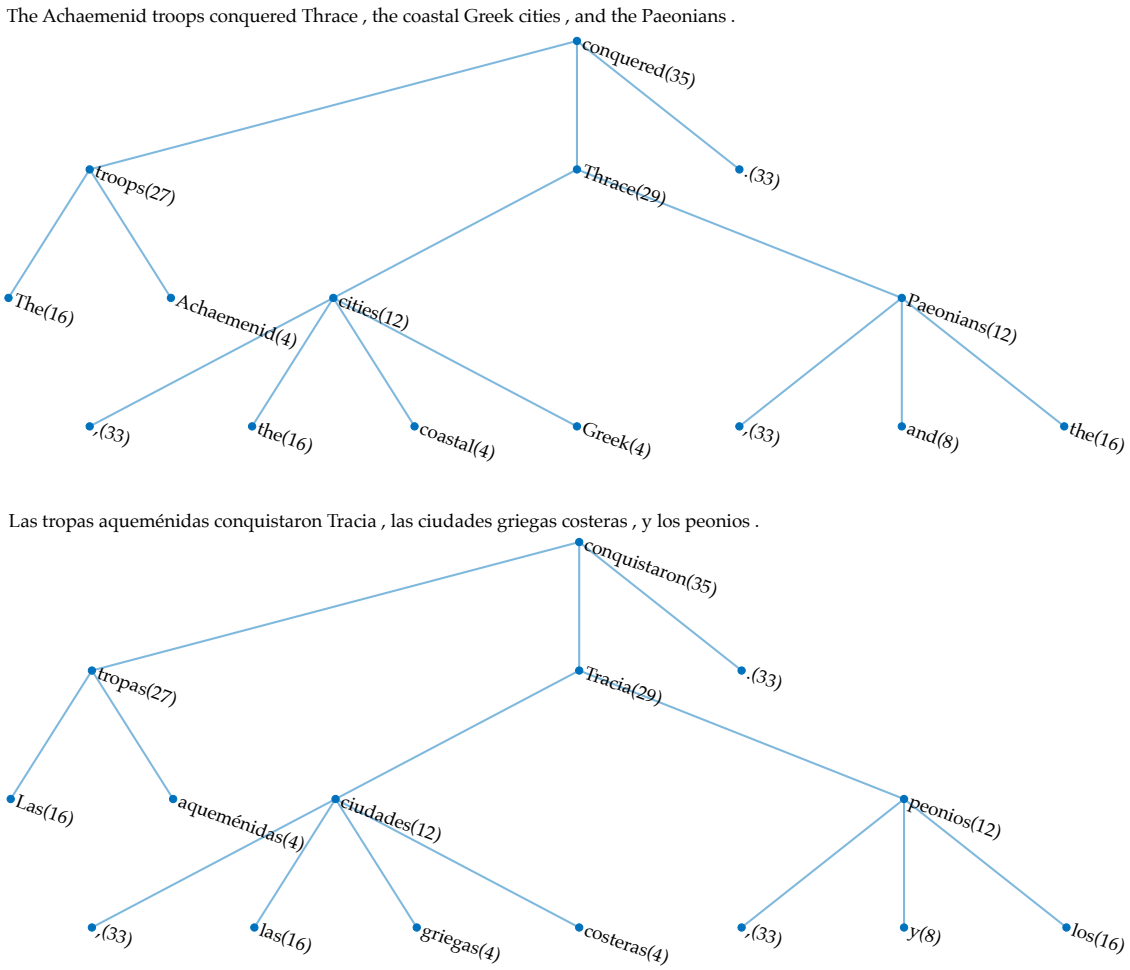
especially Russian and Polish. The pairwise language distance between Arabic and Russian is 6.02 in the ENG dataset, and the distance between Arabic and Polish is 6.07. This suggests that the syntax difference between Arabic and Russian or Polish is as large as the difference between English and Russian; see Supplementary Table 4, Figure 6 and Supplementary Figure 5-8. The farthest languages to Arabic are Japanese, Thai and Chinese.

In Supplementary Table 4, we observe that the nearest language to Finnish is Czech in the PUD treebanks. Other languages near Finnish include other Balto-Slavic languages, Indonesian and Turkish. The pairwise language distance between Finnish and Czech is 5.49 in the ENG dataset, which is as small as the distance between Italian and French; see Figure 6 and Supplementary Figure 5-8. The farthest languages to Finnish are Japanese, Thai and Hindi. For Turkish, the nearest language is Finnish in all 5 datasets, and other languages near Turkish are Balto-Slavic languages or Indonesian. The pairwise language distance between Turkish and Finnish is 5.64 in the ENG dataset, which is as small as the distance between Italian and German; see Figure 6 and Supplementary Figure 5-8. The farthest languages to Turkish are Japanese, Thai, Hindi and French. Lastly, we observe that in all 5 datasets, the nearest language to Korean is Finnish and the second nearest language to Korean is Turkish, with pairwise language distance 6.13 and 6.25 in the ENG dataset respectively; see Supplementary Table 4. Both distances are smaller than the mean value 7.73.

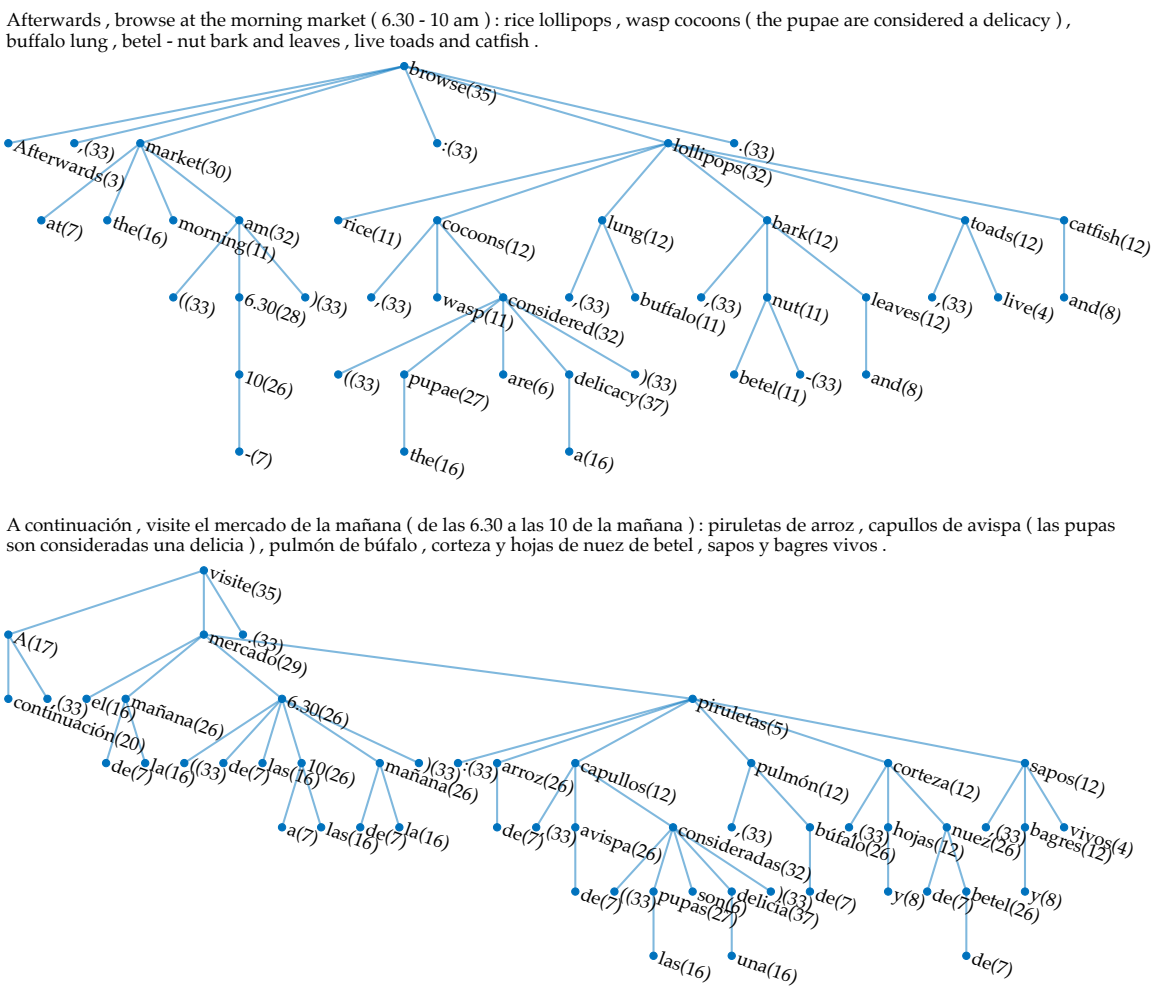
Supplementary Figure 1: A sentence in the ENG dataset with minimum polynomial distance between its English and French translations. Top: the dependency tree of the sentence's English translation (the original sentence since it is in the ENG dataset). Bottom: the dependency tree of the sentence's French translation. The polynomial distance between the dependency trees is 0.



Supplementary Figure 2: The sentence in the ENG dataset with maximum polynomial distance between its English and French translations. Top: the dependency tree of the sentence's English translation (the original sentence since it is in the ENG dataset). Bottom: the dependency tree of the sentence's French translation. The polynomial distance between the dependency trees is 20.48.



Supplementary Figure 3: A sentence in the ENG dataset with minimum polynomial distance between its English and Spanish translations. Top: the dependency tree of the sentence's English translation (the original sentence since it is in the ENG dataset). Bottom: the dependency tree of the sentence's Spanish translation. The polynomial distance between the dependency trees is 0.



Supplementary Figure 4: The sentence in the ENG dataset with maximum polynomial distance between its English and Spanish translations. Top: the dependency tree of the sentence's English translation (the original sentence since it is in the ENG dataset). Bottom: the dependency tree of the sentence's Spanish translation. The polynomial distance between the dependency trees is 15.81.

	English	German	Swedish	Icelandic	Italian	French	Portuguese	Spanish	Czech	Polish	Russian	Hindi	Arabic	Chinese	Finnish	Indonesian	Japanese	Korean	Thai	Turkish
English		4.53	4.07	7.53	4.15	5.18	3.83	4.16	5.11	5.41	5.80	6.27	5.74	7.43	6.03	5.02	9.23	6.70	7.79	6.14
German	4.53		4.74	7.57	5.04	5.77	5.01	5.20	5.37	5.44	5.69	7.35	6.10	7.48	5.84	5.59	9.80	6.58	7.91	6.11
Swedish	4.07	4.74		6.63	5.35	6.40	5.34	5.60	4.32	4.67	5.18	6.55	5.24	7.15	4.59	4.73	9.75	5.93	7.24	5.30
Icelandic	7.53	7.57	6.63		8.12	8.57	8.05	8.37	6.42	6.64	6.58	8.28	6.74	8.28	6.52	6.91	10.18	7.22	8.54	6.91
Italian	4.15	5.04	5.35	8.12		4.46	3.53	3.44	5.46	5.55	6.11	7.04	6.23	8.18	6.71	5.71	9.49	7.29	8.14	6.58
French	5.18	5.77	6.40	8.57	4.46		4.51	4.63	6.56	6.56	6.76	7.61	7.01	8.50	7.80	6.56	9.65	8.10	8.75	7.44
Portuguese	3.83	5.01	5.34	8.05	3.53	4.51		3.24	5.36	5.54	6.01	6.48	5.98	8.00	6.89	5.66	9.25	7.35	8.27	6.69
Spanish	4.16	5.20	5.60	8.37	3.44	4.63	3.24		5.76	5.81	6.31	6.91	6.07	8.26	7.15	5.94	9.35	7.56	8.28	6.83
Czech	5.11	5.37	4.32	6.42	5.46	6.56	5.36	5.76		3.46	4.15	6.63	4.64	6.85	4.32	4.78	9.29	5.93	7.38	4.82
Polish	5.41	5.44	4.67	6.64	5.55	6.56	5.54	5.81	3.46		3.89	6.70	4.55	6.93	4.69	4.91	9.16	5.95	7.46	4.95
Russian	5.80	5.69	5.18	6.58	6.11	6.76	6.01	6.31	4.15	3.89		6.81	4.81	7.29	4.97	5.14	9.29	6.03	7.53	5.21
Hindi	6.27	7.35	6.55	8.28	7.04	7.61	6.48	6.91	6.63	6.70	6.81		6.46	8.26	7.42	6.98	8.33	7.72	8.06	7.13
Arabic	5.74	6.10	5.24	6.74	6.23	7.01	5.98	6.07	4.64	4.55	4.81	6.46		7.26	5.43	4.83	8.85	6.28	7.76	5.11
Chinese	7.43	7.48	7.15	8.28	8.18	8.50	8.00	8.26	6.85	6.93	7.29	8.26	7.26		7.10	6.90	9.09	6.25	8.05	7.05
Finnish	6.03	5.84	4.59	6.52	6.71	7.80	6.89	7.15	4.32	4.69	4.97	7.42	5.43	7.10		5.01	10.08	5.21	7.55	4.52
Indonesian	5.02	5.59	4.73	6.91	5.71	6.56	5.66	5.94	4.78	4.91	5.14	6.98	4.83	6.90	5.01		8.96	5.92	7.34	5.01
Japanese	9.23	9.80	9.75	10.18	9.49	9.65	9.25	9.35	9.29	9.16	9.29	8.33	8.85	9.09	10.08	8.96		8.91	9.65	9.33
Korean	6.70	6.58	5.93	7.22	7.29	8.10	7.35	7.56	5.93	5.95	6.03	7.72	6.28	6.25	5.21	5.92	8.91		7.40	5.49
Thai	7.79	7.91	7.24	8.54	8.14	8.75	8.27	8.28	7.38	7.46	7.53	8.06	7.76	8.05	7.55	7.34	9.65	7.40		7.60
Turkish	6.14	6.11	5.30	6.91	6.58	7.44	6.69	6.83	4.82	4.95	5.21	7.13	5.11	7.05	4.52	5.01	9.33	5.49	7.60	

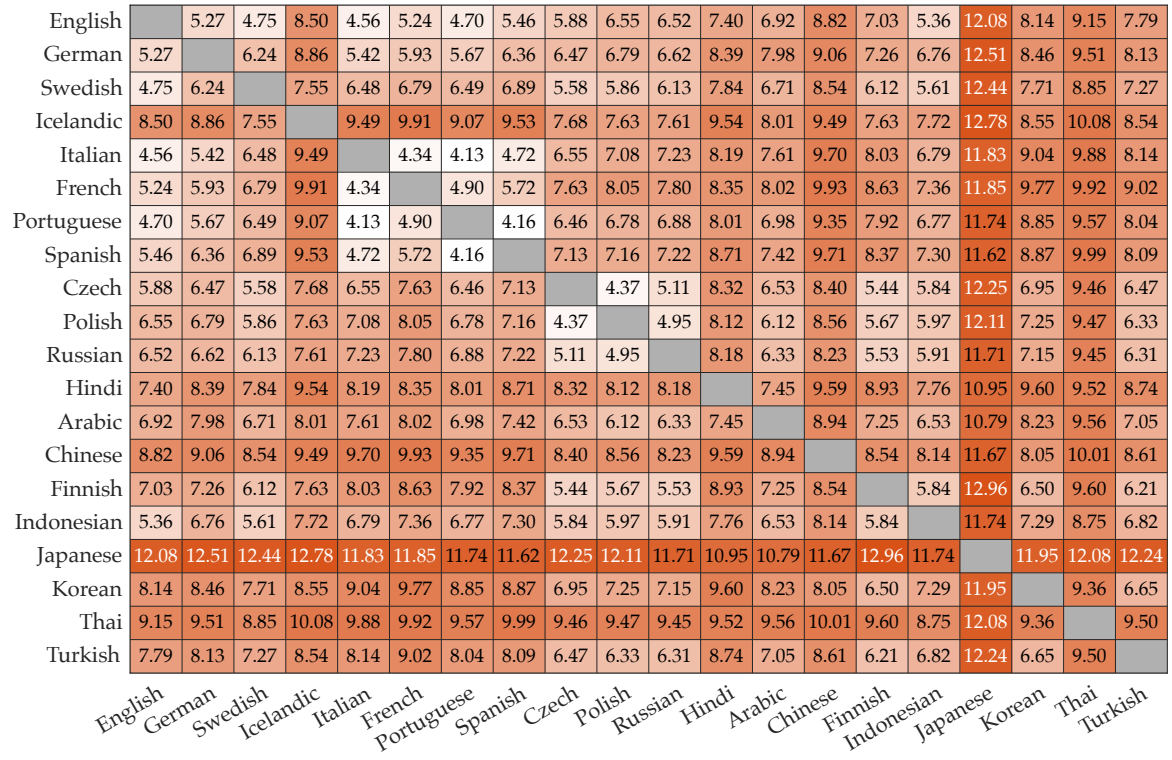
Supplementary Figure 5: The language distance matrix of the GER dataset. The languages are ordered based on Glottolog 4.6 classification: Indo-European languages are listed first and grouped according to their subclasses (Germanic, Romance, Balto-Slavic and Indo-Iranian), and other languages are following in the alphabetical order.

	English	German	Swedish	Icelandic	Italian	French	Portuguese	Spanish	Czech	Polish	Russian	Hindi	Arabic	Chinese	Finnish	Indonesian	Japanese	Korean	Thai	Turkish
English		5.09	4.84	7.86	4.66	5.94	4.78	4.64	5.85	6.11	5.63	6.94	7.05	7.62	6.72	5.09	10.72	7.32	8.92	7.07
German	5.09		5.41	8.34	5.71	6.32	5.88	5.81	5.98	6.70	6.23	8.06	7.54	8.51	6.98	6.10	11.63	7.53	9.10	7.24
Swedish	4.84	5.41		7.32	5.76	6.47	6.15	5.91	5.05	5.62	5.22	7.57	6.75	7.82	5.71	4.75	11.32	6.37	8.54	6.28
Icelandic	7.86	8.34	7.32		8.66	8.54	8.56	8.62	7.41	7.31	7.12	9.07	7.58	9.26	7.61	7.39	11.84	7.60	9.71	7.50
Italian	4.66	5.71	5.76	8.66		5.21	4.39	4.48	6.31	6.63	6.53	7.85	7.26	8.63	7.50	6.13	11.17	7.84	9.15	7.37
French	5.94	6.32	6.47	8.54	5.21		5.13	5.46	7.04	6.83	6.58	8.25	7.54	8.97	8.12	6.80	11.37	8.24	9.36	7.76
Portuguese	4.78	5.88	6.15	8.56	4.39	5.13		3.60	6.20	6.64	6.17	7.41	7.04	8.46	7.69	6.43	10.65	8.08	9.27	7.48
Spanish	4.64	5.81	5.91	8.62	4.48	5.46	3.60		6.47	6.63	6.14	7.39	7.06	8.47	7.82	6.36	10.85	8.05	9.17	7.56
Czech	5.85	5.98	5.05	7.41	6.31	7.04	6.20	6.47		4.56	4.82	7.55	6.08	8.13	5.72	5.29	11.08	6.54	8.70	5.96
Polish	6.11	6.70	5.62	7.31	6.63	6.83	6.64	6.63	4.56		4.14	7.64	5.75	8.00	5.68	5.41	11.00	6.44	8.77	5.84
Russian	5.63	6.23	5.22	7.12	6.53	6.58	6.17	6.14	4.82	4.14		7.10	5.33	7.72	5.51	5.14	10.70	6.80	8.72	6.03
Hindi	6.94	8.06	7.57	9.07	7.85	8.25	7.41	7.39	7.55	7.64	7.10		7.64	8.53	8.27	7.69	9.77	8.52	9.29	8.21
Arabic	7.05	7.54	6.75	7.58	7.26	7.54	7.04	7.06	6.08	5.75	5.33	7.64		8.51	7.20	6.08	10.59	7.44	9.05	6.19
Chinese	7.62	8.51	7.82	9.26	8.63	8.97	8.46	8.47	8.13	8.00	7.72	8.53	8.51		8.27	7.67	10.09	7.50	8.83	8.30
Finnish	6.72	6.98	5.71	7.61	7.50	8.12	7.69	7.82	5.72	5.68	5.51	8.27	7.20	8.27		5.42	12.22	5.95	9.13	5.68
Indonesian	5.09	6.10	4.75	7.39	6.13	6.80	6.43	6.36	5.29	5.41	5.14	7.69	6.08	7.67	5.42		11.10	6.60	8.52	6.04
Japanese	10.72	11.63	11.32	11.84	11.17	11.37	10.65	10.85	11.08	11.00	10.70	9.77	10.59	10.09	12.22	11.10		10.85	10.97	11.07
Korean	7.32	7.53	6.37	7.60	7.84	8.24	8.08	8.05	6.54	6.44	6.80	8.52	7.44	7.50	5.95	6.60	10.85		8.35	5.98
Thai	8.92	9.10	8.54	9.71	9.15	9.36	9.27	9.17	8.70	8.77	8.72	9.29	9.05	8.83	9.13	8.52	10.97	8.35		8.84
Turkish	7.07	7.24	6.28	7.50	7.37	7.76	7.48	7.56	5.96	5.84	6.03	8.21	6.19	8.30	5.68	6.04	11.07	5.98	8.84	

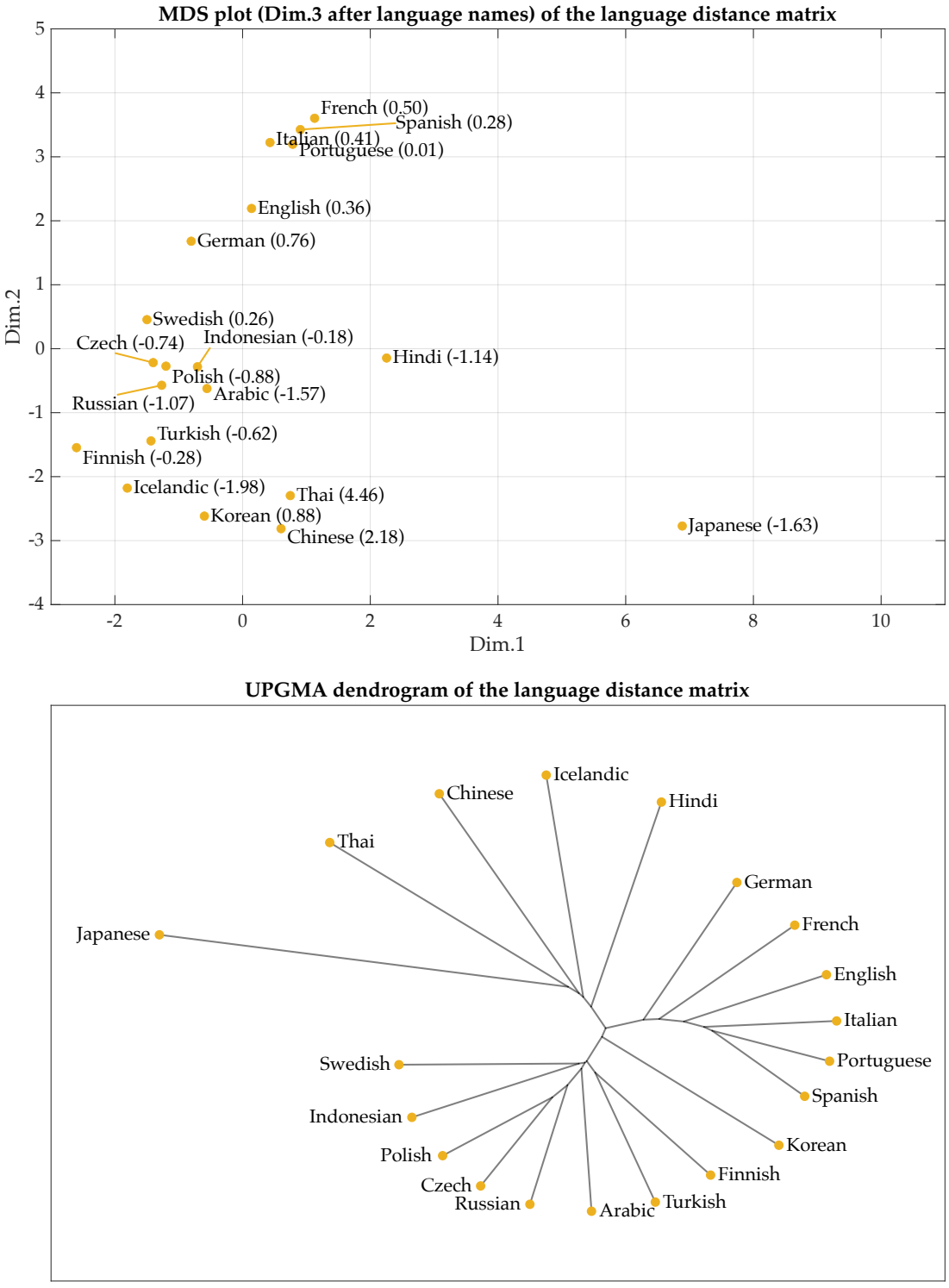
Supplementary Figure 6: The language distance matrix of the FRE dataset. The languages are ordered based on Glottolog 4.6 classification: Indo-European languages are listed first and grouped according to their subclasses (Germanic, Romance, Balto-Slavic and Indo-Iranian), and other languages are following in the alphabetical order.

	English	German	Swedish	Icelandic	Italian	French	Portuguese	Spanish	Czech	Polish	Russian	Hindi	Arabic	Chinese	Finnish	Indonesian	Japanese	Korean	Thai	Turkish
English		5.35	4.81	7.91	5.23	5.00	4.43	4.94	6.40	6.46	6.48	7.74	6.79	8.26	7.37	5.83	11.13	7.94	9.16	7.36
German	5.35		6.36	8.60	6.21	6.26	5.59	6.18	6.90	6.87	6.90	9.05	7.70	9.08	7.44	6.94	12.06	8.33	9.58	7.57
Swedish	4.81	6.36		7.25	6.52	6.62	5.89	6.59	5.79	6.08	5.99	8.40	6.35	8.47	6.47	5.88	11.98	7.38	8.91	6.72
Icelandic	7.91	8.60	7.25		8.34	9.06	7.94	8.45	7.14	7.28	7.12	9.35	7.18	9.07	7.10	7.36	12.43	7.95	9.78	7.46
Italian	5.23	6.21	6.52	8.34		5.47	4.63	5.16	6.82	6.57	6.78	8.39	7.02	9.03	7.99	6.78	11.58	8.38	9.71	7.48
French	5.00	6.26	6.62	9.06	5.47		4.74	5.06	7.31	7.05	7.27	8.45	7.83	9.51	8.41	7.31	11.66	9.06	9.92	8.37
Portuguese	4.43	5.59	5.89	7.94	4.63	4.74		3.99	6.45	6.21	6.38	7.80	6.77	8.65	7.71	6.27	10.83	8.05	9.22	7.25
Spanish	4.94	6.18	6.59	8.45	5.16	5.06	3.99		7.27	6.79	6.80	7.96	6.96	9.02	8.06	6.77	11.10	8.42	9.42	7.61
Czech	6.40	6.90	5.79	7.14	6.82	7.31	6.45	7.27		4.21	4.65	8.26	6.16	8.22	5.25	5.96	11.75	7.00	9.22	6.37
Polish	6.46	6.87	6.08	7.28	6.57	7.05	6.21	6.79	4.21		4.39	8.04	5.53	8.25	5.80	5.85	11.40	7.13	9.15	6.21
Russian	6.48	6.90	5.99	7.12	6.78	7.27	6.38	6.80	4.65	4.39		8.07	5.54	7.97	5.67	5.74	11.50	6.87	9.10	6.07
Hindi	7.74	9.05	8.40	9.35	8.39	8.45	7.80	7.96	8.26	8.04	8.07		7.91	9.42	9.02	8.15	9.83	9.16	9.67	8.75
Arabic	6.79	7.70	6.35	7.18	7.02	7.83	6.77	6.96	6.16	5.53	5.54	7.91		8.31	7.01	6.03	11.00	7.18	8.99	6.24
Chinese	8.26	9.08	8.47	9.07	9.03	9.51	8.65	9.02	8.22	8.25	7.97	9.42	8.31		8.22	7.98	11.31	7.40	9.74	8.19
Finnish	7.37	7.44	6.47	7.10	7.99	8.41	7.71	8.06	5.25	5.80	5.67	9.02	7.01	8.22		6.19	12.53	6.26	9.58	5.75
Indonesian	5.83	6.94	5.88	7.36	6.78	7.31	6.27	6.77	5.96	5.85	5.74	8.15	6.03	7.98	6.19		11.31	7.19	8.81	6.27
Japanese	11.13	12.06	11.98	12.43	11.58	11.66	10.83	11.10	11.75	11.40	11.50	9.83	11.00	11.31	12.53	11.31		11.61	11.87	11.99
Korean	7.94	8.33	7.38	7.95	8.38	9.06	8.05	8.42	7.00	7.13	6.87	9.16	7.18	7.40	6.26	7.19	11.61		8.96	6.30
Thai	9.16	9.58	8.91	9.78	9.71	9.92	9.22	9.42	9.22	9.15	9.10	9.67	8.99	9.74	9.58	8.81	11.87	8.96		9.50
Turkish	7.36	7.57	6.72	7.46	7.48	8.37	7.25	7.61	6.37	6.21	6.07	8.75	6.24	8.19	5.75	6.27	11.99	6.30	9.50	

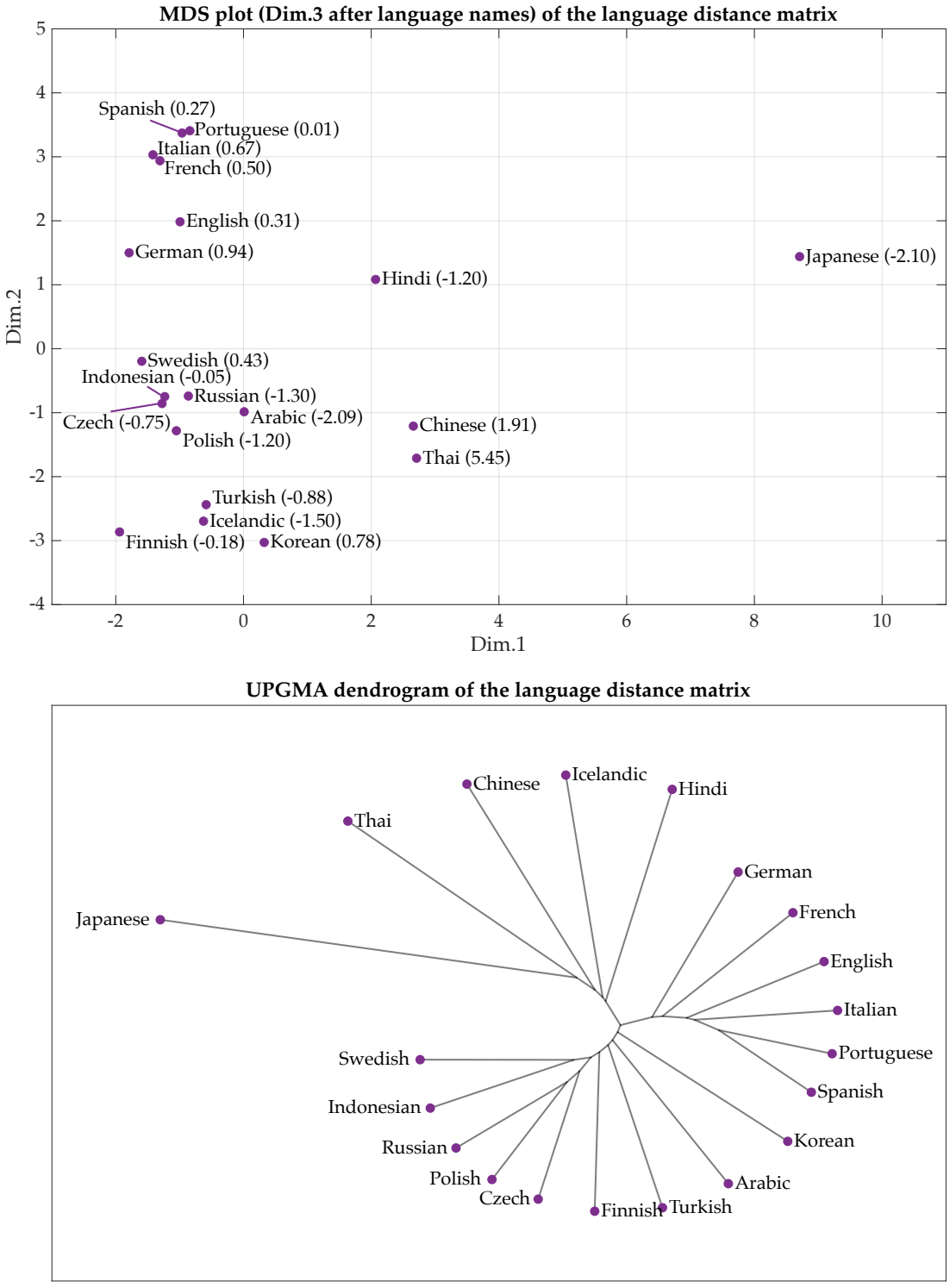
Supplementary Figure 7: The language distance matrix of the ITA dataset. The languages are ordered based on Glottolog 4.6 classification: Indo-European languages are listed first and grouped according to their subclasses (Germanic, Romance, Balto-Slavic and Indo-Iranian), and other languages are following in the alphabetical order.



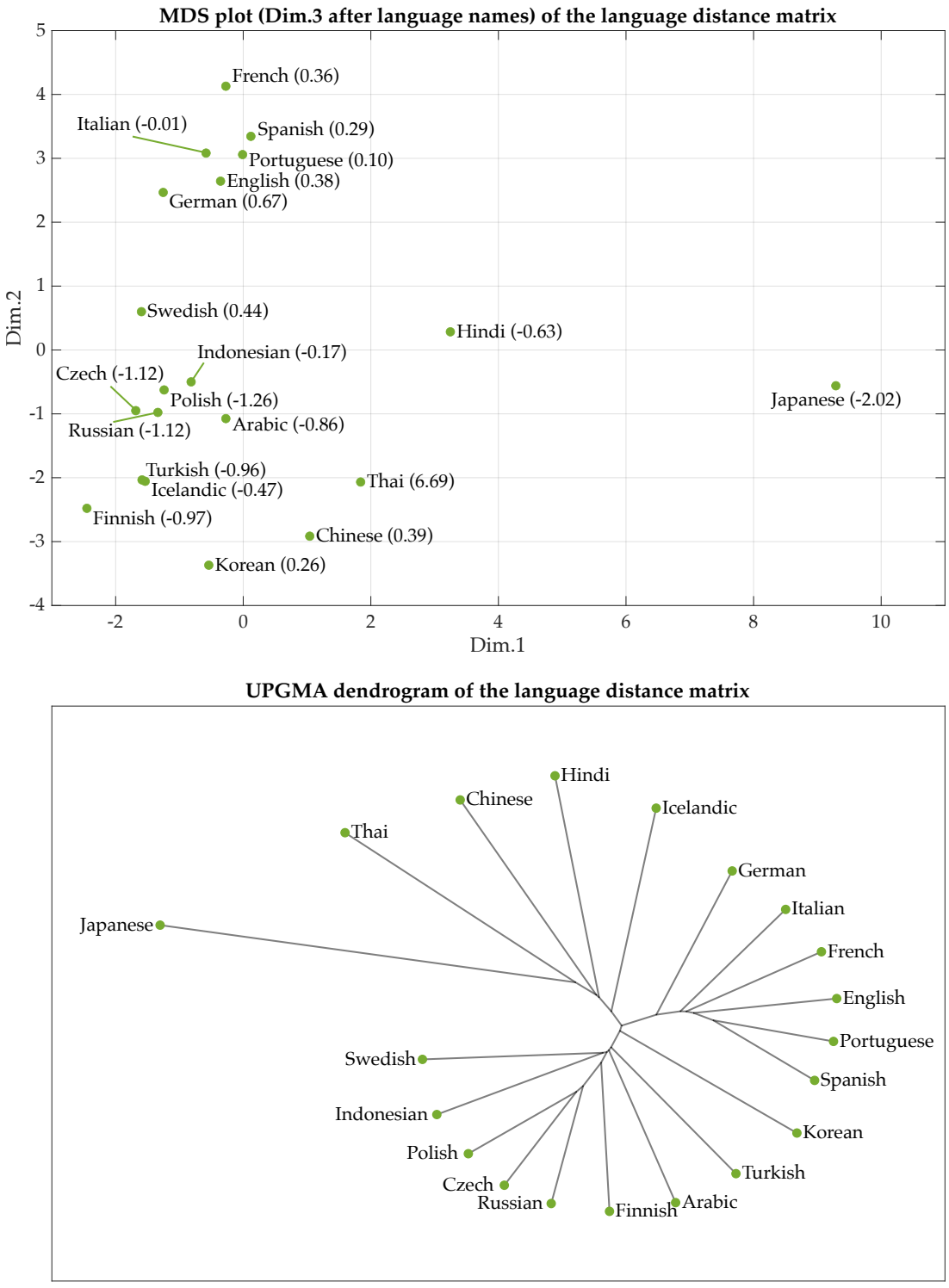
Supplementary Figure 8: The language distance matrix of the SPA dataset. The languages are ordered based on Glottolog 4.6 classification: Indo-European languages are listed first and grouped according to their subclasses (Germanic, Romance, Balto-Slavic and Indo-Iranian), and other languages are following in the alphabetical order.



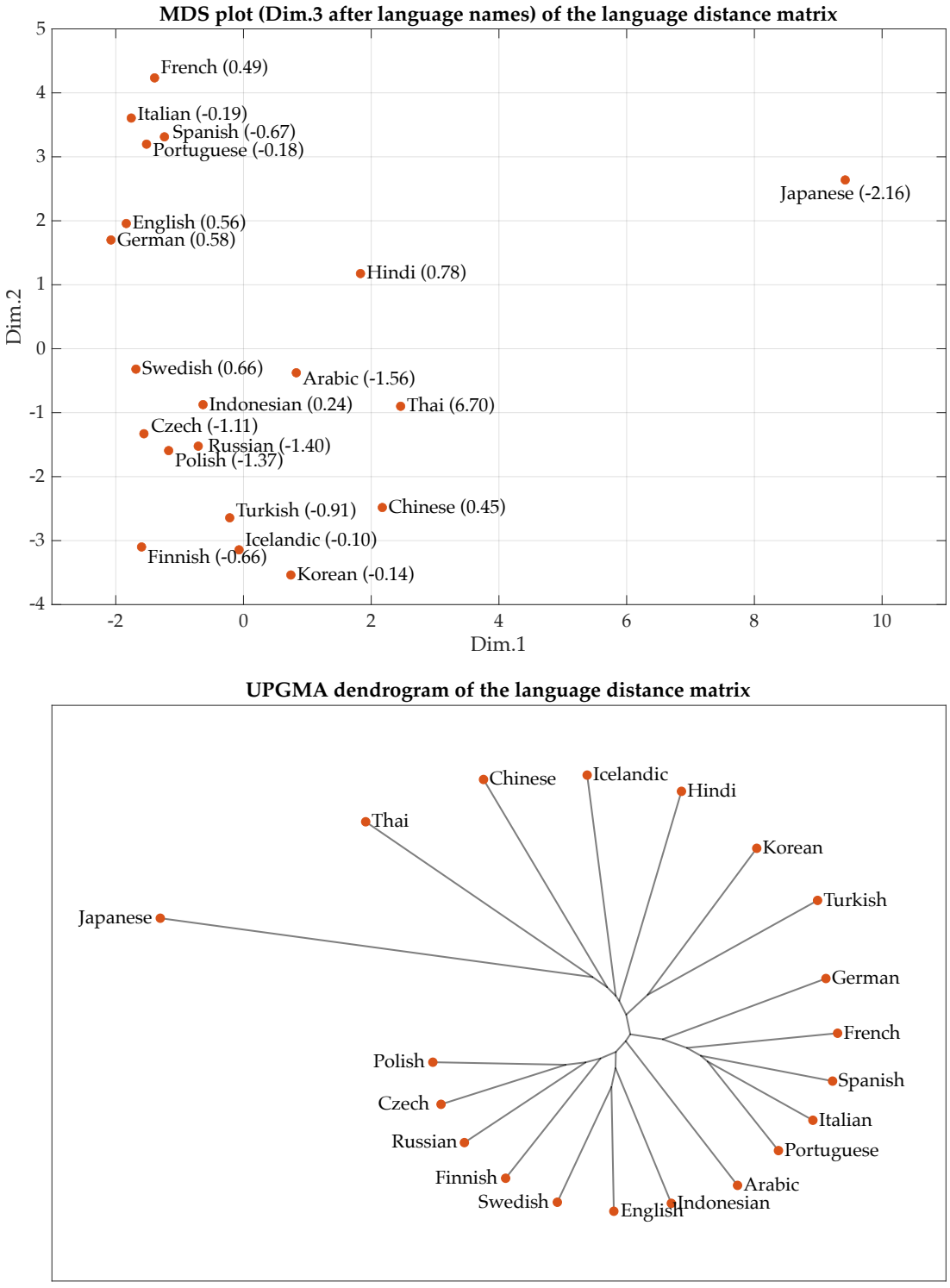
Supplementary Figure 9: Visualizations the language distance matrix of the GER dataset. Top: the multidimensional scaling plot of the language distance matrix of the GER dataset. Bottom: the UPGMA dendrogram constructed based on the language distance matrix of the GER dataset.



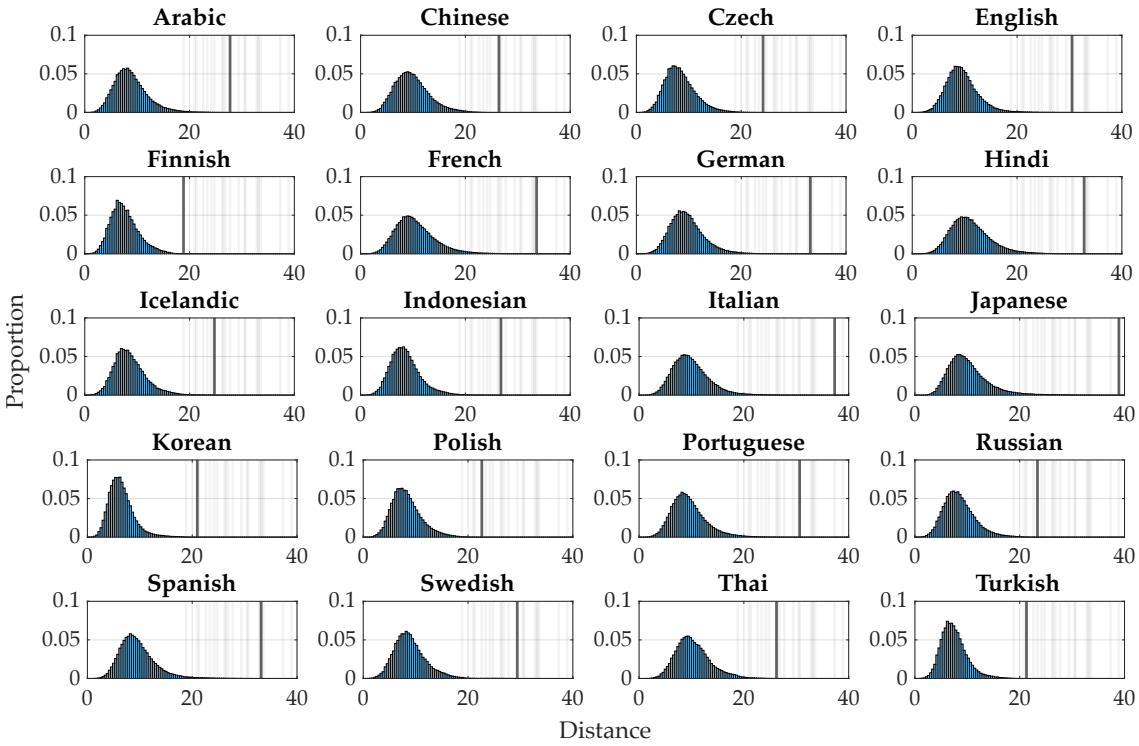
Supplementary Figure 10: Visualizations the language distance matrix of the FRE dataset. Top: the multidimensional scaling plot of the language distance matrix of the FRE dataset. Bottom: the UPGMA dendrogram constructed based on the language distance matrix of the FRE dataset.



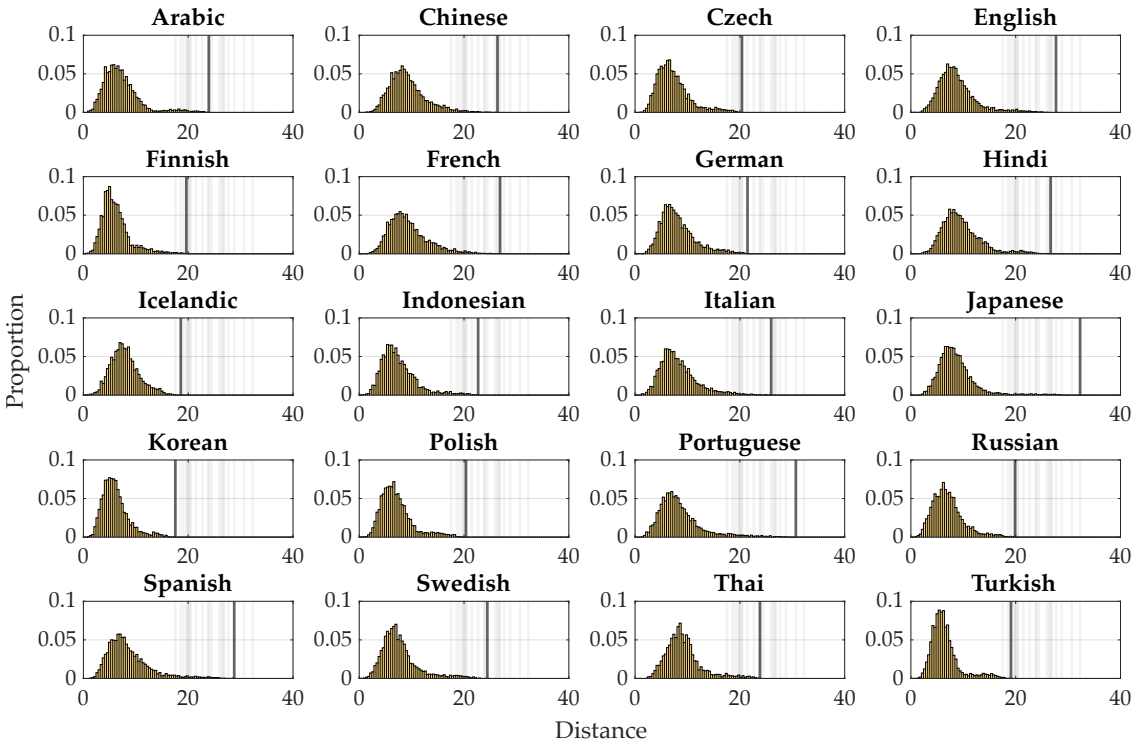
Supplementary Figure 11: Visualizations the language distance matrix of the ITA dataset. Top: the multidimensional scaling plot of the language distance matrix of the ITA dataset. Bottom: the UPGMA dendrogram constructed based on the language distance matrix of the ITA dataset.



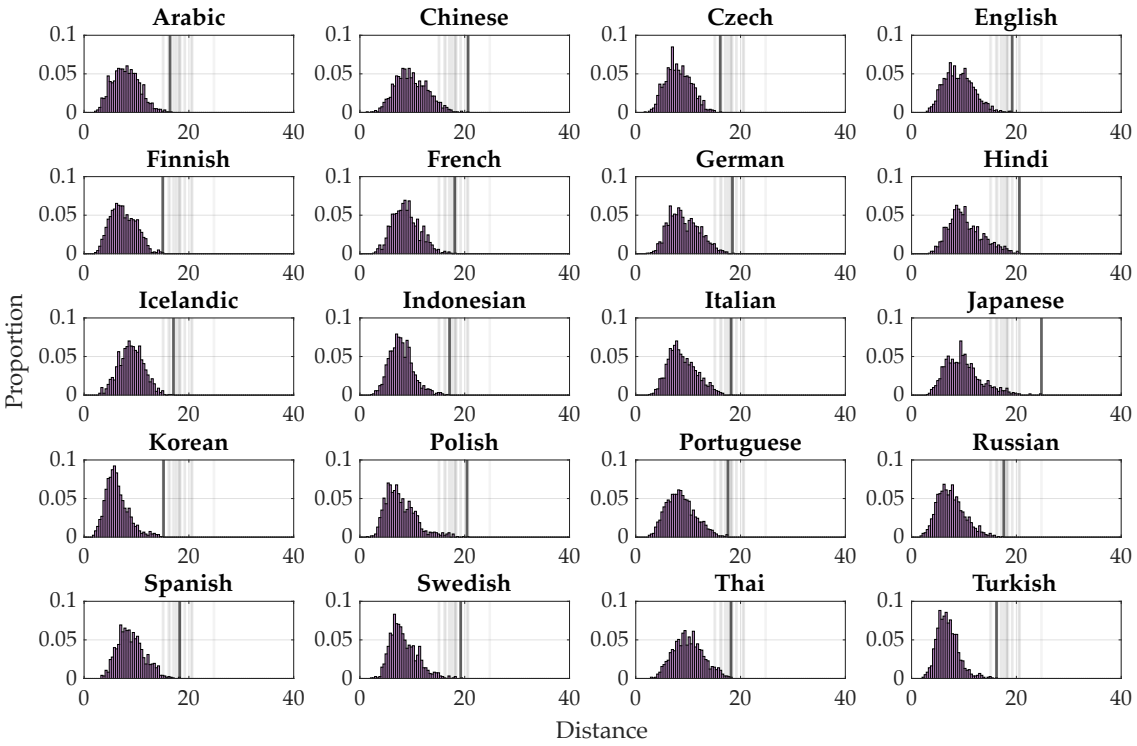
Supplementary Figure 12: Visualizations the language distance matrix of the SPA dataset. Top: the multidimensional scaling plot of the language distance matrix of the SPA dataset. Bottom: the UPGMA dendrogram constructed based on the language distance matrix of the SPA dataset.



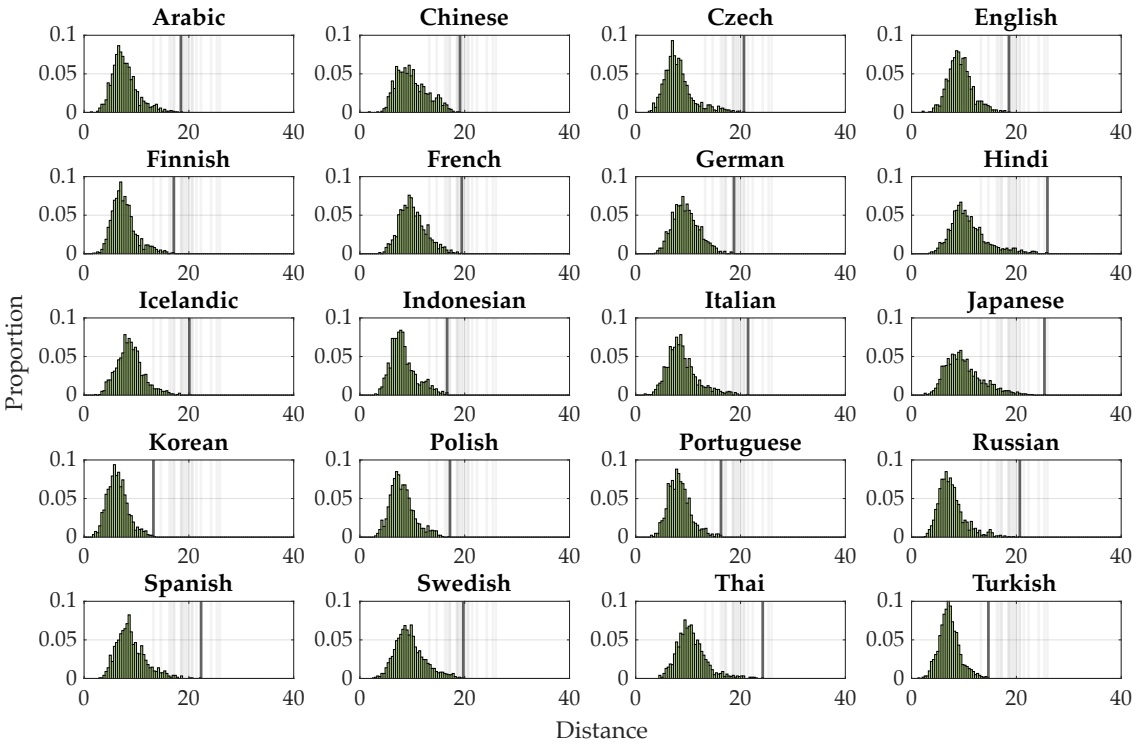
Supplementary Figure 13: The diameters and the distributions of pairwise sentence distances in 20 corpora of the ENG dataset. The diameters of 20 languages are showed as vertical lines in the panels: the solid line for the present corpus and transparent lines for other 19 corpora.



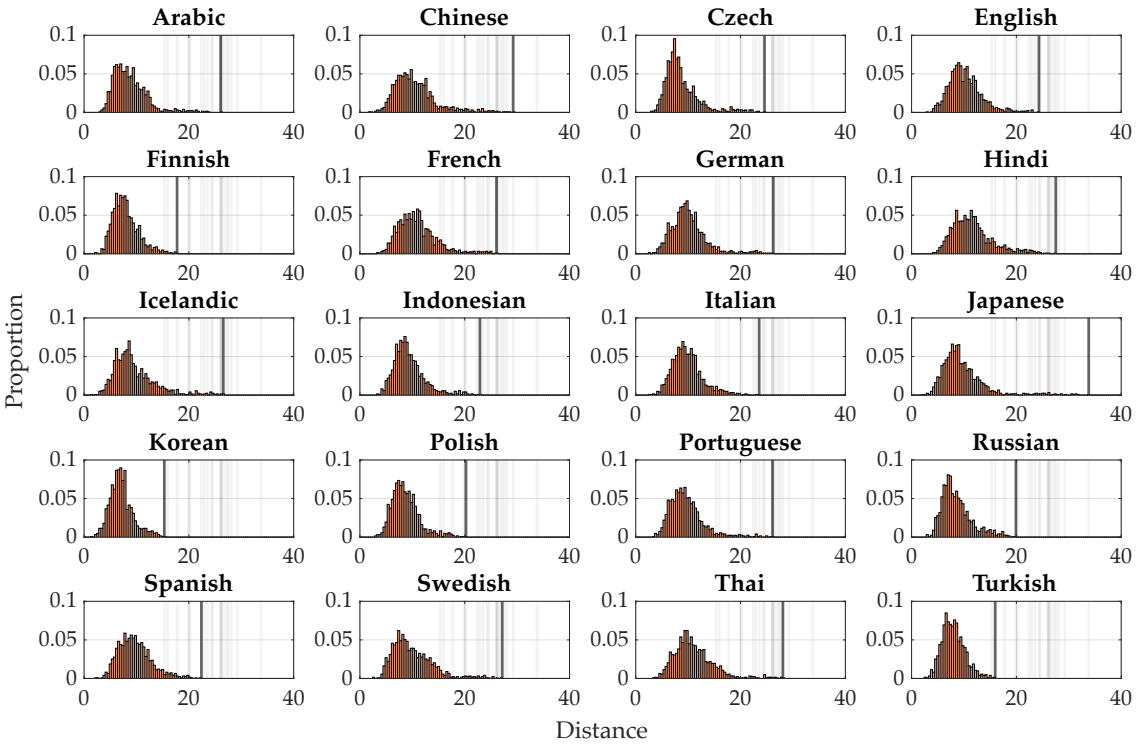
Supplementary Figure 14: The diameters and the distributions of pairwise sentence distances in 20 corpora of the GER dataset. The diameters of 20 languages are showed as vertical lines in the panels: the solid line for the present corpus and transparent lines for other 19 corpora.



Supplementary Figure 15: The diameters and the distributions of pairwise sentence distances in 20 corpora of the FRE dataset. The diameters of 20 languages are showed as vertical lines in the panels: the solid line for the present corpus and transparent lines for other 19 corpora.



Supplementary Figure 16: The diameters and the distributions of pairwise sentence distances in 20 corpora of the ITA dataset. The diameters of 20 languages are showed as vertical lines in the panels: the solid line for the present corpus and transparent lines for other 19 corpora.



Supplementary Figure 17: The diameters and the distributions of pairwise sentence distances in 20 corpora of the SPA dataset. The diameters of 20 languages are showed as vertical lines in the panels: the solid line for the present corpus and transparent lines for other 19 corpora.