UNIVERSITÄT LEIPZIG

erc European Research Council

Probabilistic semantic maps and Multidimensional Scaling

Natalia Levshina ©2017

Summer School of Linguistics

Litomyšl, August 2017

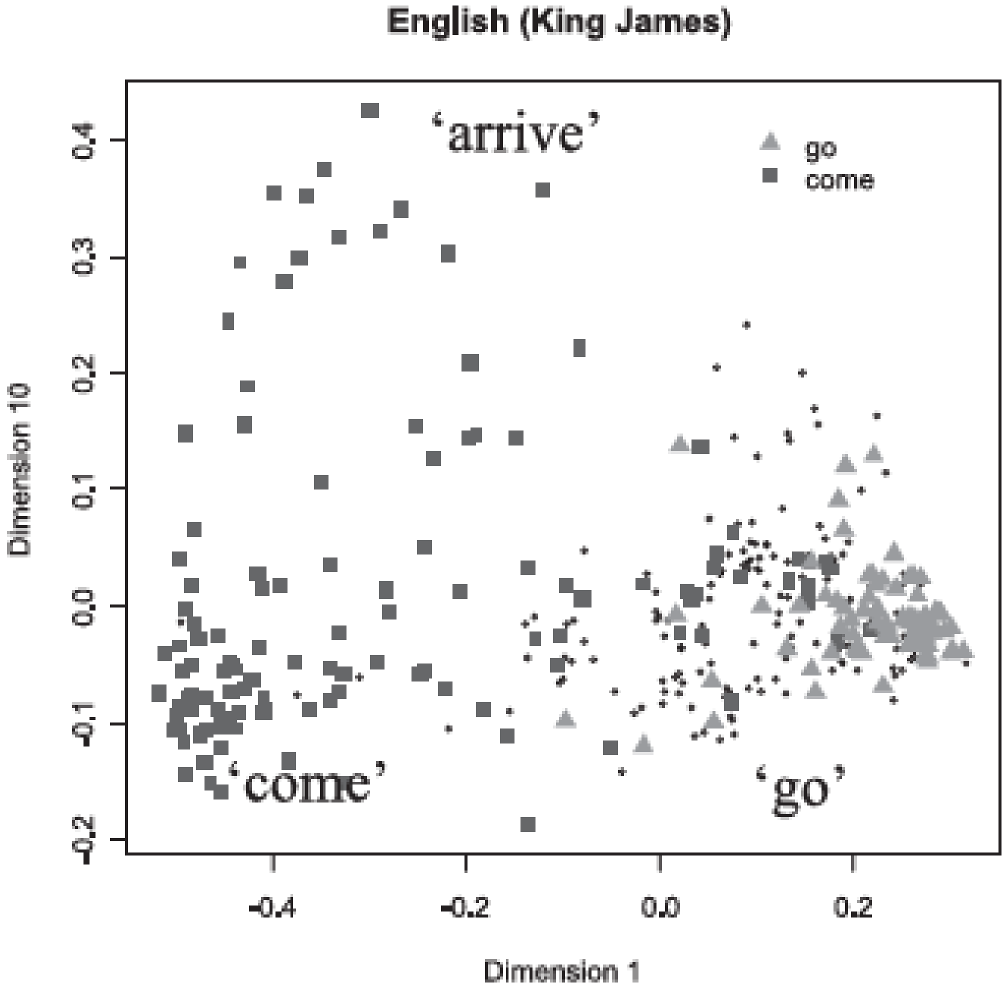
Outline

1. Introduction to probabilistic semantic maps and Multidimensional Scaling

2. Case study: analytic causatives in European languages

3. Final comments

Wälchli & Cysouw (2012): motion verbs



Algorithm for MDS: Step 1

1. Collect the data (fictitious example)

	Lang1	Lang2	Lang3	Lang4	Lang5
Sit1	bla	qu	da	nina	haha
Sit2	bla	qu	da	nana	hihi
Sit3	bla	qa	ta	nina	hehe

Algorithm for MDS: Step 1

1. Collect the data (fictitious example)

	Lang1	Lang2	Lang3	Lang4	Lang5
Sit1	bla	qu	da	nina	haha
Sit2	bla	qu	da	nana	hihi
Sit3	bla	qa	ta	nina	hehe

Comparative concepts (cf. Haspelmath 2010)

Algorithm for MDS: Step 2

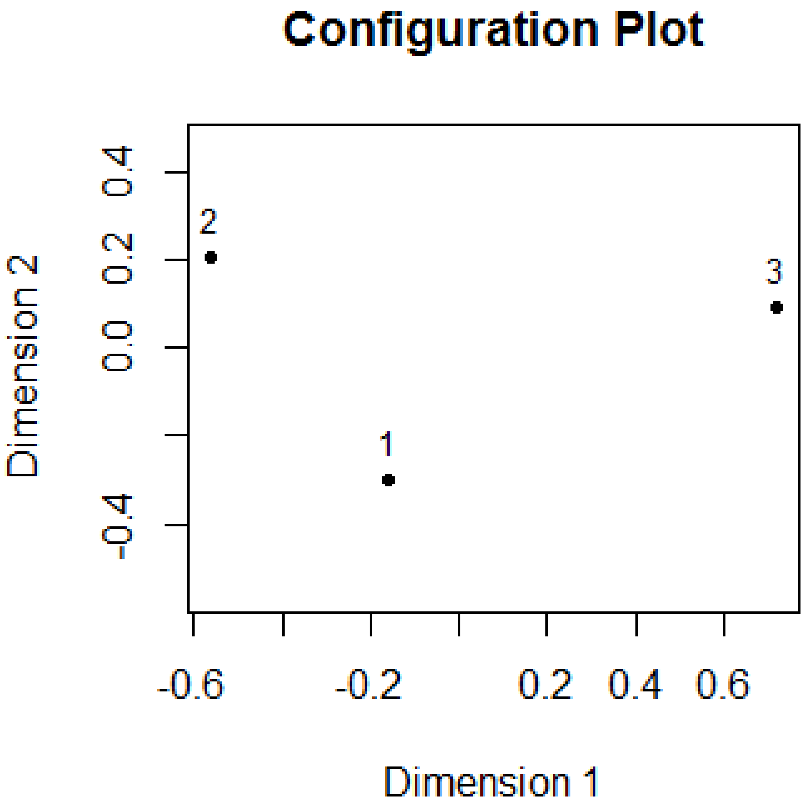
2. Compute the distances between the situations (rows) = the proportion of dissimilar values.

	Lang1	Lang2	Lang3	Lang4	Lang5
Sit1	bla	qu	da	nina	haha
Sit2	bla	qu	da	nana	hihi
Sit3	bla	qa	ta	nina	hehe

D (1,2) = 2/5 = 0.4
D (1,3) = 3/5 = 0.6
D (2,3) = 4/5 = 0.8

Algorithm for MDS: Step 3

3. Perform MDS



Interpretation

- The closer two points on the map, the more overlapping constructions they share across the languages.
- Following the isomorphism principle (same function => same form), the corresponding functions/meanings/situations are more semantically similar if more authors of the doculects chose identical constructions to represent these functions/meaning/situations.

Outline

1. Basic principles of probabilistic semantic maps and Multidimensional Scaling

2. Case study: analytic causatives in European languages

3. Final comments

Languages

- Indo-European
  - Germanic
    - Dutch, English, German, Norwegian, Swedish
  - Romance
    - French, Italian, Portuguese, Romanian, Spanish
  - Slavic
    - Bulgarian, Czech, Polish, Russian, Slovene
- Uralic
  - Finnic
    - Estonian, Finnish
  - Ugric
    - Hungarian

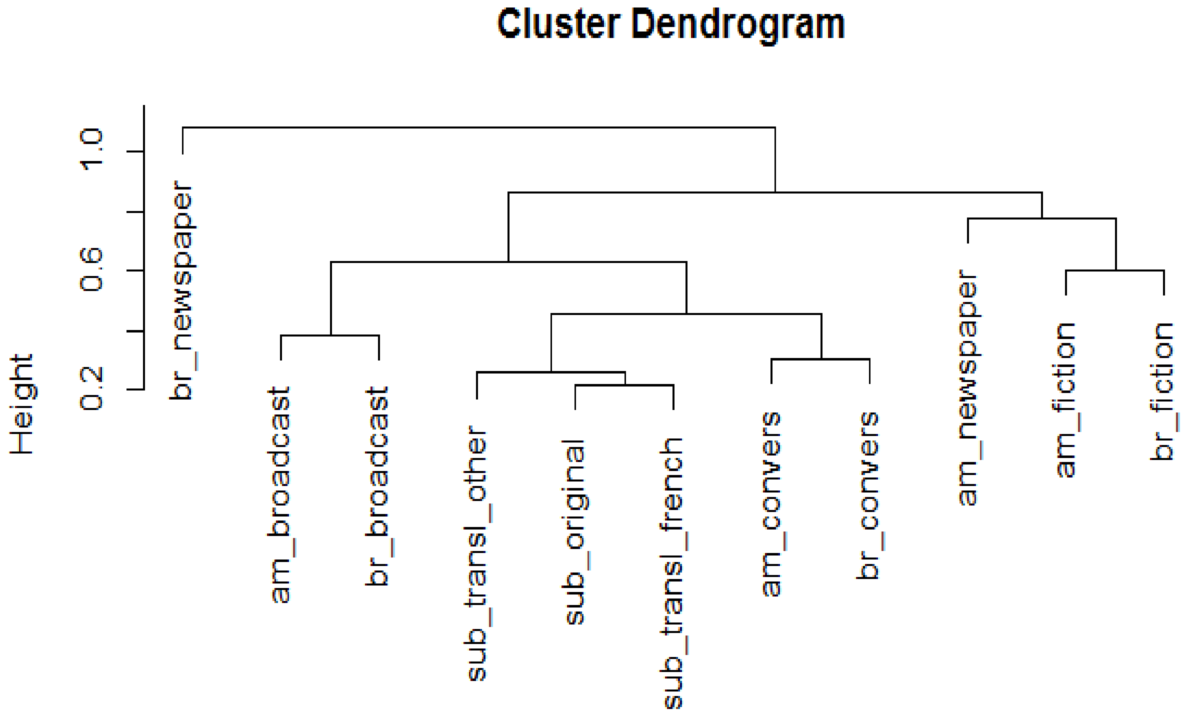
Analytic Causatives: Examples

- English:
  - *make* + Vinf, *let* + Vinf, *have* + Vinf, *cause* + *to-Vinf*
- German:
  - *lassen* + Vinf
- Dutch:
  - *laten* + Vinf, *doen* + Vinf
- Russian: *zastavljat*’ “force” + Vinf, *davat*’ “give” + Vinf
- French:
  - *faire* + Vinf, *laisser* + Vinf
- Romanian:
  - *face* + *să* + Vsubj, *lasă* + *să* + Vsubj

ParTy corpus

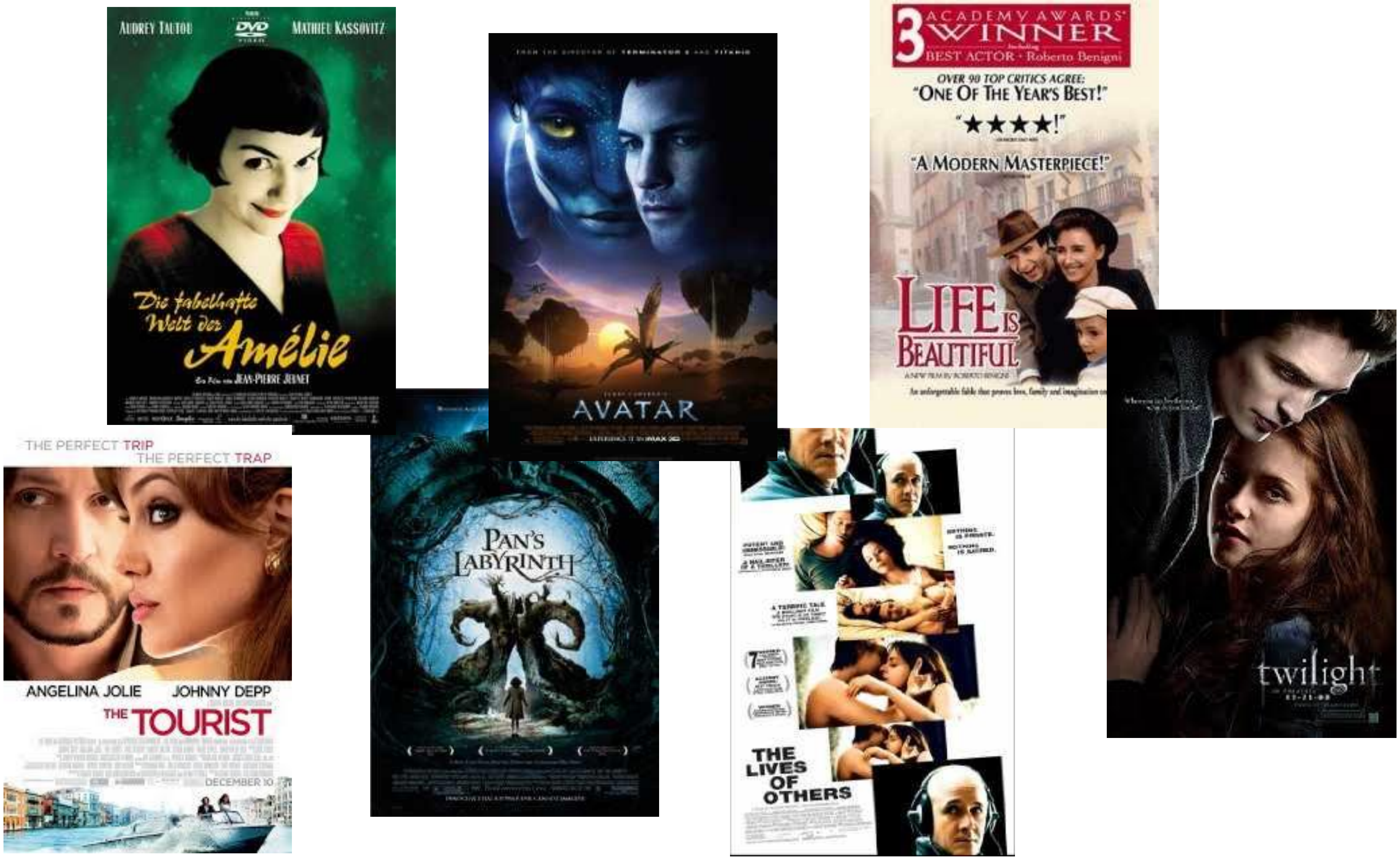
- a Parallel corpus for Typology
- subtitles of films and TED talks
- mostly European languages, but also other major languages (Chinese, Turkish, Indonesian, etc.)
- all languages aligned with English
- downloadable files at www.natalialevshina.com/corpus.html

Why subtitles?



Based on the frequencies of 3-grams (Levshina, In press)

Films for case study



An example of .srt format

```
…
646
00:51:27,880 --> 00:51:32,920
<i>For always evil will look to
find a foothold in this world.</i>

647
00:51:39,440 --> 00:51:42,603
Not good. Not good at all.

648
00:51:50,040 --> 00:51:51,326
Eww.

649
00:52:06,760 --> 00:52:09,081
Oh, no. Sebastian.

650
00:52:12,800 --> 00:52:13,847
Good gracious.

651
00:52:34,720 --> 00:52:35,767
Come on.
…
```

Data set

- Alignment: Jörg Tiedemann’s software subalign
- All contexts with of ACs in 18 languages
- Dataset: 72 contexts, in which at least 6 languages have an AC

Examples

- Situation (row) A

  ENG: And we make them do it… ...or we kill them. make

  ITA: E glielo facciamo fare ... o lo uccidiamo. fare

  CZE: Donutíme je to udělat, nebo je zabijeme. donutit

- Situation (row) B

  ENG: Pick up someone my height and build and make them believe it is me. Make

  ITA: Individua una della mia corporatura e fa credere loro che sia io. Fare

  CZE: Vyber někoho, kdo je mi podobný a přesvědč je, že jsem to já. NA

Data frame causatives

```
> str(causatives)
'data.frame':     72 obs. of  20 variables:
 $ Film: Factor w/ 8 levels "Amelie","Avatar",..:
1 1 1 1 1 1 1 1 1 1 ...
 $ Text: Factor w/ 72 levels "...and won't let the
tree thrive.",..: 20 47 48 50 62 25 5 6 54 7 ...
 $ FRA : Factor w/ 6 levels
"autoriser","faire",..: 2 2 4 NA 2 4 4 4 4 2 ...
 $ ENG : Factor w/ 6 levels "allow","force",..: 4
NA 5 6 4 5 5 5 NA 6 ...
 $ GER : Factor w/ 3 levels
"bringen","erlauben",..: 3 3 3 3 3 3 3 NA 3 3 ...
[output omitted]
```

Data frame causatives

```
> head(causatives[,3:10])
      FRA  ENG    GER   SPA   DUT  SWE      ITA    POR
1   faire have lassen  <NA>  <NA> <NA>     fare   <NA>
2   faire <NA> lassen hacer  <NA> <NA>     fare   <NA>
3 laisser  let lassen dejar laten lata     fare   <NA>
4    <NA> make lassen  <NA>  doen   fa     <NA>   <NA>
5   faire have lassen  <NA> laten <NA>     <NA>   <NA>
6 laisser  let lassen dejar laten lata lasciare deixar
```

Gower distances

```
> library(cluster)
> causatives.dist <- daisy(causatives[, 3:20])
> summary(causatives.dist)

2556 dissimilarities, summarized :
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
 0.0000  0.1429  0.6000  0.5415  0.9091  1.0000       1
Metric :  mixed ;  Types = N, N, N, N, N, N, N, N, N, N,
N, N, N, N, N, N, N, N
Number of objects : 72
```

Understanding Gower distances

```
> causatives.dist[1:3]
[1] 0.0 0.6 0.5
> causatives[1:2, 3:20]
 FRA  ENG    GER   SPA  DUT  SWE  ITA  POR  ROM  POL
1 faire have lassen  <NA> <NA> <NA> fare <NA> <NA> <NA>
2 faire <NA> lassen hacer <NA> <NA> fare <NA> face <NA>
 SLO  CZE  RUS  BUL   EST     FIN  HUN  NOR
1 dati  dat <NA> <NA> <NA>   <NA> <NA> <NA>
2 <NA> <NA> <NA> <NA> panema <NA> <NA>  la
```

D (1, 2) = 0/3 = 0

Task

- Compute the Gower distance between observations 71 and 72.

Solution

```
> causatives[71:72, 3:20]
     FRA  ENG  GER   SPA   DUT  SWE  ITA     POR  ROM
71 faire  get <NA> hacer laten <NA> fare   fazer face
72 faire <NA> <NA> hacer  <NA> <NA> fare obrigar face
    POL  SLO    CZE        RUS      BUL    EST   FIN
71 <NA> <NA> nechat zastavljat nakarvam panema saada
72 <NA> <NA>   <NA>       <NA>     <NA> panema  <NA>
    HUN  NOR
71 <NA> <NA>
72 <NA> <NA>
```

$$D\,(71, 72) = 1/6 \approx 0.167$$

Fitting MDS

```
> library(smacof)
```

Fitting a two-dimensional metric MDS (default):

```
> causatives.mds <- mds(causatives.dist)
```

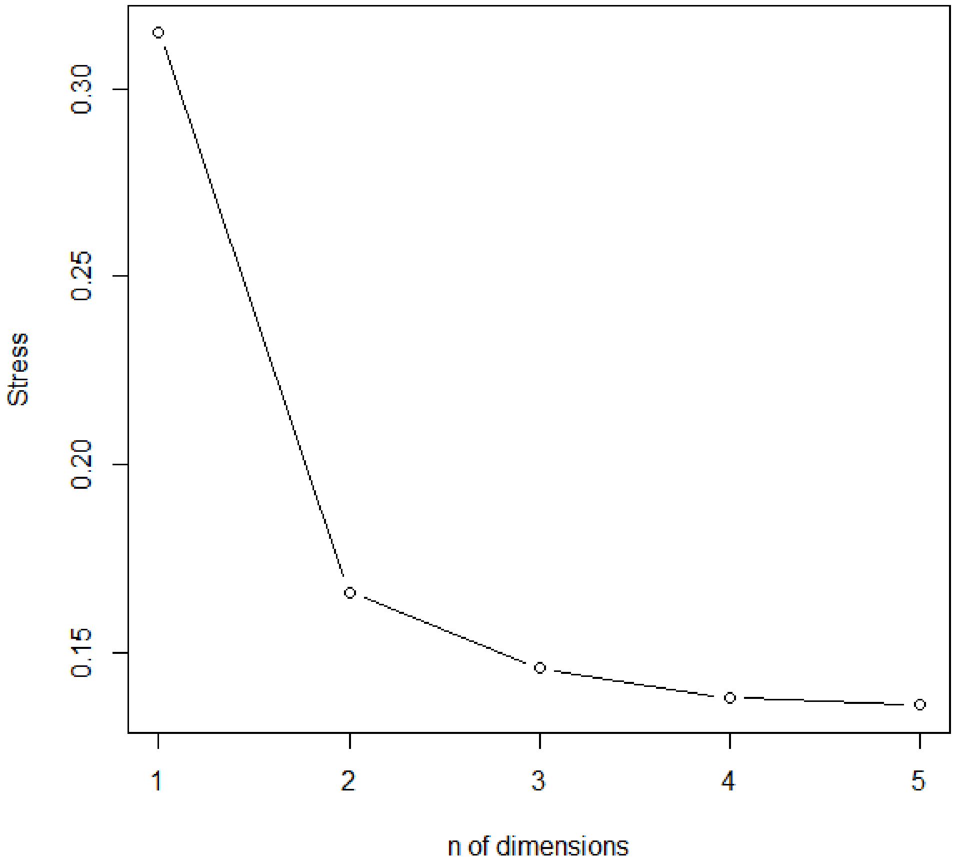
How good is the 2D solution?

```
> causatives.mds$stress
[1] 0.166 #relatively OK
> d1 <- mds(causatives.dist, ndim = 1)$stress
> d2 <- mds(causatives.dist, ndim = 2)$stress
> d3 <- mds(causatives.dist, ndim = 3)$stress
> d4 <- mds(causatives.dist, ndim = 4)$stress
> d5 <- mds(causatives.dist, ndim = 5)$stress
```

Make a scree plot:

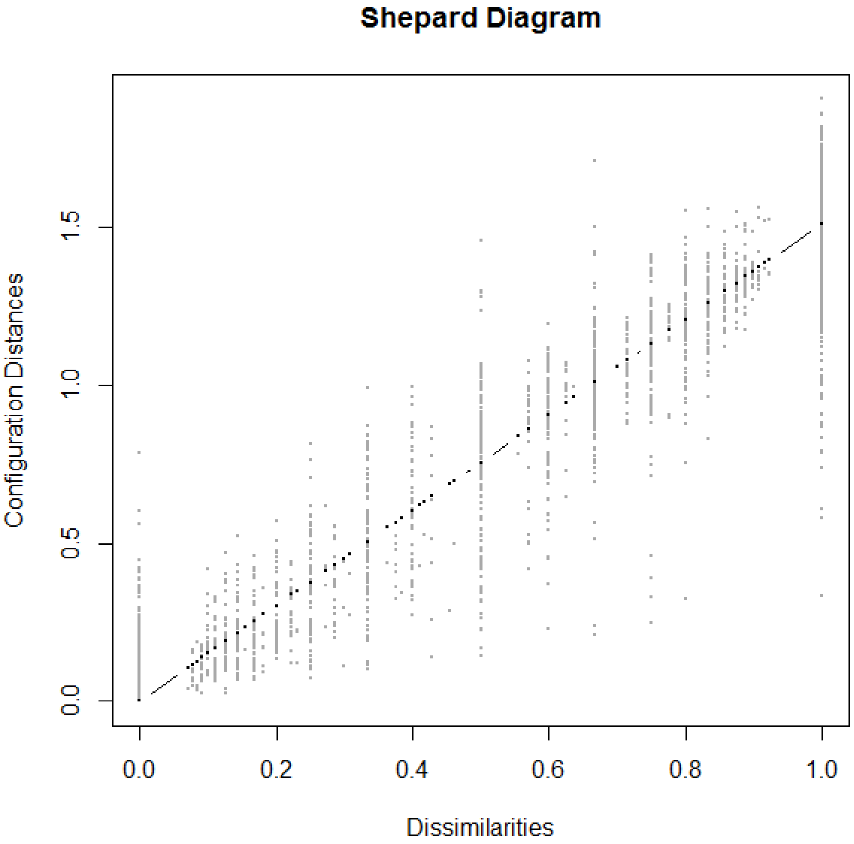
```
> plot(1:5, c(d1, d2, d3, d4, d5), type = "b",
xlab = "n of dimensions", ylab = "Stress")
```

Watch the ‘elbow’



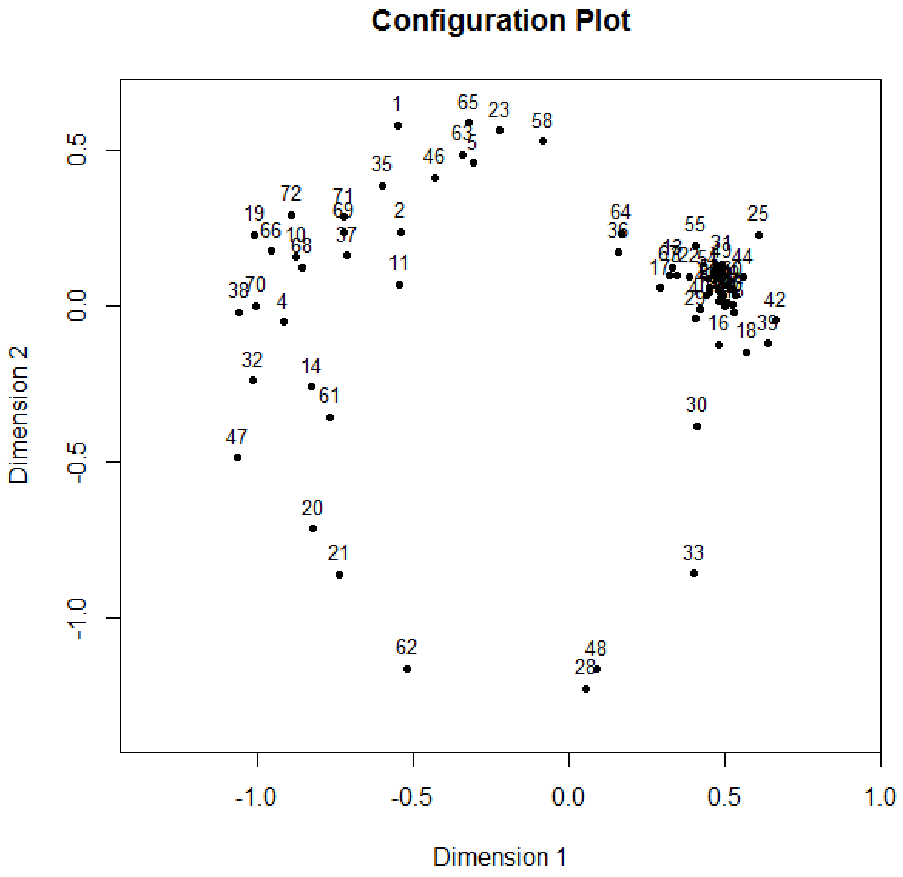
Stress and individual distances

```
> plot(causatives.mds,
"Shepard")
```



Interpreting the solution

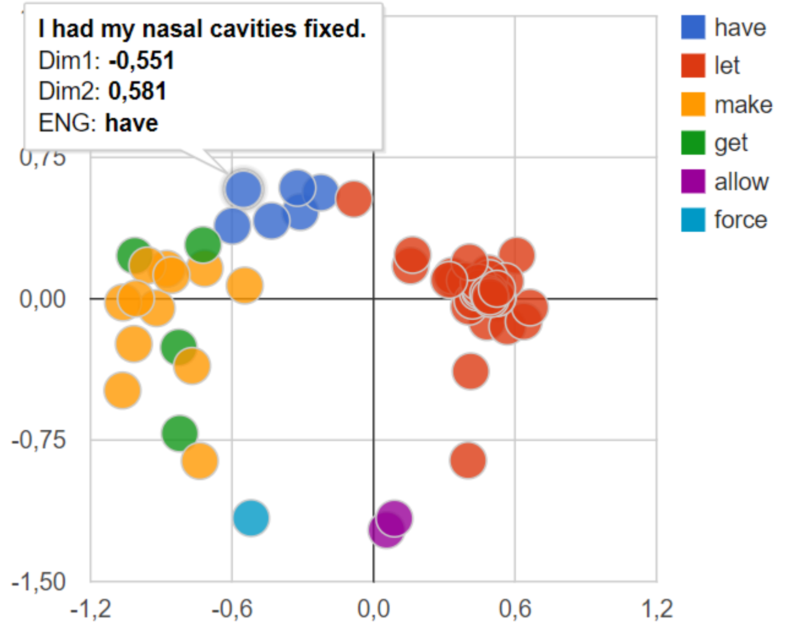
```
> plot(causatives.mds,
"conf")
```



Exploring the contexts: bubbles

```
> library(googleVis)
> text.df <- data.frame(Text = causatives$Text,
Dim1 = causatives.mds$conf[, 1], Dim2 =
causatives.mds$conf[, 2], ENG = causatives$ENG)


> bubbles <- gvisBubbleChart(text.df, idvar =
"Text", xvar = "Dim1", yvar = "Dim2", colorvar =
"ENG", options = list(sizeAxis = '{maxSize:
10}', vAxis = '{minValue:-0.8, maxValue:0.8}',
height = 500, width = 500,
bubble="{textStyle:{color: 'none'}}"))


> plot(bubbles)
```

Bubble chart

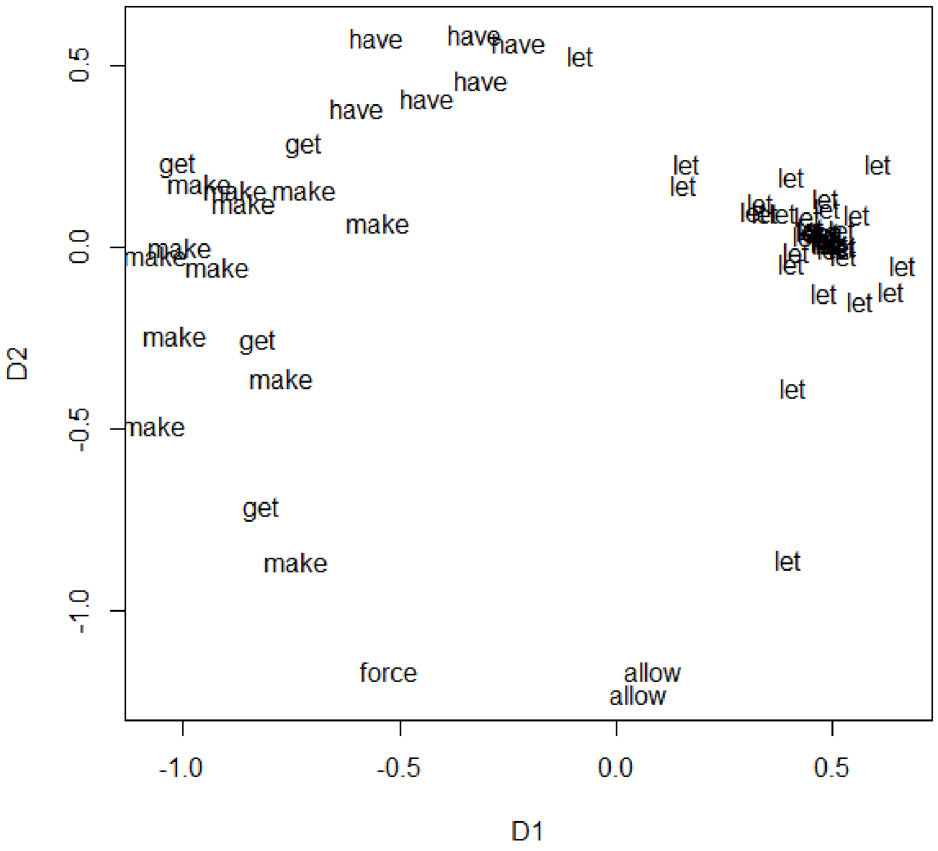


I had my nasal cavities fixed.
Dim1: **-0,551**
Dim2: **0,581**
ENG: **have**

Data: text.df • Chart ID: BubbleChartID1d6472c66b06 • googleVis-0.6.2
R version 3.4.0 (2017-04-21) • Google Terms of Use • Documentation and Data Policy

Exploring form-meaning mapping

```
> plot(causatives.mds$conf,
type = "n")

> text(causatives.mds$conf,
labels = causatives$ENG)
```



“Make” in Romance

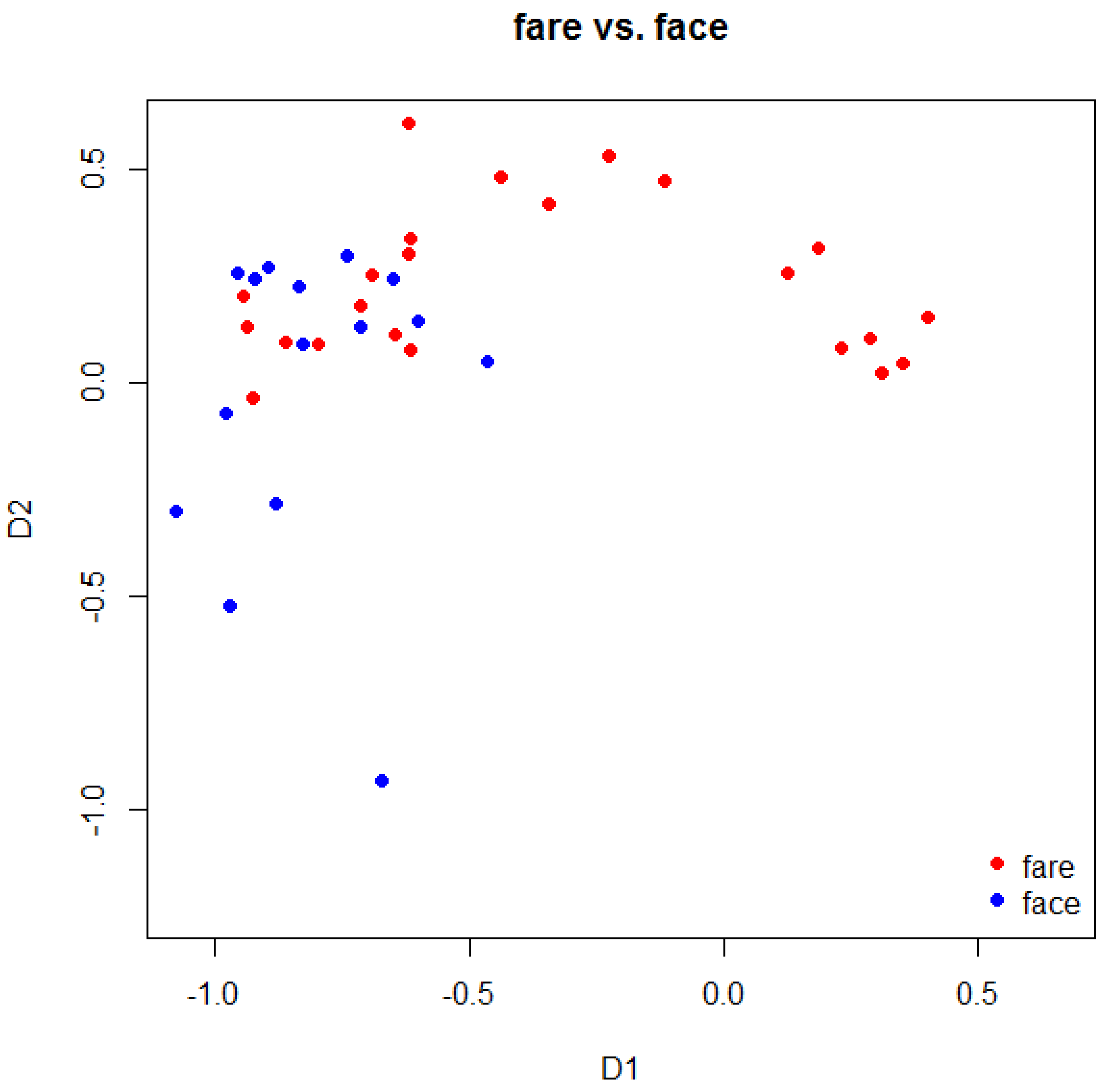
- Let us compare the semantics of cognate causal auxiliaries in Romance:
  - FRA faire
  - ITA fare
  - POR fazer
  - ROM (a) face
  - SPA hacer
- Are there semantic differences?
- For theoretical background, see Levshina (2015).

ITA fare vs. ROM (a) face

```
> plot(causatives.mds$conf, type = "n", main =
"fare vs. face")
```

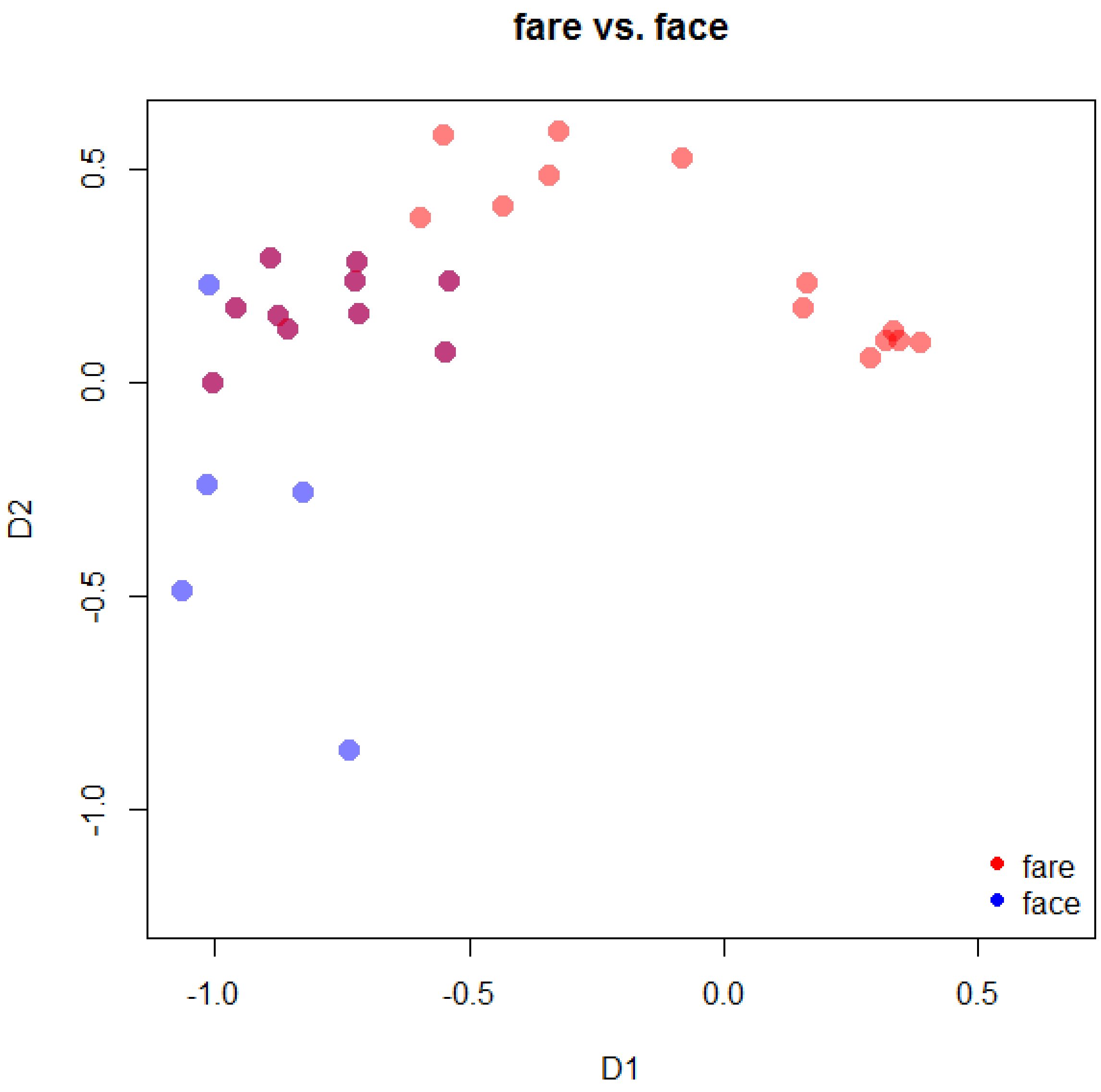
Add some jitter to avoid overplotting:

```
> points(jitter(causatives.mds$conf[causatives$ROM
== "face",], amount = 0.1), col = "blue", pch =
16)

> points(jitter(causatives.mds$conf[causatives$ITA
== "fare",], amount = 0.1), col = "red", pch = 16)

> legend("bottomright", legend = c("fare",
"face"), col = c("red", "blue"), pch = 16, bty =
"n")
```


fare vs. face
D2
0.5
0.0
-0.5
-1.0
D1
-1.0
-0.5
0.0
0.5
● fare
● face

Transparent colours

```
> library(grDevices)

> plot(causatives.mds$conf, type = "n", main =
"fare vs. face")

> points(causatives.mds$conf[causatives$ROM ==
"face",], col = adjustcolor("blue", alpha.f =
0.5), pch = 16, cex = 1.5)

> points(causatives.mds$conf[causatives$ITA ==
"fare",], col = adjustcolor("red", alpha.f = 0.5),
pch = 16, cex = 1.5)

> legend("bottomright", legend = c("fare",
"face"), col = c("red", "blue"), pch = 16, bty =
"n")
```


fare vs. face
D2
0.5
0.0
-0.5
-1.0
D1
-1.0
-0.5
0.0
0.5
● fare
● face

Kriging: preparation

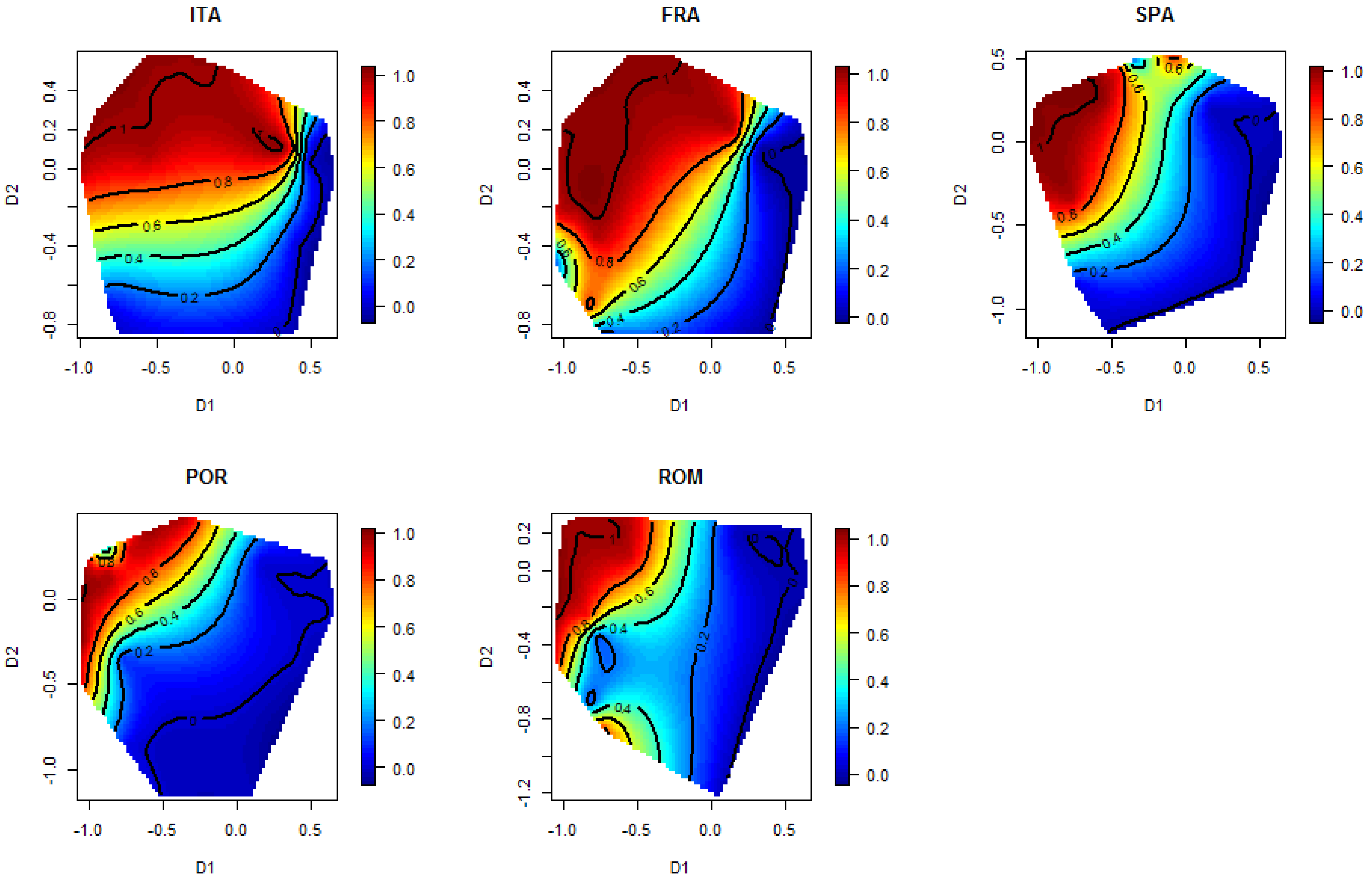
```
> y.ita <- ifelse(causatives$ITA == "fare", 1, 0)
> y.fra <- ifelse(causatives$FRA == "faire", 1, 0)
> y.spa <- ifelse(causatives$SPA == "hacer", 1, 0)
> y.por <- ifelse(causatives$POR == "fazer", 1, 0)
> y.rom <- ifelse(causatives$ROM == "face", 1, 0)
```

Kriging

```
> Krig.rom <- Krig(causatives.mds$conf, y.rom,
lambda = 0.05) #try different lambda values

> Krig.ita <- Krig(causatives.mds$conf, y.ita,
lambda = 0.05)

> Krig.fra <- Krig(causatives.mds$conf, y.fra,
lambda = 0.05)

> Krig.spa <- Krig(causatives.mds$conf, y.spa,
lambda = 0.05)

> Krig.por <- Krig(causatives.mds$conf, y.por,
lambda = 0.05)
```

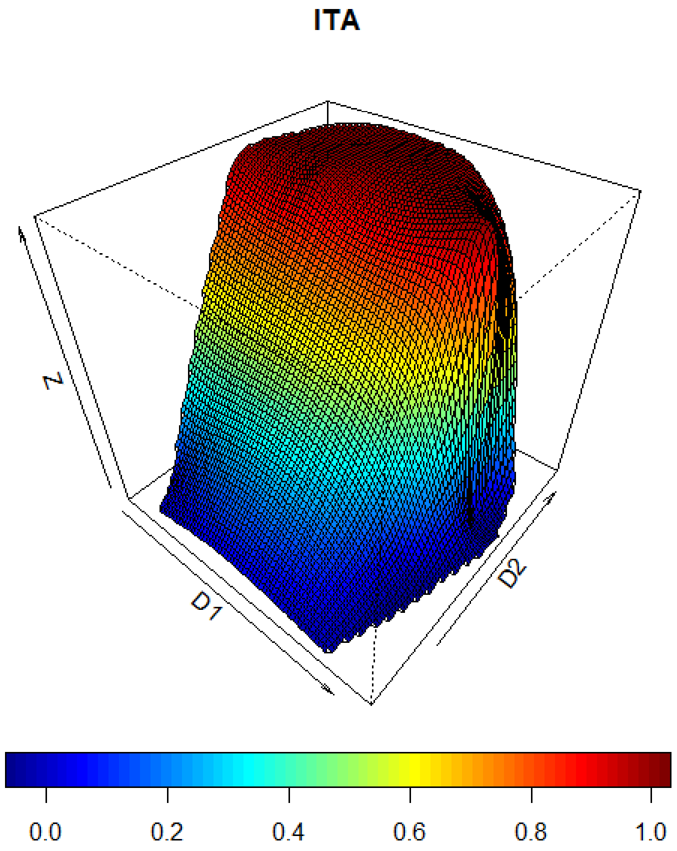
Fitted surface plots

```
> surface(Krig.ita, main = "ITA")
> surface(Krig.fra, main = "FRA")
> surface(Krig.spa, main = "SPA")
> surface(Krig.por, main = "POR")
> surface(Krig.rom, main = "ROM")
```


ITA
FRA
SPA
POR
ROM
D1
D2
-1.0
-0.5
0.0
0.5
-1.2
-0.8
-0.4
0.2
0.4
1.0
0.8
0.6
0.4
0.2
0.0
1
0.8
0.6
0.4
0.2
0

Perspective plot

```
> surface(Krig.ita,
main = "ITA", type =
"p", theta = 40, phi
= 40)
```



Romance causatives: conclusion

- The Italian causative verb fare is the most semantically bleached with regard to the distinction between letting and marking, and the Romanian face is the least bleached.
- The other languages are in-between.
- A scale of grammaticalization:
  - ITA > FRA > SPA > POR > ROM
- This is reflected in the different levels of syntactic integration of the auxiliary and the second predicate:
  - The Italian *fare* and French *faire* are normally followed immediately by an infinitive (VV)
  - Portuguese *fazer* and Spanish *hacer* are often used in the pattern V + NP + V
  - Romanian *a face* is followed by the complementizer *să* and a subjunctive clause (finite).

Outline

1. Basic principles of probabilistic semantic maps and Multidimensional Scaling

2. Case study: analytic causatives in European languages

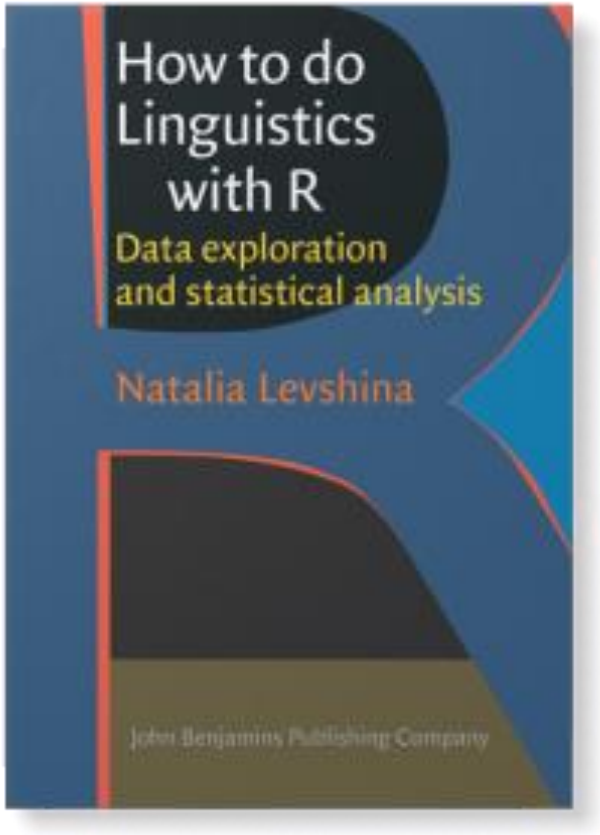
3. Final comments

Take-home messages

1. Exploratory methods are only exploratory: their purpose is to help you understand your data better and form hypotheses, which you can test with the help of confirmatory methods.

2. While performing your analysis, play around with different exploratory methods and visualization techniques. If the results converge, this means that the pattern is robust. If they diverge, your task is to try to understand why they do.

More ideas and examples here:



How to do Linguistics with R

Data exploration and statistical analysis

Natalia Levshina

John Benjamins Publishing Company

References

- Haspelmath, M. (2010) Comparative concepts and descriptive categories in cross-linguistic studies. *Language* 86(3). 663-687
- Levshina, N. (2015) European analytic causatives as a comparative concept. Evidence from a parallel corpus of film subtitles. *Folia Linguistica* 49(2). 487-520.
- Wälchli, B. & Cysouw, M. (2012). Lexical typology through similarity semantics: Toward a semantic map of motion verbs. *Linguistics* 50(3). 671–710.