Reading as Statistical Learning (in spite of language arbitrariness?)

Davide Crepaldi
[davide.crepaldi@sissa.it]
[http://lrlac.sissa.it]

International School for Advanced Studies (SISSA), Trieste
Language, Learning and Reading lab

ILC CNR, Pisa, 5 October 2017

Reading is a human wonder

Reading is outside of our genetic endowment:

- Not observed universally
- Not learned spontaneously

Nearly all readers are astonishingly efficient:

- 8–letter words in ~35ms (Forster and Davis, 1984)
- ~20 letters every ~250ms (Rayner, 1998)

Arbitrariness

- elephant
- table
- heat
- drum

Arbitrariness. Really?

- elephant
- table
- heat
- drum

- preheat
- juicer

- bioweapon
- guesstimate

The core idea

- Morphology* has created probabilistic regularities in language form . . .
- . . . and in form–to meaning mapping.
- The brain codes for these regularities . . .
- . . . and uses them during processing.

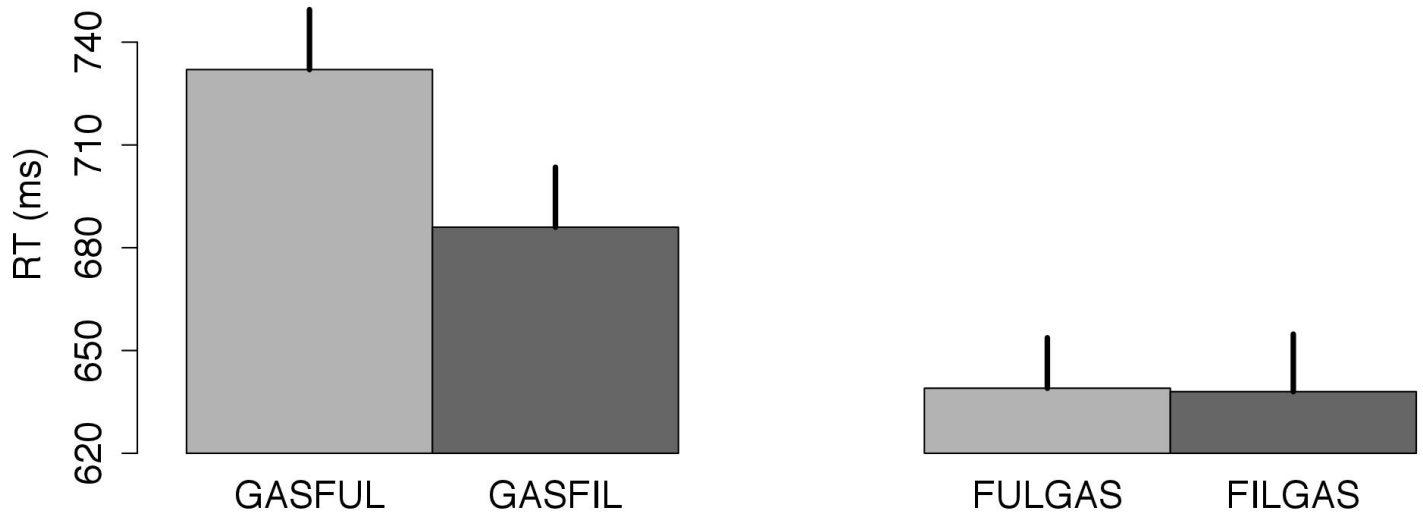
Positional constraints

Morpheme positional constraints

- KINDNESS and NESSKIND
- PREHEAT and HEATPRE
- CATWALK and WILDCAT
- OVERHANG and HANGOVER

Blind to suffixes

- (GASFUL vs. GASFIL) vs. (FULGAS vs. FILGAS)



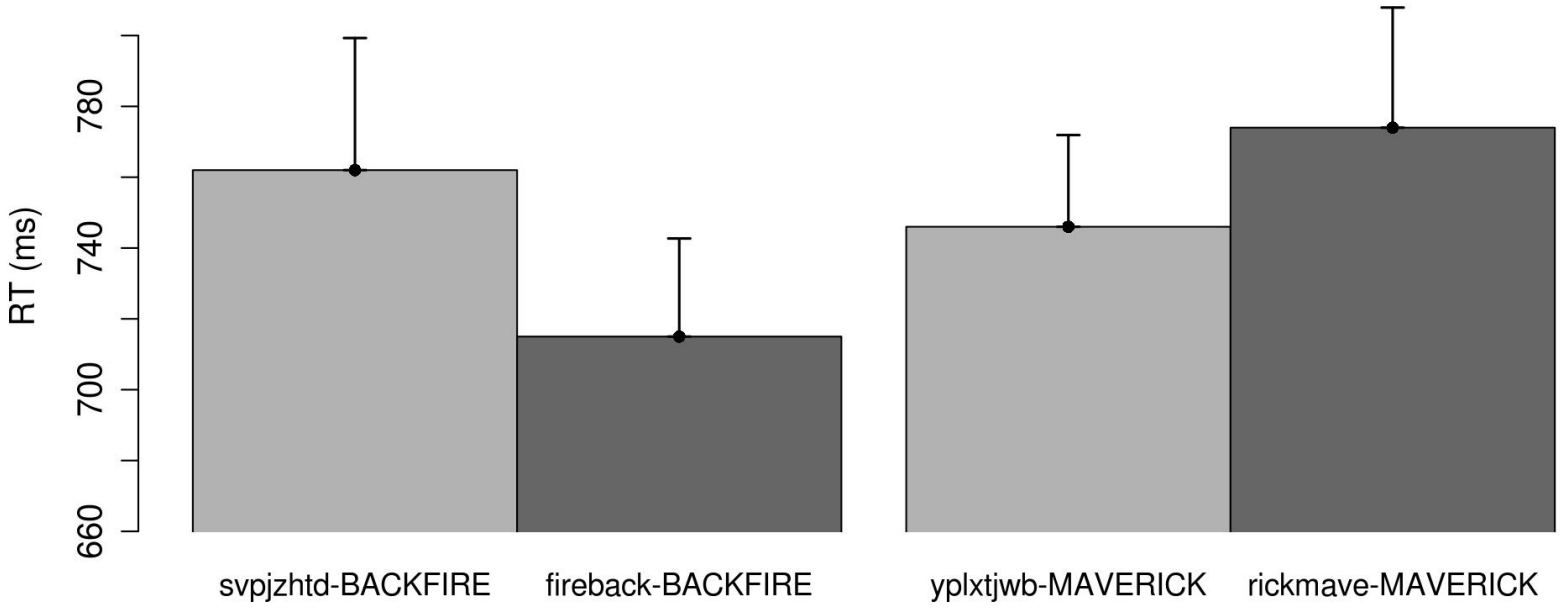
(Crepaldi et al., 2010)

Blind to prefixes

- (PREHOSE vs. PLEHOSE) vs. (HOSEPRE vs. HOSEPLE)

Stems everywhere

- (fishgold–GOLDFISH vs. kacnvrqw–GOLDFISH) vs. (tonebari–BARITONE vs. suyzchmw–BARITONE)



(Crepaldi et al., 2013)

How far do these constraints go?

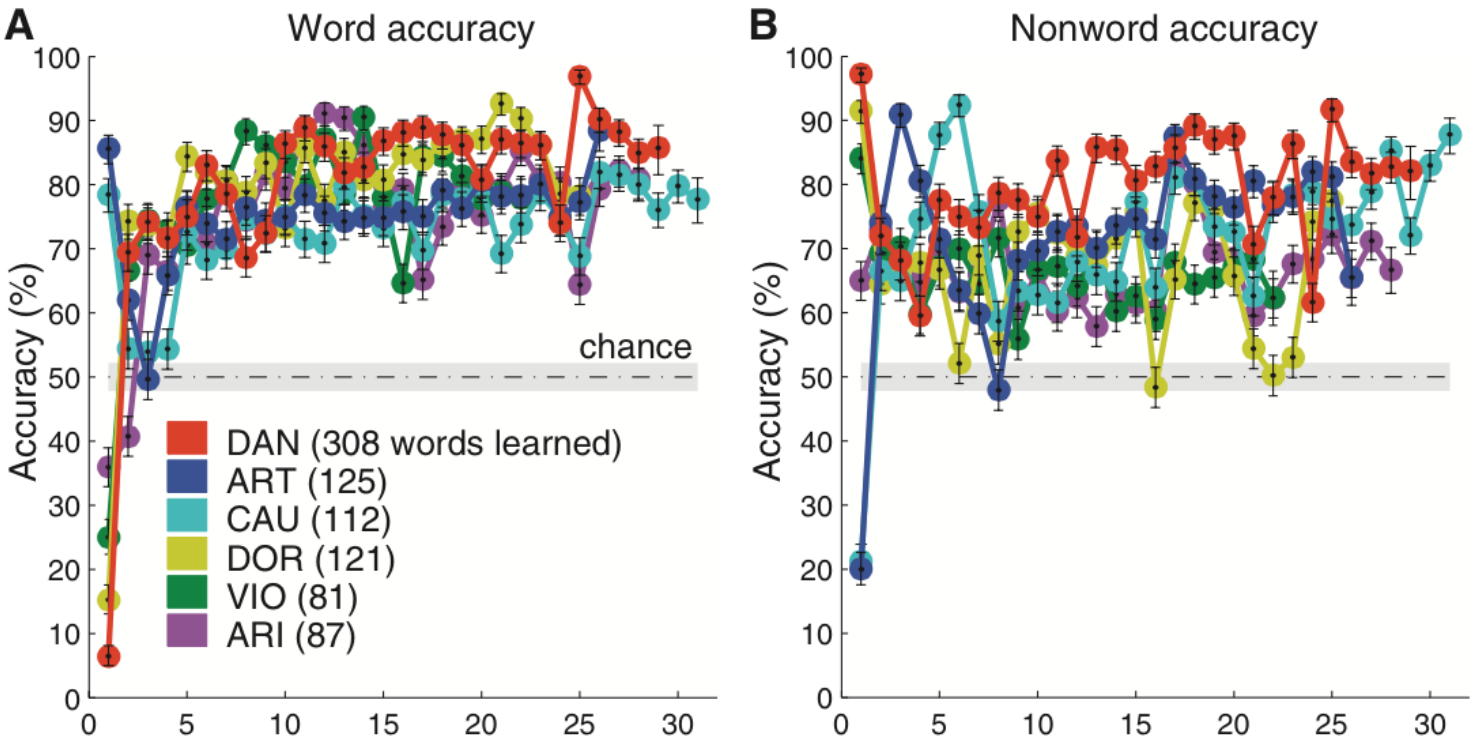
- Word boundaries vs. local constraints (in preparation, with Kathy Rastle and Colin Davis)
- All–or–none vs. graded constraints (current work, with Maria Ktori and Jana Hasenäcker)

Orthography in Baboons

Reading (!?) without language

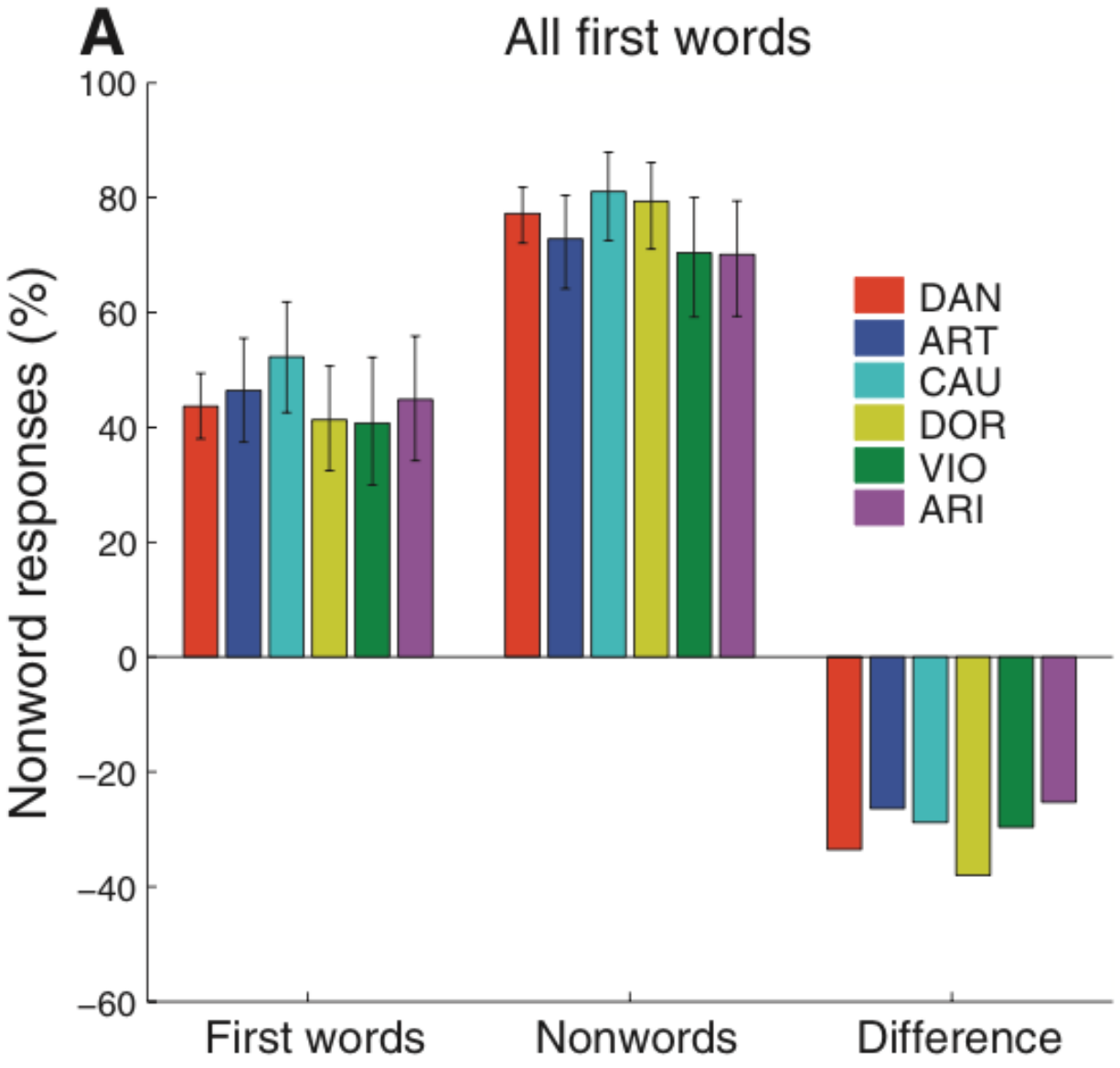
- Baboons can learn visually English words
- Baboons have no human–like language

Baboons learn words

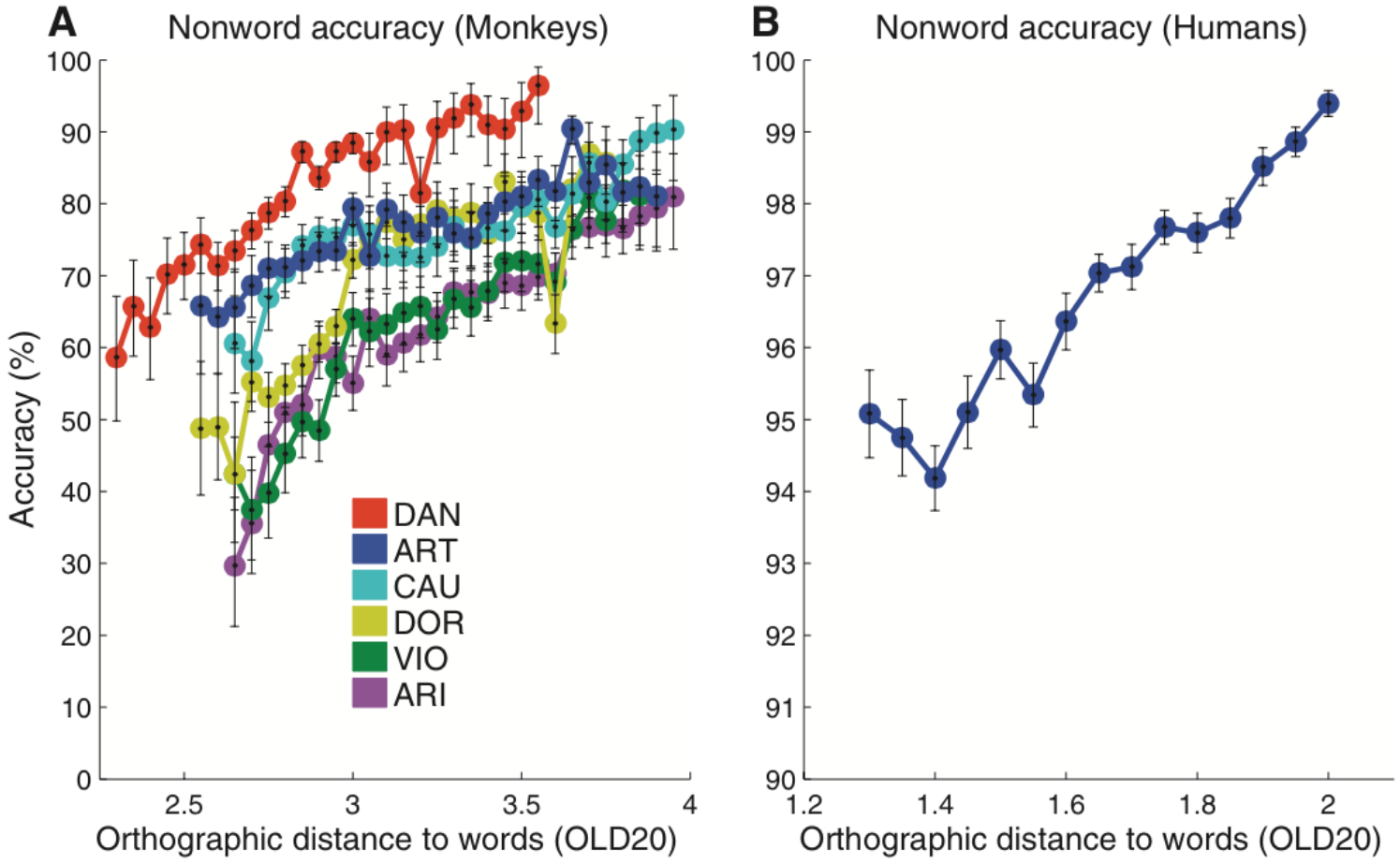


(Grainger et al., 2012)

Baboons extract knowledge about letter stats



Baboons extract knowledge about letter stats



Eye Tracking in Children Learning to Read

An experiment, but not so much of

- Natural reading
  - Stories (=connected text)
  - Just read and understand (=no strange task to carry out)

- Many children, create a database to share
- Across a natural spectrum of age
- Across a natural spectrum of reading proficiency

- Check sensitivity to statistical regularities

Eye tracking

For today

- Data from 22 kids (out of the 80 tested so far)

nGrams

- ALBERO:
  - 2grams: AL, LB, BE, ER, RO
  - 3grams: ALB, LBE, BER, ERO
  - 4grams: ALBE, LBER, BERO

- Average nGram frequency across whole words

Brains At Work

Brains At Work

- School trip
- The scientist gathers data, the kids gather experience
- SISSA Medialab

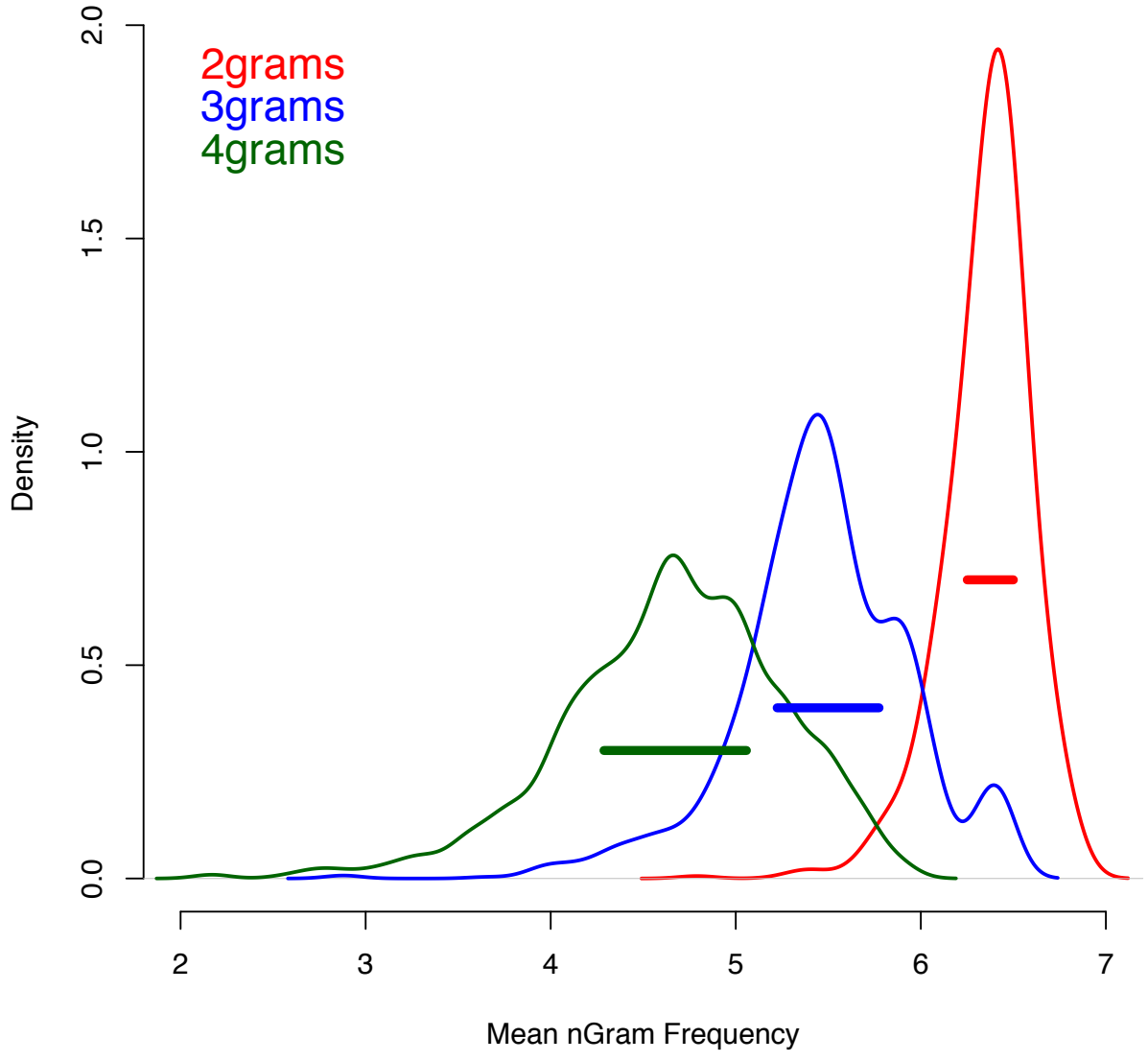
- 7 sessions, 140 kids in total

Brains At Work

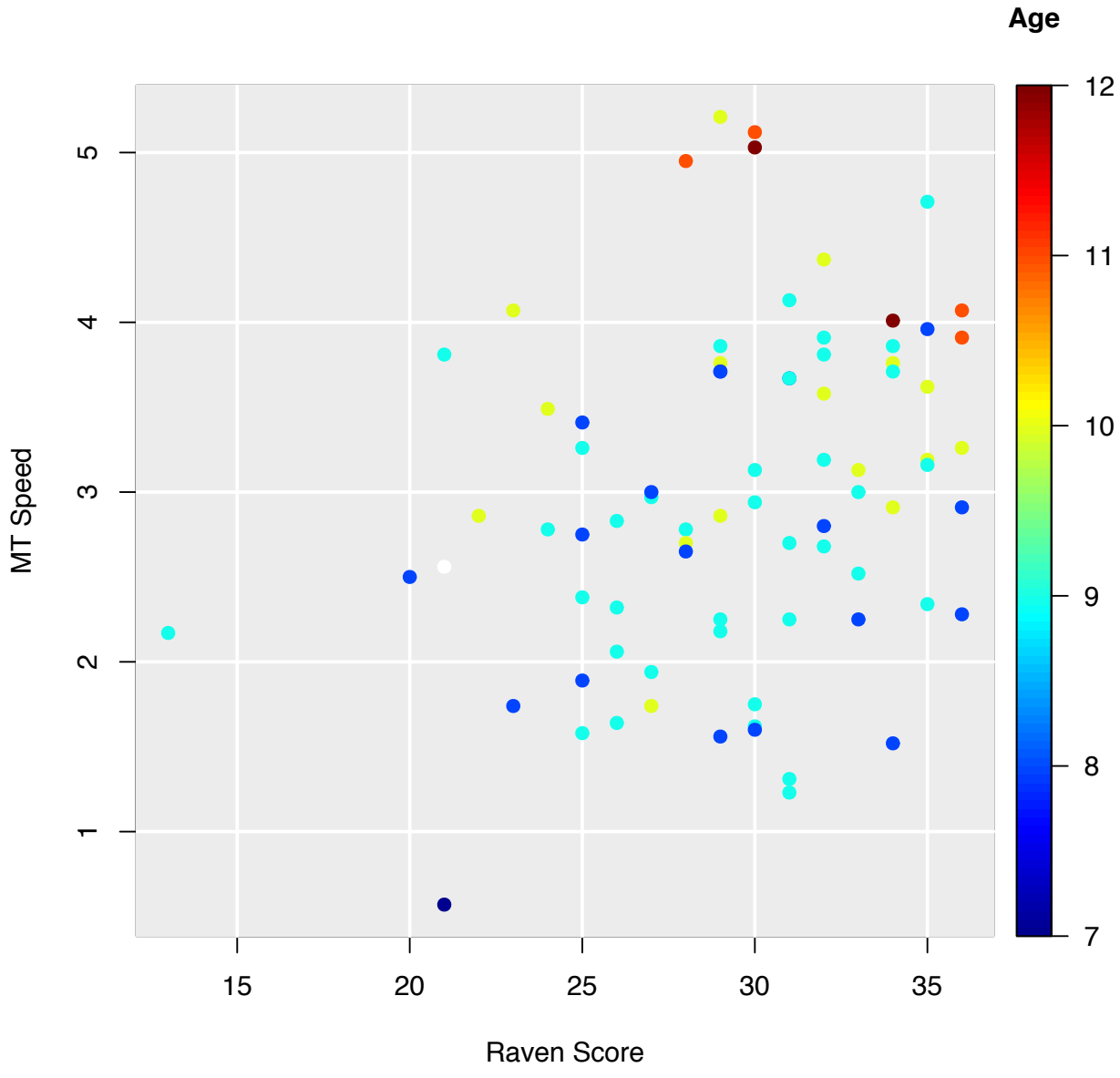
Word sample

- 1745 tokens, from 728 different words, across 12 short stories

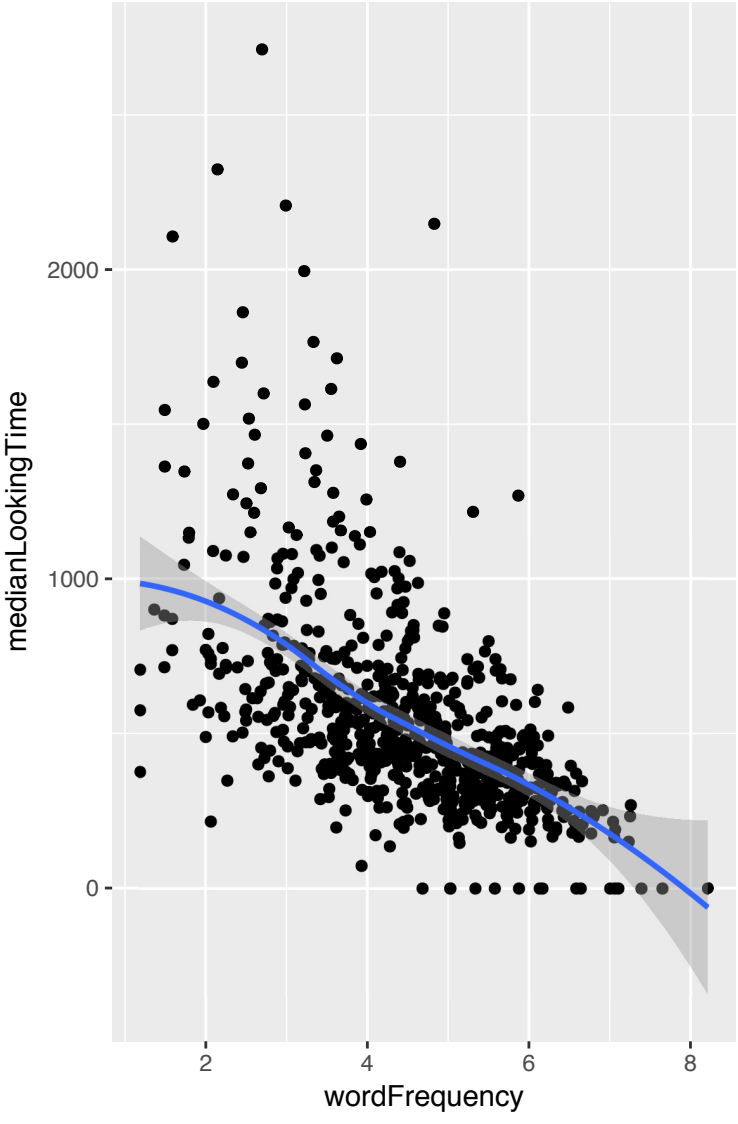
nGrams distribution

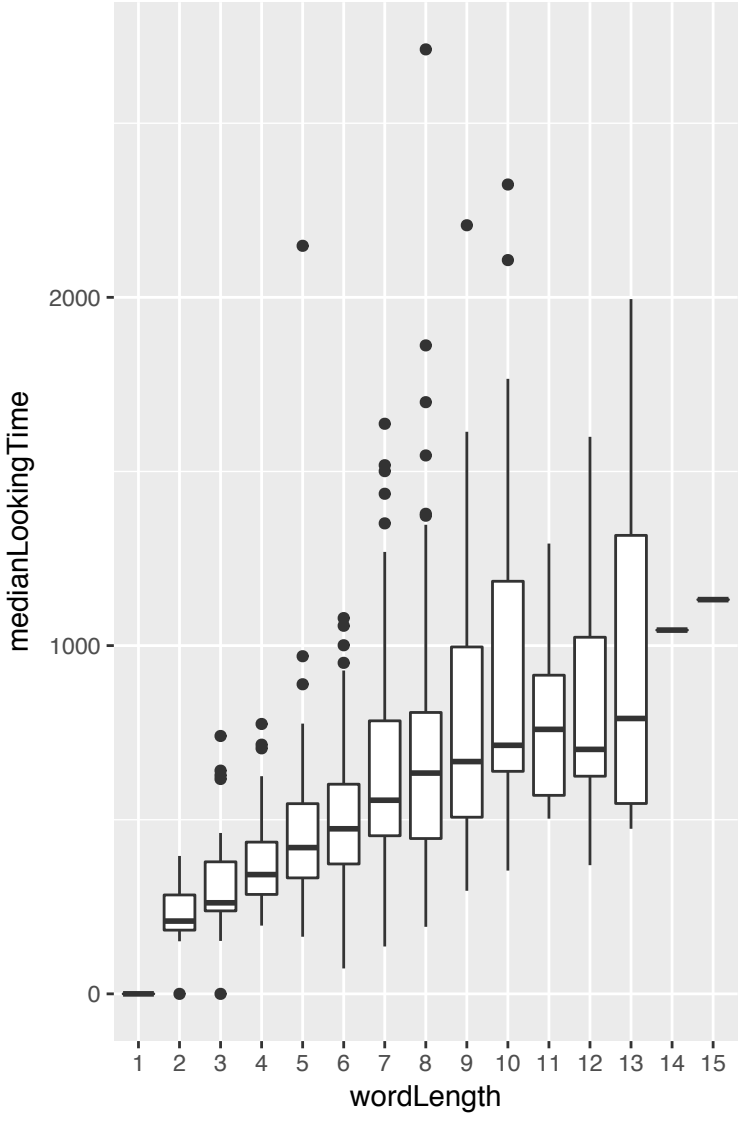


Participant sample

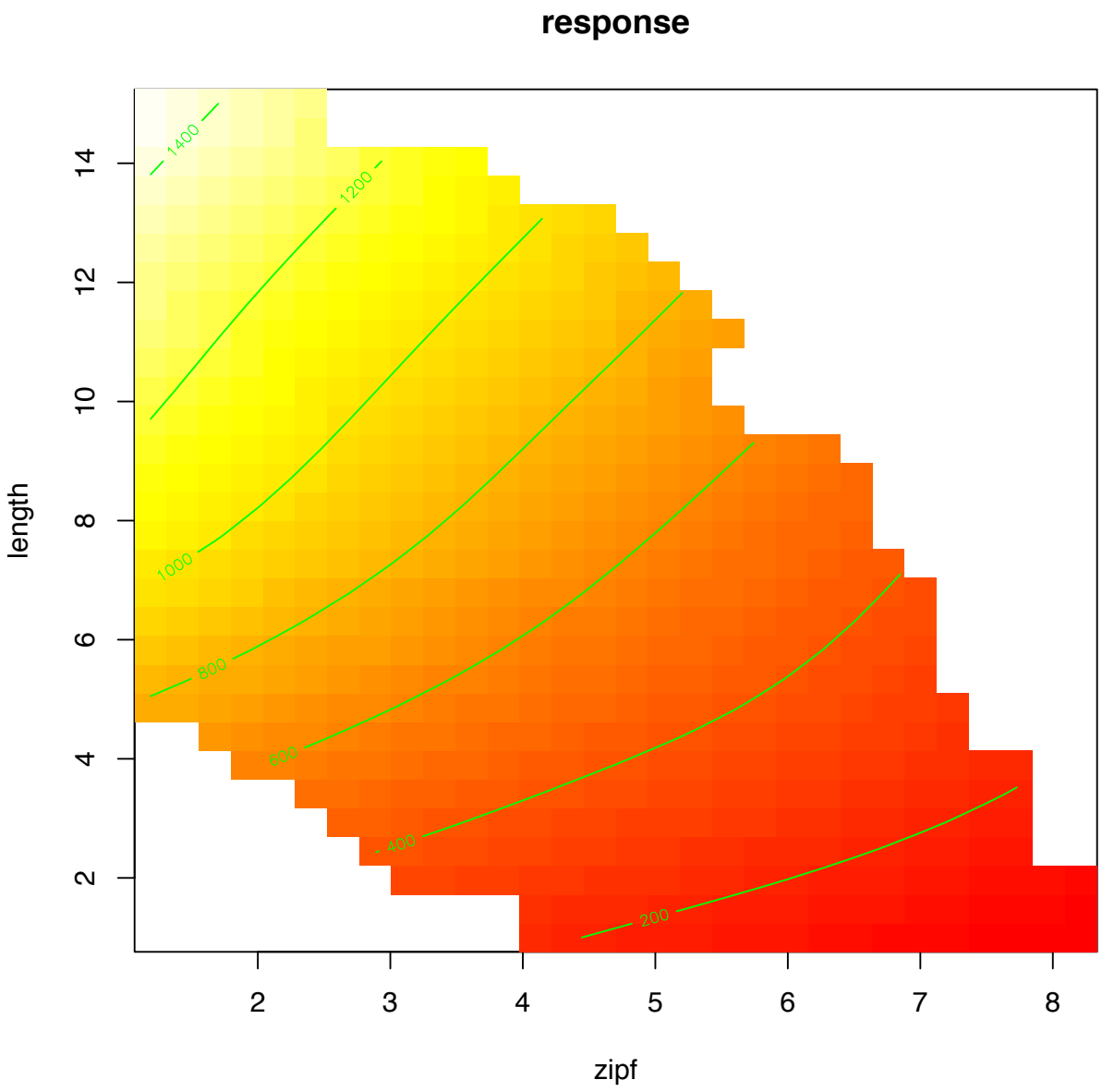


Frequency and length

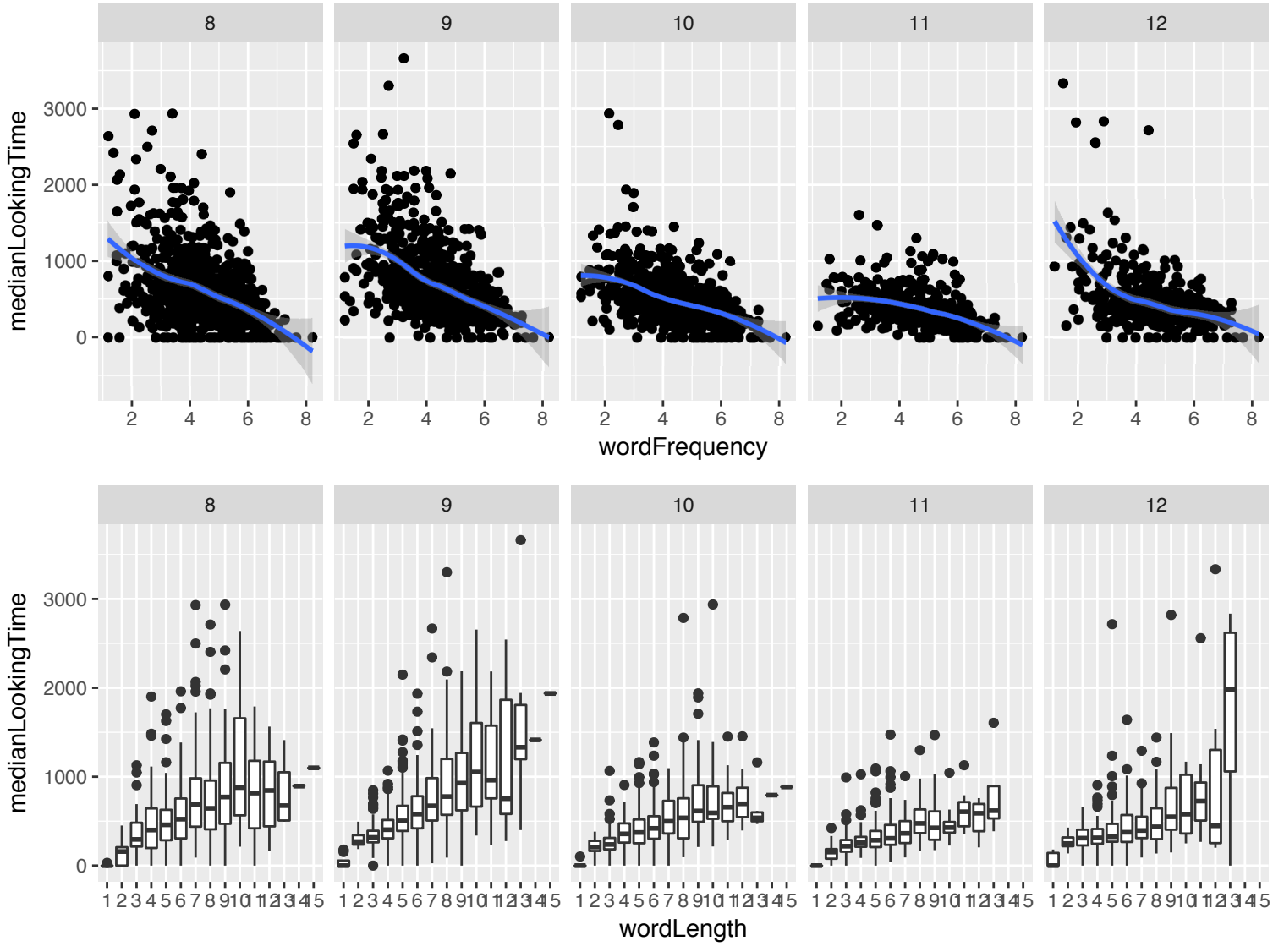




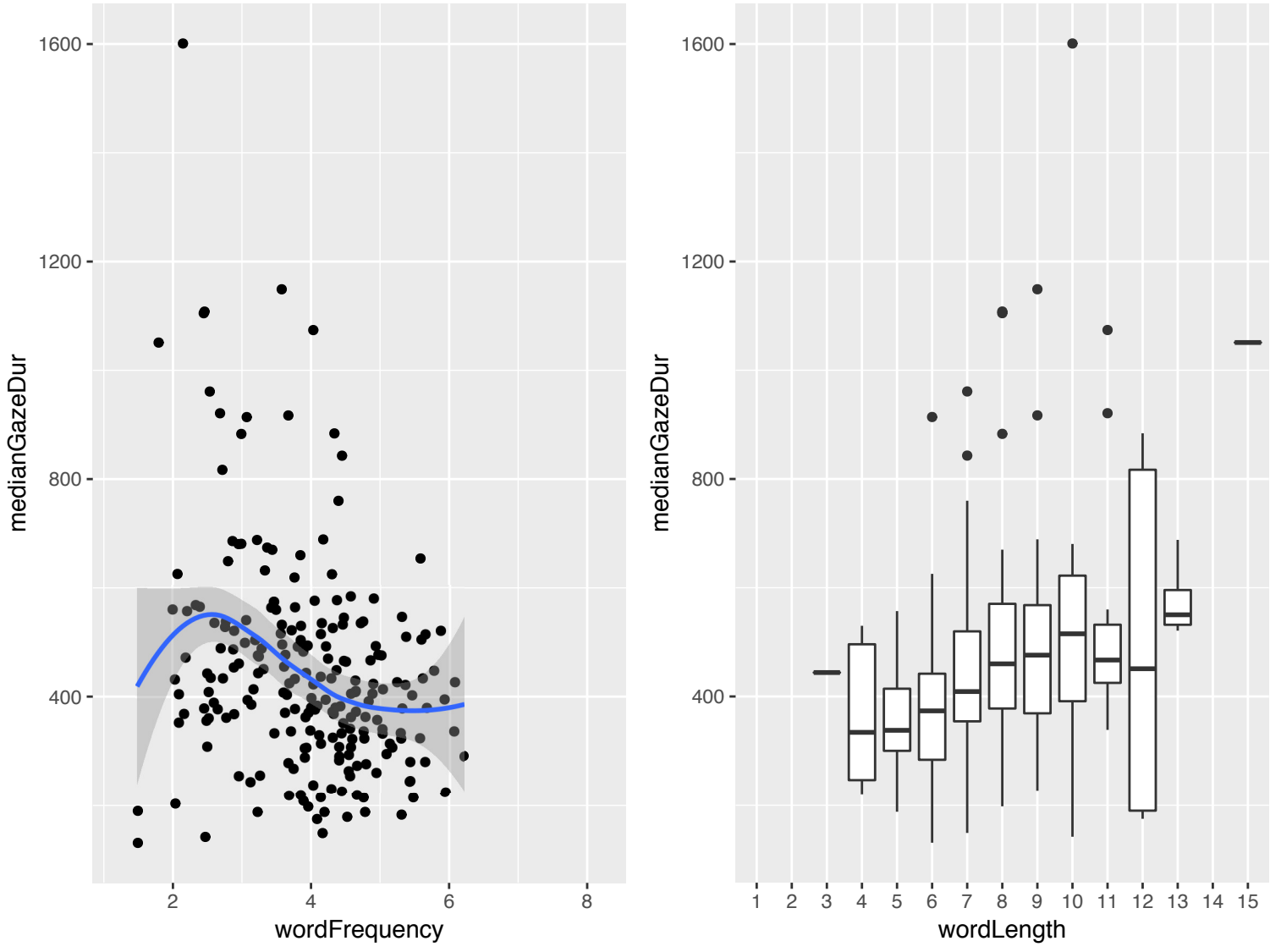
Frequency and length



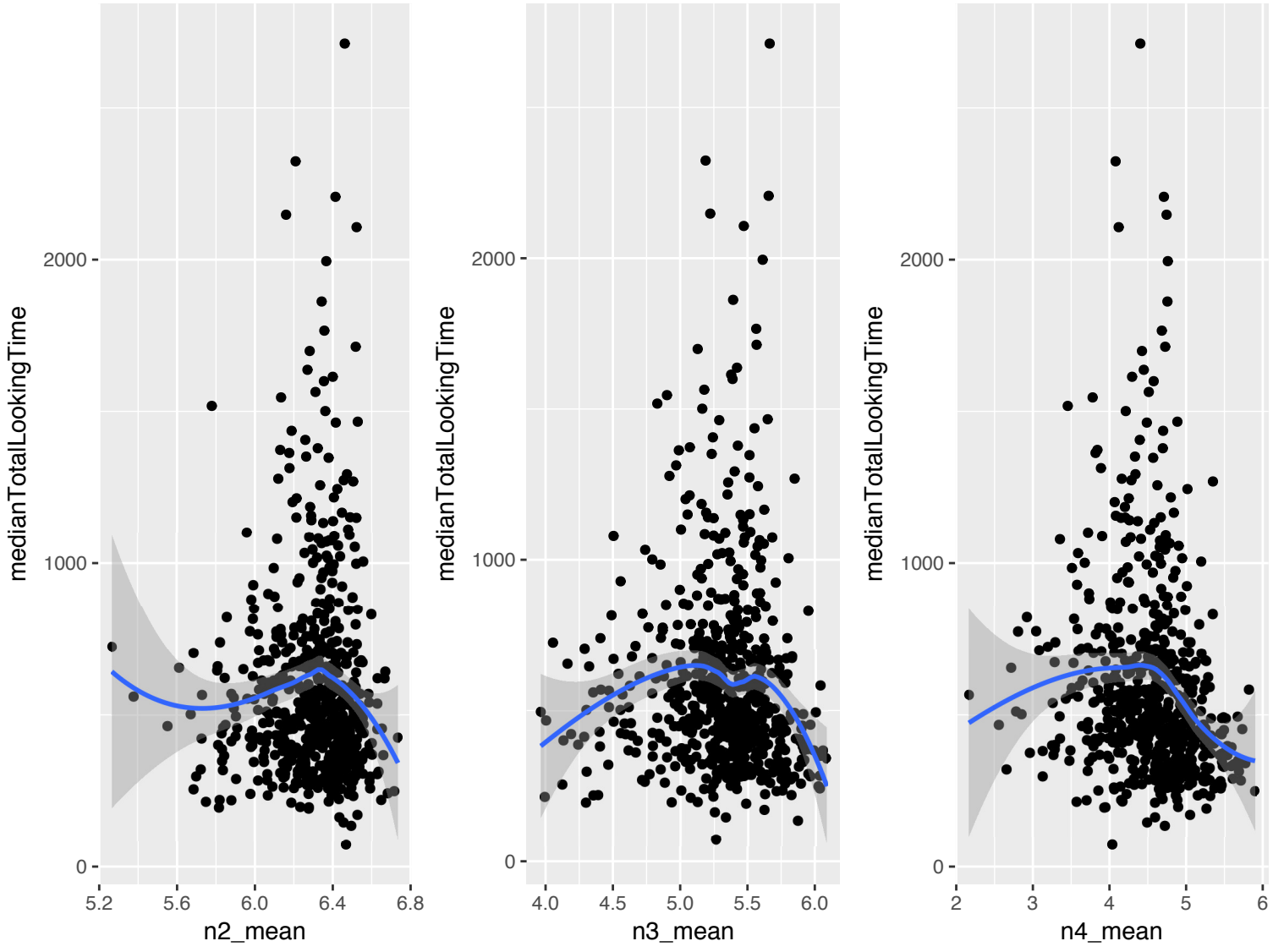
Age effects


8
9
10
11
12
medianLookingTime
3000
2000
1000
0
2
4
6
8
wordFrequency
wordLength
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15

Early processing?



nGrams effects



To sum up

- 2grams more characteristics of words, thus good to **distinguish words from non–existing strings**; but also less variable across words, thus **ineffective to identify specific words**.
- Frequency effects **(which is statistical learning)** in very young kids, and in early measures of processing.
- nGram frequency seems to affect eye movements in children.
- Children seem to track better the stats of larger chunks (jumping to lexicality?).

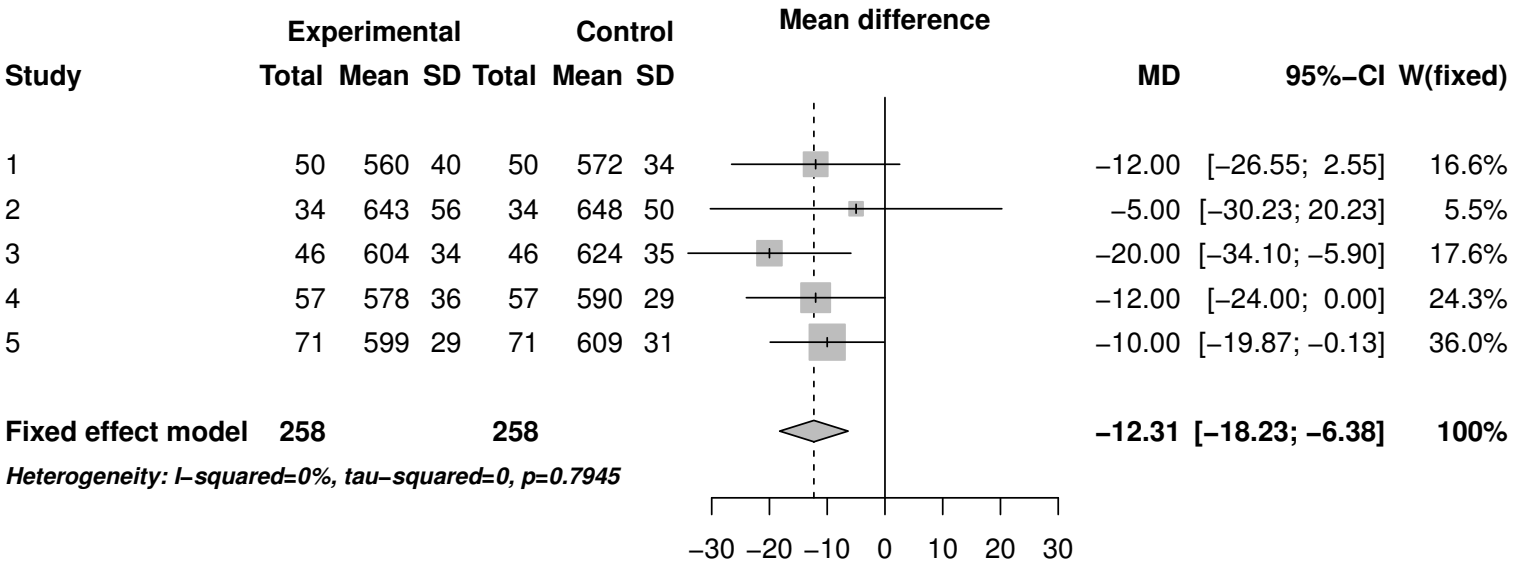
- The logic behind the experiment seems to work
- The logistics behind the experiment seem to work

Stepping outside form

Transparent stems?

	Transparent	Opaque	Orthographic
Related primes	dealer–DEAL	corner–CORN	dialog–DIAL
Control primes	poetry–DEAL	folder–CORN	prudish–DIAL
	DEAL	CORN	DIAL

Transparent stems?



(Marelli et al., 2015)

Orthography–Semantic Consistency (OSC)

CORN

- Get all words that start with CORN
- Take their semantic representations
- Compute their similarity
- Take the mean

$$OSC(t) = \frac{\sum_{j=1}^{k} f_{r_x} \cos(\vec{t}, \vec{r_x})}{\sum_{j=1}^{k} f_{r_x}}$$

- How good is form as a cue to meaning

OSC gets unique variance

Table 6. *Results of the regression analysis on the lexical decision latencies extracted from the BLP for a large set of random words*

	Estimate	*Std error*	t *value*	p *value*
Intercept	6.5922	.0109	602.89	.0001
Word frequency	−0.0308	.0009	33.41	.0001
Word FS	−0.0041	.0021	1.97	.0495
Word length	0.0035	.0013	2.74	.0061
OSC	−0.0254	.0066	3.84	.0002

(Marelli et al., 2015)

OSC gets further

- OSC modulates morphological priming (in preparation, with Simona Amenta and Marco Marelli)
- OSC modulates brain electrophysiology (in preparation, with Simona Amenta, Marco Marelli, and Leo Budinich)
- PSC (Amenta et al., 2016)

Wrap up

A new approach to reading

- Scripts can be seen as fully–fledged visual systems
- They can be studied as such (without language)
- The way we learn to deal with them can be captured through statistical learning
- The way we learn to map them onto language can be captured through statistical learning

A new approach to reading

- Scripts can be seen as fully–fledged visual systems
- They can be studied as such (without) language
- The way we learn to deal with them can be captured through statistical learning
- The way we learn to map them onto language can be captured through statistical learning

Acknowledgments

- Valentina Pescuma, Eva Viviani, Maria Ktori, Marijana Sjekloća, Francesca Franzon (SISSA); Benedetta Cevoli (now at RHUL), Eleonora Lomi (now at UCL).
- Kathy Rastle (Royal Holloway), Colin Davis (Bristol), Steve Lupker (Western).
- Simona Amenta (Gent), Marco Marelli (Milano Bicocca)
- Valentina Parma (SISSA) and Simona Cerrato (SISSA Medialab).

Reading as Statistical Learning (in spite of language arbitrariness?)

Davide Crepaldi
[davide.crepaldi@sissa.it]
[http://lrlac.sissa.it]

International School for Advanced Studies (SISSA), Trieste
Language, Learning and Reading lab

ILC CNR, Pisa, 5 October 2017

Refrences I

Amenta, S., Marelli, M., and Sulpizio, S. (2016). From sound to meaning: Phonology-to-semantics mapping in visual word recognition. *Psychonomic Bulletin & Review*. Published online ahead of print.

Crepaldi, D., Rastle, K., and Davis, C. (2010). Morphemes in their place: Evidence for position-specific identification of suffixes. *Memory and Cognition*, 38(3):312–321.

Crepaldi, D., Rastle, K., Davis, C. J., and Lupker, S. J. (2013). Seeing stems everywhere: Position-independent identification of stem morphemes. *Journal of Experimental Psychology: Human Perception and Performance*, 39:510–525.

Forster, K. I. and Davis, C. (1984). Repetition priming and frequency attenuation in lexical access. *Journal of Experimental Psychology: Learning Memory and Cognition*, 10:680–698.

Grainger, J., Dufau, S., Montant, M., Ziegler, J., and Fagot, J. (2012). Orthographic processing in baboons (papio papio). *Science*, 336(6078):245–248.

Marelli, M., Amenta, S., and Crepaldi, D. (2015). Semantic transparency in free stems: The effect of Orthography–Semantics Consistency on word recognition. *Quarterly Journal of Experimental Psychology*, 68(8):1571–1583.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124:372–422.

nGrams correlation