Examining strains and symptoms of the ‘Literacy Virus’

The effects of orthographic transparency on phonological processing in a connectionist model of reading

Alastair Smith\textsuperscript{1,2,*}, Padraic Monaghan\textsuperscript{3} & Falk Huettig\textsuperscript{1,4}

\textsuperscript{1}. Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
\textsuperscript{2}. International Max Planck Research School for Language Sciences, Nijmegen, The Netherlands
\textsuperscript{3}. Department of Psychology, Lancaster University, Lancaster, United Kingdom
\textsuperscript{4}. Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, The Netherlands

* Corresponding Author: alastair.smith@mpi.nl
The ‘Literacy Virus’

“The virus infects all speech processing, as now whole word sounds are automatically broken up into sound constituents.” (Frith, 1998)

Phonological awareness (explicit tasks):

- **Children**: Pre-literate $<$ Literate
  - e.g. Alcock et al. 2010; De Jong & Van Der Leij, 2003; Hulme et al. 2005; Morrison et al. 1995; Trieman & Zukowski, 1991

- **Adult**: Illiterate $<$ Literate
  - e.g. Adrian et al. 1995; Loureiro et al. 2004; Morais et al. 1979; Sciliar-Cabral et al. 1997
The ‘Literacy Virus’

“The virus infects all speech processing, as now whole word sounds are automatically broken up into sound constituents.” (Frith, 1998)

Phonological awareness (explicit tasks):

– **Children: Pre-literate < Literate**
  » e.g. Alcock et al. 2010; De Jong & Van Der Leij, 2003; Hulme et al. 2005; Morrison et al. 1995; Trieman & Zukowski, 1991

– **Adult: Illiterate < Literate**
  » e.g. Adrian et al. 1995; Loureiro et al. 2004; Morais et al. 1979; Sciliar-Cabral et al. 1997

Effects on ‘Online’ speech processing?

– **Categorical Perception** (Serniclaes et al. 2005)
  • Sharper boundary precision (ba-da) shown by literates
  • Correlates with reading (Hoonhorst et al. 2011)

– **Visual World Paradigm** (Huettig, Singh & Mishra, 2011)
  • Illiterates eye gaze NOT time locked to phonological overlap in speech signal
  • Compatible with literacy resulting in finer grained phonological processing (Smith, Monaghan & Huettig, 2013)
Phonological Restructuring Hypothesis

- Phonological processing regions restructured as a result of learning orthographic <-> phonological mappings

- Orthographic effects observed in time windows classically viewed as preceding lexical access
  - Lexical Decision: Perre et al. 2008; 2009a; 2009b

- Effects of literacy on speech processing localised to phonological processing regions:
  - TMS: Pattamadilok et al. 2010
Harm & Seidenberg, 1999

Phonological restructuring captured explicitly in computational model of reading
Harm & Seidenberg, 1999

Phonological restructuring captured explicitly in computational model of reading

3 versions of the model

1. Pre-Literate  2. Illiterate  3. Literate

Fig: Model Architecture
Harm & Seidenberg, 1999

Phonological restructuring captured explicitly in computational model of reading

3 versions of the model

1. Pre-Literate                         2. Illiterate                               3. Literate

Fig: Model Architecture
Harm & Seidenberg, 1999

Phonological restructuring captured explicitly in computational model of reading

3 versions of the model

1. Pre-Literate  2. Illiterate  3. Literate

**Phonetic Feature Restoration Error:** Pre-Literate > Illiterate > Literate
Harm & Seidenberg, 1999

Phonological restructuring captured explicitly in computational model of reading

3 versions of the model

1. Pre-Literate 2. Illiterate 3. Literate

Phoneme Segment Restoration: Pre-Literate > Illiterate > Literate
Phonological restructuring captured explicitly in computational model of reading

3 versions of the model

1. Pre-Literate
2. Illiterate
3. Literate

Connection Weights:

Literate model resulted in stronger connections within phonemes and weaker connections between phonemes.

=> The granularity of processing was affected by orthographic training.

Difference between weights in literate and illiterate network averaged over phoneme segment.
Effect of Orthographic Transparency

Orthographic transparency predicted to modulate the effects of literacy on phonological processing:

- **Psycholinguistic Grain Size Theory (Ziegler & Goswami, 2005)**
  - Nature of correspondence between orthographic units and phonological units determines the grain size of processing for the given language

- **Harm & Seidenberg, 1999**
  - Weights/processing within the phonological attractor network will be sensitive to regularities/irregularities in the orthographic -> phonological mapping
Orthographic Systems

• Orthographic Systems:
  – Alphabetic
  – Alphasyllabic
  – Consonantal
  – Syllabic
  – Logographic
Orthographic Systems

• Orthographic Systems:
  – Alphabetic
    • Shallow
      – e.g. Serbian: each letter perfectly corresponds to a speech sound
  – Consonantal
  – Alphasyllabic
  – Syllabic
  – Logographic

Orthographic Systems

• Orthographic Systems:
  – Alphabetic
  – Consonantal
  – Alphasyllabic
  – Syllabic
  – Logographic
    – basic unit = the morpheme
    e.g. Mandarin
      人  木  湖  壶
      rén  mù  hú  hú
      ‘person’ ‘tree’ ‘lake’ ‘pot’
Orthographic Systems

Empirical evidence for effects of orthographic transparency on phonological processing:

- e.g. Shu et al. 2008
  - phoneme onset and tone awareness in Chinese children coincides with Pinyin tuition

Research Question:

Is phonological restructuring affected by the structure of the orthographic system?

Study 1: Does transparency modulate phonological restructuring in Harm & Seidenberg (1999)?

1) Model trained on semi-transparent corpus (English)
2) Model trained on non-transparent corpus (randomised orthographic ↔ phonological mappings)

Study 2: How do restructuring effects vary across the world’s major orthographic systems?

- Harm & Seidenberg (2004) model of reading
- Model trained on 7 orthographic systems

Orthographic Systems

1. Deep Alphabetic
2. Shallow Alphabetic
3. Alphasyllabic
4. Consonantal
5. Syllabic
6. Logographic (semantically non-transparent)
7. Logographic (semantically semi-transparent)
Does transparency modulate phonological restructuring in Harm & Seidenberg (1999)?

**Corpus:** 6,188 monosyllabic English words.

**Training:**

- **Pre-literacy:** Phonological Retention trials
- **Literacy:** Orthography -> Phonology (p=0.8)
  Phonological Retention (p=0.2)

**Literate Networks:**

a) **Transparent:**

  *Trained on English orthographic to phonological mapping*

a) **Non-transparent:**

  *Orthographic representations randomly reassigned to phonological representations*

<table>
<thead>
<tr>
<th>Network</th>
<th>Phonology Retention</th>
<th>Reading</th>
</tr>
</thead>
<tbody>
<tr>
<td>Literate (Transparent)</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Literate (Non-Transparent)</td>
<td>0.999</td>
<td>0.987</td>
</tr>
</tbody>
</table>
Phonetic Feature Restoration

Networks trained on a transparent corpus re-activated corrupted phonological features more accurately than networks trained on non-transparent corpora [$\mu = -0.013$, $\sigma = 0.005$, $t(3) = -4.747$, $p = 0.018$].
Segment Restoration
Phoneme

The Euclidian distance between the phoneme segment produced and the nearest phoneme neighbour.

Networks trained on a transparent corpus restored corrupted phonemes marginally better than those trained on a non-transparent corpus [$\mu = -0.045, \sigma = 0.029, t(3) = -3.054, p = 0.055]$. 
Segment Restoration
Sub-syllabic Units

**Measure:**

*Restored segment (onset, vowel or coda) examined to assess whether it exists as a ‘legal’ unit within the corpus.*

**Overall**

Networks trained on a transparent corpus made more illegal segment restorations than those trained on a non-transparent corpus 
[$\mu = -0.102$, $\sigma = 0.012$, $t(3) = -16.492$, $p < 0.001$].
Connection Weights
Literate (Transparent) - Literate (Non-transparent)
Connection Weights

Literate (Transparent) - Literate (Non-transparent)
Study 1: Summary & Conclusions

Results Summary:

- Phonetic feature restoration: Transparent > Non-transparent
- Phoneme restoration: Transparent > Non-transparent
- Valid segment restoration: Transparent < Non-transparent

Conclusions:

Each systems phonological processing is affected differently by literacy training:

- Transparent model: Phoneme level
- Non-transparent model: Word level
Study 2: How may effects vary across the world’s major orthographic systems?

- Orthographic systems vary in the degree to which they encode a language's phonological (and semantic) structure.
- Model trained on 7 orthographic systems
  - Systems differed only in the structure of their orthographic representations
  - Identical phonological and semantic representations and mappings used across orthographic systems.

Fig: Model derived from Harm & Seidenberg’s 2004 implementation of the triangle model of reading.
Study 2: How may effects vary across the world’s major orthographic systems?

- Orthographic systems vary in the degree to which they encode a language’s phonological (and semantic) structure.

- Model trained on 7 orthographic systems:
  - Systems differed only in the structure of their orthographic representations.
  - Identical phonological and semantic representations and mappings used across orthographic systems.

![Diagram](Fig: Model derived from Harm & Seidenberg’s 2004 implementation of the triangle model of reading.)
Study 2: How may effects vary across the world’s major orthographic systems?

- Orthographic systems vary in the degree to which they encode a language’s phonological (and semantic) structure.
- Model trained on 7 orthographic systems
  - Systems differed only in the structure of their orthographic representations
  - Identical phonological and semantic representations and mappings used across orthographic systems

Fig: Model derived from Harm & Seidenberg’s 2004 implementation of the triangle model of reading.
Study 2: How may effects vary across the world’s major orthographic systems?

- Orthographic systems vary in the degree to which they encode a language’s phonological (and semantic) structure.

- Model trained on 7 orthographic systems:
  - Systems differed only in the structure of their orthographic representations.
  - Identical phonological and semantic representations and mappings used across orthographic systems.

Fig: Model derived from Harm & Seidenberg’s 2004 implementation of the triangle model of reading.
Study 2: How may effects vary across the world’s major orthographic systems?

- Orthographic systems vary in the degree to which they encode a language’s phonological (and semantic) structure.

- Model trained on 7 orthographic systems
  - Systems differed only in the structure of their orthographic representations.
  - Identical phonological and semantic representations and mappings used across orthographic systems.

Fig: Model derived from Harm & Seidenberg’s 2004 implementation of the triangle model of reading.
Study 2: How may effects vary across the world’s major orthographic systems?

- Orthographic systems vary in the degree to which they encode a language’s phonological (and semantic) structure.
- Model trained on 7 orthographic systems:
  - Systems differed only in the structure of their orthographic representations.
  - Identical phonological and semantic representations and mappings used across orthographic systems.

![Diagram of the model](Fig: Model derived from Harm & Seidenberg’s 2004 implementation of the triangle model of reading.)
Artificial Corpora

Artificial corpora contain 500 unique items each assigned:

A Unique Semantic Representation
• Semantic representations encoded by 150 unit binary feature vectors

Phonological Representation
• Monosyllabic word
• 5 phonemes in length (CCVCC)
• 50 unit binary feature vector

Orthographic Representation
• Description of each orthographic system is taken from Comrie (2013).
• 50 unit binary feature vectors
Orthographic Representations

1. **Alphabetic:** The basic unit of representation is the phoneme.
   
   a) *Alphabetic Shallow (e.g. Finnish):* Complete correspondence between grapheme and phoneme.
   
   b) *Alphabetic Deep (e.g. English):* 20% of mappings for each vowel will be irregular.

2. **Consonantal:** Only consonants are represented (e.g. Hebrew*)

3. **Alphasyllabic:** Basic grapheme indicates a consonant with diacritic added to indicate a combination of given vowel with a given consonant. (e.g. Thai)

4. **Syllabic:** Distinct grapheme for each syllable. (e.g. Japanese hiragana)

5. **Logographic (e.g. Chinese):**
   
   a) *Logographic Opaque:* Basic representational unit is the morpheme.
   
   b) *Logographic Semantically Transparent:* Orthographic features are more likely to be shared with items within the same semantic category.

* In Hebrew it is possible to add diacritics to indicate vowels
Reading Task
(Orthography -> Semantics)
Reading Task
(Orthography -> Phonology)
Variation in Consonant Representation

Measure: Mean distance between all instances of a given consonant
Variation in Vowel Representation

Measure: Mean distance between all instances of a given vowel

Orthographic System

Literacy Training Onset
Variation in Vowel Representation

Measure: Mean distance between all instances of a given vowel
Study 2: Summary

• Model captures graded effects of transparency on reading acquisition

• Results predict effects of orthographic structure on phonological representation
  – Networks trained on logographic & syllabic systems display greater variation in consonant representation
  – Networks trained on consonantal systems display greater variation in vowel representation
Conclusions & Predictions

• Study 1:
  – Effects of literacy training on phonological processing modulated by orthographic transparency

• Study 2:
  – The consequences of phonological restructuring vary across the world’s major orthographic systems

• Predictions:
  – Behavioural and neural consequences of literacy training for phonological processing dependant on the orthographic structure on which the system is trained.
Does training on orthographic mappings lead to richer phoneme level representation?

- Validity of illiterate (baseline) measure
  - Illiterate effects in Harm & Seidenberg (1999) may result from lack of exposure to noisy phonological input

- Do we potentially underestimate effects?
  - Noisy speech signal (e.g. co-articulation, elision, assimilation, reductions)

  - Emergent phonological representations required
Artificial Corpora

Artificial corpora to contain 500 unique items each assigned:

A Unique Semantic Representation
• Semantic representations encoded by 150 unit binary feature vectors, with $p(active) = 0.1$.
• Prototypes used (cf. Dilkina, McClelland & Plaut, 2010) to construct an artificial semantic taxonomy
• 2 high-level semantic categories, each with 5 sub-categories [50 items per sub category].
• Items more likely to share semantic features with items within the same semantic category than those outside their category.

Phonological Representation
• Monosyllabic word
• 5 phonemes in length
• Of the form CCVCC.
• The phonological layer therefore consists of 5 phoneme slots, organised CCVCC, each 10 units in length. 25 homophones included in the training corpus.

Orthographic Representation
• Description of each orthographic system is taken from Comrie (2013).
• 50 unit binary feature vectors
• feature $p( active ) = 0.5$. 
**Systematicity:**

- The basic unit of representation is the phoneme.

**Patterns:**

- The orthographic layer is defined in terms of 5 letter slots, organised CCVCC, each 10 units in length.
- All words consist of 5 letters taken from an alphabet of 15 letters, 5 letters representing vowels and 10 representing consonants.

**Alphabetic Shallow (e.g. Finnish)**

- Complete correspondence between grapheme and phoneme.
- Each letter within the alphabet is assigned a phoneme. Orthographic representations are created using this fixed mapping.

**Alphabetic Deep (e.g. English)**

- 20% of mappings for each vowel will be irregular.
- For each vowel, letter -> vowel mappings will be randomly reassigned for 20% of items.
Consonantal

Systematicity:
• Only consonants are represented.
• Same consonant sequence can be several different words.

Patterns:
• 4 letters, organised CCCC
• An alphabet of 10 letters constructed, each representing one of the 10 consonants in the phoneme inventory.
• Each letter encoded by a 12 unit binary vector.

e.g. Hebrew

 Arabic  b.  hebrew  c.  hebrew
 hp       hop      hep
Alphasyllabic

Systematicity:

• Basic grapheme indicates a consonant with diacritic added to indicate a combination of given vowel with a given consonant.

Patterns:

• 4 slots, C[CV]CC
• C slots defined by 12 units
• [CV] slots defined by 14 units
• Prototypes produced for each consonant and used to create 5 unique 20 unit vectors for each consonant, each of which represents the consonant in combination with a given vowel.

e.g. Thai

ม มี มุ ม่า เม เมะ

m mi mu ma: me: mə:
Syllabic

Systematicity:
- Distinct grapheme for each syllable.
- No part can be identified as denoting distinct phonemes.

Patterns:
- Single slot defined by 50 units.
- Unique 50 unit binary feature for each syllable/word (CCVCC).

e.g. Japanese (hiragana)

ひ へ に ね

hi  he  ni  ne
Logographic

Logographic Semantically Non-Transparent:

Systematicity: Basic representational unit is the morpheme.

Patterns: Unique 50 unit binary feature vector randomly constructed for each item in the corpus.

Logographic Semantically Semi-transparent:

Systematicity: Same as ‘logographic semantically opaque’ implementation yet also includes subtle relationship between orthography and semantics

Patterns: Orthographic representation consists of unique 50 unit binary feature vector First 30 features are more likely to be shared with items within the same semantic category.

e.g. Mandarin

人 木 湖 壺
rén mù hú hú

‘person’ ‘tree’ ‘lake’ ‘pot’
Training

*(Similar procedure to Harm & Seidenberg, 2004)*

Systems differed only in the structure of their orthographic representations

- Identical phonological and semantic representations and mappings used across orthographic systems

- Pre-literate training (initial 150000 trials)
  - phonology retention task [10% of trials]
  - semantic retention task [10% of trials]
  - phonology to semantics task [“comprehension”, 40% of trials]
  - semantics to phonology task [“production”, 40% of trials].

- Literacy training (450000 trials):
  - phonology retention task [5% of trials]
  - semantic retention task [5% of trials]
  - phonology to semantics task [“comprehension”, 25% of trials]
  - semantics to phonology task [“production”, 25% of trials]
  - orthography -&gt; semantics and phonology [40% of trials].
Variation in Consonant Representation

Measure: Mean distance between all instances of a given consonant

![Graphs showing consonant distance](image)
Variation in Vowel Representation

Measure: Mean distance between all instances of a given vowel
To be continued....