Databases and corpora 3

Quantitative Methods in Historical Linguistics
Dr. Henri Kauhanen / University of Konstanz

27 June 2018

First things first

- Questions?
- Workshop timing — is it all right now?
- Please download the following dataset from ILIAS:
  - ‘Data‘ folder > `german_fortition.csv`
- And the following piece of code:
  - ‘Code‘ folder > `cretest.R`

From last time: *Do*-support

```
elleg_full <- read.csv("ellegard_full.csv")
```

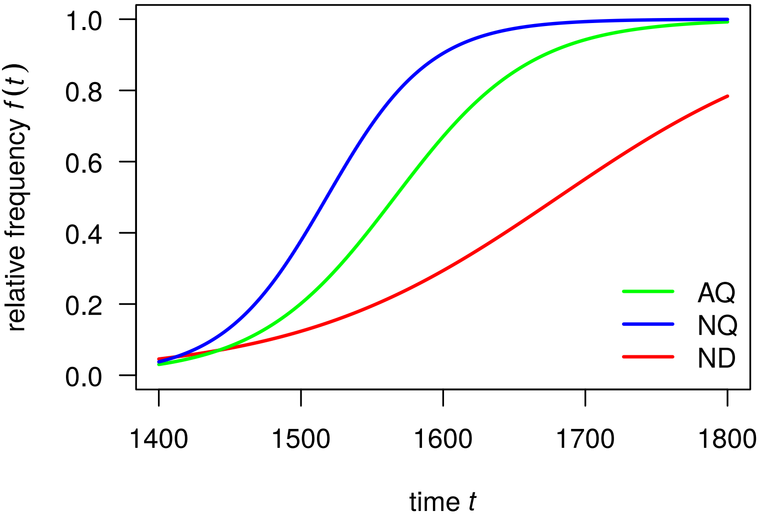
- Dataset has columns of the form `X_do` and `X_freq`:

`X`	**context**
AQ	affirmative questions
NQ	negative questions
ND	negative declaratives

Exercise

Use the Gauss–Newton algorithm (`nls`) to find the best-fitting logistic curves (s and k parameters) for each context.

X	**context**	*s*	*k*
AQ	affirmative questions	0.021	1565.967
NQ	negative questions	0.027	1518.13
ND	negative declaratives	0.011	1680.91



Final fortition in Early New High German

- In (Northern) German, a **FINAL FORTITION RULE** applies:

Final fortition

A voiced obstruent becomes voiceless if it appears at the end of a syllable.

- E.g. /taːg/ > [taːk] (but /taː.ge/ > [taː.gə])
- The rule was lost in a number of (mainly Southern) dialects around 1400
- Let’s now take a look at how this happened

Grammar competition

- Grammar: the abstract representation of (in this case, phonological) knowledge of a speaker
- Grammar G_1 has the final fortition rule
- Grammar G_2 is identical to G_1 except that it doesn't have the final fortition rule
- We have change $G_1 > G_2$
- During this period of change, the two grammars **COMPETE**, each being used with some probability
- At the end of the change, $\text{Prob}(G_1) = 0$ and $\text{Prob}(G_2) = 1$
- The grammar probabilities are reflected in corpus data as the relative frequencies of final fortition (G_1) and no final fortition (G_2)

Glaser's data

- Elvira Glaser[1] provides the following data for the stops /b,d,g/ based on an analysis of spelling:

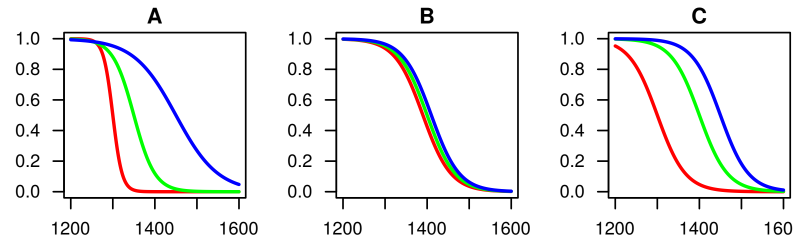
	/b/		/d/		/g/	
year	[p]	[b]	[t]	[d]	[k]	[g]
1276	18	0	29	0	54	19
1373	10	8	24	5	17	59
1483	2	16	2	22	0	78
1523	2	14	3	6	0	73

- We have three contexts, the phonemes /b/, /d/ and /g/

[1]Glaser, E. 1985. *Graphische Studien zum Schreibsprachwandel vom 13. bis 16. Jahrhundert*. Heidelberg: Carl Winter Universitätsverlag.

Possible shapes of the competition

- How exactly is final fortition lost?
- Three possibilities:

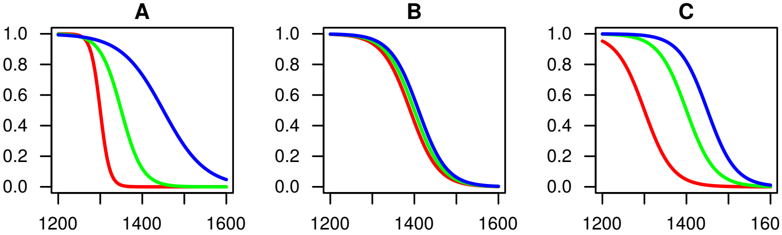


A The contexts change independently (different *s*, different *k*)

B The contexts change together (same *s*, same *k*)

C The contexts change at slightly different times but at the same rate (same *s*, different *k*)

Possible shapes of the competition



A Competition is at the level of phonemes: individual competitions in /b/, /d/ and /g/

- Or possibly even at the level of individual words, and just reflected in Glaser's data at the phoneme level

B Competition is at the level of the entire grammar: /b/, /d/ and /g/ change "in sync"

C Competition is at the level of the entire grammar, but some external factors cause a time difference between the contexts

Fitting a logistic model to Glaser's data

- Let's find out!
- I.e. let's fit a logistic curve to each context and see what the result looks like

Exercise

1. Download `german_fortition.csv` and load it into R
2. Make columns that give the relative frequency of fortition in each context
3. Use `nls` to fit a logistic curve to each context separately
4. Make note of the s and k parameters found by `nls` for each context

Adding relative frequency columns

```
gf <- read.csv("german_fortition.csv")
```

```
gf$p_freq <- gf$p/(gf$p + gf$b)
gf$t_freq <- gf$t/(gf$t + gf$d)
gf$k_freq <- gf$k/(gf$k + gf$g)
```

```
gf

##   date  p  b  t  d  k  g    p_freq     t_freq    k_freq
## 1 1276 18  0 29  0 54 19 1.0000000 1.00000000 0.7397260
## 2 1373 10  8 24  5 17 59 0.5555556 0.82758621 0.2236842
## 3 1483  2 16  2 22  0 78 0.1111111 0.08333333 0.0000000
## 4 1523  2 14  3  6  0 73 0.1250000 0.33333333 0.0000000
```

Fitting the curves

```r
p_model <- nls(p_freq~1/(1 + exp(s*(k-date))), gf,
               start=list(s=-0.01, k=1400))

t_model <- nls(t_freq~1/(1 + exp(s*(k-date))), gf,
               start=list(s=-0.01, k=1400))

k_model <- nls(k_freq~1/(1 + exp(s*(k-date))), gf,
               start=list(s=-0.01, k=1400))
```

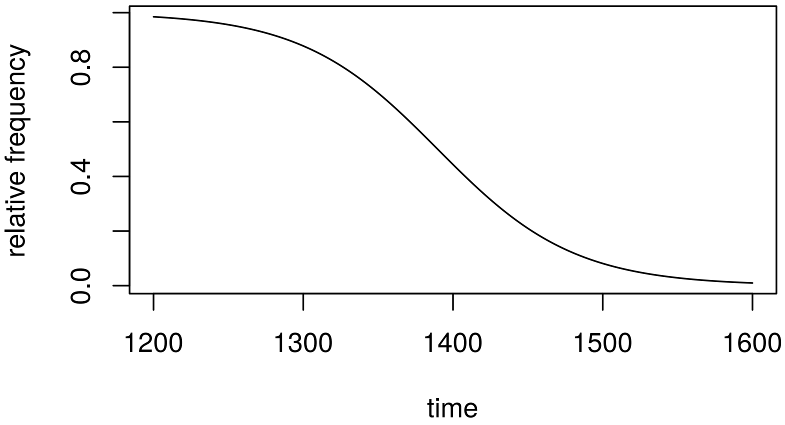
Examining the s and k parameters

```
coef(p_model)

##               s             k
##     -0.02204335 1389.71780632

coef(t_model)

##               s             k
##     -0.02223679 1432.22617617

coef(k_model)

##               s             k
##     -0.02406459 1319.97726598
```

⇒ Does this represent scenario **A**, **B** or **C**?

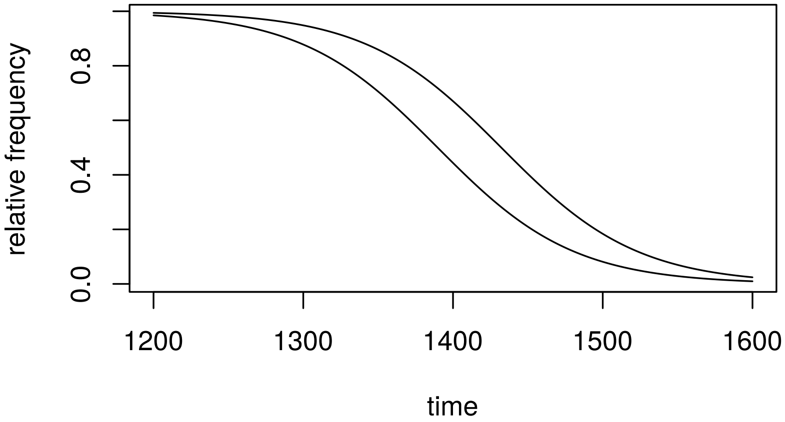
Plotting the curves

```
t <- seq(from=1200, to=1600, length.out=1000)
plot(t, 1/(1 + exp(-0.022*(1389.718 - t))), type="l",
     xlab="time", ylab="relative frequency")
```



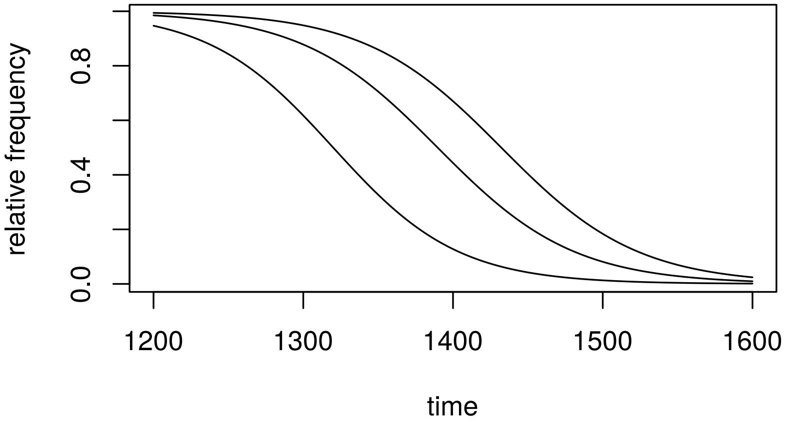
Plotting the curves

```
points(t, 1/(1 + exp(-0.022*(1432.226 - t))), type="l")
```

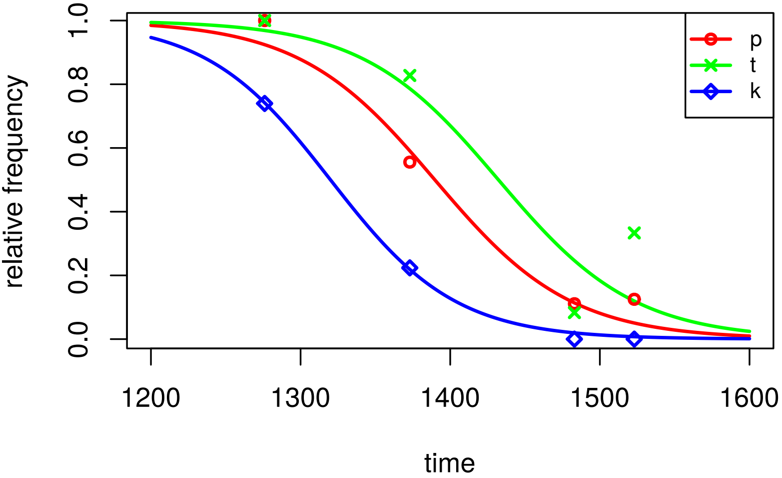


Plotting the curves

```
points(t, 1/(1 + exp(-0.024*(1319.977 - t))), type="l")
```

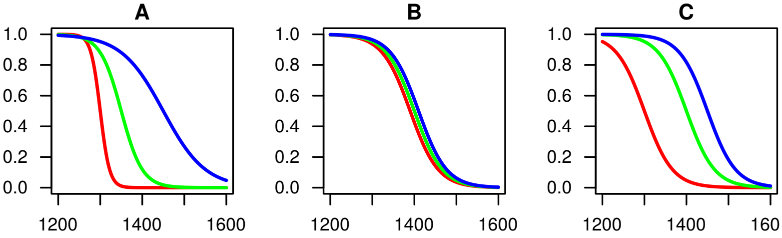


Plotting the curves



⇒ This looks like scenario **C** (same *s*, different *k*).

Constant Rate Effect (CRE)



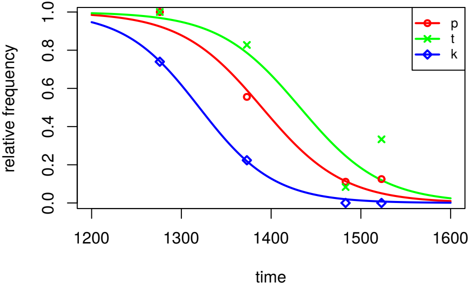
- Scenario **C** is known as a **Constant Rate Effect** (CRE)
  - the rate (*s*) is constant across contexts
  - but the value of *k* may be different
- First identified by Anthony Kroch in the 1980s[2]
- In our present case study, the observation of a CRE means that
  - the competition is at the level of the entire grammar
  - and not at the level of phonemes or words

---

[2] Kroch, A. S. (1989). Reflexes of grammar in patterns of language change. *Language Variation and Change*, 1, 199–244.

Constant Rate Effect (CRE)

- However, something causes a difference in the probability of final fortition between the phonemes **DURING** the change (but not after it)
  - /g/ has the least fortition, /d/ the most
- These effects are external to the grammatical competition itself and could arise from different sources
  - articulatory/perceptual facts
  - sociolinguistic facts



Constant Rate Effect (CRE)

- **But**, how do we know it is really a CRE?
- The rates could be the same just by chance…
- …especially as the database is very small (the smaller your sample size, the less reliable your statistics!)
- Techniques have been developed to answer this question
- We will look at just one of them (and only superficially)

Testing for a CRE

- The idea: fit a competing model to the data that **FORCES** the *s* parameters to be the same (call this the **CRE MODEL**)
- In the original model (call it the **ALTERNATIVE MODEL**) both *s* and *k* are free to vary across contexts
- If alternative model does **NOT** fit the data any better than CRE model, we diagnose a CRE
- If alternative model **DOES** fit the data better than CRE model, we conclude there is no CRE
- (You don't need to know the details, but technically this statistical test is known as the **LIKELIHOOD RATIO TEST** if you want to google it up)

Testing for a CRE

- We will use `cretest.R` for this:

```
source("cretest.R")
```

- The first thing we need to do is to put our context curves in a list:

```
alt_model <- list(p_model, t_model, k_model)
```

- Next, we take the average *s* from the three models:

```
avg_s <- mean(c(-0.022, -0.022, -0.024))
avg_s
```

```
## [1] -0.02266667
```

Testing for a CRE

- Next we make the CRE model for each context using `avg_s` as the value of *s*:

```
p_CRE <- nls(p_freq~1/(1 + exp(avg_s*(k-date))),
             gf, start=list(k=1400))

t_CRE <- nls(t_freq~1/(1 + exp(avg_s*(k-date))),
             gf, start=list(k=1400))

k_CRE <- nls(k_freq~1/(1 + exp(avg_s*(k-date))),
             gf, start=list(k=1400))
```

- Put these in a list, too:

```
CRE_model <- list(p_CRE, t_CRE, k_CRE)
```

Testing for a CRE

- Finally, run the test:

```
cretest(alt_model, CRE_model)

## Likelihood ratio test
##
##      L-ratio: 0.032
##   chi-square: 0.064
##           df: 3
##      p-value: 0.996
```

- (NB You **MUST** give the arguments in this order!)
- The important thing for us is the `p-value`
- This is the probability of observing the kind of variation in s that we see in the data, **IF** the CRE model is true

Testing for a CRE

- In this case, the p-value is high (0.996), so we believe in the CRE model
- In other words, there is no reason to assume that the value of *s* changes from context to context
- **But!** This is not to say that we have **proved** that we have a CRE
- Technically, we have only **failed to reject** the hypothesis of a CRE
- (If the p-value was very small (close to zero), we would reject the CRE hypothesis and conclude that *s* varies between the contexts)

Credits

- The Early New High German fortition case study is originally from:
    - Fruehwald, Josef, Gress-Wright, Jonathan & Wallenberg, Joel C. (2013). Phonological rule change: the constant rate effect. In S. Kan, C. Moore-Cantwell & R. Staubs (Eds.), *NELS 40: Proceedings of the 40th Annual Meeting of the North East Linguistic Society* (pp. 219–230). GLSA Publications.

  who use Glaser's data but also explore other datasets

Lastly...

- Any questions?
- A new Portfolio Exercise on ILIAS (on S-curves and CREs)
- Next week, we will move on to the third part of the course