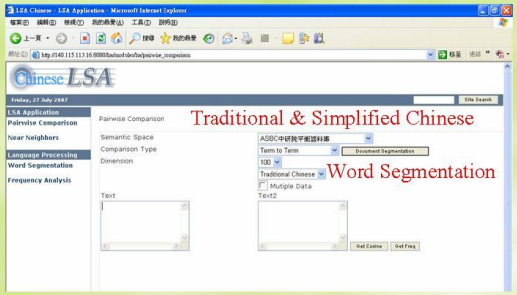
Estimating Word's Predictability on Lexical Processing using Latent Semantic Analysis – Verification from Eye Movement Data

Hsueh-Cheng Wang[1], Minglei Chen[1], Hwawei Ko[1], and Walter Kintsch[2]
[1] National Central University & [2] Colorado University at Boulder

Word Predictability & Lexical Processing

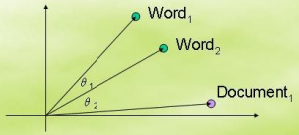
- E-Z Reader Model (Reichle, et al., 2003)
  - Early Stage (L1)
    - Orthographic Form
    - Familiar Check
    - $t(L_1)=[\beta_1 - \beta_2 \ln(\text{frequency})](1-\theta\ \textbf{predictability})$
  - Late Stage (L2)
    - Phonological / Semantic Form
    - Completion of Lexical Access
    - $t(L_2)=\Delta[\beta_1 - \beta_2 \ln(\text{frequency})](1-\textbf{predictability})$
  - Predictability Effect is stronger on L2 than on L1
- SWIFT Model (Engbert, et al., 2002)
  - $L_n = (1-p_n) I_n,\ I_n = \alpha - \beta \log(f_n)$
- High-predictable words are read faster than low-predictable words.
  - First Fixation Duration
  - Gaze Duration
  - Total Time
- High-predictable words are skipped more and re-fixated less than low-predictable words.
  - Skip Rate
  - Regression
- Chinese reader exploit target word predictability during reading. (Rayner, et al. 2005)

Latent Semantic Analysis (LSA)



Pairwise Comparison Application

- From ASBC (5M Chinese Words Corpus)
- Term-to-Document Co-occurrence Matrix
- 49021 Terms (Words)
- 40463 Documents (Paragraphs)
- Local & Global Weighting
- Singular Value Decomposition (SVD)
- High-Dimension Semantic Space
- Mean:0.067 Std: 0.115



Estimates of Predictability by LSA, Cloze Task and Transitional Probability

LSA vs. Cloze Task

- Re-analysis of Materials for P/U targets in Rayner, et al., 2004.
- Semantic Space of General English Reading (http://lsa.colorado.edu/)

Context_Pre	Context_Post	Target Word	Predictability	Freq	LSA
Most cowboys know how to ride a	if necessary.	horse	P	H	0.62
Most cowboys know how to ride a	if necessary.	camel	U	L	0.1
In the desert, many Arabs ride a	to get around.	horse	U	H	0.29
In the desert, many Arabs ride a	to get around.	camel	P	L	0.64

LSA CAN distinguish High- / Low- Predictability

LSA & Transitional Probability

- Previous Content Word vs. Target (LSA)
- Contextual Constraint (by LSA and TP)

他 在 沙漠 裡 騎 _____ 。
He desert ride

candidate	explanation	TP	LSA (沙漠)
著	[zhao1] catch/receive/suffer	0.231	0.35
馬	[ma3] horse/horse chess piece/Surname	0.139	0.77
腳踏車	[jiao3 ta4 che1] bicycle	0.134	0.07
⋮	⋮	⋮	⋮
駱駝	[luo4 tuo5] camel	0.004	**0.83**

Eye Movement Data vs. Estimated Predictability

Correlations Among Predictors

Variable	InFreq	WordLength	Strokes	LSA	TP
InFreq	—	-.639	-.570	.180	.306
WordLength		—	.044	-.128	-.318
Strokes			—	-.116	-.278
LSA				—	.301
TP					—

Frequency
Complexity
Predictability

Unstandardized Regression Coefficients

Variable	FFD Mean	FFD Std. D	GD Mean	GD Std. D	TotalTime Mean	TotalTime Std. D
WFreq	-5.2**	3.3	-6.6 **	3.9	-10.9**	9.7
WordLength	24.5**	8.7	35.1**	13.6	113.8**	31.5
Strokes	1.1**	0.7	1.5**	0.9	3.7**	1.7
LSA	*-5.2*	*14.4*	*-14.1*	*29.2*	-76.2**	55.5
TP	-56.1**	46.0	-43.5**	47.0	*36.6*	*174.5*

- LSA might be adopted to calculate contextual constraint and represent word predictability on lexical processing.

Poster and Oral Presentation in 14th European Conference of Eye Movements

ECEM2007
14th European Conference on Eye Movements
August 19-23, Potsdam, Germany