UNIVERSITÄT LEIPZIG

erc European Research Council

Correspondence Analysis and type-based semantic maps

Natalia Levshina ©2017

Corpus Linguistics Summer School

Birmingham, July 2017

Outline

1. Correspondence Analysis: introduction

2. Simple Correspondence Analysis of verbs of speaking in COCA

3. Multiple Correspondence Analysis of analytic causatives in Germanic languages

4. MDS vs. MCA: When to use which method?

Introduction to CA

- CA is used to explore associations between the values of two and more categorical variables (usually represented as factors in R),
    - e.g. Do upper middle-class people prefer to play tennis and listen to opera?
    - Do languages with the Adj + N order also tend to have Num + N and Gen + N?
- Similar to MDS, CA allows to see structure in the data and identify which variables are associated and which of their values tend to co-occur.

The main idea behind CA

- CA is based on comparison of row profiles and column profiles, e.g.

	Birds	Music	Games	Total
M	20	30	50	100
F	10	70	20	100
Total	30	100	70	200

row profiles

	Birds	Music	Games	Total
M	0.2	0.3	0.5	1
F	0.1	0.7	0.2	1

column profiles

	Birds	Music	Games
M	0.67	0.3	0.71
F	0.33	0.7	0.29
Total	1	1	1

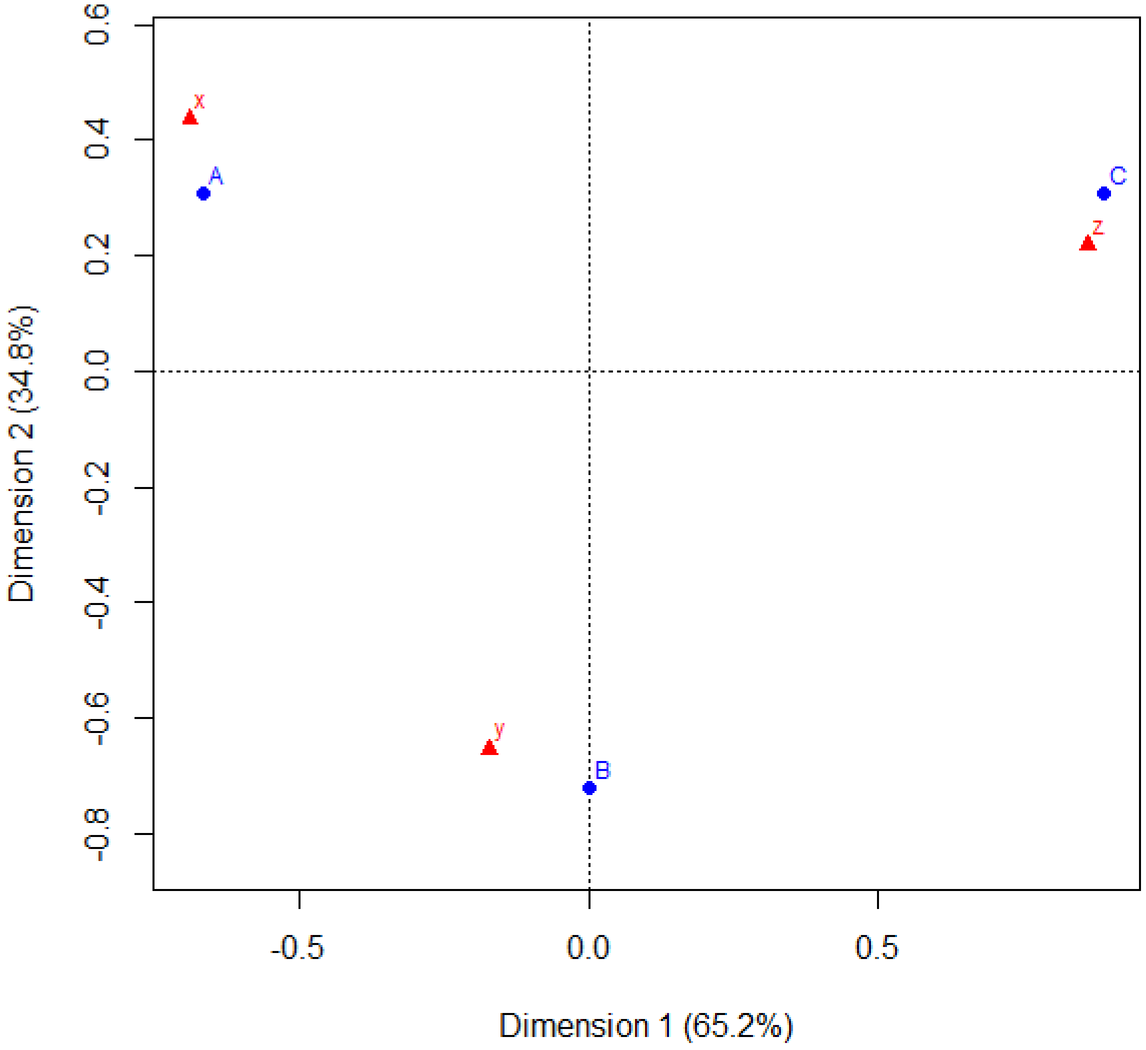
The main idea behind CA

- If two row or column profiles are similar, their labels will be closely located in a semantic map.
- If two row or column profiles are dissimilar, their labels will be located far from each other.

Strong association (all profiles are dissimilar)

	x	y	z
A	80	30	10
B	10	60	20
C	10	10	70

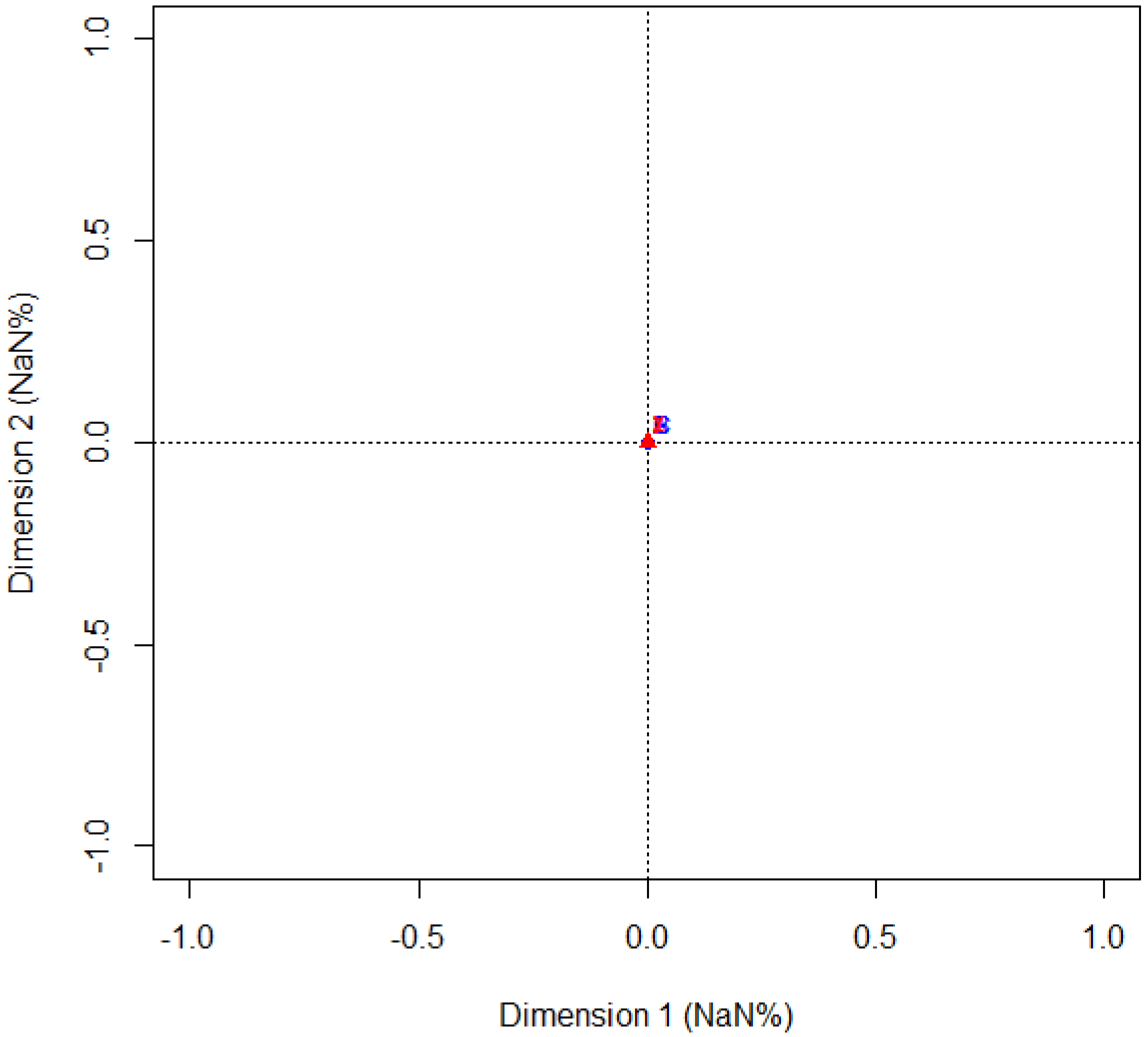
```
> chisq.test(example)

	Pearson's Chi-squared test

data:  example
X-squared = 191.67, df = 4, p-value < 2.2e-16
```


x
A
C
z
y
B
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
Dimension 2 (34.8%)
-0.5
0.0
0.5
Dimension 1 (65.2%)

Lack of association (all profiles are similar)

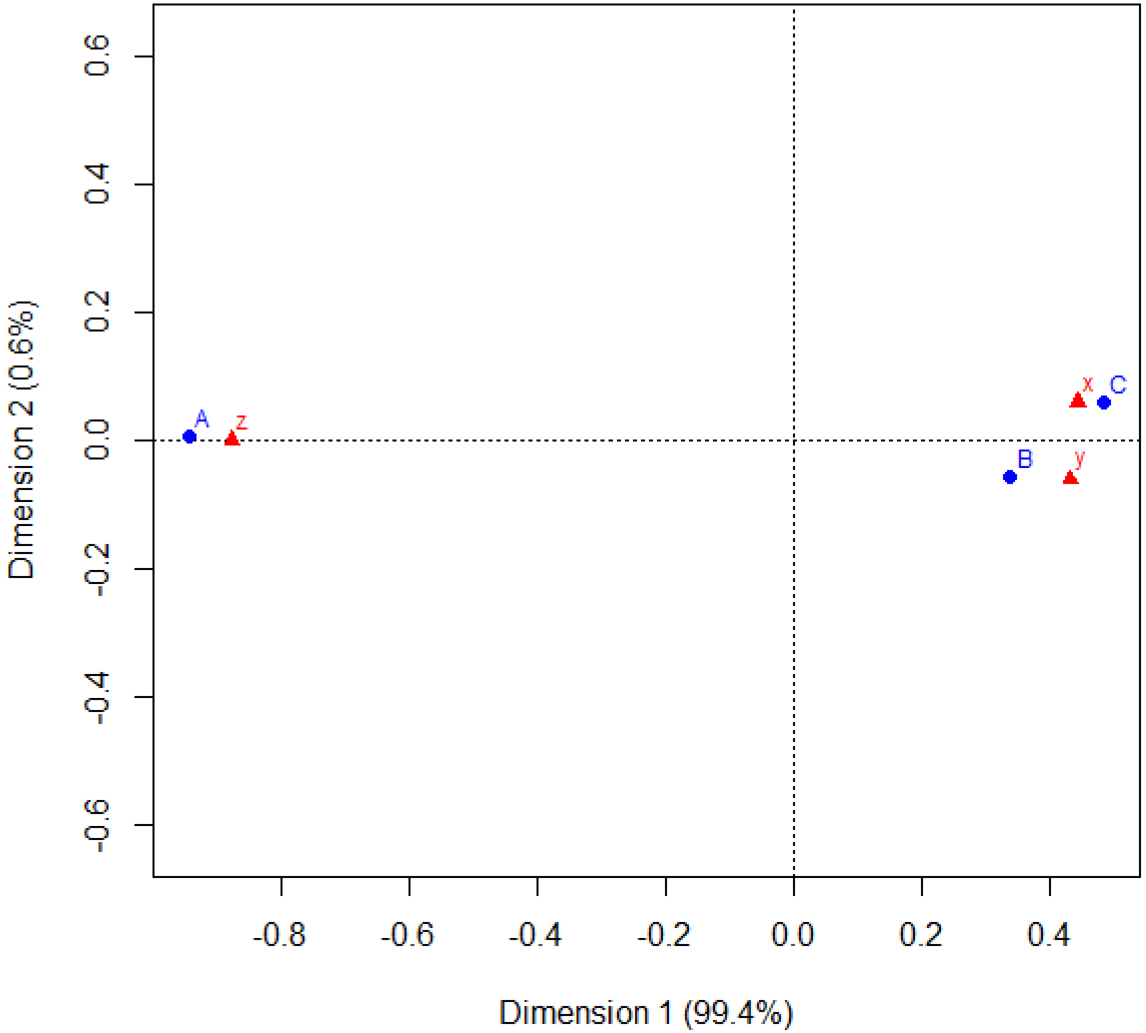
	x	y	z
A	10	10	10
B	80	80	80
C	10	10	10

```
> chisq.test(example1)

        Pearson's Chi-squared test

data:  example1
X-squared = 0, df = 4, p-value = 1
```


1.0
0.5
0.0
-0.5
-1.0
Dimension 2 (NaN%)
-1.0
-0.5
0.0
0.5
1.0
Dimension 1 (NaN%)

Some profiles are similar

	x	y	z
A	10	10	70
B	45	50	20
C	45	40	10


A
z
x
C
B
y
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-0.6
-0.4
-0.2
0.0
0.2
0.4
Dimension 1 (99.4%)
Dimension 2 (0.6%)

Simple and multiple CA

- If there are two variables, which are cross-tabulated, perform a simple CA on the table with counts.
- If there are more than two variables, perform a multiple CA on the data frame with variables as columns.

Outline

1. Correspondence Analysis: introduction

2. Simple Correspondence Analysis of verbs of speaking in COCA

3. Multiple Correspondence Analysis of analytic causatives in Germanic languages

4. MDS vs. MCA: When to use which method?

Verbs of communication

- announce
- assert
- babble
- blab
- chat
- chatter
- comment
- communicate
- converse
- declare
- discuss
- enunciate
- gab
- mumble
- murmur
- notify
- proclaim
- schmooze
- speak
- talk
- utter
- verbalize
- whisper
- yap

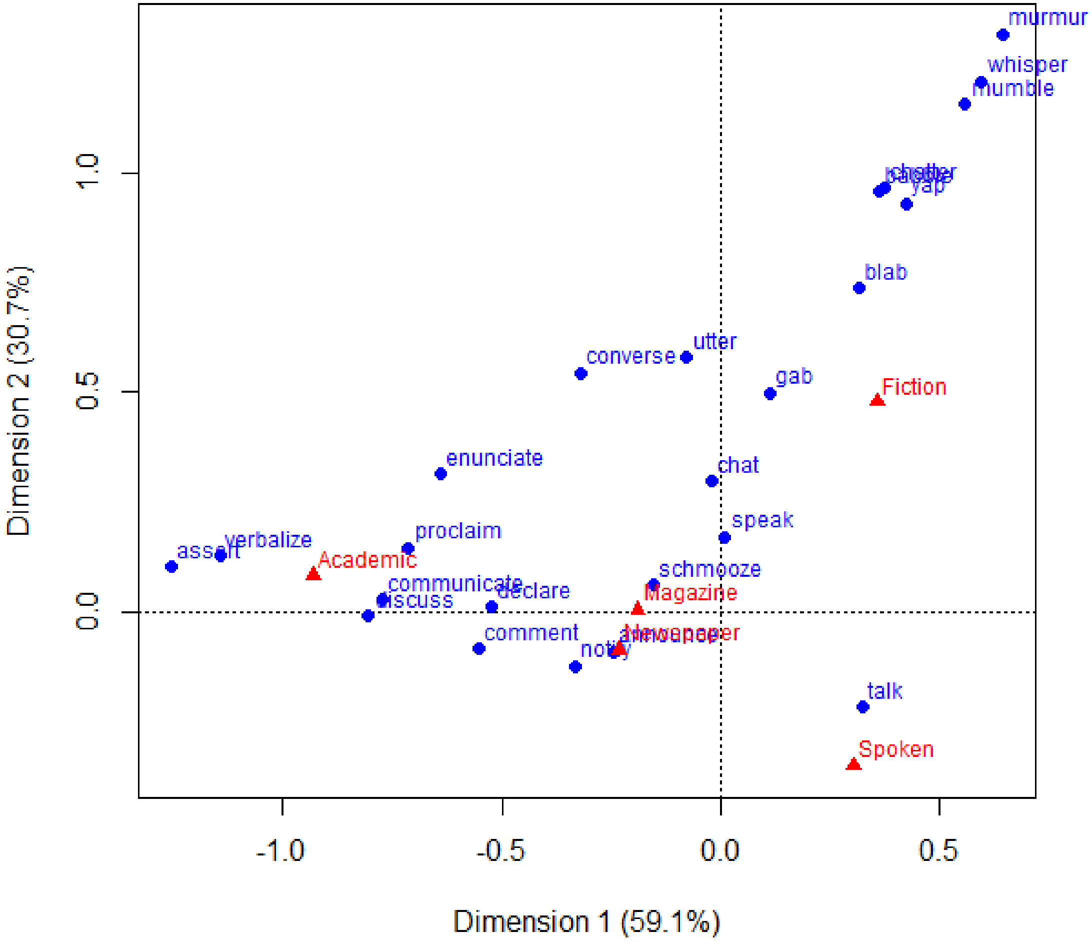
Cross-tabulated data with counts

```
> head(speak)
            Spoken Fiction Magazine Newspaper Academic
communicate   2327    1393     2664      1825     5147
chat           684    1672     1335      1155      354
declare       3449    2762     5335      5413     5167
utter          249    1336      595       397      484
whisper        465   13668     1445       779      273
assert         577     445     2259      1654     5784
```

Performing a simple CA

```
> library(ca)

> speak.ca <- ca(speak)

> plot(speak.ca) #the first two dimensions, by
default

> plot(speak.ca, dim = 2:3) #dimensions 2 and 3
```

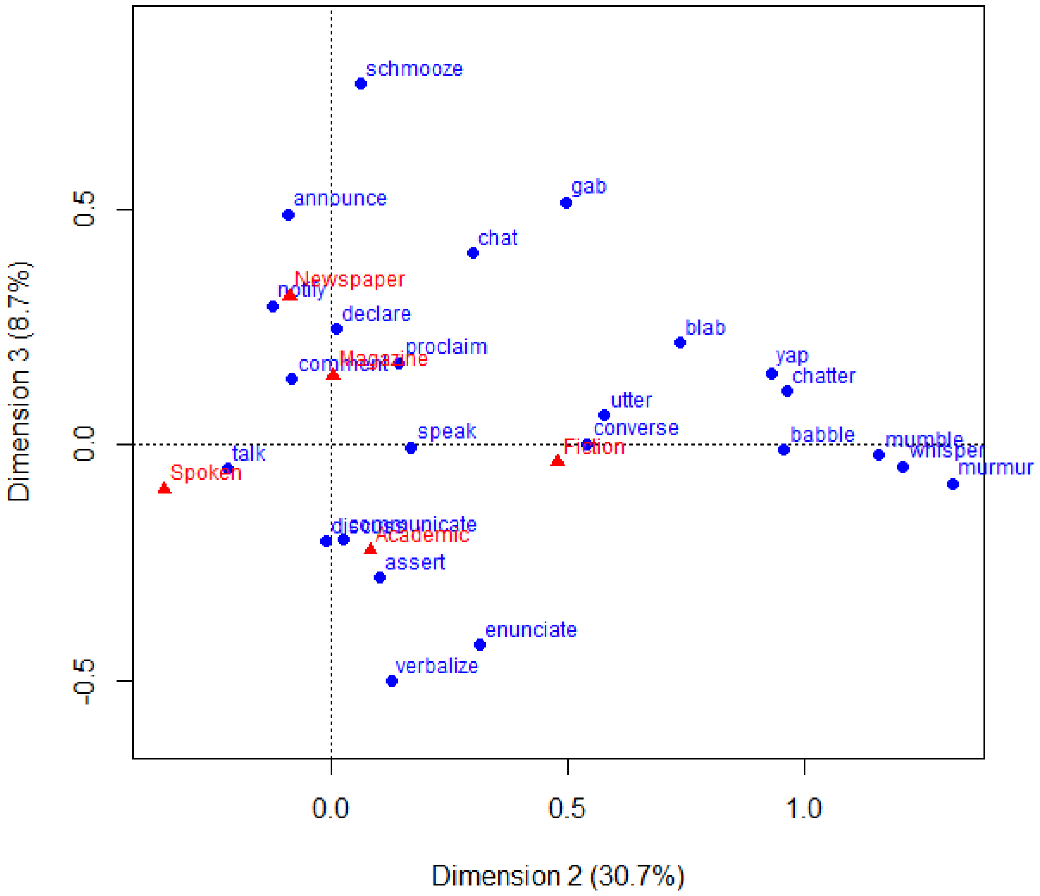
Dimensions 1 & 2



How to read a simple CA map

- If two row values are located close to each other, they have similar profiles.
  - Here, the rows are individual verbs. Their proximity means that their relative frequencies of occurrence in the subcorpora (registers) are similar.
- If two column values are close, this means that they have similar profiles, too.
  - Here, the columns are the subcorpora (registers). Their proximity means that they share similar proportions of the verbs.
- If the row and column labels are located in the same area regarding the origin, this means they co-occur frequently in the data. But the absolute distance should not be taken as a representation of association!

Dimensions 2 & 3



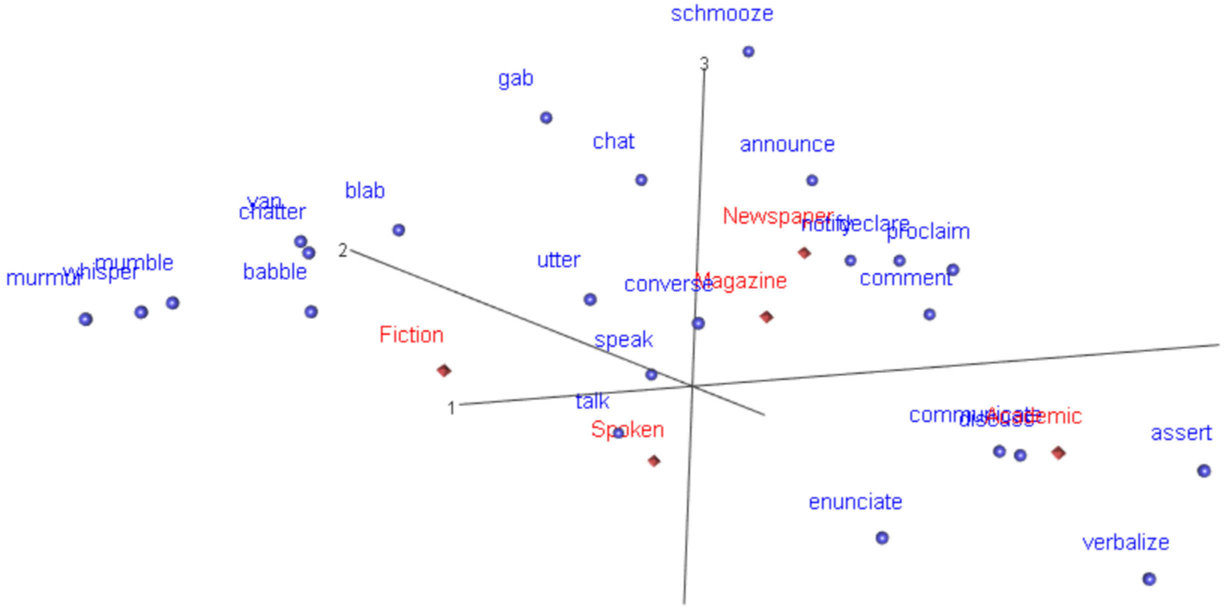
Creating an interactive 3D plot

- Important: you’ll need to install package rgl first!

```
> plot3d.ca(speak.ca)
```

- The plot is interactive. You can use your mouse or touchpad to rotate the axes and zoom in/out.

Interactive 3D plot



Quality of the 3D solution

- How much information do we lose if we take the three-dimensional solution?

```
> summary(speak.ca)

Principal inertias (eigenvalues):

 dim    value      %   cum%   scree plot
 1      0.190546  59.1  59.1  ***************
 2      0.099111  30.7  89.9  ********
 3      0.028158   8.7  98.6  **
 4      0.004538   1.4 100.0
        -------- -----      3 dimensions explain 98.6%!
 Total: 0.322352 100.0
```

Interpretation

- The spoken subcorpus is associated with the verb *talk*.
- The verbs of manner of saying (onomatopoeic) are associated with fiction, e.g. *murmur, whisper, chatter, babble*.
- Some Latinate verbs of argumentation and verbal expression (*discuss, assert, enunciate, verbalize*) are associated with the academic prose.
- Some neutral verbs of sharing information (*notify, announce, declare, comment*) are more associated with newspapers and magazines.

Outline

1. Correspondence Analysis: introduction

2. Simple Correspondence Analysis of verbs of speaking in COCA

3. Multiple Correspondence Analysis of analytic causatives in Germanic languages

4. MDS vs. MCA: When to use which method?

Prepare the data

```
> causatives_germ <- causatives[, c(1, 4, 5, 7, 8, 20)]

> causatives_germ$ENG <- as.factor(y.eng)
> causatives_germ$GER <- as.factor(y.ger)
> causatives_germ$DUT <- as.factor(y.dut)
> causatives_germ$SWE <- as.factor(y.swe)
> causatives_germ$NOR <- as.factor(y.nor)

> levels(causatives_germ$ENG) <- c("Other", "let")
> levels(causatives_germ$GER) <- c("Other", "lassen")
> levels(causatives_germ$DUT) <- c("Other", "laten")
> levels(causatives_germ$SWE) <- c("Other", "lata")
> levels(causatives_germ$NOR) <- c("Other", "la")
```

Assign “Other” to NA

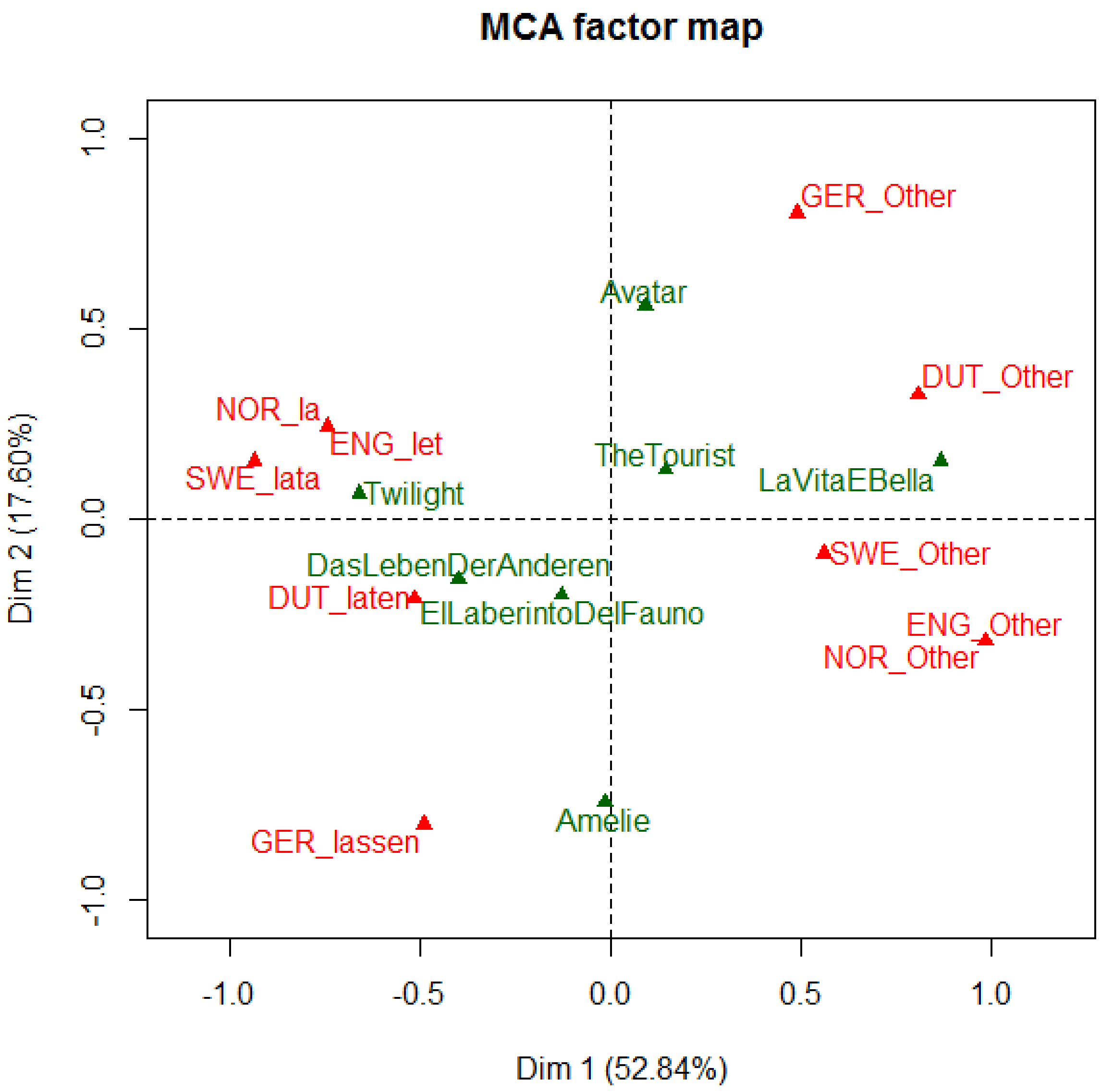
```
> causatives_germ$ENG[is.na(causatives_germ$ENG)]
<- "Other"

> causatives_germ$GER[is.na(causatives_germ$GER)]
<- "Other"

> causatives_germ$DUT[is.na(causatives_germ$DUT)]
<- "Other"

> causatives_germ$SWE[is.na(causatives_germ$SWE)]
<- "Other"

> causatives_germ$NOR[is.na(causatives_germ$NOR)]
<- "Other"
```

Run MCA and plot a map

```
> library(FactoMineR)

> causatives.mca <- MCA(causatives_germ, quali.sup =
1) #one supplementary qualitative variable (Film)
```

Supplementary variables (or points) do not influence the position of the main variables on the plot (i.e. the Germanic verbs). Supplementary points are plotted later depending on their co-occurrence with the values of the main variables. They usually provide us with additional information about the data.

```
> plot(causatives.mca)

> plot(causatives.mca, invisible = "ind") #make the
row names invisible
```


MCA factor map
GER_Other
Avatar
DUT_Other
NOR_la
ENG_let
TheTourist
SWE_lata
LaVitaEBella
Twilight
SWE_Other
DasLebenDerAnderen
DUT_laten
ElLaberintoDelFauno
ENG_Other
NOR_Other
Amelie
GER_lassen
1.0
0.5
0.0
-0.5
-1.0
-1.0
-0.5
0.0
0.5
1.0
Dim 1 (52.84%)
Dim 2 (17.60%)

How to read a multiple CA map

- If two labels are located closely, they tend to co-occur in the data.
  - Here: if two letting verbs are close, they are used frequently by the translators to convey the same causative situations.

Interpretation

- As on the MDS maps, German *lassen* and Dutch *laten* are different from the other letting verbs.
- The English and Scandinavian verbs are very close.
- As can be seen from the distribution of the supplementary points, *lassen* and *laten* occur particularly frequently in The Lives of Others and Pan’s Labyrinth, but infrequently in Life is Beautiful.

Checking the fit

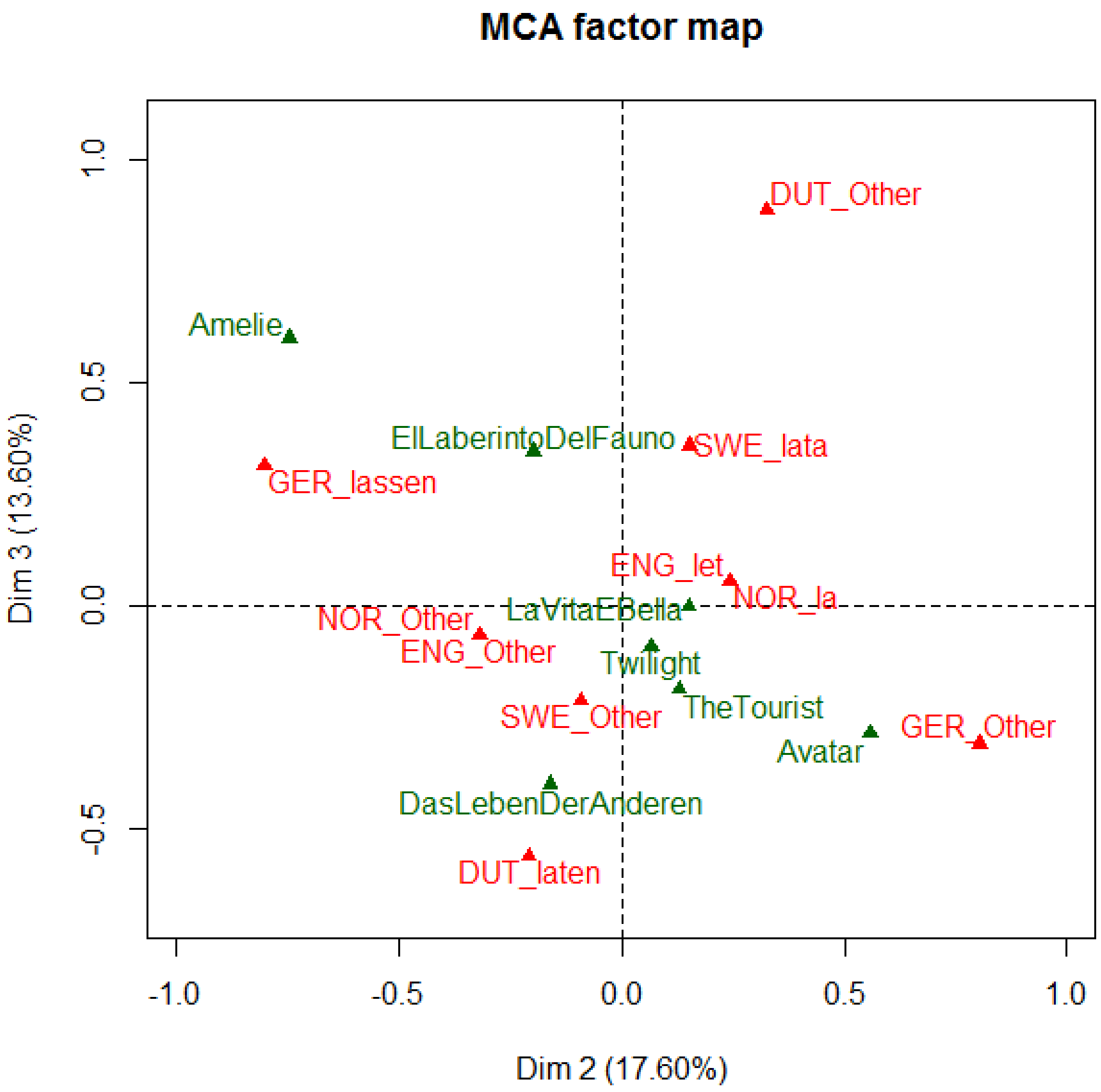
```
> summary(causatives.mca)
```

```
Call:
MCA(X = causatives_germ, quali.sup = 1)
Eigenvalues
                       Dim.1   Dim.2   Dim.3   Dim.4   Dim.5
Variance               0.528   0.176   0.136   0.114   0.045
% of var.             52.836  17.601  13.601  11.431   4.532
Cumulative % of var.  52.836  70.437  84.037  95.468 100.000
```

Plotting Dim 3

```
> plot(causatives.mca, axes = 2:3,  inv = "ind")
```

Unfortunately, not easy to interpret…


MCA factor map
DUT_Other
Amelie
ElLaberintoDelFauno
SWE_lata
GER_lassen
ENG_let
NOR_la
LaVitaEBella
NOR_Other
ENG_Other
Twilight
TheTourist
SWE_Other
GER_Other
Avatar
DasLebenDerAnderen
DUT_laten
Dim 3 (13.60%)
1.0
0.5
0.0
-0.5
-1.0
-0.5
0.0
0.5
1.0
Dim 2 (17.60%)

Outline

1. Correspondence Analysis: introduction

2. Simple Correspondence Analysis of verbs of speaking in COCA

3. Multiple Correspondence Analysis of analytic causatives in Germanic languages

4. MDS vs. MCA: When to use which method?

Some informal guidelines

- If you have only two categorical variables or a table with counts, then use simple CA.
- If you have many categorical variables:
  - If you have many missing values (NA), use MDS.
  - If you have many rare values with very small counts, use MDS.
  - If you want to represent the density of certain values (e.g. by Kriging), use MDS.
  - Else:
    - If you need the information about the variables and their levels, use MCA.
    - Else: feel free to choose!

Reference

- Greenacre, M. 2016. *Correspondence Analysis in Practice*. 3rd edn. Boca Raton, FL: CRC Press.