

Reading Time and Vocabulary Rating in the Japanese Language: Large-Scale Reading Time Data Collection Using Crowdsourcing

National Institute for Japanese Language and Linguistics, Japan
Tokyo University of Foreign Studies
Masayuki Asahara

Reading time × Vocabulary rating in the Japanese language

How do vocabulary rating scores affect reading time in the Japanese language?

Word Familiarity Survey

<https://github.com/masayu-a/WLSP-familiarity>

Word Familiarity rating using crowdsourcing

Rating five perspectives

KNOW, WRITE, READ, SPEAK, LISTEN

Stimuli: 84,114 surface forms from the Word List by Semantic Principles (WLSP)

Modeling using the Bayesian Linear Mixed Model

Lexical Random Factors → Word Familiarity

Subject Random Factors → Vocabulary Rating

以下の単語についてお答えください

意欲

単語の意味は知っていますか?

全く知らない あまり知らない
 どちらともいえない 何となく知っている
 よく知っている

どのくらい普段書いているものに出現しますか?

全く出現しない あまり出現しない
 どちらともいえない たまに出現する
 よく出現する

どのくらい普段読んでいるものに出現しますか?

全く出現しない あまり出現しない
 どちらともいえない たまに出現する
 よく出現する

どのくらい普段話すときに出現しますか?

全く出現しない あまり出現しない
 どちらともいえない たまに出現する
 よく出現する

どのくらい普段聞くときに出現しますか?

全く出現しない あまり出現しない
 どちらともいえない たまに出現する
 よく出現する

【参考情報: 026127-体・活動・心・欲望・期待・失望】

Survey date	Number of subjects	Number of answers
2018/11/15-21	3,391	1,617,215
2019/11/14-22	2,421	288,000
2020/10/09-12	2,372	943,295
2021/09/12-14	2,396	385,380

Reading Time Collection

<https://github.com/masayu-a/BCCWJ-SPR2>

Self-paced reading using crowdsourcing

Stimuli:

Whitepaper, Textbooks, Books in the Balanced Corpus of Contemporary Written Japanese (BCCWJ)

Modeling using:

Generalized Linear Mixed Model
Bayesian Linear Mixed Model

Reading Time ⇔ Vocabulary Rating

Stimuli and Participants

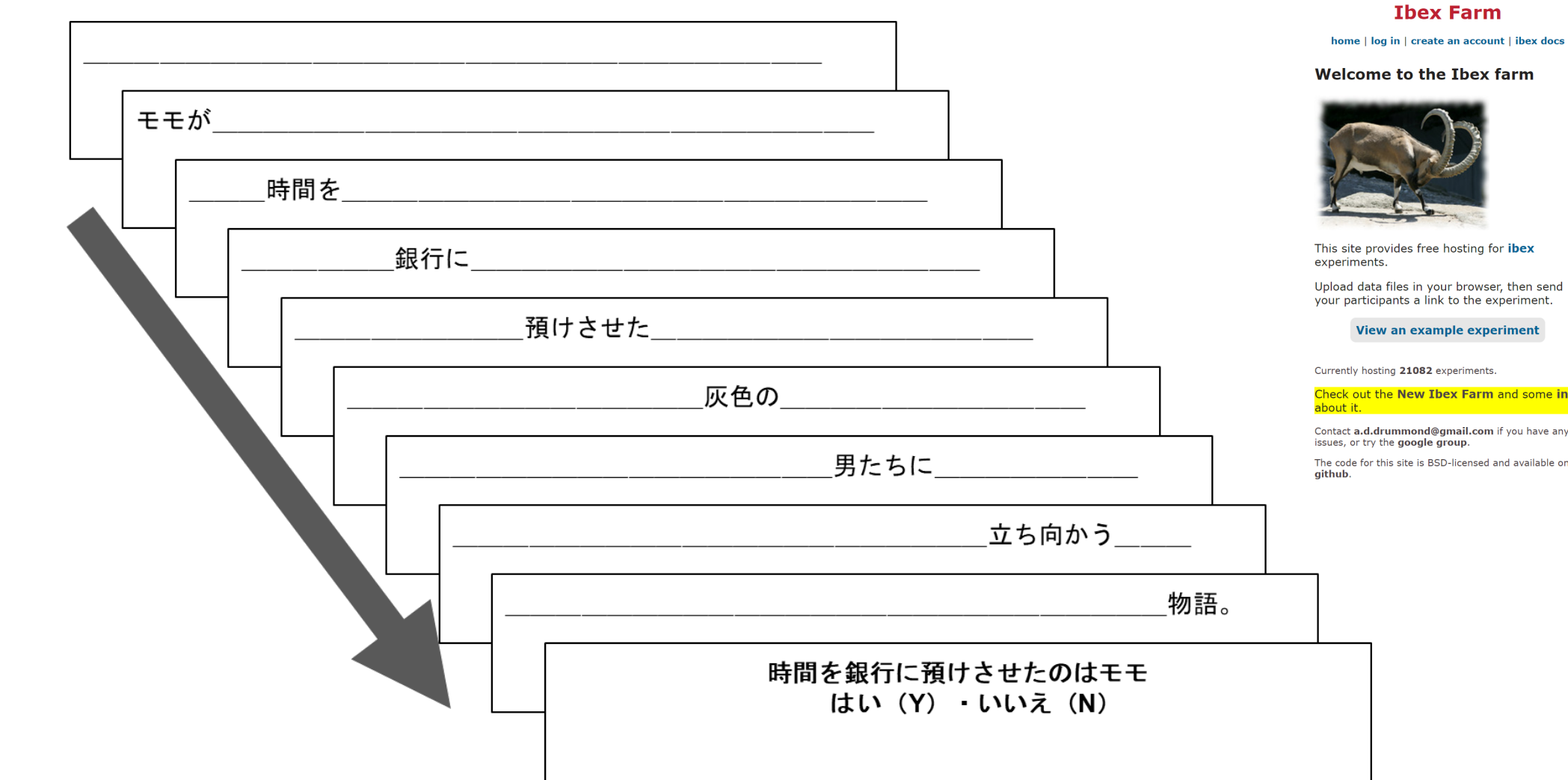
Register	Samples	Sentences	Phrases	Participants
OW	1	36	462	277
Whitepaper				
OT	38	9,521	50,606	422
Textbooks	(average)	250.6	1,331.7	
PB	83	10,075	84,736	388
Books	(average)	121.4	1,020.9	

Reading Time

Sample x Participants	OW	OT	PB
BCCWJ-SPR2	308	4,865	11,325
w/ vocab. rating	277	4,685	10,932

Data points	OW	OT	PB
BCCWJ-SPR2	136,797	5,704,898	10,769,380
w/ vocab. rating	124,502	5,490,977	10,484,300

Vocab. Rating 200 >= answers
Reading Time 5 >= samples



Statistical Analysis

Generalized Linear Mixed Model: Reading Time

		OW	Whitepaper	OT	Textbooks	PB	Books
SPR_sentence_ID	Order	-6.087***	(0.051)	-0.127***	(0.0004)	-0.142***	(0.001)
SPR_bunsetsu_ID	Order	-1.501***	(0.049)	-2.046***	(0.011)	-0.856***	(0.006)
SPR_word_length	Length	24.820***	(0.170)	5.170***	(0.021)	6.798***	(0.015)
SPR_trial	Order			-0.757***	(0.005)	0.382***	(0.006)
DepPara_depnum	Dependent	-15.310***	(0.591)			-5.258***	(0.034)
WFR_subj_rate	Vocab	-81.239***	(21.227)	-16.169*	(9.087)	-18.405**	(8.731)
Constant		558.984***	(12.936)	353.723***	(6.548)	306.631***	(5.425)
Data Points		121,769		5,407,252		10,321,560	
< -3SD or 3SD <	Deleted	2,732	(0.0219)	83,724	(0.0152)	162,740	(0.0155)
Log Likelihood		-818,815.1		-32,796,021		-62,393,234	

Statistical Analysis

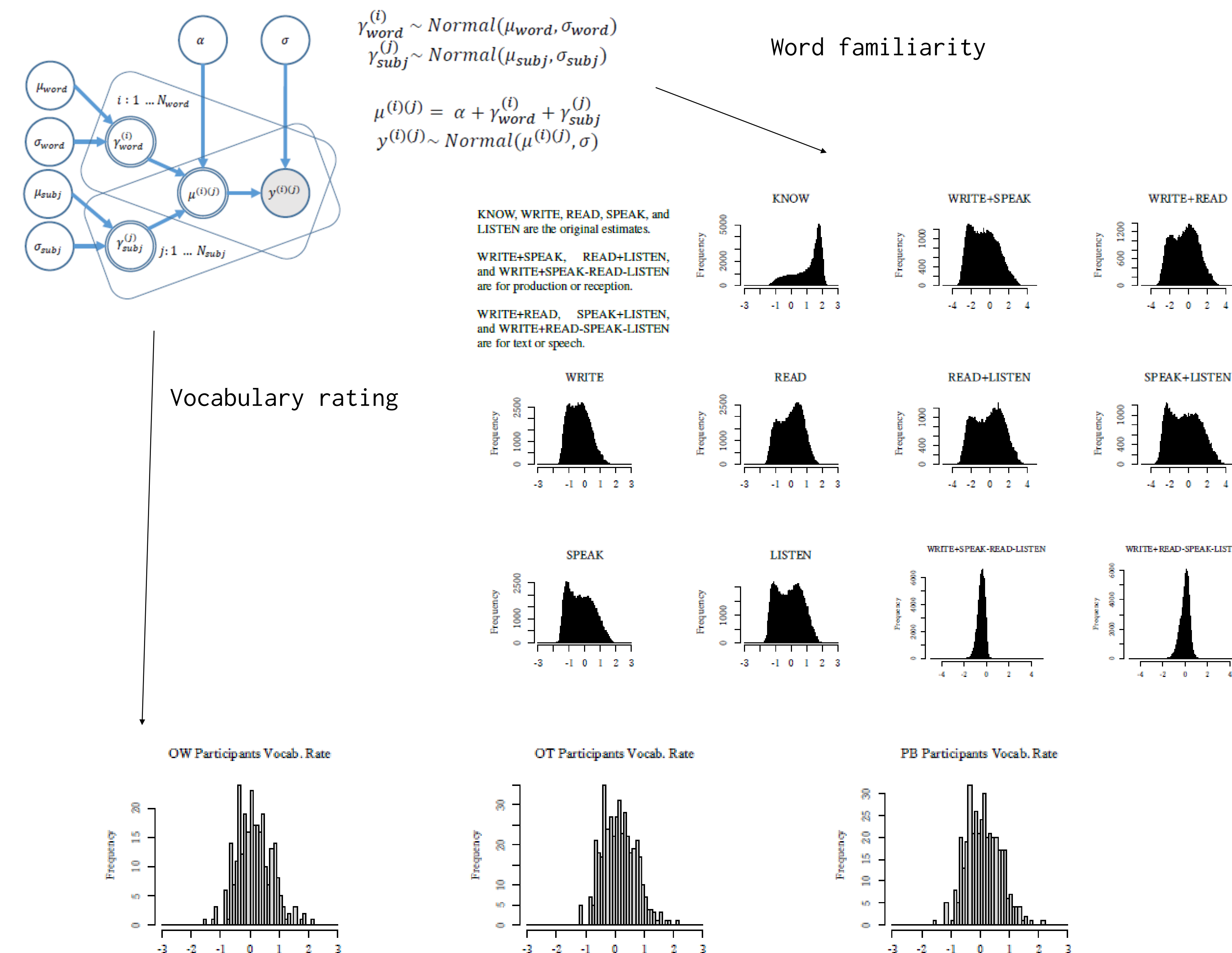
Generalized Linear Mixed Model: Logarithms of Reading Time

		OW	Whitepaper	OT	Textbooks	PB	Books
SPR_sentence_ID	Order	-0.012***	(0.0001)	-0.0004***	(0.0000)	-0.0004***	(0.0000)
SPR_bunsetsu_ID	Order	-0.0003***	(0.0001)	-0.006***	(0.00003)	-0.003***	(0.00002)
SPR_word_length	Length	0.036***	(0.0002)	0.011***	(0.0001)	0.014***	(0.00004)
SPR_trial	Order			-0.002***	(0.00001)	0.001***	(0.00002)
DepPara_depnum	Dependent	-0.022***	(0.001)			-0.012***	(0.0001)
WFR_subj_rate	Vocab	-0.180***	(0.040)	-0.052**	(0.025)	-0.052**	(0.026)
Constant		6.255***	(0.025)	5.826***	(0.020)	5.664***	(0.016)
Data Points		135,070		5,412,398		10,327,584	
< -3SD or 3SD <	Deleted	1,559	(0.0125)	78,578	(0.0143)	156,716	(0.0149)
Log Likelihood		-38,816.7		-598,180.4		-743,585.5	

Comparison with participants' vocabulary rating results
Higher vocabulary rating ⇔ Shorter reading time
Whitepaper > Books > Textbooks

Future directions

Working memory estimation ⇔ reading time



SPR_reading_time ~ SPR_sentence_ID + SPR_bunsetsu_ID + DepPara_depnum + SPR_trial +
Reading Time Sentence order Phrase order Number of Dependent Trial Order
+ SPR_word_length + WFR_subj_rate + (1 | SPR_subj_ID_factor) + (1 | BCCWJ_Sample_ID)
Word length Vocab. Rating Subject Sample

Data points with values outside 3SD were eliminated

Statistical Analysis

Bayesian Linear Mixed Model

```
model {
  real mu;
  gamma_subj ~ normal(0, sigma_subj); // prior
  for (k in 1:N) { //
    mu = alpha + beta_length * length[k] +
      beta_dependent * dependent[k] +
      beta_sentid * sentid[k] +
      beta_bid * bid[k] +
      beta_subjrate * subjrate[k] +
      gamma_subj[subj_id[k]];
    time[k] ~ lognormal(mu, sigma);
  }
}
```

Whitepaper	(OW)	mean	se_mean	sd
α	Constants	6.217	0.064	0.123
β_{sentid}	Order	-0.012	0.000	0.002
β_{bid}	Order	-0.003	0.000	0.001
β_{length}	Word length	0.037	0.000	0.006
β_{dep}	Dependency	-0.024	0.001	0.009
β_{subj}	Vocab. Rating	-0.118	0.094	0.116
σ		0.988	0.075	0.097
σ_{subj}		2.603	1.020	1.254