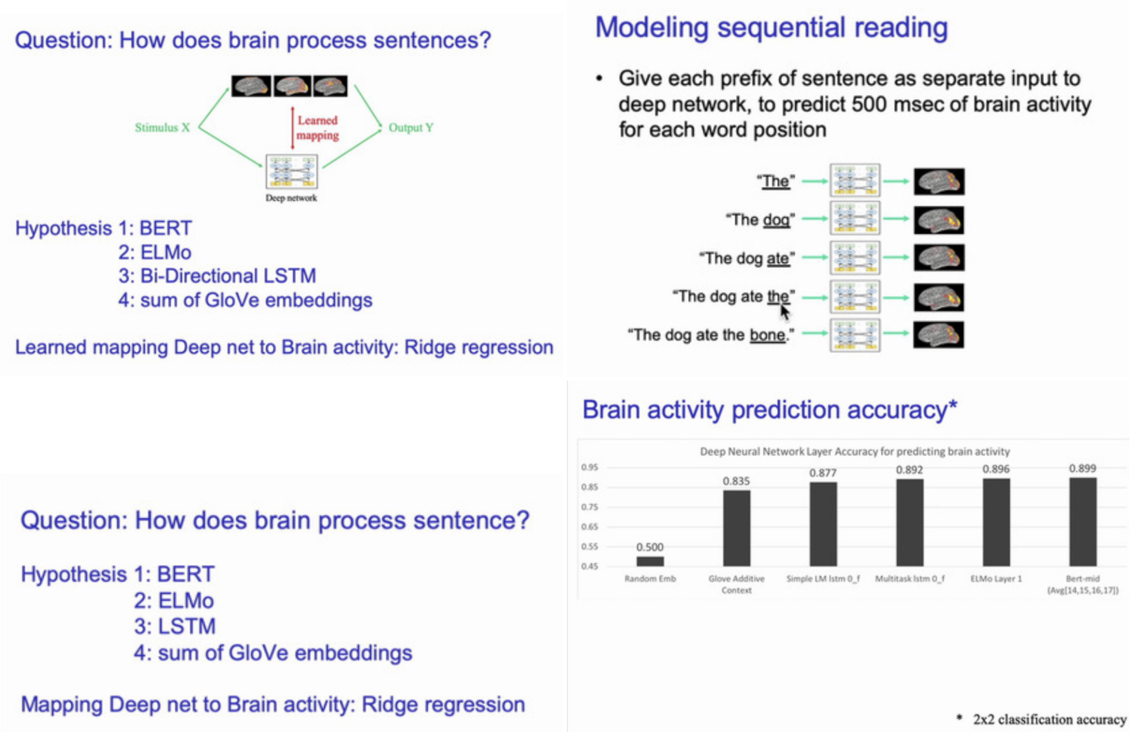
- Date: 2020-0128
- Author: 浅川伸一
- FileName: 2020-0128cnps_video.md
- Note: 勉強会ビデオ会議資料，BERT その他

cnps 勉強会資料 2020-0128

導入

- Googleの検索エンジンに「過去5年で最大の飛躍」。新たな言語処理モデル「BERT」の秘密 (Original)
- Context and Compositionality in Biological and Artificial Neural Systems in neurips2019
  1. Tom Mitchell
     - 2019年は2018年までの世界とは異なる (45:40 くらいから)
     - (1:01:40くらい) おそらく我々(人類)は今特別な時点にいる。脳が数千年に渡って実行してきた機能を実行する深層学習モデルを持ったからだ。



1. Yoshua Bengio (20:20 くらいから) システム 1 とシステム 2 とを結びつける鍵は注意である。

用語集

- SD (Semantic Differential), LSA (Latent Semantic Analysis), SVD (Singular Value Decompositon), LDA (Latent Direchlet Allocation), NMF (Non-negative Matrix Factorization)
- seq2seq, BERT
  - BERT = masked language model + transformer (self attention) + position encoder,
- Language model, SRN (Simple recurrent networks), BiRNN (bidirectiornla RNN), LSTM (Long short-term memory), VAE (variational auto-encoder)

実習用資源:

- TensorFlow Hub
- Seedbank colab のサンプル集

資料

- 2019 GAUSS G 検定講座講演資料
- 2019シンギュラリティサロン
- Pytorch 版 colab

リンク

- The Annotated Transformer
- Illustrated transformer
- The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning
  - T2T notebook
  - Deconstructing BERT: Distilling 6 Patterns from 100 Million Parameters
  - Deconstructing BERT, Part 2: Visualizing the Inner Workings of Attention
    - 注意の視覚化ツール
    - Visualizing Attention in Transformer-Based Language Representation Models, YouTube
    - A Multiscale Visualization of Attention in the Transformer Model, video
    - Analyzing the Structure of Attention in a Transformer Language Model
- https://github.com/tensorflow/tensor2tensor
- https://github.com/huggingface/pytorch-pretrained-BERT

自然言語処理諸課題における各モデルの性能比較 GLUE leaderboad

- GLUE
- super GLUE

Noris (2013)

2013Norris table 1

Trends in Cognitive Sciences October 2013, Vol. 17, No. 10

Table 1. Major computational models of reading organised in terms of their primary focus[a,b]

Model	Style	Task	Phenomena	Large lexicon
Models of visual word recognition				
IA [11,22]	IA	PI	Word-superiority effect	
Multiple read-out [3]	IA	PI, LD	Word-superiority effect	
SCM [2]	IA	LD, MP	Letter order	
BR [4–6]	Math/comp	LD, MP	Word frequency, letter order, RT distribution	√
LTRS [8]	Math/comp	MP, PI	Letter order	
Overlap [66]	Math/comp	PI	Letter order	
Diffusion model [30]	Math/comp	LD	RT distribution, word frequency	
SERIOL [7]	Math/comp	LD, MP	Letter order	
Models of reading aloud				
CDP++ [13]	Localist/symbolic	RA	Reading aloud	√
DRC [12]	IA	RA, LD	Reading aloud	
Triangle [24,25]	Distributed connectionist	RA	Reading aloud	
Sequence encoder [15]	Distributed connectionist	RA	Reading aloud	√
Junction model [50]	Distributed connectionist	RA	Reading aloud	√
Models of eye-movement control in reading				
E-Z reader [17,18]	Symbolic	R	Eye movements	
SWIFT [19]	Symbolic	R	Eye movements	
Model of morphology				
Amorphous discriminative learning [16]	Symbolic network	Self-paced reading, LD	Morphology	√

[a]The table also indicates the modelling style or framework, the main task that the model simulates, the main phenomena that the model simulates (not exhaustive), and whether the model uses a realistically sized lexicon. Note that the review concentrates on 'Models of visual word recognition'.

[b]Abbreviations: Math/comp, mathematical or computational; LD, lexical decision; PI, perceptual identification; RA, reading aloud; MP, masked priming; R, natural reading.

Logogen and Triangle models

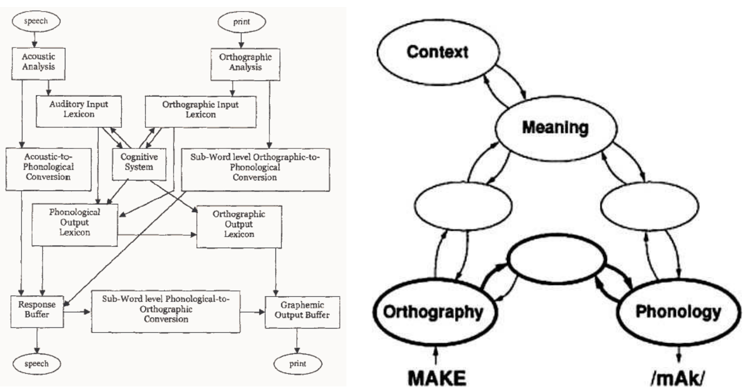
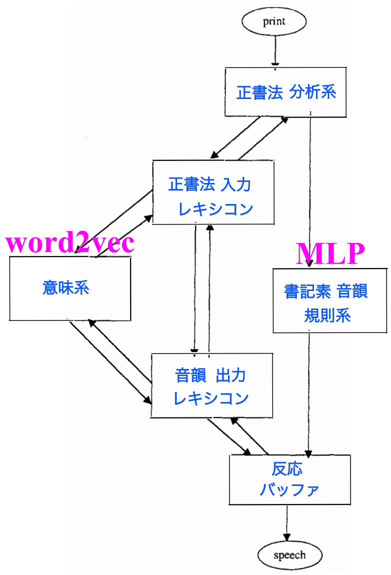


Figure 1. Two diagrams of the lexical processing system. The one on the left is taken from Patterson and Shewell (1987), and the one on the right is taken from Seidenberg and McClelland (1989). Lexical and sublexical pathways of processing can be found in both diagrams.



CDP++

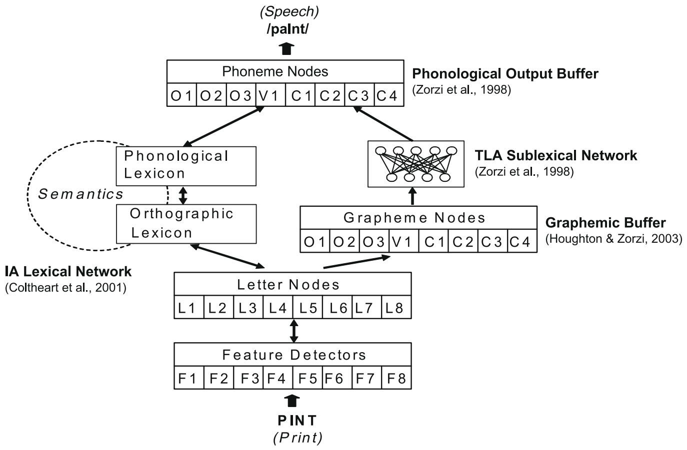


Fig. 2. The overall architecture of CDP+. Note: Numbers shown inside the various layers index slot positions, whereas letters indicate the type of representation (f = features, l = letter, o = onset, v = vowel, c = coda).

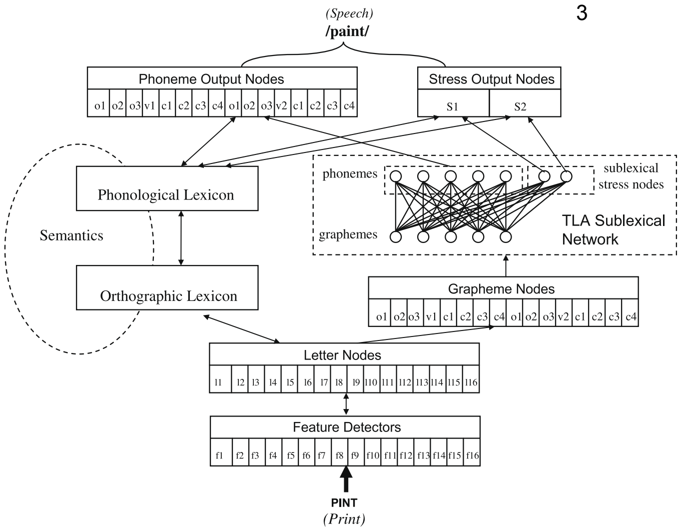


Fig. 3. The overall architecture of CDP++. Note: Numbers shown inside the various layers index slot positions, whereas letters indicate the type of representation (f = feature, l = letter, o = onset, v = vowel, c = coda). S1 = first syllable stress; S2 = second syllable stress.

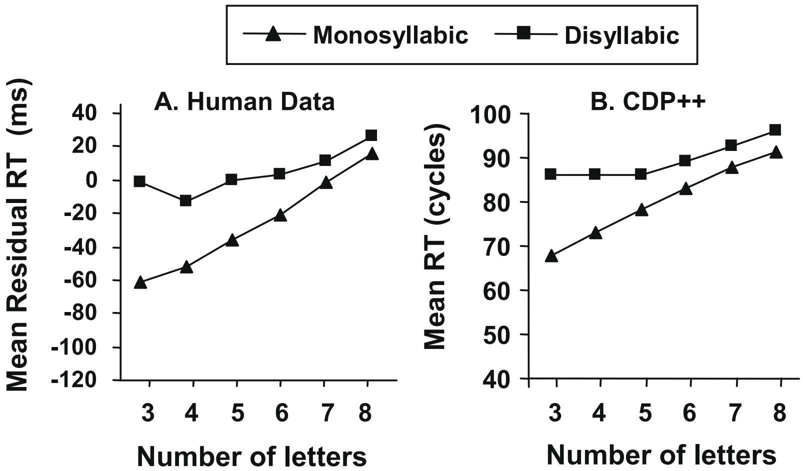


Fig. 4. Mean human and CDP++ reaction times (RTs) of monosyllabic and disyllabic words on the full ELP (2007) database.

Table 1
List of monosyllabic benchmark effects (from Perry et al. (2007)). Tick marks indicate successful simulations (for details, see Appendix D).

Name of effect	Description	CDP+	CDP++
Frequency	High-frequency words are faster/more accurate than low-frequency words	✓	✓
Lexicality	Words are faster/more accurate than pseudowords	✓	✓
Length × lexicality	Nonword naming latencies increase linearly with each additional letter	✓	✓
Frequency × regularity	Irregular words are slower/less accurate than regular words. This effect is typically larger for low-frequency words (Paap and Noel, 1991) but has also been reported for high-frequency words (Jared, 2002)	✓	✓
Word consistency	Inconsistent words are slower/less accurate than consistent words. The size of the effect depends on the friend–enemy ratio	✓	✓
Nonword consistency	Nonword pronunciations show graded consistency effects; that is, people do not always use the most common grapheme–phoneme correspondences	✓	✓
Position of irregularity	The size of the regularity effect is bigger for words with first position irregularities (e.g., *chef*) than for words with second or third position irregularities	✓	✓
Body neighborhood	Words with many body neighbors are faster/more accurate than words with few body neighbors	✓	–
Pseudohomophone advantage	Nonwords that sound like real words (e.g., *bloo*) are faster/more accurate than orthographic controls	✓	✓
Surface dyslexia	Patient MP showed a specific irregular word reading impairment that was modulated by the consistency ratio of the words as well as their frequency	✓	✓
Phonological dyslexia	Patient LB showed a specific nonword reading impairment which was reduced with pseudohomophones orthographically similar to their base words	✓	✓
Masked priming	Words preceded by a masked onset prime are faster/more accurate than words preceded by unrelated primes	✓	✓

Table 2
Percentage of variance accounted for (R^2) by CDP++, CDP+ (Perry et al., 2007), CDP (Zorzi et al., 1998a), the Triangle model (Plaut et al., 1996), and the DRC (Coltheart et al., 2001) on the Spieler and Balota (SB, 1997), Balota and Spieler (BS, 1998), Treiman et al. (1995), and Seidenberg and Waters (SW, 1989) databases.

Database	Models				
	CDP++	CDP+	CDP	Triangle	DRC
SB (1997)	19.5	17.3	5.9	3.3	3.7
BS (1998)	24.0	21.6	6.7	2.9	5.5
Treiman	18.1	15.9	6.5	3.3	4.8
SW	10.9	9.6	2.7	3.0	6.1

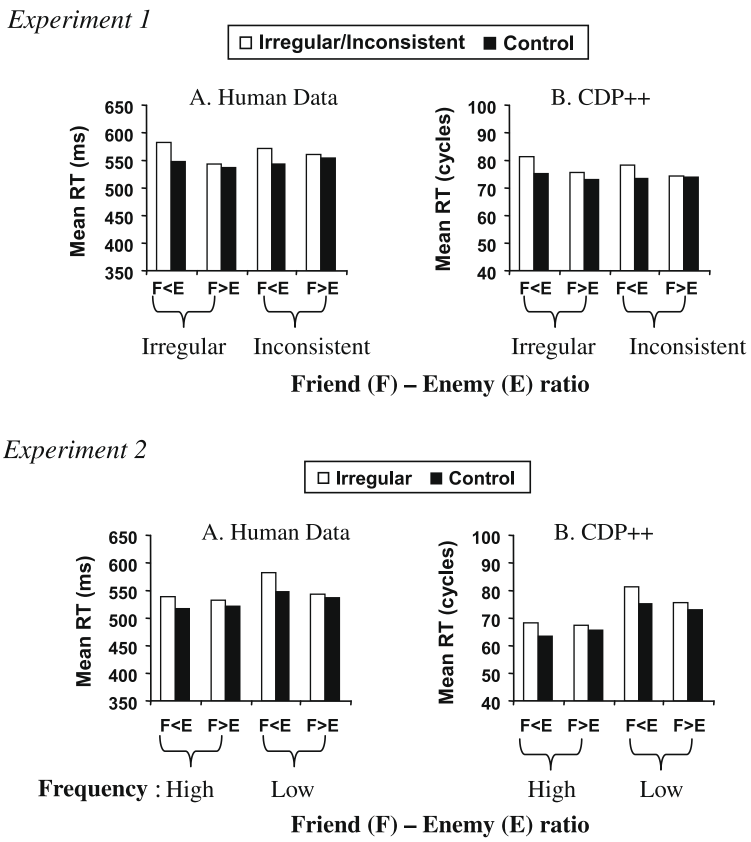


Fig. D1. Human data (milliseconds) and CDP++ simulations (cycles) of Jared's (2002) Experiment 1 and Experiment 2.

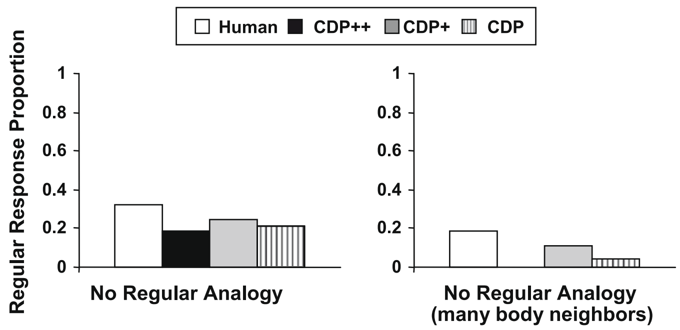


Fig. D2. Human data (response probabilities for regular pronunciations) and simulations of different models for the "no regular analogy nonwords" (Experiment 1) and the "no regular analogy with many body neighbors nonwords" (Experiment 2) of Andrews and Scarratt (1998).

Junction model

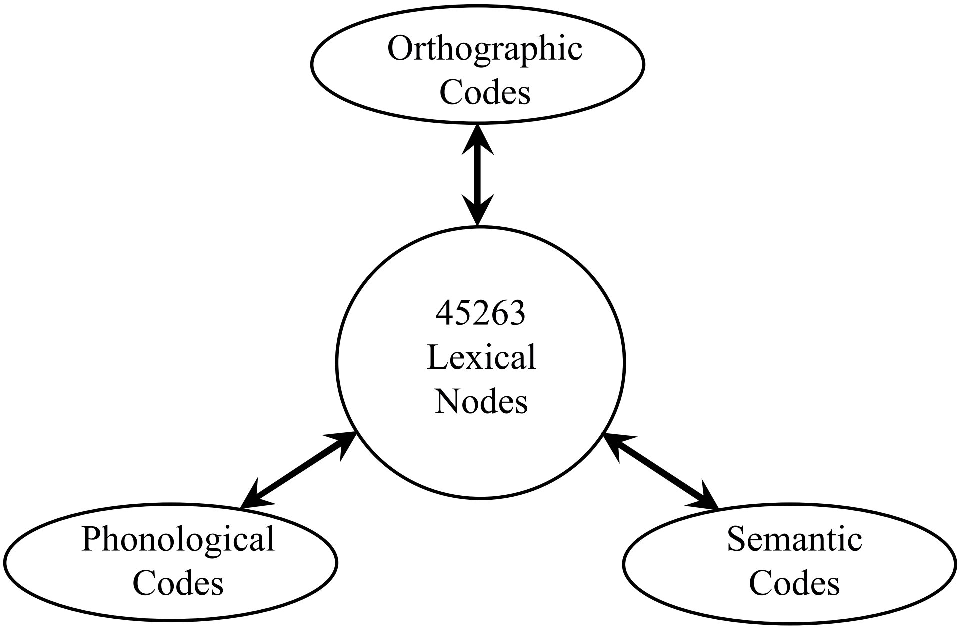


Figure 5. Basic architecture of the large-scale junction model.

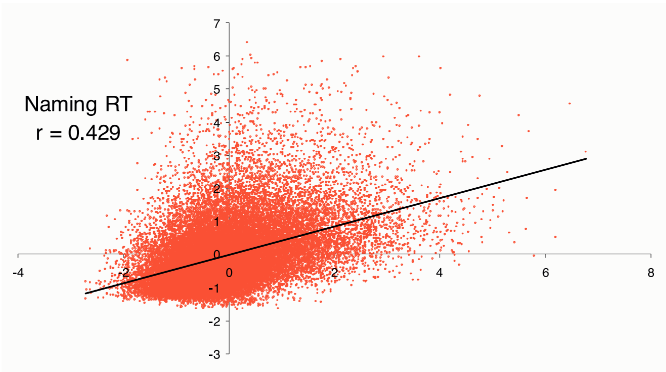


Figure 9. Model mean response times plotted against the naming response time residuals from the Elexicon database, in normalized coordinates.

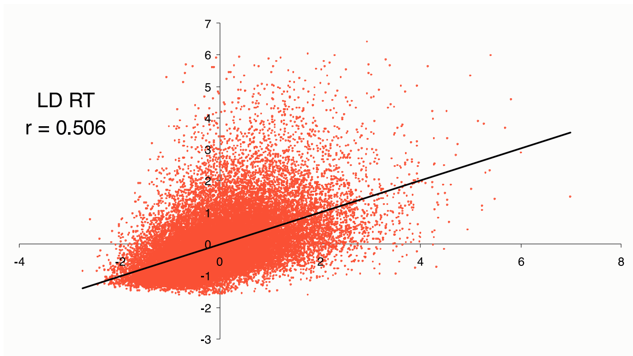


Figure 10. Model mean response times plotted against the lexical decision response time residuals from the Elexicon database, in normalized coordinates.

	DRC comparison N = 5190		PMSP comparisons N = 2808				
	Junction	DRC	Junction	Sim 1	Sim 2	Sim 3	Sim 4
R^2	12.2%	5.1%	14.7%	5.2%	4.1%	2.1%	11.9%

Table 1. Proportions of variance in naming response times accounted for by the junction model, compared with the DRC and PMSP models

Sequence encoder

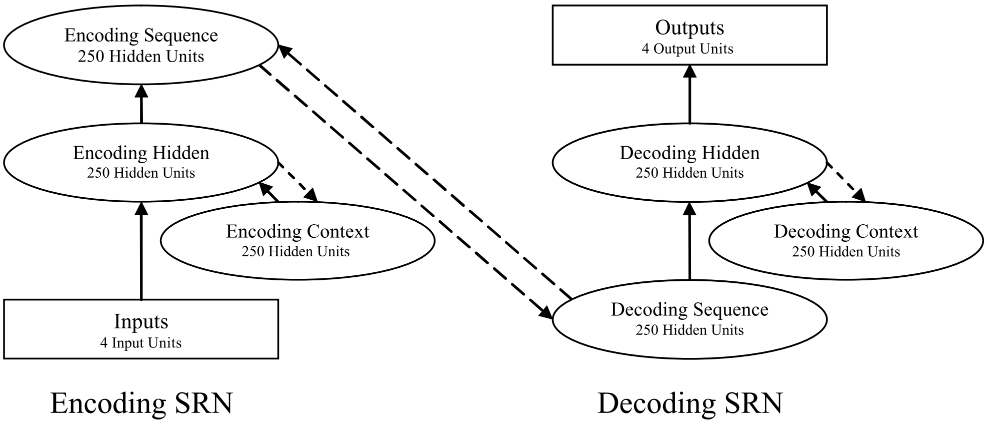


Fig. 1. Diagram of the sequence encoder architecture, with numbers of units used in Simulation 1 shown for each group. *Note:* These numbers were determined by trial and error to be sufficient to support near asymptotic performance on the training sequences. Solid arrows denote full connectivity and learned weights, and dashed arrows denote one-to-one copy connections. Rectangular groupings denote external (prescribed) representations coded over localist units, and oval groupings denote internal (learned) representations distributed over hidden units. SRN = simple recurrent network.

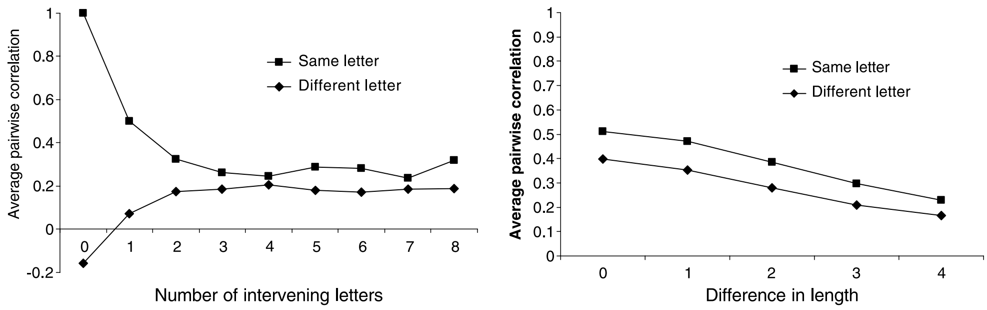


Fig. 2. Average pairwise correlations between conjunction patterns, plotted as a function of intervening letters (left) or difference in wordform length (right). *Note:* For intervening letters, the effect of wordform length was partialled out before correlations were computed.

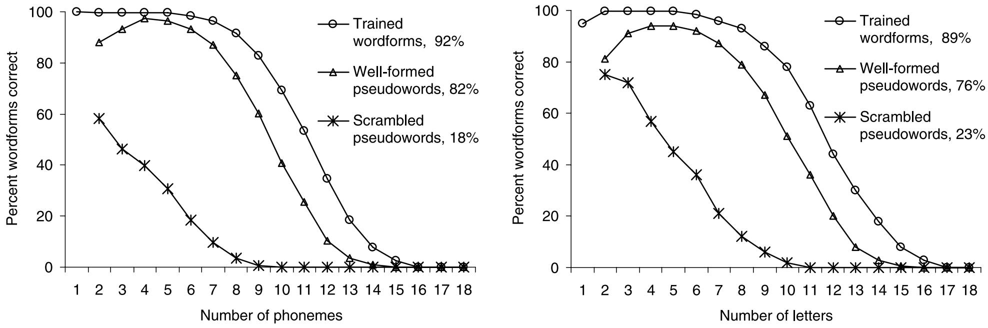


Fig. 3. Correct percentages for Simulations 2a and 2b, plotted as a function of wordform length and type.