

Anything special about words?

(Statistical learning and written language)

Davide Crepaldi

SISSA

Trieste, Italy

davide.crepaldi@sissa.it

Tweets at @CrepaldiDavide

Universitat Pompeu Fabra

11 November 2022



The reading paradox

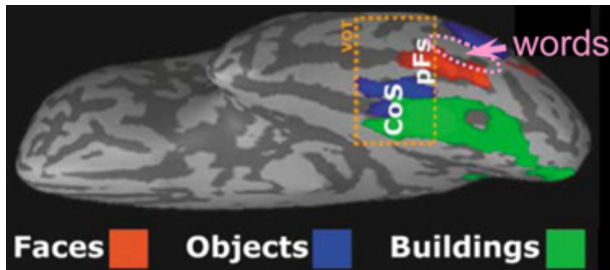
We're fantastic readers. . .

- ▶ We can identify words in ~35ms and without awareness (e.g., Forster and Davis, 1984)
- ▶ We read ~260 words per minute (e.g., Brysbaert, 2019)

. . . and yet, no direct genetic endowment

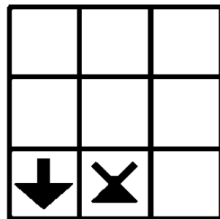
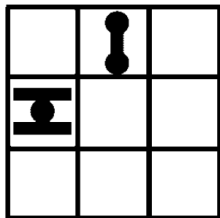
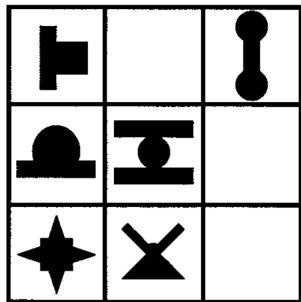
- ▶ Written language is a recent invention (~5.5K ya)
- ▶ Written language isn't observed universally
- ▶ Literacy isn't acquired spontaneously

A solution to the paradox



(Dehaene and Cohen, 2007)

Statistical learning in visual scenes



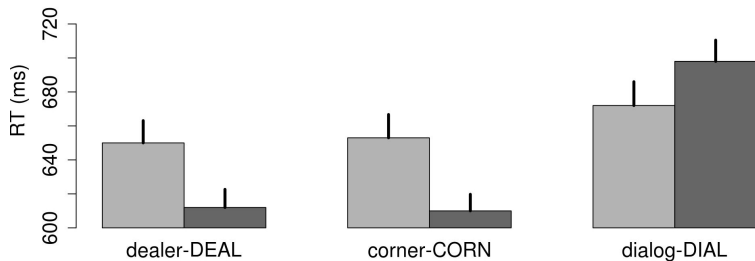
(Fiser and Aslin, 2001)

Statistical learning in the lexicon

- ▶ Words as ordered chunks of letters
- ▶ Morphemes (e.g., *deal-er*) as recurring chunks of letters

There's corn in the corner

dealer-DEAL vs. corner-CORN vs. dialog-DIAL



(Longtin et al., 2003; Rastle et al., 2004)

Morphemes in their place

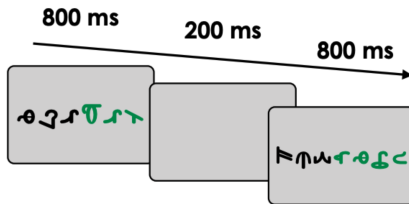


(Crepaldi et al., 2010)

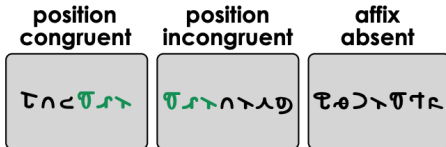
The statistical learning of affixes

Artificial affixes

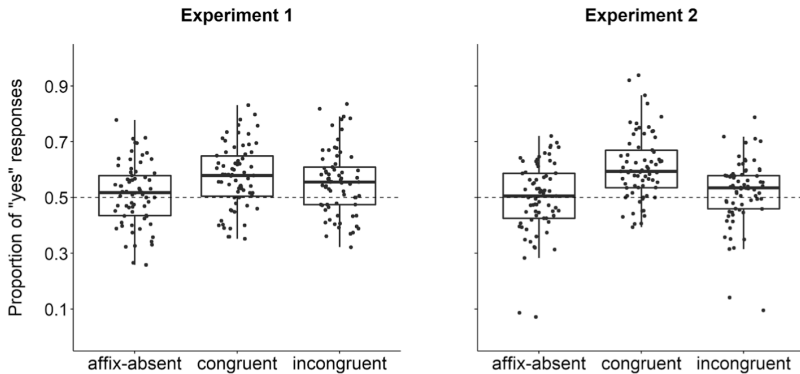
Training



Testing



Results



(Lelonkiewicz et al., 2020)

Visual affixes

Our chunks:

- ▶ clusters of pseudo-letters that occur together frequently across different items
- ▶ no contact with phonology, semantics or syntax

And yet:

- ▶ sensitivity to these chunks
- ▶ sensitivity to their position (e.g., Crepaldi et al., 2010)

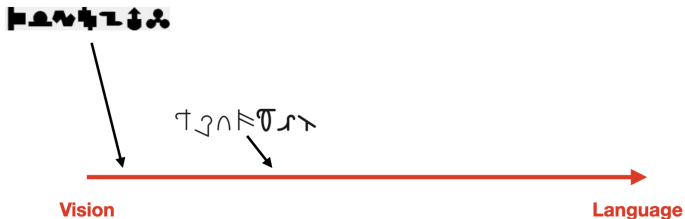
Stairway to language



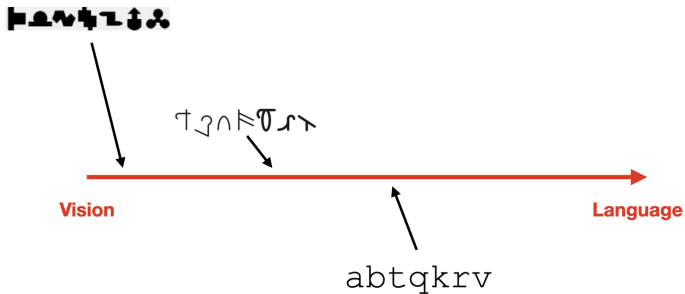
Stairway to language



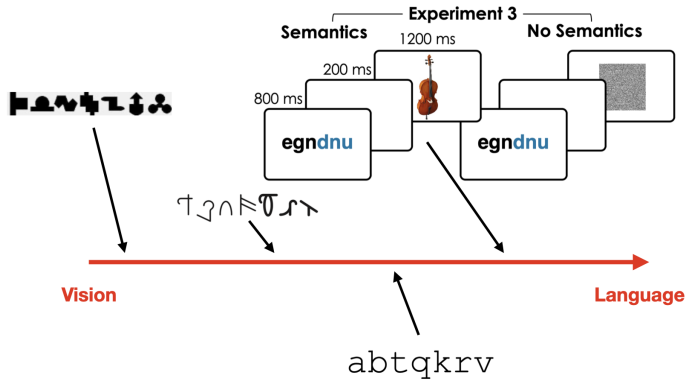
Stairway to language



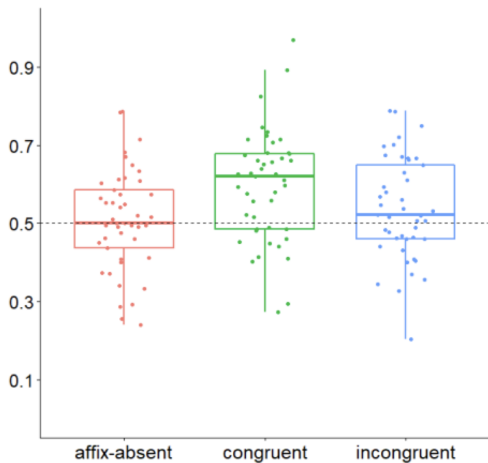
Stairway to language



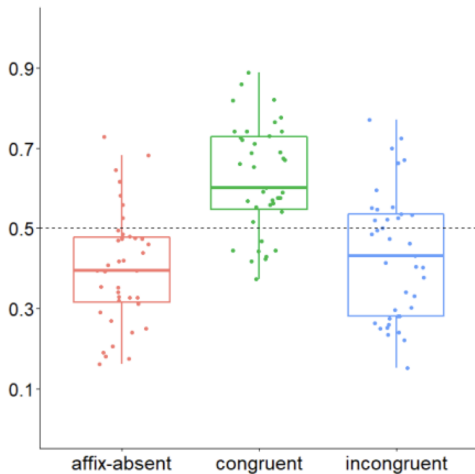
Stairway to language



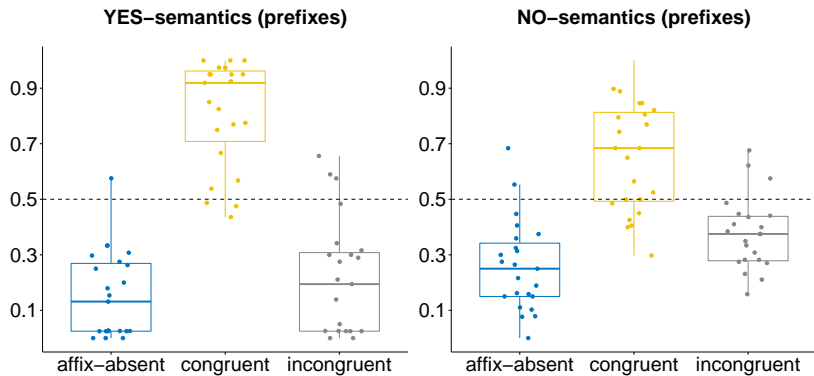
One step down: abstract shapes



One step up: real letters



Two steps up: meaning



Take home message

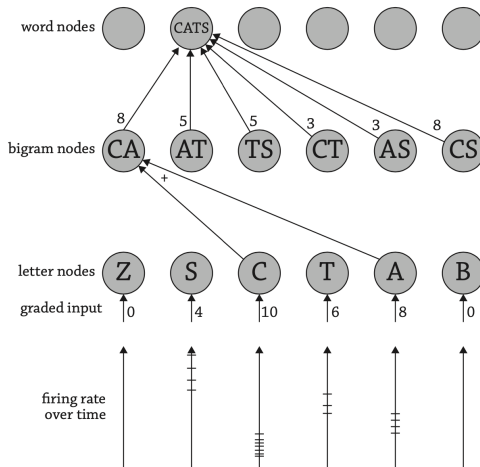
- ▶ Readers spontaneously extract visual statistical regularities and use them to identify affix-like character chunks
- ▶ Such learning occurs in purely visual and language-like material, showing that the core computational engine is not language-specific
- ▶ Such learning is, however, enhanced by the availability of linguistic information (phonology and meaning)

More about this?

- ▶ Lelonkiewicz, J., Ktori, M. and Crepaldi, D. (2020). Morphemes as letter chunks: Discovering affixes through visual regularities. *Journal of Memory and Language*, 115, 104152
- ▶ Lelonkiewicz, J., Ktori, M. and Crepaldi, D. (2021). Morphemes as letter chunks: Linguistic information enhances the learning of visual regularities. In revision at *Journal of Memory and Language*

Does this go beyond morphology?

Bigrams



ABF

DBC

AEC

Phantom words (Endress and Mehler, 2009)

ABF

DBC

AEC

ABC

ABF

DBC

AEC

DEF

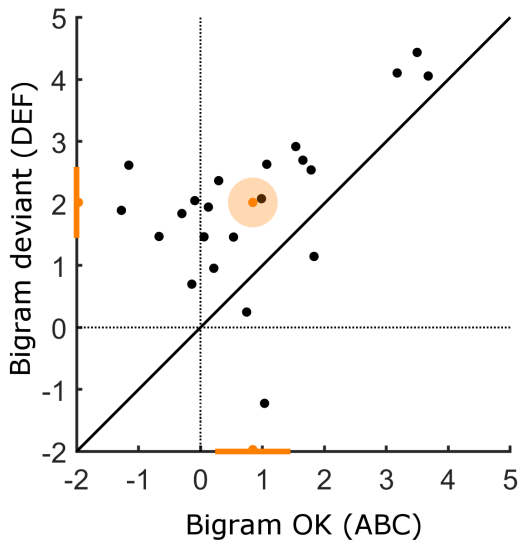
ABF
DBC
AEC

ABC
DEF

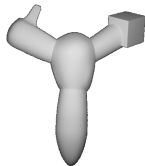
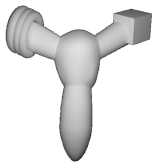
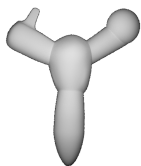
Pseudo-letters (Vidal et al., 2017)



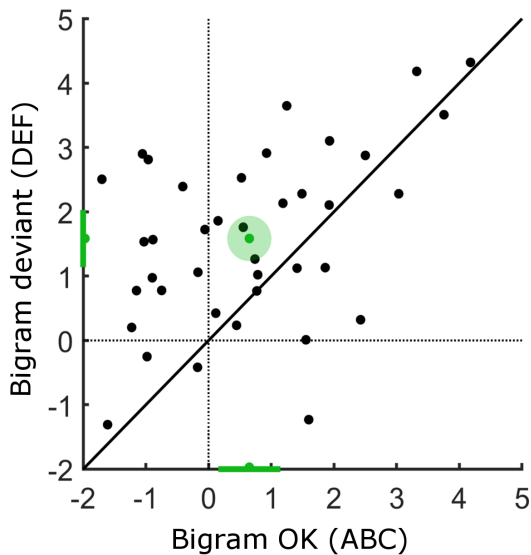
Results



Phantom tripods



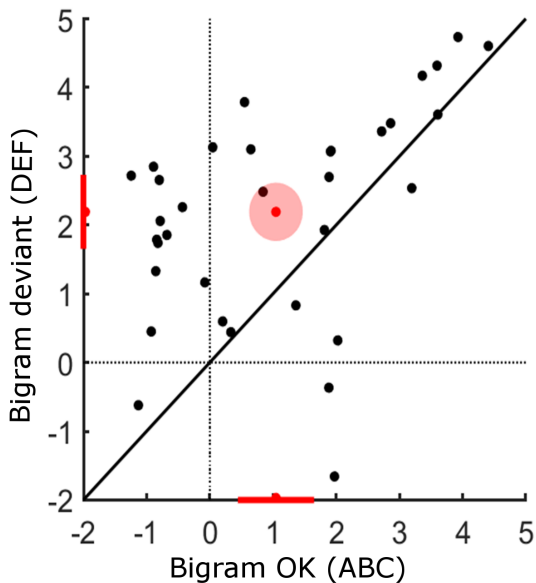
Phantom tripods



Phantom Gabors



Phantom Gabors



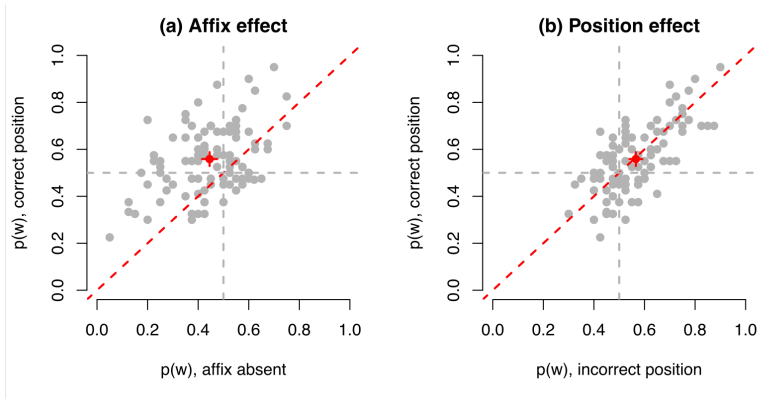
Take-home message

- ▶ We code for nGrams/letter transition stats while learning novel words
- ▶ We use the same mechanism while learning novel objects, where the lower-level units are:
 - ▶ not arranged horizontally, and very different visually from letters
 - ▶ not even spatially segregated
- ▶ Word learning shares (part of) its computational core with vision

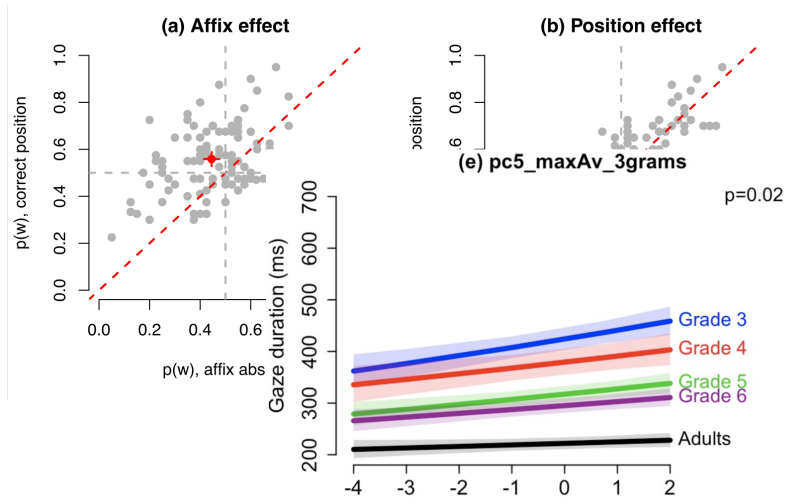
More about this?

- ▶ Vidal, Y., Viviani, E., Zoccolan, D. and Crepaldi, D. (2021). A general-purpose mechanism of visual feature association in visual word identification and beyond. *Current Biology*, 31, 1261-1267.

Children interlude

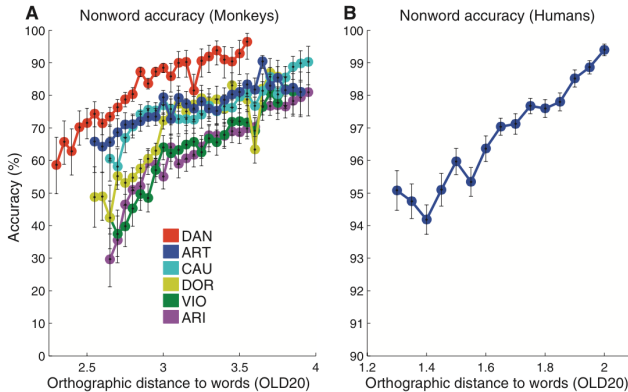


Children interlude



Do we see orthographic in non-linguistic animals?

Orthographic coding in Baboons



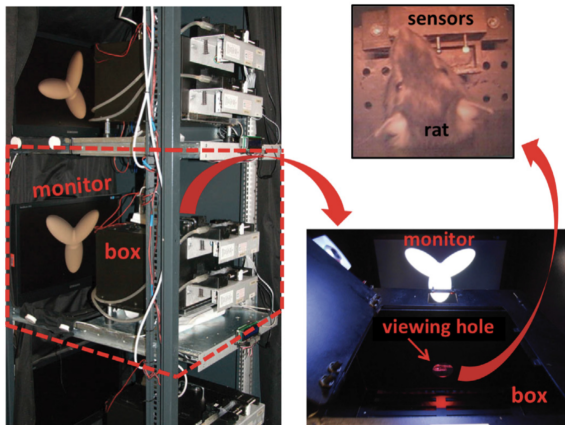
(Grainger et al., 2012)

Proof-of-principle

Long-Evans rats

- ▶ A jump back in evolution
- ▶ Low acuity
- ▶ Object invariance is not entirely clear (e.g., Tafazoli et al., 2017; Vinken and Op De Beeck, 2021)

The setup



Letter identification

Go-left letters



Go-right letters



Letters as abstract objects

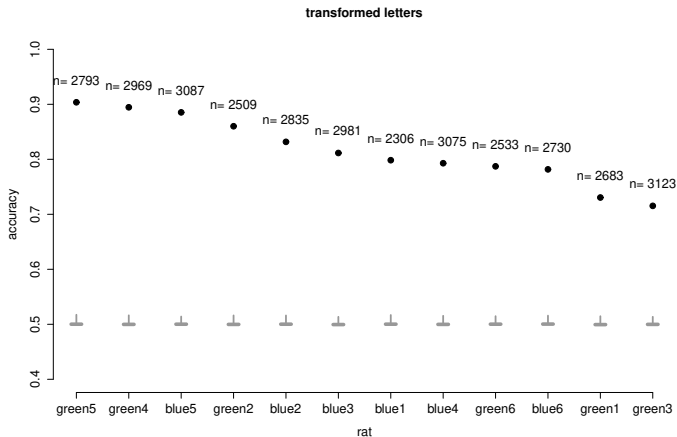
B

G

J

U

Letters as abstract objects



A rodent model of visual word identification

- ▶ Learn to identify letters with at least some degree of invariance

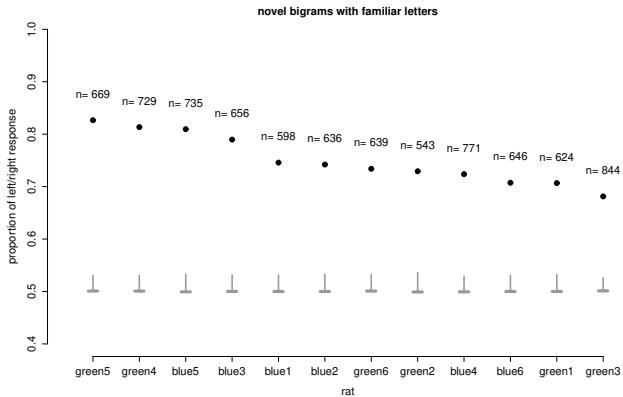
Letters in bigrams

One trained letter, one distractor letter

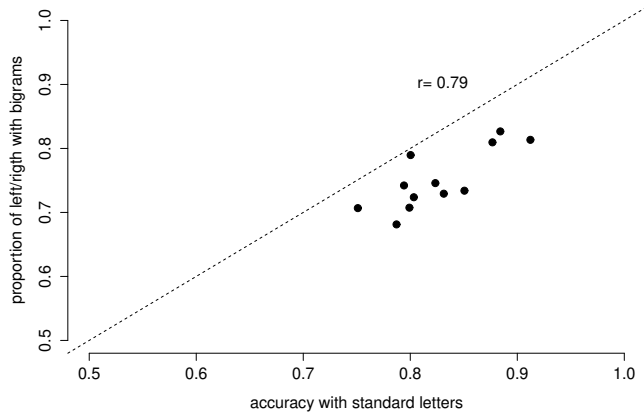


B M

Letters in bigrams



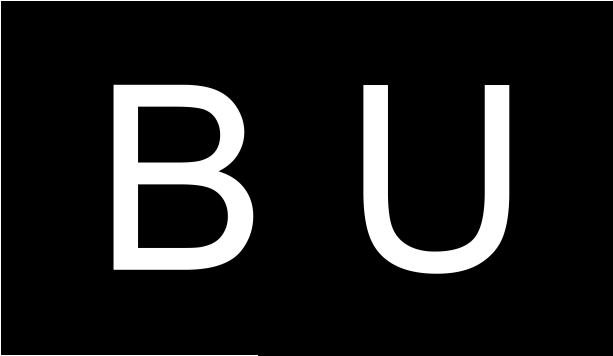
Letters in bigrams



A rodent model of visual word identification

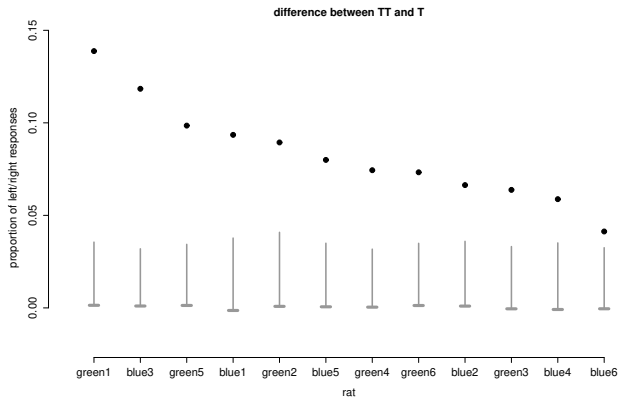
- ▶ Learn to identify letters with at least some degree of invariance
- ▶ Identify individual letters as independent objects within bigrams

Two familiar letters

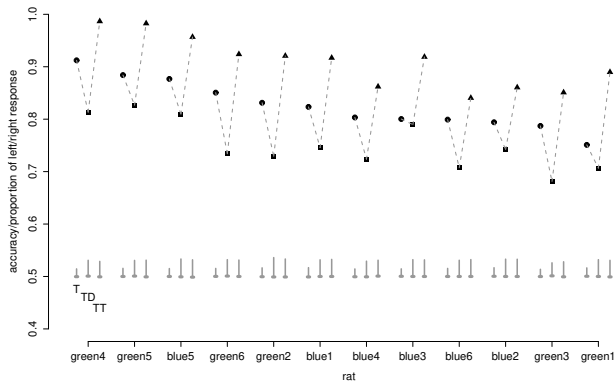


B U

Information from multiple letters



Information from multiple letters



A rodent model of visual word identification

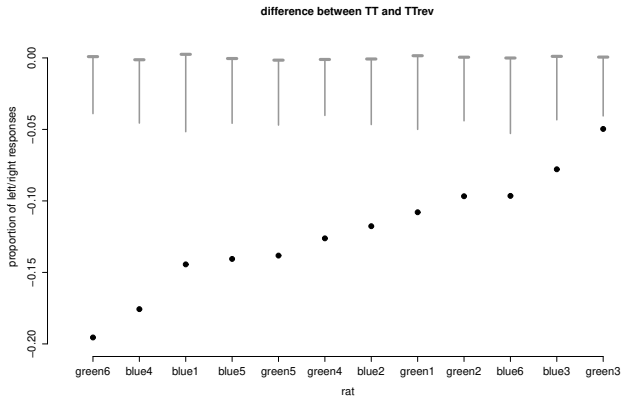
- ▶ Learn to identify letters with at least some degree of invariance
- ▶ Identify individual letters as independent objects within bigrams
- ▶ Use information from multiple letters within bigrams

Transposed letter effects

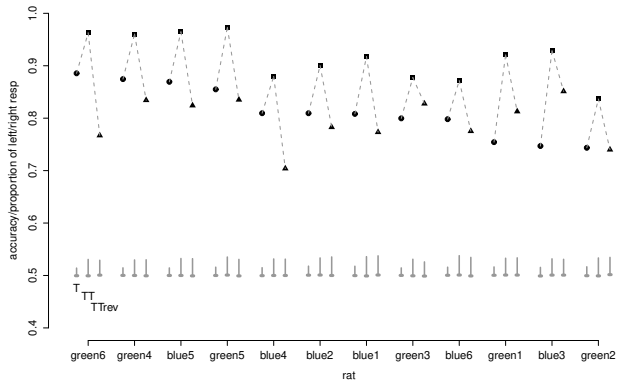
Two trained letters, but in the 'wrong' position



Letters are identified, still



But position is coded, too



A rodent model of visual word identification

- ▶ Learn to identify letters with at least some degree of invariance
 - ▶ Identify individual letters as independent objects within bigrams
 - ▶ Use information from multiple letters within bigrams
 - ▶ Identify letters in a position-invariant way
 - ▶ Code for letter position within bigrams
-
- ▶ Rats **spontaneously** process bigrams orthographically, given familiarity with letters

Take-home message

- ▶ A general-purpose computational core behind visual word learning and processing
- ▶ This computational core is based on statistical regularities, of different forms (item frequency; De Rosa et al., 2022) (form to meaning mapping; Dessì et al., 2019; Marelli et al., 2015)

Is that it?

- ▶ Nope! This is perhaps a bootstrapping system

Acknowledgments

- ▶ Jarosław Lełonekiewicz
- ▶ Maria Ktori
- ▶ Yamil Vidal
- ▶ Eva Viviani
- ▶ Antonella Torrisi
- ▶ Angelina Tadić
- ▶ Davide Zoccolan



Anything special about words?

(Statistical learning and written language)

Davide Crepaldi

SISSA

Trieste, Italy

davide.crepaldi@sissa.it

Tweets at @CrepaldiDavide

Universitat Pompeu Fabra

11 November 2022



References I

- Brysbaert, M. (2019). How many words do we read per minute? A review and meta-analysis of reading rate. *Journal of Memory and Language*, 109.
- 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.
- De Rosa, M., Ktori, M., Vidal, Y., Bottini, R., and **Crepaldi**, t. (2022). Frequency-based discrimination in fast periodic visual stimulation. *Cortex*, 148:193–203.
- Dehaene, S. and Cohen, L. (2007). Cultural recycling of cortical maps. *Neuron*, 56:384–398.
- Dessì, R., Bouchacourt, D., **Crepaldi**, t., and Baroni, M. (2019). Focus on what's informative and ignore what's not: Communication strategies in a referential game. *NeurIPS Workshop on Emergent Communication*.
- Endress, A. D. and Mehler, J. (2009). The surprising power of statistical learning: When fragment knowledge leads to false memories of unheard words. *Journal of Memory and Language*, 60:351–367.
- Fiser, J. and Aslin, R. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science*, 12(6):499–504.
- 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.

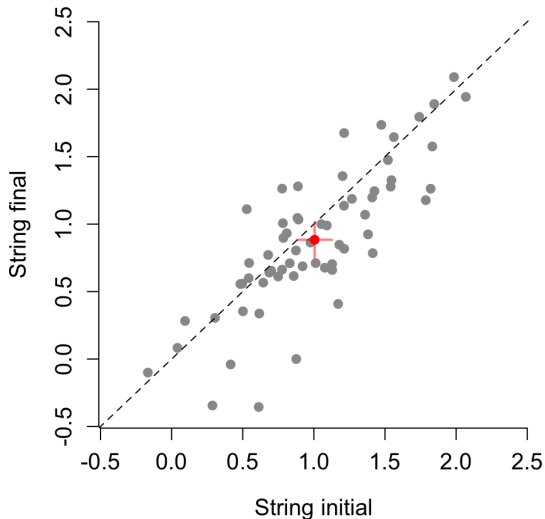
References II

- Grainger, J. (2018). Orthographic processing: a “mid-level” vision of reading. *Quarterly Journal of Experimental Psychology*, 71:335–359.
- Grainger, J., Dufau, S., Montant, M., Ziegler, J., and Fagot, J. (2012). Orthographic processing in baboons (*papio papio*). *Science*, 336(6078):245–248.
- Leclonkiewicz, J., Ktori, M., and Crepaldi, D. (2020). Morphemes as letter chunks: Discovering affixes through visual regularities. *Journal of Memory and Language*, page 104152.
- Longtin, C.-M., Segui, J., and Hallé, P. (2003). Morphological priming without morphological relationship. *Language and Cognitive Processes*, 18(3):313–334.
- Marelli, M., Amenta, S., and **Crepaldi**, t. (2015). Semantic transparency in free stems: the effect of Orthography–Semantics Consistency in word recognition. *Quarterly Journal of Experimental Psychology*, 68(1571–1583).
- Rastle, K., Davis, M., and New, B. (2004). The broth in my brother’s brothel: Morpho-orthographic segmentation in visual word recognition. *Psychonomic Bulletin and Review*, 11(6):1090–1098.
- Tafazoli, S., Safaai, H., De Franceschi, G., Rosselli, F. B., Vanzella, W., Riggi, M., Buffolo, F., Panzeri, S., and Zoccolan, D. (2017). Emergence of transformation-tolerant representations of visual objects in rat lateral extrastriate cortex. *ELife*, (e22794).

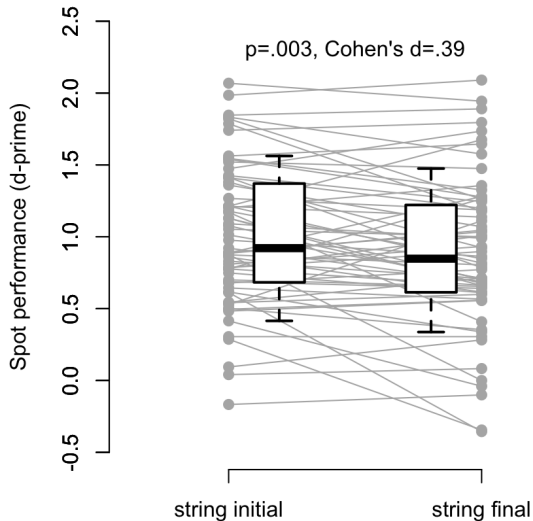
References III

- Vidal, C., Content, A., and Chetail, F. (2017). BACS: The Brussel Artificial Character Set for studies in Cognitive Psychology and Neuroscience. *Behavior Research Methods*, 49:2093–2112.
- Vinken, K. and Op De Beeck, H. (2021). Using deep neural networks to evaluate object vision tasks in rats. *PLoS Computational Biology*, 17(e1008714).

String initial chunks are easier to detect



String initial chunks are easier to detect



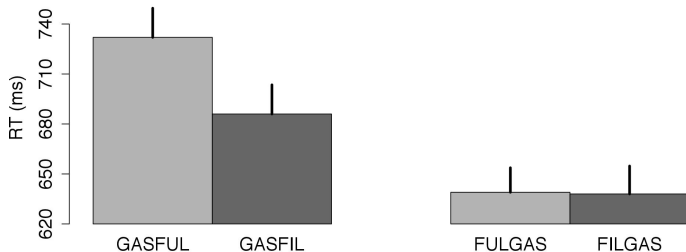
Orthographic coding

"I will assume that most of the information used by skilled readers to silently read words for meaning concerns information about abstract (i.e., case and font independent) letter identities, plus information about letter positions – in other words, orthographic information"

(Grainger, 2018)

Blind to suffixes?

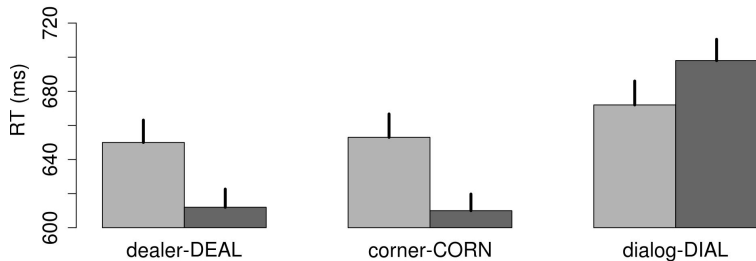
- ▶ (GASFUL vs. GASFIL) vs. (FULGAS vs. FILGAS)



(Crepaldi et al., 2010)

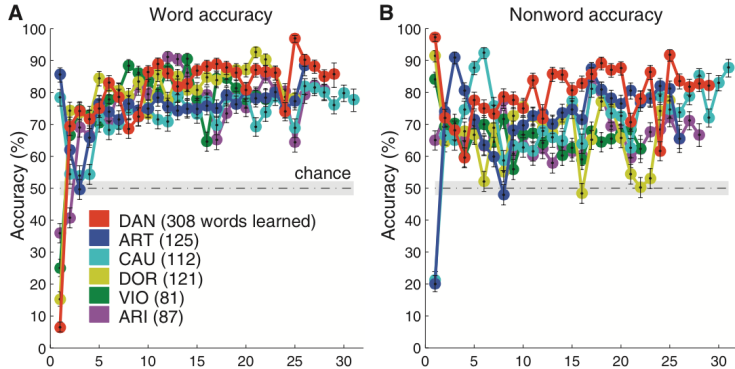
Corners that corn

dealer-DEAL vs. corner-CORN vs. dialog-DIAL



(Rastle et al., 2004)

Orthographic coding in Baboons



(Grainger et al., 2012)

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

The Statistical Learning principle

- ▶ Find regularities in low-level objects. . .
- ▶ . . . and build higher-level units based on this regularities

