Cultural distances: Common crawl data

Contents

Introduction

This file replicates the tests for the main wikipedia data on the subtitles data.

Load libraries

```
library(ape)
library(ecodist)
library(lme4)
library(sjPlot)
library(ggplot2)
library(igraph)
library(lattice)
```

Parameters (using data from Northuralex and subtitles, k=100, unfiltered):

```
datasetName = "subs"
lingDistancesFile = "../data/FAIR/nel-k100-subs-alignments-by-language-pair.csv"
lingDistancesFileNK = "../data/FAIR/nel-k100-subs-alignments-by-language-pair-without-kinsip.csv"
lingDistancesByDomainFile = "../results/EA_distances/nel-k100-subs_with_ling.csv"
# (generated by ../processing/combineCultAndLingDistances.R)
```

All domains

Load data

Read the cultural distances:

```r
cult = read.csv("../results/EA_distances/CulturalDistances_Long.csv", stringsAsFactors = F)
names(cult) = c("l1","l2","cult.dist")
cultLangs = unique(c(cult$Var1,cult$Var2))
```

Add language family:

```r
l = read.csv("../data/FAIR_langauges_glotto_xdid.csv", stringsAsFactors = F)
g = read.csv("../data/glottolog-languoid.csv/languoid.csv", stringsAsFactors = F)
l$family = g[match(l$glotto,g$id),]$family_pk
l$family = g[match(l$family,g$pk),]$name
```

Read the semantic distances

```r
ling = read.csv(lingDistancesFile, stringsAsFactors = F)
```

Combine the lingusitic and cultural distances. Note that we flip the cultural measure from a distance measure to a similarity measure.

```r
cult$l1.iso2 = l[match(cult$l1,l$Language2),]$iso2
cult$l2.iso2 = l[match(cult$l2,l$Language2),]$iso2

fairisos = unique(c(ling$l1,ling$l2))
cultisos = unique(c(cult$l1.iso2, cult$l2.iso2))

cult = cult[(cult$l1.iso2 %in% fairisos) & (cult$l2.iso2 %in% fairisos),]
ling = ling[(ling$l1 %in% cultisos) & (ling$l2 %in% cultisos),]

matches = sapply(1:nrow(ling), function(i){
  which(cult$l1.iso2==ling$l1[i] & cult$l2.iso2==ling$l2[i])
})

ling$cult.dist = cult[matches,]$cult.dist
# Flip
ling$cult.dist = 1 - ling$cult.dist
# Scale
ling$cult.dist.center = scale(ling$cult.dist)
cdc.s = attr(ling$cult.dist.center,"scaled:scale")
cdc.c = attr(ling$cult.dist.center,"scaled:center")
ling$cult.dist.center = as.numeric(ling$cult.dist.center)
ling$comparison_count.center =
  scale(ling$comparison_count)

ling$family1 = l[match(ling$l1, l$iso2),]$family
ling$family2 = l[match(ling$l2, l$iso2),]$family
ling$area1 = l[match(ling$l1, l$iso2),]$autotyp.area
ling$area2 = l[match(ling$l2, l$iso2),]$autotyp.area


fgroup = cbind(ling$family1,ling$family2)
fgroup = apply(fgroup,1,sort)
```

```
ling$family.group = apply(fgroup,2,paste,collapse=":")
agroup = cbind(ling$area1,ling$area2)
agroup = apply(agroup,1,sort)
ling$area.group = apply(agroup,2,paste,collapse=":")

ling$rho.center = scale(ling$local_alignment)
```

Each observation is now assocaited with a language family pair:

```
head(ling[,c("l1","l2","local_alignment",'family.group')])
```

```
##    l1 l2 local_alignment                family.group
## 18 hy sq      0.01014219 Indo-European:Indo-European
## 23 hy ko      0.01862729      Indo-European:Koreanic
## 32 hy is      0.03000878 Indo-European:Indo-European
## 48 hy ja      0.05326118       Indo-European:Japonic
## 49 et hy      0.05535150        Indo-European:Uralic
## 52 hy nl      0.05835286 Indo-European:Indo-European
```

And the same is true for area:

```
tail(ling[,c("l1","l2","local_alignment",'area.group')])
```

```
##     l1 l2 local_alignment        area.group
## 651 cs es       0.3339303     Europe:Europe
## 653 bg cs       0.3353707     Europe:Europe
## 656 el bg       0.3378855     Europe:Europe
## 659 el ru       0.3446584 Europe:Inner Asia
## 661 el cs       0.3625137     Europe:Europe
## 664 cs ru       0.3790251 Europe:Inner Asia
```

Number of observations:

```
# Number of datapoints:
nrow(ling)
```

```
## [1] 190
```

```
# Number of unique languages:
length(unique(unlist(ling[,c("l1","l2")])))
```

```
## [1] 20
```

```
# Number of unique langauge families:
uniqueFamilies = unique(unlist(ling[,c("family1","family2")]))
length(uniqueFamilies)
```

```
## [1] 6
```

```
# Number of unique areas:
uniqueAreas = unique(unlist(ling[,c("area1","area2")]))
length(uniqueAreas)
```

```
## [1] 4
```

Cross-over between language famlies and areas:

```
tx = data.frame(lang= c(ling$l1,ling$l2),
          fam = c(ling$family1,ling$family2),
          area= c(ling$area1,ling$area2))
```

```r
tx = tx[!duplicated(tx),]
table(tx$fam,tx$area)
```

```
##
##                  Europe Greater Mesopotamia Inner Asia N Coast Asia
##   Afro-Asiatic       0                   1          0            0
##   Indo-European      9                   1          4            0
##   Japonic            0                   0          0            1
##   Koreanic           0                   0          0            1
##   Turkic             0                   1          0            0
##   Uralic             1                   0          1            0
```

LMER models

Mixed effects model, predicting Linguistic similaritys from cultural similarity, with random intercept for family and area and random slope for cultural similarity for family and area.

We start with a null model with random intercepts for family and area, and random slopes for cultural similarity by both. We add a fixed effect of the number of comparisons made for each datapoint (number of concepts that were available to compare). Then we add a fixed effect of cultural similarity.

Note that due to convergence issues, these models do not include a random slope for area.

```
m0 = lmer(
  rho.center ~ 1 +
    (1 + comparison_count.center | family.group) +
    (1 | area.group),
  data = ling
)
m0.5 = lmer(
  rho.center ~ 1 +
    comparison_count.center +
    (1 + comparison_count.center| family.group) +
    (1 | area.group),
  data = ling
)
m1 = lmer(
  rho.center ~ 1 +
    comparison_count.center +
    cult.dist.center +
    (1 + comparison_count.center| family.group) +
    (1 | area.group),
  data = ling
)
anova(m0,m0.5,m1)
```

```
## refitting model(s) with ML (instead of REML)

## Data: ling
## Models:
## m0: rho.center ~ 1 + (1 + comparison_count.center | family.group) +
## m0:     (1 | area.group)
## m0.5: rho.center ~ 1 + comparison_count.center + (1 + comparison_count.center |
## m0.5:     family.group) + (1 | area.group)
## m1: rho.center ~ 1 + comparison_count.center + cult.dist.center +
## m1:     (1 + comparison_count.center | family.group) + (1 | area.group)
##      Df    AIC    BIC  logLik deviance   Chisq Chi Df Pr(>Chisq)
## m0    6 343.15 362.63 -165.57   331.15
## m0.5  7 301.74 324.47 -143.87   287.74 43.4106      1  4.438e-11 ***
## m1    8 295.71 321.69 -139.86   279.71  8.0247      1   0.004614 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Cultural similarity is significantly correlated with Linguistic similarity. Here are the model estimates:

```
summary(m1)
```

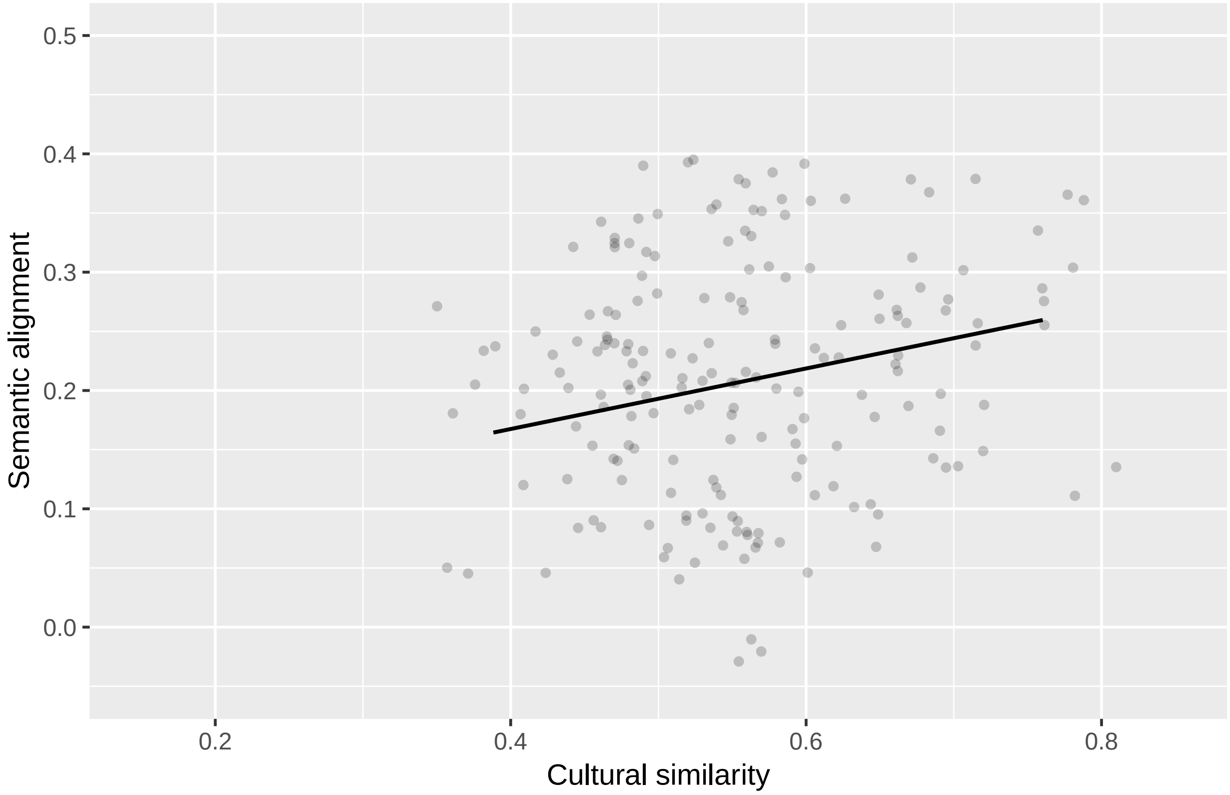
```
## Linear mixed model fit by REML ['lmerMod']
## Formula: rho.center ~ 1 + comparison_count.center + cult.dist.center +
##     (1 + comparison_count.center | family.group) + (1 | area.group)
##    Data: ling
##
## REML criterion at convergence: 293.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.8748 -0.8141  0.0081  0.6054  3.4942
##
## Random effects:
##  Groups       Name                    Variance  Std.Dev. Corr
##  family.group (Intercept)             0.0017280 0.04157
##               comparison_count.center 0.0006107 0.02471  -1.00
##  area.group   (Intercept)             0.0000000 0.00000
##  Residual                             0.2575746 0.50752
## Number of obs: 190, groups:  family.group, 17; area.group, 10
##
## Fixed effects:
##                          Estimate Std. Error t value
## (Intercept)             -0.005957   0.041664  -0.143
## comparison_count.center  0.833474   0.039964  20.855
## cult.dist.center         0.108251   0.038542   2.809
##
## Correlation of Fixed Effects:
##             (Intr) cmpr_.
## cmprsn_cnt. -0.145
## clt.dst.cnt  0.053 -0.231
```

Plot the estimates, rescaling the variables back to the original units:

```
gx = sjp.lmer(m1,'pred','cult.dist.center',
              prnt.plot = F)
```

```
## Warning: package 'bindrcpp' was built under R version 3.3.2
```

```
gx$plot$data$y = gx$plot$data$y *
  attr(ling$rho.center,"scaled:scale") +
  attr(ling$rho.center,"scaled:center")
gx$plot$data$resp.y = gx$plot$data$resp.y *
  attr(ling$rho.center,"scaled:scale") +
  attr(ling$rho.center,"scaled:center")
gx$plot$data$x = gx$plot$data$x *
  cdc.s +cdc.c
gx = gx$plot + coord_cartesian(ylim=c(-0.05,0.5),
                        xlim=c(0.15,0.85)) +
  xlab("Cultural similarity") +
  ylab("Semantic alignment") +
  ggtitle("")
gx
```



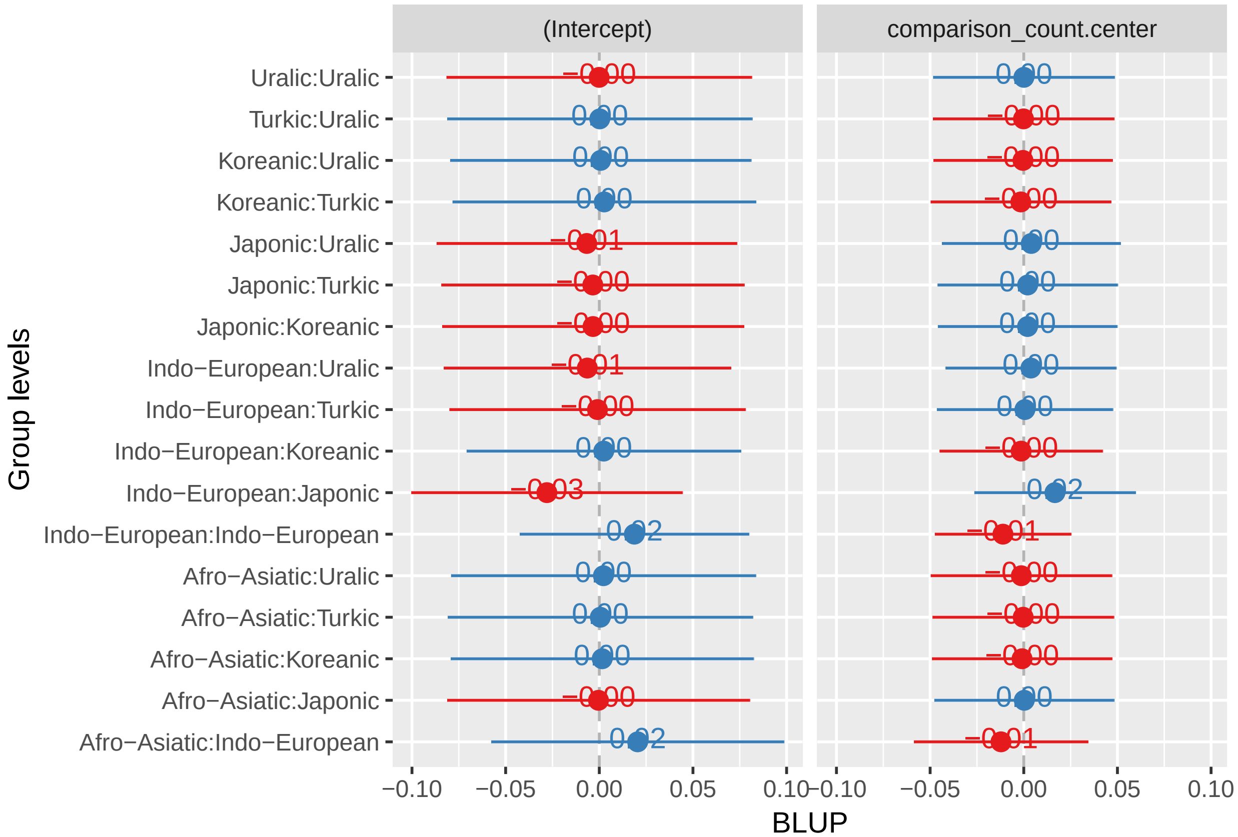
```
pdf(paste0("../results/stats/",datasetName,"/CulturalDistance_Rho_Graph.pdf"),
    height=2.5, width=2.5)
gx
dev.off()
```

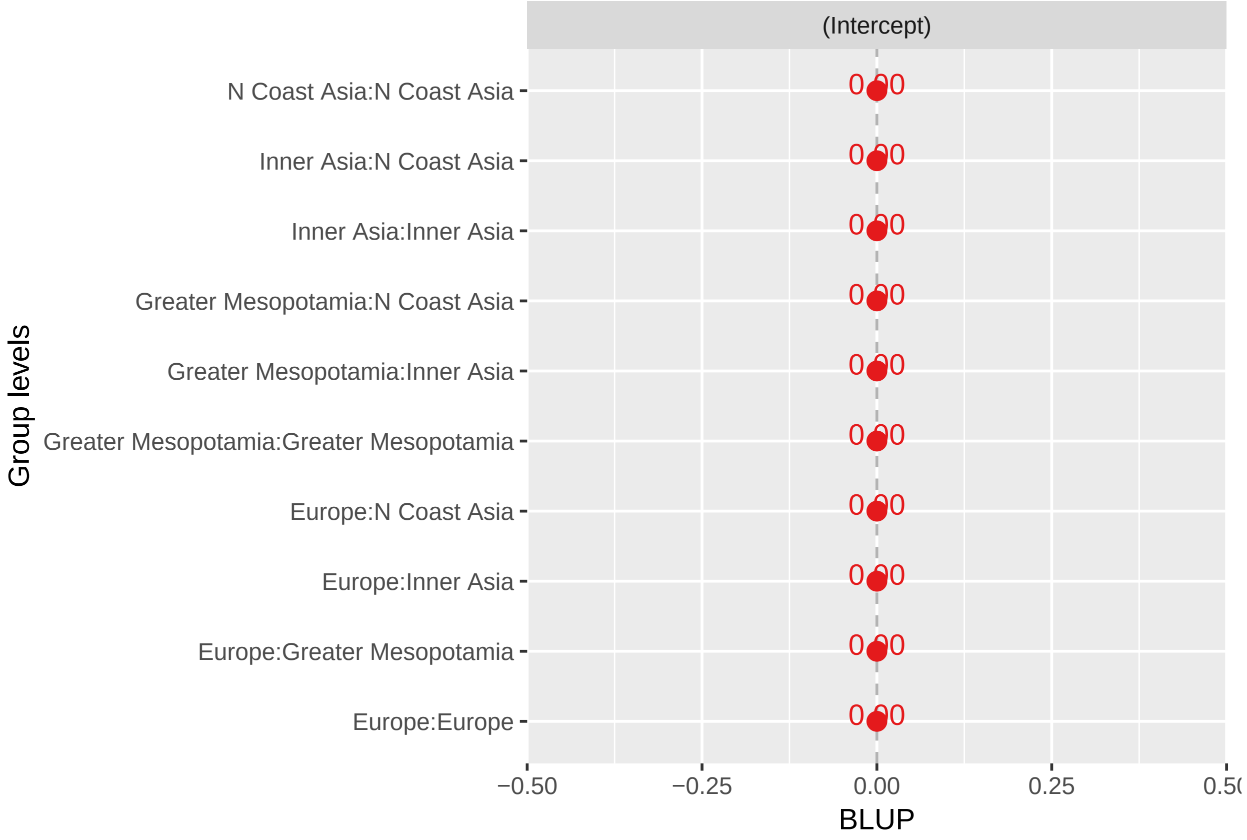
```
## pdf
##   2
```

Plot the random effects:

```
sjp.lmer(m1,'re')
```

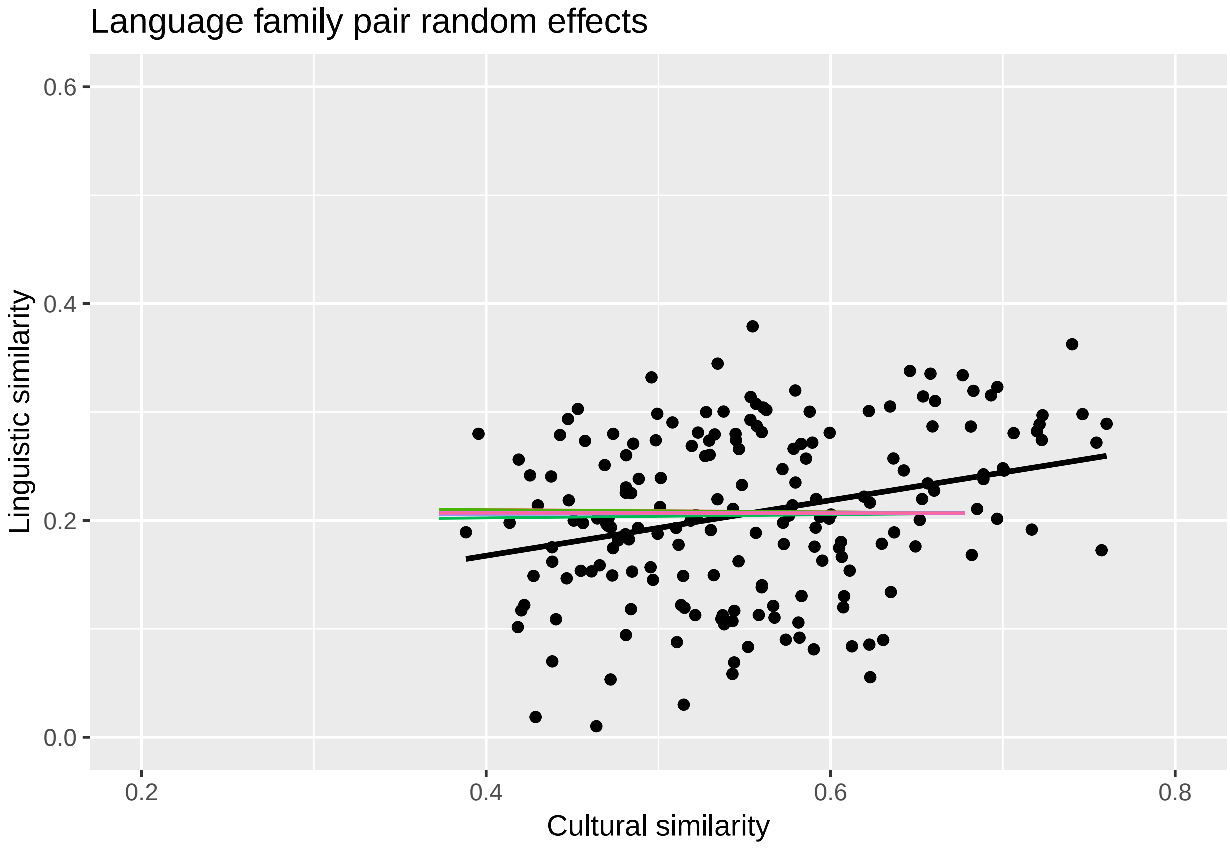
```
## Plotting random effects...
## Plotting random effects...
```


(Intercept)
comparison_count.center
Group levels
Uralic:Uralic
Turkic:Uralic
Koreanic:Uralic
Koreanic:Turkic
Japonic:Uralic
Japonic:Turkic
Japonic:Koreanic
Indo−European:Uralic
Indo−European:Turkic
Indo−European:Koreanic
Indo−European:Japonic
Indo−European:Indo−European
Afro−Asiatic:Uralic
Afro−Asiatic:Turkic
Afro−Asiatic:Koreanic
Afro−Asiatic:Japonic
Afro−Asiatic:Indo−European
−0.00
0.00
0.00
0.00
−0.01
−0.00
−0.00
−0.01
−0.00
0.00
−0.03
0.02
0.00
0.00
0.00
−0.00
0.02
0.00
−0.00
−0.00
−0.00
0.00
0.00
0.00
0.00
0.00
−0.00
0.02
−0.01
−0.00
−0.00
−0.00
0.00
−0.01
−0.10
−0.05
0.00
0.05
0.10
−0.10
−0.05
0.00
0.05
0.10
BLUP


(Intercept)
Group levels
N Coast Asia:N Coast Asia
Inner Asia:N Coast Asia
Inner Asia:Inner Asia
Greater Mesopotamia:N Coast Asia
Greater Mesopotamia:Inner Asia
Greater Mesopotamia:Greater Mesopotamia
Europe:N Coast Asia
Europe:Inner Asia
Europe:Greater Mesopotamia
Europe:Europe
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
0.00
−0.50
−0.25
0.00
0.25
0.5
BLUP

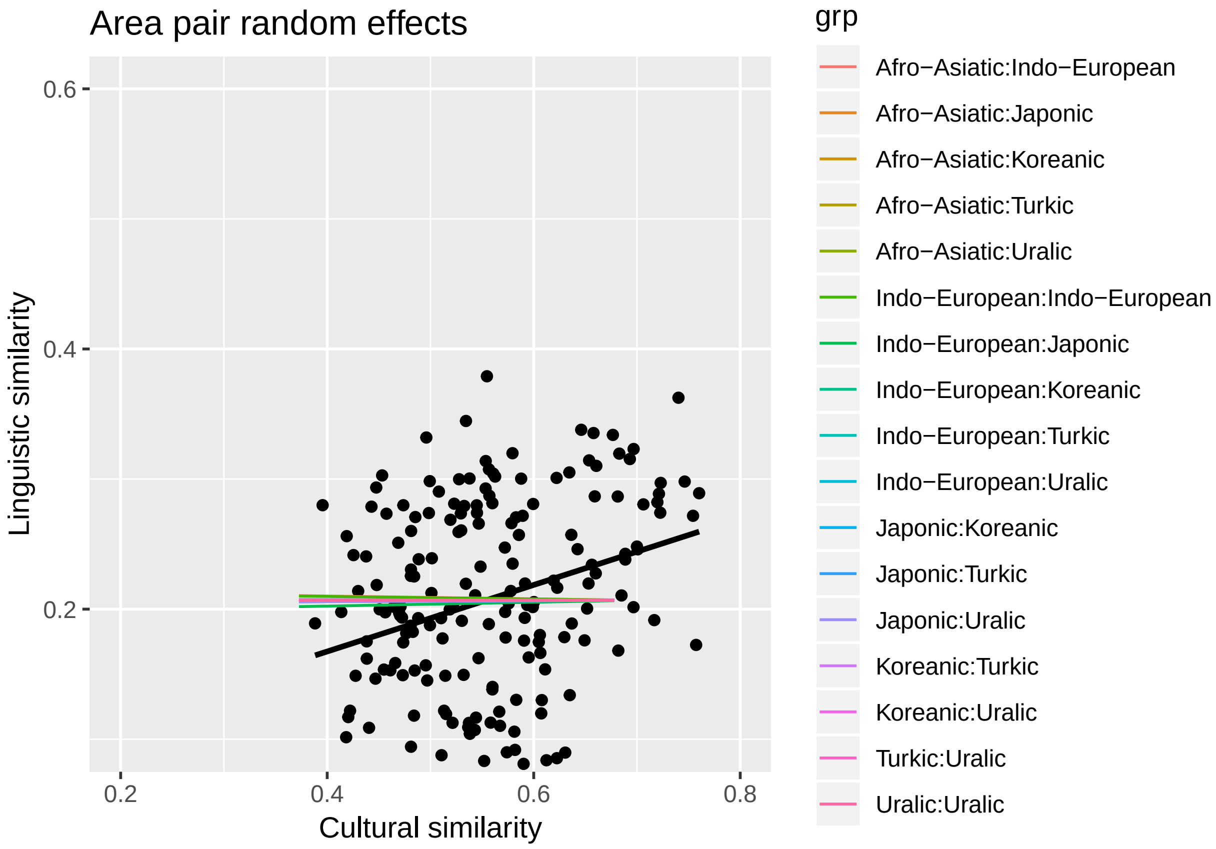
```
px = sjp.lmer(m1,'rs.ri', prnt.plot = F)
dx = px$plot[[1]]$data
dx$x = dx$x * cdc.s + cdc.c
dx$y = dx$y *
  attr(ling$rho.center,"scaled:scale") +
  attr(ling$rho.center,"scaled:center")

ggplot(dx,aes(x,y)) +
  geom_point(data=ling,
   mapping=aes(x=as.numeric(cult.dist),
               y=as.numeric(local_alignment))) +
  stat_smooth(data=gx$data,method="lm",colour="black",
              se=F)+
  geom_line(mapping = aes(colour=grp)) +
  xlab("Cultural similarity")+
  ylab("Linguistic similarity") +
  ggtitle("Language family pair random effects") +
  coord_cartesian(ylim=c(0.0,0.6),
                  xlim=c(0.2,0.8)) +
  theme(legend.position = "none")
```



```
dx = px$plot[[1]]$data
dx$x = dx$x * cdc.s + cdc.c
dx$y = dx$y *
  attr(ling$rho.center,"scaled:scale") +
  attr(ling$rho.center,"scaled:center")
```

```
ggplot(dx,aes(x,y)) +
  geom_point(data=ling,
             mapping=aes(x=as.numeric(cult.dist),
                         y=as.numeric(local_alignment))) +
  stat_smooth(data=gx$data,method="lm",colour="black",
              se=F)+
  geom_line(mapping = aes(colour=grp)) +
  xlab("Cultural similarity")+
  ylab("Linguistic similarity") +
  ggtitle("Area pair random effects") +
  coord_cartesian(ylim=c(0.1,0.6),
                  xlim=c(0.2,0.8))
```



Without Kinship data

The analyses below show that the strongest relationship is with Kinship. Here we run the analysis as above, but using semantic distances computed without concepts that relate to kinship. Note that the local alignment values correlate with r > 0.99.

Code for constructing the data is hidden, but it is the same as above and available in the Rmd file:

Run the lmer models:

```
m0NK = lmer(
  rho.center ~ 1 +
    (1 + cult.dist.center | family.group) +
    (1 | area.group),
  data = lingNK
)
m0.5NK = lmer(
  rho.center ~ 1 +
    comparison_count.center +
    (1 + cult.dist.center | family.group) +
    (1 | area.group),
  data = lingNK
)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control
## $checkConv, : unable to evaluate scaled gradient

## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control
## $checkConv, : Model failed to converge: degenerate Hessian with 1 negative
## eigenvalues
```

```
m1NK = lmer(
  rho.center ~ 1 +
    comparison_count.center +
    cult.dist.center +
    (1 + cult.dist.center | family.group) +
    (1 | area.group),
  data = lingNK
)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control
## $checkConv, : unable to evaluate scaled gradient

## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control
## $checkConv, : Model failed to converge: degenerate Hessian with 1 negative
## eigenvalues
```

```
anova(m0NK,m0.5NK,m1NK)
```

```
## refitting model(s) with ML (instead of REML)

## Data: lingNK
## Models:
## m0NK: rho.center ~ 1 + (1 + cult.dist.center | family.group) + (1 |
## m0NK:     area.group)
## m0.5NK: rho.center ~ 1 + comparison_count.center + (1 + cult.dist.center |
## m0.5NK:     family.group) + (1 | area.group)
## m1NK: rho.center ~ 1 + comparison_count.center + cult.dist.center +
## m1NK:     (1 + cult.dist.center | family.group) + (1 | area.group)
##        Df    AIC    BIC  logLik deviance    Chisq Chi Df Pr(>Chisq)
## m0NK    6 418.46 437.94 -203.23   406.46
## m0.5NK  7 294.70 317.43 -140.35   280.70 125.7551      1     <2e-16 ***
## m1NK    8 294.81 320.79 -139.41   278.81   1.8895      1     0.1693
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(m1NK)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: rho.center ~ 1 + comparison_count.center + cult.dist.center +
##     (1 + cult.dist.center | family.group) + (1 | area.group)
##    Data: lingNK
##
## REML criterion at convergence: 292.1
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.0282 -0.8202  0.0017  0.6505  3.5837
##
## Random effects:
##  Groups       Name             Variance Std.Dev. Corr
##  family.group (Intercept)      0.00000  0.0000
##               cult.dist.center 0.00988  0.0994    NaN
##  area.group   (Intercept)      0.00000  0.0000
##  Residual                      0.25259  0.5026
## Number of obs: 190, groups:  family.group, 17; area.group, 10
##
## Fixed effects:
##                         Estimate Std. Error t value
## (Intercept)             -0.01788    0.03949  -0.453
## comparison_count.center  0.82854    0.03794  21.837
## cult.dist.center         0.08093    0.05994   1.350
##
## Correlation of Fixed Effects:
##             (Intr) cmpr_.
## cmprsn_cnt.  0.003
## clt.dst.cnt  0.173 -0.134
## convergence code: 0
## unable to evaluate scaled gradient
## Model failed to converge: degenerate  Hessian with 1 negative eigenvalues
```

Mantel tests

Read the historical distances for Indo-European, based on the phylogenetic distances.

Data prep

Load historical distances:

```
hist = read.csv("../data/trees/IndoEuropean_historical_distances.csv", stringsAsFactors = F)
hist = hist[!duplicated(hist[,1]),!duplicated(hist[,1])]
rownames(hist) = hist[,1]
hist = hist[,2:ncol(hist)]
hist.m = as.matrix(hist)
colnames(hist.m) = rownames(hist.m)
hist.m = hist.m/max(hist.m)
```

Read the cultural distance as a matrix:

```
cult.m = read.csv("../results/EA_distances/CulturalDistances.csv", stringsAsFactors = F)
rownames(cult.m) = cult.m[,1]
cult.m = cult.m[,2:ncol(cult.m)]
```

Flip the cultural distance into a cultural similarity measure:

```
cult.m = 1-cult.m
```

Convert the linguistic similarities to a matrix. This uses `igraph` to make an undirected graph from the long format with `local_alignment` as the edge weights, then output a matrix of adjacencies.

```
grph <- graph.data.frame(ling[,c("l1",'l2','local_alignment')], directed=FALSE)
# add value as a weight attribute
ling.m = get.adjacency(grph, attr="local_alignment", sparse=FALSE)
rownames(ling.m) = l[match(rownames(ling.m),l$iso2),]$Language2
colnames(ling.m) = l[match(colnames(ling.m),l$iso2),]$Language2
```

Load the geographic distances:

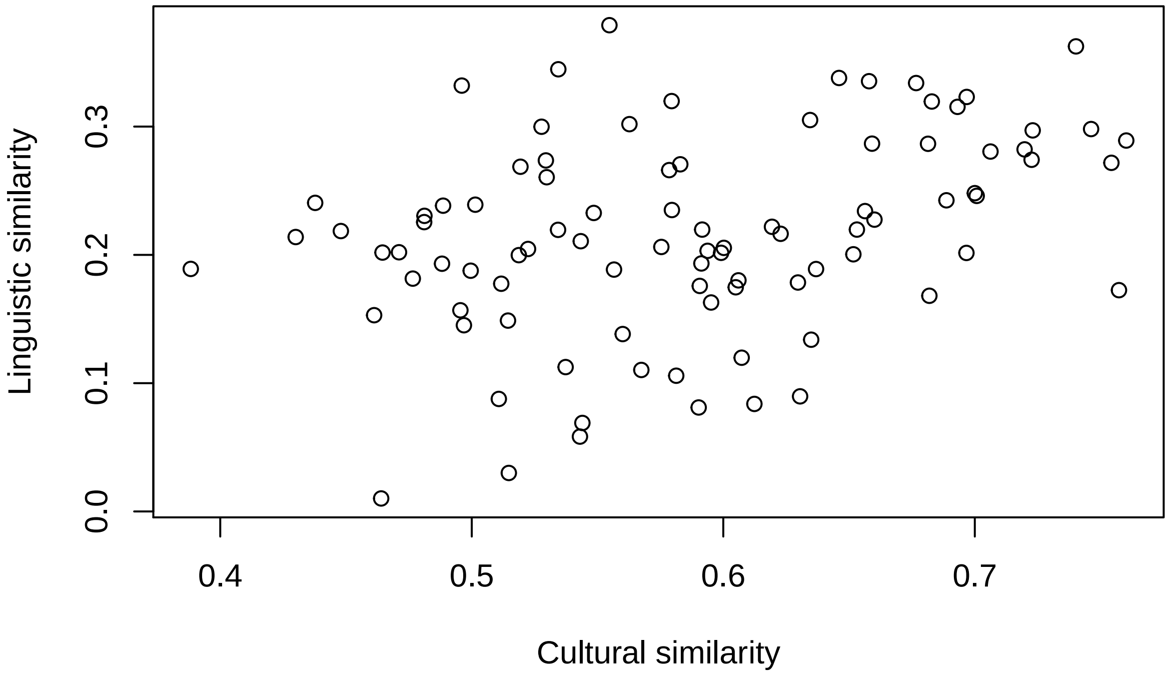
```
geoDist = read.csv("../data/GeographicDistances.csv",stringsAsFactors = F)
geoDist.m = as.matrix(geoDist)
# Convert to log distance
geoDist.m = log(geoDist.m)
geoDist.m[is.infinite(geoDist.m)] = 0
rownames(geoDist.m) = colnames(geoDist.m)
```

Match the distance matrices

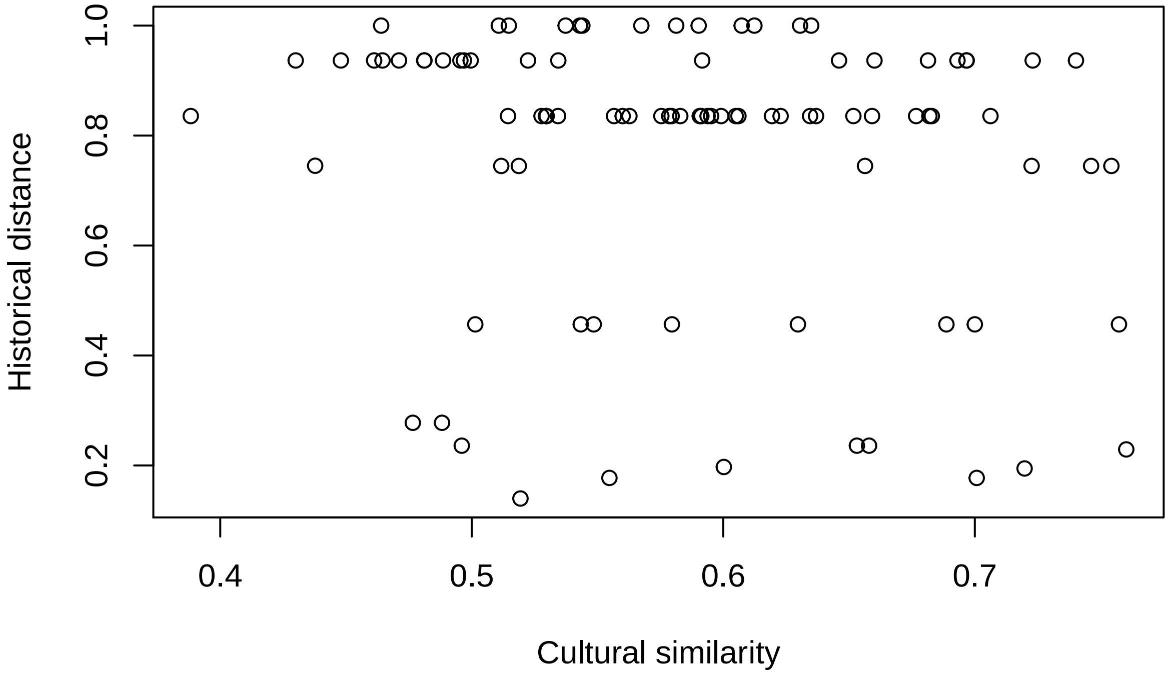
```
in.analysis = intersect(rownames(ling.m),rownames(cult.m))
in.analysis = intersect(in.analysis, rownames(hist.m))
cult.m2 = cult.m[in.analysis,in.analysis]
ling.m2 = ling.m[in.analysis,in.analysis]
hist.m2 = hist.m[in.analysis,in.analysis]
geo.m2 = geoDist.m[in.analysis,in.analysis]
```

Note that there are only 14 languages with data on lingusitic, cultural and historical distance.

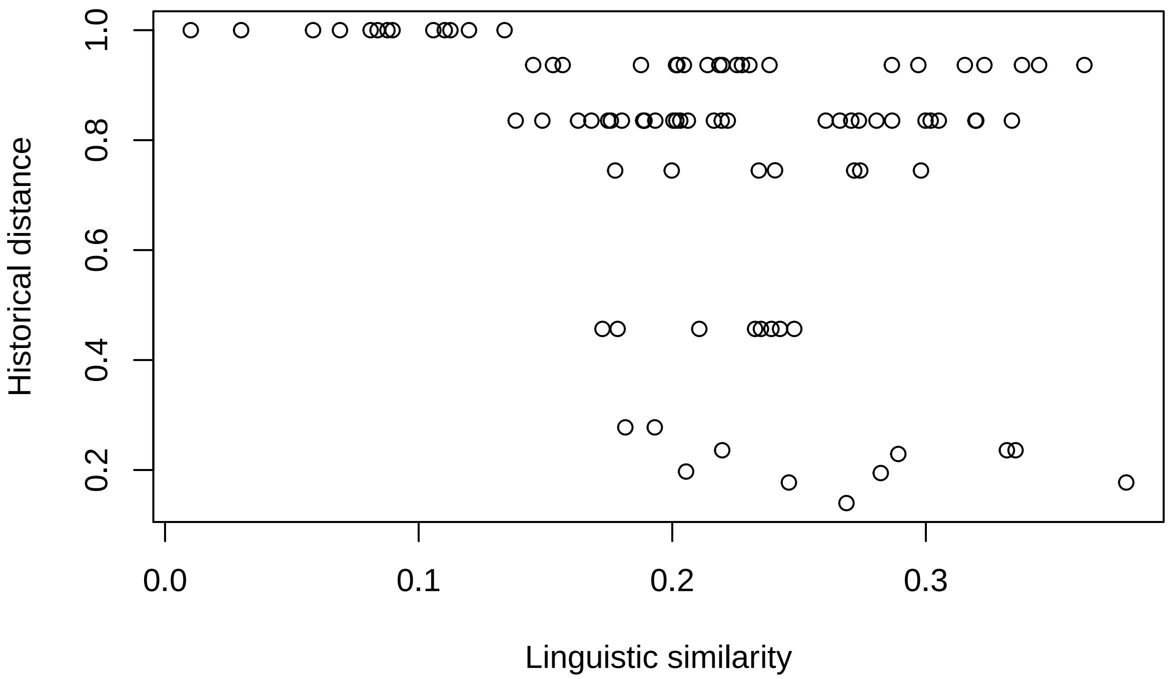
```
plot(as.dist(cult.m2),as.dist(ling.m2),
     xlab="Cultural similarity",
     ylab="Linguistic similarity")
```

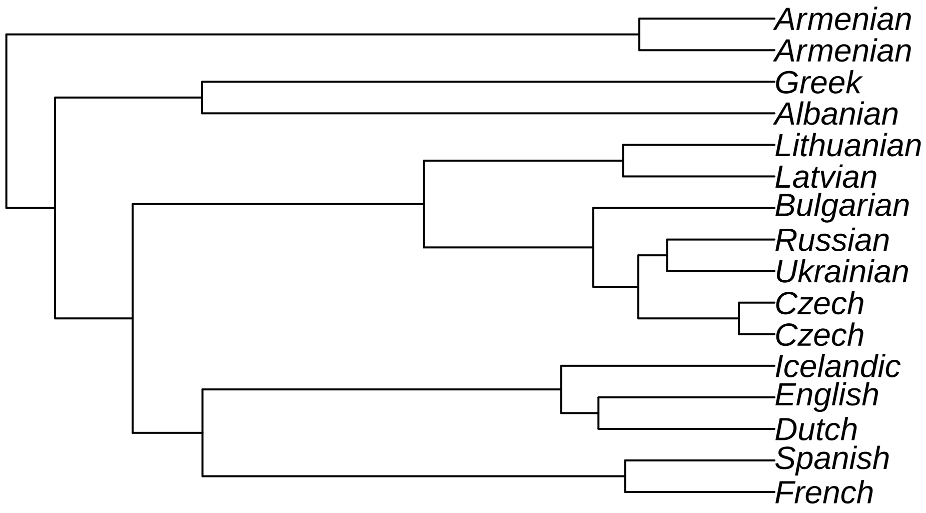


```
plot(as.dist(cult.m2),as.dist(hist.m2),
     xlab="Cultural similarity",
     ylab="Historical distance")
```



```
plot(as.dist(ling.m2),as.dist(hist.m2),
     xlab="Linguistic similarity",
     ylab="Historical distance")
```


Historical distance
1.0
0.8
0.6
0.4
0.2
0.0
0.1
0.2
0.3
Linguistic similarity


Armenian
Armenian
Greek
Albanian
Lithuanian
Latvian
Bulgarian
Russian
Ukrainian
Czech
Czech
Icelandic
English
Dutch
Spanish
French

Tests

The results of the test list the following measures:

- mantelr: Mantel correlation coefficient.
- pval1: one-tailed p-value (null hypothesis: r <= 0).
- pval2: one-tailed p-value (null hypothesis: r >= 0).
- pval3: two-tailed p-value (null hypothesis: r = 0).
- llim: lower confidence limit for r.
- ulim: upper confidence limit for r.

```
set.seed(1498)
```

```
distms = list("Cultrual"= cult.m2,
              "Linguistic" = ling.m2,
              "Historical" = hist.m2,
              "Geographic" = geo.m2)
for(i in 1:3){
  for(j in (i+1):4){
    print(paste("Correlation between",
                names(distms)[i],"and",names(distms)[j]))
    print(ecodist::mantel(as.dist(distms[[i]]) ~
                as.dist(distms[[j]]),
                nperm = 100000))
  }
}
```

```
## [1] "Correlation between Cultrual and Linguistic"
##    mantelr      pval1      pval2      pval3  llim.2.5% ulim.97.5%
##  0.3509606  0.0883600  0.9116500  0.1382200  0.2413895  0.5377689
## [1] "Correlation between Cultrual and Historical"
##     mantelr       pval1       pval2       pval3   llim.2.5%  ulim.97.5%
## -0.16966425  0.84537000  0.15464000  0.30858000 -0.28569297 -0.01526378
## [1] "Correlation between Cultrual and Geographic"
##    mantelr      pval1      pval2      pval3  llim.2.5% ulim.97.5%
## -0.3397939  0.9723900  0.0276200  0.0308800 -0.5704463 -0.1718367
## [1] "Correlation between Linguistic and Historical"
##     mantelr       pval1       pval2       pval3   llim.2.5%  ulim.97.5%
## -0.35203243  0.96988000  0.03013000  0.03453000 -0.50321827 -0.08218648
## [1] "Correlation between Linguistic and Geographic"
##    mantelr      pval1      pval2      pval3  llim.2.5% ulim.97.5%
## -0.2734487  0.9193900  0.0806200  0.1277600 -0.4617151 -0.0818242
## [1] "Correlation between Historical and Geographic"
##    mantelr      pval1      pval2      pval3  llim.2.5% ulim.97.5%
##  0.3457543  0.0100300  0.9899800  0.0100300  0.1811727  0.5217647
```

Run a mantel test comparing the Linguistic alignment to the cultural similarity, controlling for the historical distance between languages:

```
ecodist::mantel(as.dist(ling.m2)~
                  as.dist(cult.m2) +
                  as.dist(hist.m2),
                nperm = 100000)
```

```
##    mantelr      pval1      pval2      pval3  llim.2.5% ulim.97.5%
##  0.3157282  0.1119100  0.8881000  0.1894500  0.1495754  0.5004242
```

Main Test: Run a mantel test comparing the Linguistic alignment to the cultural similarity, controlling for the historical distance and geographic distance between languages:

```r
ecodist::mantel(as.dist(ling.m2)~
                  as.dist(cult.m2) +
                  as.dist(hist.m2) +
                  as.dist(geo.m2),
                nperm = 100000)
```

```
##    mantelr      pval1      pval2      pval3  llim.2.5% ulim.97.5%
##  0.2805001  0.1354200  0.8645900  0.2399900  0.1287435  0.4663710
```

Comparison between domains

The code that produce the results of this section can be found in `analysis/compareDomains.R`.

Part 1: Compare each linguistic domain to the overall cultural similarity

We fit a mixed effects model to compare the linguistic similarity in a given domain to the overall cultural distance. The linguistic similarity for the given domain is the dependent variable. There are random intercepts for language family and area pairs, and random slopes for overall cultural similarity by language family and by area. The `comparison_count` variable is added as a fixed effect. This null model is compared to a model with an additional fixed effect for the overall cultural similairty.

The full results are in the file:

`../results/stats/subs/Cor_LingAlignmentByDomains_vs_OverallCulturalSimilarity.csv`

Summary:

```
p1res
```

```
##                            Domain         Beta         p Adjusted p sig
## 5    Basic actions and technology  0.072027944 0.1813122          1
## 16     Agriculture and vegetation  0.086016716 0.1918220          1
## 1                  Food and drink  0.090849087 0.2565827          1
## 4    Miscellaneous function words  0.113890876 0.2792560          1
## 11            Emotions and values  0.061974724 0.3417402          1
## 6                    Modern world  0.067283668 0.3443726          1
## 10                       The body  0.060347923 0.3970814          1
## 8                        Quantity  0.152277174 0.4165094          1
## 9                         Animals  0.072117731 0.4850463          1
## 2                      Possession  0.037895649 0.6108568          1
## 19               Sense perception  0.081567002 0.6859150          1
## 3              The physical world  0.027793204 0.6969550          1
## 20                        Kinship -0.070823366 0.7184490          1
## 12                      Cognition -0.043080692 0.7644507          1
## 7                       The house  0.019379501 0.7930119          1
## 15 Social and political relations  0.021324237 0.8201265          1
## 14                         Motion  0.009042785 0.8408053          1
## 21          Clothing and grooming  0.008810033 0.8779180          1
## 17              Spatial relations -0.010572491 0.9441853          1
## 18                           Time  0.006212859 0.9926897          1
## 13            Speech and language  0.062847541 1.0000000          1
```

Part 2: Compare each linguistic domain to the cultural similarity of each original D-PLACE domain

The method is the same as for part 1, except the cultural distance for a particular cultural domain is used instead of the overall cultural distance.

The full results are in the file:

`../results/stats/subs/Cor_LingAlignmentByDomains_vs_DPlaceCulturalDomains.csv`

The graph below shows the mixed effects model coefficient estimate for the relationship between each linguistic domain and each cultural domain. Pink colours indicate positive correlations and blue colours indicate negative correlations. Stronger colours indicate stronger correlations. An asterisk indicates that the correlation is stronger than would be expected by chance, when adjusting the p-value for multiple comparisons.

The insert in the top left shows the distribution of Beta values.

The domains are clustered using higherarchical clustering. This is for visualisaiton and reflects similarity in the numeric relations, not history or conceptual hierarchies.

List of significant correlations (after adjusting p-value for multiple comparisons):

```
## [1] Ling Domain Cult Domain Beta        Adjusted p
## <0 rows> (or 0-length row.names)
```

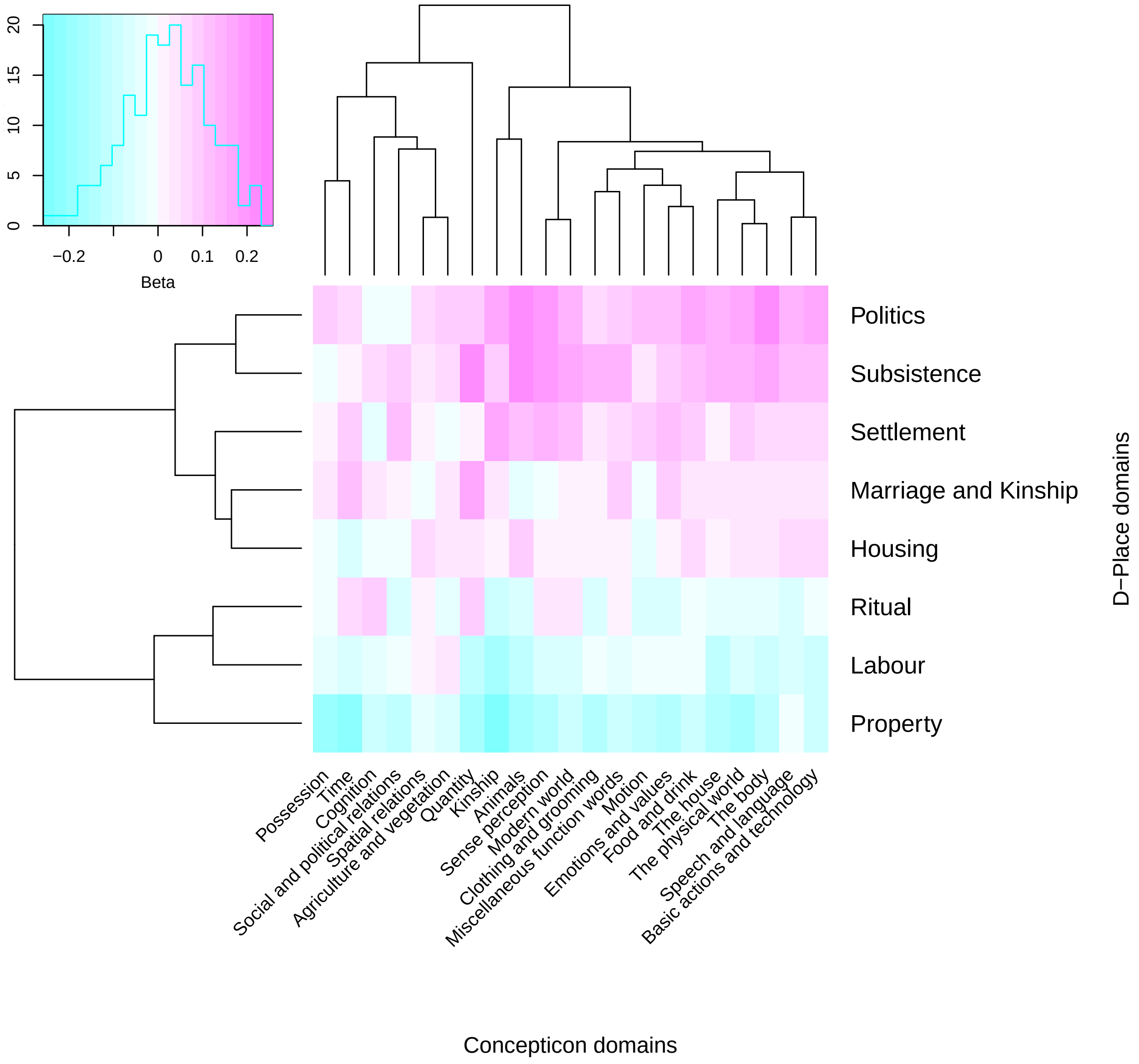

20
15
10
5
0
−0.2
0
0.1
0.2
Beta
Politics
Subsistence
Settlement
Marriage and Kinship
Housing
Ritual
Labour
Property
D−Place domains
Possession
Time
Cognition
Social and political relations
Spatial relations
Agriculture and vegetation
Quantity
Kinship
Animals
Sense perception
Modern world
Clothing and grooming
Miscellaneous function words
Motion
Emotions and values
Food and drink
The house
The physical world
The body
Speech and language
Basic actions and technology
Concepticon domains

Figure 1:

Part 3: Compare each linguistic domain to the phylogenetic and geographic distance

This test compares each linguistic similarity scores to each of three target distances: the cultural distance, the historical distance and the geographic distance. We use a partial Mantel test (from the package `ecodist`) to estimate the strength of the relationship between the linguistic domain and the target distance, while controlling for the other two distances. The test uses 100,000 permutations.

The full results are in the file:

`Cor_LingAlignmentByDomains_vs_HistoricalAndGeographicalDistance.csv`

The graph below shows the results. Point estimates are the estimated Mantel R. The error bars show the 95% confidence intervals from the permutation test.

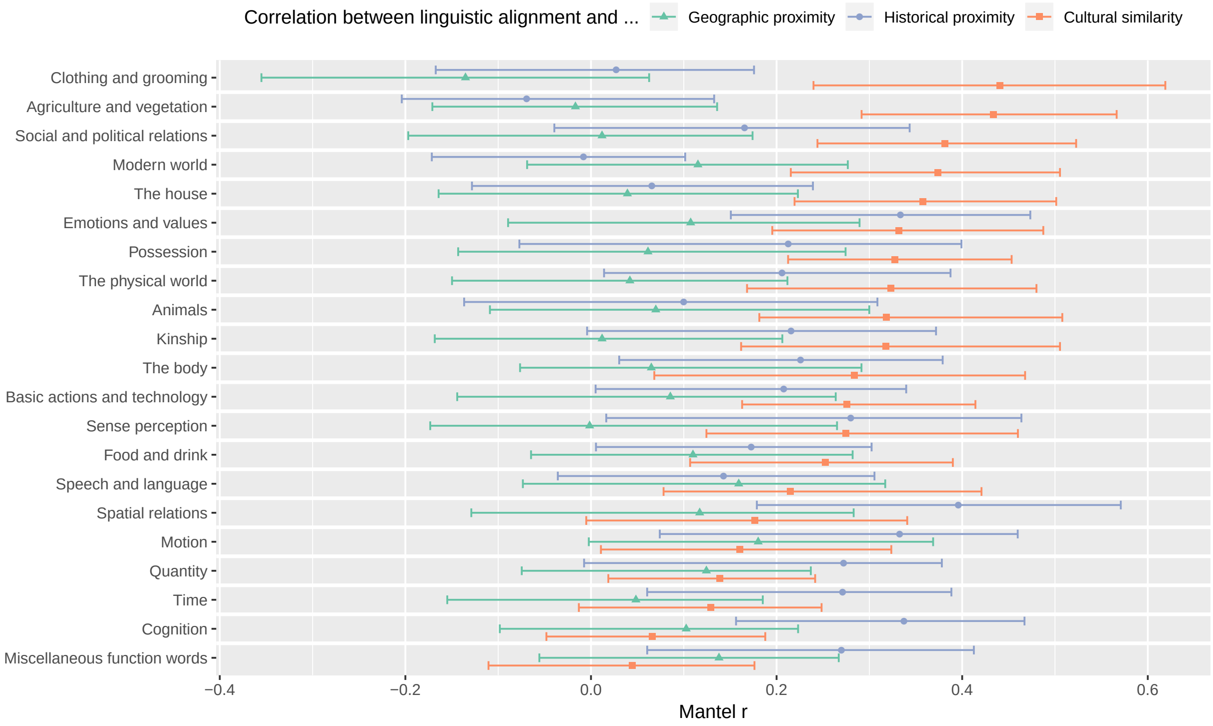


Figure 2:

References

Bouckaert, Remco, Philippe Lemey, Michael Dunn, Simon J. Greenhill, Alexander V. Alekseyenko, Alexei J. Drummond, Russell D. Gray, Marc A. Suchard, and Quentin D. Atkinson (2012). Mapping the origins and expansion of the Indo-European language family. Science, 337(6097), 957-960.