A Corpus Analysis of Frequency Effects on Eye-Movements in Sentence Context

Philip Dilts[1] Gary Libben[1,2] Harald Baayen[1]

1-University of Alberta

2-University of Calgary

American Association for Corpus Linguistics Conference, 2009

1. The Story
   - Do frequency effects disappear in sentence context?
2. The Data
   - The Dundee Corpus
   - Types of Predictor Variables
   - Exploring the Predictors
3. Linear Regression Modelling
   - A Large Model
   - A Useful Model
   - A Better Model
4. Generalized Additive Models
   - Why GAMs
   - The Best Model
5. Conclusions

The Story

The Story

Do frequency effects disappear in sentence context?

- We know a lot about how people process words presented one at a time, but what happens in context?

The Story

Do frequency effects disappear in sentence context?

- We know a lot about how people process words presented one at a time, but what happens in context?
- Van Petten and Kutas (1990) looked at the effects of word frequency on the N400 event-related brain potentials
  - Word frequency had no effect later in the sentence.

The Story

Do frequency effects disappear in sentence context?

- We know a lot about how people process words presented one at a time, but what happens in context?
- Van Petten and Kutas (1990) looked at the effects of word frequency on the N400 event-related brain potentials
    - Word frequency had no effect later in the sentence.
- That's crazy!

The Story

Do frequency effects disappear in sentence context?

- We know a lot about how people process words presented one at a time, but what happens in context?
- Van Petten and Kutas (1990) looked at the effects of word frequency on the N400 event-related brain potentials
  - Word frequency had no effect later in the sentence.
- That's crazy!
- The frequency effect is everywhere!

The Story

Do frequency effects disappear in sentence context?

- We know a lot about how people process words presented one at a time, but what happens in context?
- Van Petten and Kutas (1990) looked at the effects of word frequency on the N400 event-related brain potentials
  - Word frequency had no effect later in the sentence.
- That's crazy!
- The frequency effect is everywhere!
- Maybe later in a sentence, context is more important than word properties?

The Story

Do frequency effects disappear in sentence context?

- We know a lot about how people process words presented one at a time, but what happens in context?
- Van Petten and Kutas (1990) looked at the effects of word frequency on the N400 event-related brain potentials
  - Word frequency had no effect later in the sentence.
- That's crazy!
- The frequency effect is everywhere!
- Maybe later in a sentence, context is more important than word properties?
- Can we test this using a corpus of eye-tracking?

The Story

Do frequency effects disappear in sentence context?

- Eye movement is affected by frequency
  - Readers spend longer looking at less frequent words
- Does this effect disappear at the end of a sentence?
- We can use the Dundee corpus to investigate
- We make and compare statistical models:
  - Trying to predict how long people first look at each word (first fixation duration)
  - Using models of two different types (Linear Models and GAMs)
- Does adding a frequency-by-sentence-position interaction help us predict this duration?

The Story

Do frequency effects disappear in sentence context?

- Eye movement is affected by frequency
  - Readers spend longer looking at less frequent words
- Does this effect disappear at the end of a sentence?
- We can use the Dundee corpus to investigate
- We make and compare statistical models:
  - Trying to predict how long people first look at each word (first fixation duration)
  - Using models of two different types (Linear Models and GAMs)
- Does adding a frequency-by-sentence-position interaction help us predict this duration?
- In short, No. Frequency effects don't go away or get smaller.

The Data

The Data
The Dundee Corpus

- The Dundee Corpus of Eye-movement data
  Kennedy et al. (2003)
- 10 native English speakers in Dundee, Scotland
- Each read 20 editorials from *The Independent* newspaper
- Editorials were presented on 40 screens of 5 lines each
- Some sentences were split across two screens

The Data
Types of Variables

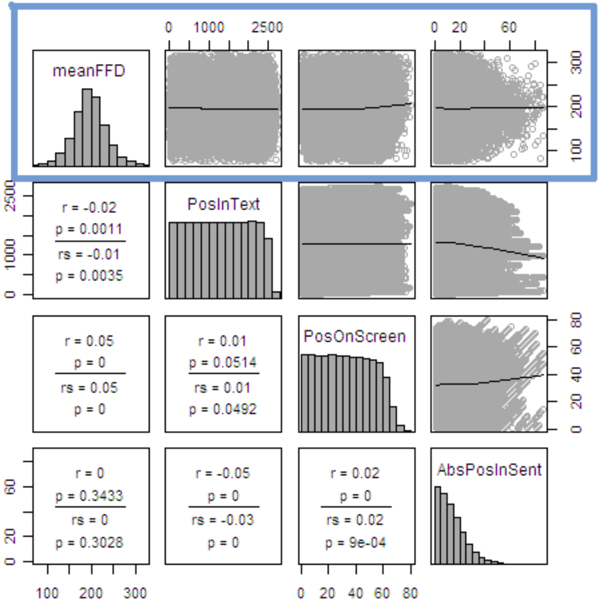
- 10 predictors considered
  - 4 measures of word properties:
    - CobLog Log-transformed CELEX word-form frequency
    - WLen Length of the current word in characters
    - PostMarks Punctuation marks following the word
    - PreMarks Punctuation marks preceding the word
  - 6 measures of context and position:
    - PosOnLineInChars How far right on a line the word is
    - PosInText How far along in the text the word is
    - PosOnScreen How far along in the screen the word is
    - AbsPosInSent How far along in the sentence the word is
    - SentLength Length of the current sentence
    - SentAllOneScreen Is sentence all on one screen?
  - Collinearity is high (condition number 23) but not fatally high
- Dependent variable: First Fixation Duration

The Data

Exploring the Predictors

- Look at each predictor's relationships with fixation duration
- For example....

Predictor Variables
Exploring the Predictors



Predictor Variables

Exploring the Predictors

- The average line for PosOnScreen looked curvy
  - (PosOnLineInChars, WLen and CobLog do too)
  - We'll model those 4 predictors with cubic splines
- We also consider every possible 2-way interaction between predictors
  - Our main question is about an interaction! (Sentence position x Frequency)

Linear Regression Modelling

Linear Regression Modelling
A Large Model

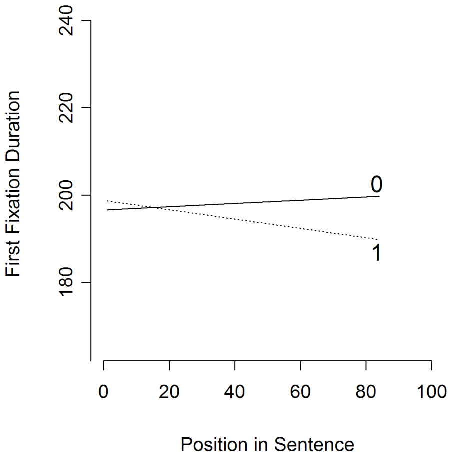
- Linear modelling using ordinary least squares regression (`ols` in R)
- Large initial model:
  - All 10 predictors, 6 straight and 4 curvy
  - 39 2-way interactions
    (interactions between curvy predictors are too hard)
- Use fast backwards elimination (`fastbw`) to cut out non-significant or inefficient predictors

Linear Regression Modelling

A Useful Model

- New model with only the factors remaining after `fastbw`
- 10 of the original 49 predictors remain:
- 6 single predictors, 4 interactions
- Our variables of interest:
  - Frequency is a significant predictor ($p < .0001$)
  - Position in sentence is NOT a significant predictor! (p appx. 0.2)
  - But, the sentence-position by sentence-all-on-one-screen interaction IS significant ($p < .0001$)
  - Here's what the interaction looks like:

Position-by-Split-screen interaction



0: Two screens
1: One screen

Linear Regression Modelling

A Better Model

- Let's cut out split-screen sentences and words with punctuation
  - We shouldn't expect them to be typical
  - They participate in all interactions in the model!
- Let's use at a better modelling tool, too...

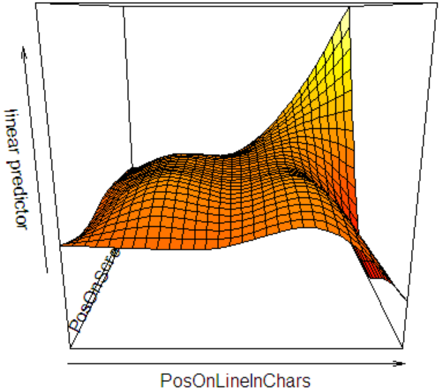
Generalized Additive Models

Generalized Additive Models

Why GAMs?

- Two advantages of Generalized Additive Models:
  - They choose how curvy the predictors are for you
    - Arbitrary "Smooth functions" of the dependent variables instead of splines
  - They can look at interactions between two curvy predictors
    - For example...

Generalized Additive Models
Why GAMs?



Generalized Additive Models
The Best Model

Predictor	Estimated degrees of freedom	F	p
s(PosInText)	3.1	3.357	0.01
s(WLen)	5.4	6.397	1.15×10^{-6}
s(CobLog)	8.6	45.454	$< 2.0 \times 10^{-16}$
s(AbsPosInSent)	4.5	8.742	2.44×810^{-6}
te(PosOnLineInChars,PosOnScreen)	21.2	85.529	$< 2.0 \times 10^{-16}$
	Total R-squared for this model:		9.98%

- (`te(,)` is a way of investigating interactions)
- Frequency (CobLog) and sentence position are very significant!
- What if we add their interaction?

Generalized Additive Models

The Target Model

Predictor	Estimated degrees of freedom	F	p
s(PosInText)	3.1	3.331	0.01
s(WLen)	5.4	5.984	3.35×10^{-6}
te(AbsPosInSent, CobLog)	7.6	52.497	$< 2.0 \times 10^{-16}$
te(PosOnLineInChars, PosOnScreen)	21.2	85.240	$< 2.0 \times 10^{-16}$
	Total R-squared for this model:		9.95%

- The interactive term is significant
- ... but it replaced two significant factors
- Does the interaction make the MODEL better?

No!

Predictor	Estimated degrees of freedom	F	p
s(PosInText)	3.1	3.357	0.01
s(WLen)	5.4	6.397	1.15×10^{-6}
s(CobLog)	8.6	45.454	$< 2.0 \times 10^{-16}$
s(AbsPosInSent)	4.5	8.742	2.44×810^{-6}
te(PosOnLineInChars,PosOnScreen)	21.2	85.529	$< 2.0 \times 10^{-16}$
Total R-squared for this model:			9.98%

Predictor	Estimated degrees of freedom	F	p
s(PosInText)	3.1	3.331	0.01
s(WLen)	5.4	5.984	3.35×10^{-6}
te(AbsPosInSent, CobLog)	7.6	52.497	$< 2.0 \times 10^{-16}$
te(PosOnLineInChars, PosOnScreen)	21.2	85.240	$< 2.0 \times 10^{-16}$
Total R-squared for this model:			9.95%

Generalized Additive Models

Comparing Models

- The higher R-squared means more predictive power
- The model WITH the interaction has LESS power
- The AICs tells us that it is slightly less EFFECIENT, too:

```
with interaction:     AIC = 335,644
without interaction:  AIC = 335,640
```

Conclusions

Conclusions

- Frequency effects on eye movement DO NOT go away later in a sentence
  - Linear regression model shows no significant interaction
  - Generalized additive models get worse if we add an interaction
- Context plays a role in word reading, but not the only role

Selected References

- A Few References
  - ERP Study:
    Van Petten, C. and Kutas, M. (1990). Interactions between sentence context and word frequency in event-related brain potentials. Memory and Cognition 18(4), 380-393.
  - Some frequency effects:
    J Bybee, P Hopper - Frequency and the emergence of linguistic structure, 2001.
  - Dundee Corpus:
    Kennedy, A., Hill, R., and Pynte, J. (2003). The Dundee corpus. Poster presented at ECEM12: 12th European Conference on eye movements., Dundee, August 2003.
  - Linear Modelling of Linguistic Data:
    Baayen, R. H. (2007). Analyzing Linguistic Data: A practical introduction to statistics using R. Cambridge: Cambridge University Press.
  - Generalized Additive Modelling:
    Wood, S. N. (2006). Generalized additive models: An introduction with R. Chapman and Hall CRC Press.

Conclusions

- Frequency effects on eye movement DO NOT go away later in a sentence
  - Linear regression model shows no significant interaction
  - Generalized additive models get worse if we add an interaction
- Context plays a role in word reading, but not the only role

Factors in useful linear model

Factor	*t (linear)*	*p (linear)*	*F (non-lin.)*	*p (non-lin.)*	*F(all)*	*p (all)*
PostMarks	5.96	2.57×10^{-9}			12.8	$< .0001$
SentAllOneScreen	-4.04	5.26×10^{-5}			24.1	$< .0001$
AbsPosInSent	1.29	0.196			15.2	$< .0001$
PosOnLineInChars	-2.14	0.033	99.67	$< .0001$	69.2	$< .0001$
PosOnScreen	5.76	8.43×10^{-9}	108.67	$< .0001$	122.8	$< .0001$
CobLog	-13.43	0.000	11.21	$< .0001$	127.7	$< .0001$
WLen	-0.38	0.704	4.21	.0149	26.4	$< .0001$
SentAllOneScreen * AbsPosInSent	-4.25	2.18×10^{-5}			18.0	$< .0001$
PostMarks * PosOnLineInChars	-1.23	0.22	26.42	$< .0001$	17.6	$< .0001$
SentAllOneScreen * PosOnLineInChars	7.36	5.44×10^{-12}	26.16	$< .0001$	35.5	$< .0001$
PostMarks * CobLog	-3.50	4.72×10^{-4}	2.43	0.0880	9.9	$< .0001$

Total R-squared for this model: 7.82%

Smooth 1-d predictors in Best GAM