# 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

Cornell University, 3 November 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 (among other, less important factors) has created probabilistic regularities in language form and in form-to meaning mapping.
- The brain codes for these regularities and uses them during language processing/reading.

#### Orthography in Baboons

#### Baboons learn words



(Grainger et al., 2012)

#### Baboons extract knowledge about letter stats



#### Baboons extract knowledge about letter stats



### 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)

#### Eye Tracking in Children Learning to Read

#### An experiment, but not so much of

#### Natural reading

- Connected text
- Just read and understand (=no strange task to carry out)
- Many children, create a database to share
- Across a wide spectrum of age
- Across a wide spectrum of reading proficiency

# Eye tracking



### Brains At Work



#### Brains At Work



### Brains At Work



### For today

- Data from 39 kids (out of the 80 tested so far)
- nGrams
  - ► ALBERO:
    - 2grams: AL, LB, BE, ER, RO
    - 3grams: ALB, LBE, BER, ERO
    - Agrams: ALBE, LBER, BERO
  - Average nGram frequency across whole words

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

#### nGrams distribution



Mean nGram Frequency

#### Participant sample



Raven Score

#### Frequency and length



## Early processing?



#### nGrams effects



#### To sum up

- Frequency effects 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

#### Down the line

- Check morphology
- A `sliding window' analysis
- Word predictability in context (corpora, cloze task with kids)
- Explore other types of statistical regularities (e.g., transitional probabilities, long-distance relationships)
- Consider spaces, which may be critical for its perceptual salience
- Takes care of predictor correlation more seriously

#### 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 <i>value</i>
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)
- OSC effect grows with proficiency in L2 (in preparation, with Eva Viviani). Talk at Psychonomics, Sunday 9.20AM.

# 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)





# 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

Cornell University, 3 November 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

Average nGram Frequency



#### OSC tracks language learning



#### OSC explains brain potentials

