



zipfR

Baroni & Evert

zipfR

A guided tour

Playtime

Counting Words: The zipfR Toolkit

Marco Baroni & Stefan Evert

Málaga, 10 August 2006



Outline

zipfR

Baroni & Evert

zipfR

zipfR

A guided tour

A guided tour

Playtime

Playtime



zipfR

zipfR

Baroni & Evert

zipfR

A guided tour

Playtime

- ▶ <http://purl.org/stefan.evert/zipfR>
- ▶ <http://www.r-project.org/>



Outline

zipfR

Baroni & Evert

zipfR

zipfR

A guided tour

A guided tour

Playtime

Playtime



Loading

zipfR

Baroni & Evert

```
library(zipfR)
```

zipfR

A guided tour

```
?zipfR
```

Playtime

```
data(package="zipfR")
```



Importing data

zipfR

Baroni & Evert

zipfR

A guided tour

Playtime

```
data(ItaRi.spc)
```

```
data(ItaRi.emp.vgc)
```

```
my.spc <- read.spc("my.spc.txt")
```

```
my.vgc <- read.vgc("my.vgc.txt")
```

```
my.tfl <- read.tfl("my.tfl.txt")
```

```
my.spc <- tfl2spc(my.tfl)
```



Looking at spectra

zipfR

Baroni & Evert

zipfR

A guided tour

Playtime

```
summary(ItaRi.spc)
print(ItaRi.spc)

N(ItaRi.spc)
V(ItaRi.spc)
Vm(ItaRi.spc,1)
Vm(ItaRi.spc,1:5)

# Baayen's P
Vm(ItaRi.spc,1) / N(ItaRi.spc)

plot(ItaRi.spc)
plot(ItaRi.spc, log="x")
```



Looking at vgcs

zipfR

Baroni & Evert

zipfR

A guided tour

Playtime

```
summary(ItaRi.emp.vgc)
print(ItaRi.emp.vgc)
```

```
N(ItaRi.emp.vgc) # NB!
```

```
plot(ItaRi.emp.vgc, add.m=1)
```




Creating vgc's with binomial interpolation

zipfR

Baroni & Evert

```
# interpolated vgc
```

zipfR

A guided tour

```
ItaRi.bin.vgc <- vgc.interp(ItaRi.spc,  
N(ItaRi.emp.vgc), m.max=1)
```

Playtime

```
summary(ItaRi.bin.vgc)
```

```
# comparison
```

```
plot(ItaRi.emp.vgc, ItaRi.bin.vgc,  
legend=c("observed", "interpolated"))
```



Estimating LNRE models

zipfR

Baroni & Evert

```
# ZM model
```

zipfR

A guided tour

Playtime

```
ItaRi.zm <- lnre("zm", ItaRi.spc)  
summary(ItaRi.zm)
```

```
# ZM estimated fitting V and V_1 only
```

```
ItaRi.mmax1.zm <- lnre("zm", ItaRi.spc, m.max=1)  
summary(ItaRi.mmax1.zm)
```

```
# fZM model
```

```
ItaRi.fzm <- lnre("fzm", ItaRi.spc, exact=F) # NB!  
summary(ItaRi.fzm)
```



Observed/expected spectra at estimation size 1

zipfR

Baroni & Evert

```
# expected spectra
```

zipfR

A guided tour

```
ItaRi.zm.spc <- lnre.spc(ItaRi.zm, N(ItaRi.zm))
```

Playtime

```
ItaRi.mmax1.zm.spc <- lnre.spc(ItaRi.mmax1.zm,  
N(ItaRi.mmax1.zm))
```

```
ItaRi.fzm.spc <- lnre.spc(ItaRi.fzm, N(ItaRi.fzm))
```



Observed/expected spectra at estimation size 2

zipfR

Baroni & Evert

```
# compare
```

zipfR

A guided tour

Playtime

```
plot(ItaRi.spc, ItaRi.zm.spc,  
ItaRi.mmax1.zm.spc, ItaRi.fzm.spc,  
legend=c("observed", "zm", "zm1", "fzm"))
```

```
# plot first 10 elements only
```

```
plot(ItaRi.spc, ItaRi.zm.spc, ItaRi.mmax1.zm.spc,  
ItaRi.fzm.spc, legend=c("observed", "zm", "zm1", "fzm")  
m.max=10)
```



Expected spectra at 10 times the estimation size

zipfR

Baroni & Evert

zipfR

A guided tour

Playtime

```
# extrapolated spectra
```

```
ItaRi.zm.spc <- lnre.spc(ItaRi.zm, 10*N(ItaRi.zm))
```

```
ItaRi.fzm.spc <- lnre.spc(ItaRi.fzm,  
10*N(ItaRi.fzm))
```

```
# compare
```

```
plot(ItaRi.zm.spc, ItaRi.fzm.spc,  
legend=c("zm", "fzm"))
```



Evaluating extrapolation quality 1

zipfR

Baroni & Evert

zipfR

A guided tour

Playtime

```
# taking a subsample and estimating a model (if you  
# repeat you'll get different sample and different  
# model!)
```

```
ItaRi.sub.spc <- sample.spc(ItaRi.spc, N=700000)
```

```
ItaRi.sub.fzm <- lnre("fzm", ItaRi.sub.spc,  
exact=F)
```

```
ItaRi.sub.fzm
```



Evaluating extrapolation quality 2

zipfR

Baroni & Evert

```
# extrapolate vgc up to original sample size
```

zipfR

A guided tour

Playtime

```
ItaRi.sub.fzm.vgc <- lnre.vgc(ItaRi.sub.fzm,  
N(ItaRi.emp.vgc))
```

```
# compare
```

```
plot(ItaRi.bin.vgc, ItaRi.sub.fzm.vgc,  
NO=N(ItaRi.sub.fzm), legend=c("interpolated", "fZM"))
```



Compare growth of two categories 1

zipfR

Baroni & Evert

zipfR

A guided tour

Playtime

```
# the ultra- prefix
```

```
data(ItaUltra.spc)
```

```
summary(ItaUltra.spc)
```

```
# cf.
```

```
summary(ItaRi.spc)
```

```
# estimating model
```

```
ItaUltra.fzm <- lnre("fzm",ItaUltra.spc,exact=F)
```

```
ItaUltra.fzm
```




Compare growth of two categories 2

zipfR

Baroni & Evert

```
# extrapolation of V to ri- sample size
```

zipfR

A guided tour

```
ItaUltra.ext.vgc <- lnre.vgc(ItaUltra.fzm,  
N(ItaRi.emp.vgc))
```

Playtime

```
# compare
```

```
plot(ItaUltra.ext.vgc, ItaRi.bin.vgc,  
NO=N(ItaUltra.fzm), legend=c("ultra-","ri-"))
```

```
# zooming in
```

```
plot(ItaUltra.ext.vgc, ItaRi.bin.vgc,  
NO=N(ItaUltra.fzm), legend=c("ultra-","ri-"),  
xlim=c(0,1e+5))
```



Outline

zipfR

Baroni & Evert

zipfR

zipfR

A guided tour

A guided tour

Playtime

Playtime



Now, try it yourself

zipfR

Baroni & Evert

zipfR

A guided tour

Playtime

- ▶ Pick comparable datasets
- ▶ Explore spc, empirical vgc, interpolated vgc
- ▶ Compute LNRE model(s)
- ▶ Compare vgc and spectra of classes at different sample sizes



Data

zipfR

Baroni & Evert

zipfR

A guided tour

Playtime

- ▶ `data(package="zipfR")`
- ▶ E.g.:
 - ▶ Brown adjectives vs. verbs
 - ▶ Tiger NP vs. PP rules
 - ▶ Great Expectations vs. Oliver Twist
 - ▶ ...
- ▶ Or import your own frequency lists



Explore

zipfR

Baroni & Evert

zipfR

A guided tour

Playtime

- ▶ Remember: ?zipfR
- ▶ Summaries, spectrum plots
- ▶ Empirical and interpolated vgcs
- ▶ Plot vgcs of two classes together



LNRE modeling

zipfR

Baroni & Evert

zipfR

A guided tour

Playtime

- ▶ Try more than one model
- ▶ Play with `exact` and `m.max` arguments
- ▶ Look at goodness of fit, expected V and V_m
- ▶ Comparative spc plots at estimation size and larger sizes



Class comparison

zipfR

Baroni & Evert

zipfR

A guided tour

Playtime

- ▶ Extrapolate class with shorter sample
- ▶ Extrapolate both classes to very large sample size
- ▶ Look at spectra for matching sample sizes



Already done?

zipfR

Baroni & Evert

Try Case Study 2 from the tutorial (or go to get some lunch!)

zipfR

A guided tour

Playtime