Current status report

Contents

1 Points/topics we could potentially want to cover in the paper

1.1 High-level motivation

- VSMs are popular for representing words, and in particular, a strength that they have is their ability to capture relations between words.

  It's difficult to know how best to evaluate VSMs. Datasets based on human judgments (of similarity, synonymy, analogy, etc) are a typical evaluation standard. Since these judgments are arrived at, presumably, by some human cognitive process or combination of cognitive processes, this suggests a tendency to compare VSMs with aspects of human cognition. This is reasonable, but since cognition is complex, we should think about what we are getting from any given measure.

- Questions that we should be asking:

  1) i. What exactly is being captured (what kinds of relations, what other kinds of information) by VSMs? How does this vary across VSMs?

     ii. Do cognitive processes/measures reflect the same kind of relations/information? If so, which cognitive processes/measures, and under which circumstances?

  2) i. What do we *want* representations in a VSM to reflect (possibly task-dependent)?

ii. Which cognitive measures (if any) best reflect what we want?

- By careful examination of correspondences between cognitive measures and VSMs, we can attack the questions in 1. And in doing so we stand to learn both about what is reflected by the relevant VSMs, as well as what is reflected by the relevant cognitive measures (since we have different types of information about each).

- The question of what we want VSMs to reflect is a complex one, which we may need to evaluate based on task performance. But it's one that needs to be considered when choosing an evaluation metric. And once we have an answer to it, then answers to the questions in 1 can help to inform construction/choice of human-based evaluation metrics.

1.2 Motivations for these particular results

- Single-word priming data from psycholinguistic experiments is an obvious candidate for something we might want VSMs to emulate, since it also quantifies relations between words. Additionally, it measures responses in a task that is irrelevant to explicit determination of word relations, which plausibly means that it taps into relations existing at a less conscious level than is presumably involved in similarity rating tasks. This could be a good thing, if we don't want the noise introduced by people needing to decide what is meant by "similar" (and where things should fall on their scale once they decide what it is).

- Literature has shown successful simulation of priming by VSMs (Section 7 provides more detail on what has been done, for your reference). But these studies have typically been evaluating with the rather low bar of comparing condition means by hypothesis testing, and as far as I can tell they have not done any systematic consideration of experimental factors such as task and SOA.

- A finer-grained metric for comparing VSM values to priming would be itemwise correlation (provides possible continuous measure of model performance—only one study seems to try this, observing that correlations are generally weaker than performance on classification task). It would be helpful if we could use a finer-grained continuous measure like this, and also if we could examine effects of experimental factors in a controlled fashion, rather than mixing different experimental factors together.

- The Semantic Priming Project (SPP) is a large database of priming data for a large number of pairs (1661), with RTs measured systematically for all pairs at two different SOAs (200ms, 1200ms) and in two different task types (lexical decision, naming). The database also includes information about relation type and lexical characteristics (such as frequency). This allows us a much more controlled and comprehensive source of data for comparison of VSM predictions with priming measures.

- Our study leverages the information made available in the SPP to begin exploring correspondences between the relations and information reflected in VSMs, and the relations and information reflected in the cognitive processes that give rise to priming effects.

1.3 Observations from current results

In the results (tables below), I think we can see a few potentially interesting trends. (CAVEAT: I am trying to be very careful about using and interpreting these statistical tests, but I could absolutely have missed important information or misinterpreted something. There are also a lot of tables, and I could certainly have misread patterns that I report here. If you notice something wrong just tell me!)

I'm showing results from three different word2vec models (trained on different amounts of data, and one without lowercasing the training data), one LSA model (from the web interface, included in SPP), and two GloVe models (one trained on Wikipedia+Gigaword, one trained on Twitter).

1) Section 3 shows performance at the two different latencies (200, 1200) on the two different tasks (lexical decision, naming), when **correlating priming magnitude with the difference between cosine similarities**.

   Section 4 shows performance at the two different latencies (200, 1200) on the two different tasks (lexical decision, naming), when **correlating cosine similarities with raw RTs**.

In both cases, I show results from the Spearman rank correlation (rho, p) and the Pearson correlation (r, p). There is also the t-value from the corresponding linear regression, and the number of items included in the correlation (n).

- Within these two sections, there looks to be a pretty consistent decline in correlation strengths as a function of both latency and task. The strongest correlations seem to be in LDT with latency 200, with weaker correlations in naming and at a 1200ms latency. (Though within the weak naming correlations, it looks like the 1200 latency is sometimes better than 200.)
- Comparing the sections, it looks like correlations with raw RTs are consistently stronger than correlations with priming magnitudes. This is a bit of a surprise in the context of the conversation Colin and I were having (where we agreed that it seemed wise to use the priming magnitudes, which are meant to subtract out other factors contributing to RT).
- Also a bit of a surprise: word2vec is being consistently outperformed by both LSA and GloVe. That said, the GloVe vectors appear to have been trained on much larger datasets than I used for w2v, so that might explain that. I'm not sure how large the corpus was for this particular LSA model.

2) Section 5 uses just LDT, 200ms, raw RT measures (where we seem to have the best correlations overall) and shows performance within the different prime-target relation types. Some quick summaries, based primarily on p-values. (I only mention a model if I saw that the p-values were at least marginal.)

   - synonym: LSA and GloVeTwitter both decent
   - antonym: LSA decent, both GloVe better
   - forward phrasal associate: LSA and both GloVe decent
   - script: all nonsignificant p values
   - unclassified: LSA and both GloVe decent
   - category: LSA and GloVeWG decent, GloveTW better
   - supraordinate: LSA decent, both Glove better
   - instrument: only w2v below p = .05, LSA and GloveWG marginal (small number of items)
   - functional property: only w2vbig and GloveTW marginal
   - backward phrasal associate: all bad (check whether there is in fact priming in this category)
   - perceptual property: LSA and GloVe good
   - action: LSA good, w2v and GloveTW all below p=.05, GloveWG marginal
   - norel: LSA and Glove all good

   So there is definitely some variation which I think is potentiall interesting/meaningful, but given the difference in corpus sizes/types, I should probably train GloVe (and LSA?) on the same corpora I used for w2v so that we can determine to what extent corpus is a factor.

3) Section 6 shows the performance of the different models on several standard similarity and synonymy evaluations sets.

   [One issue is that I don't have LSA values for these datasets because the LSA values I have been using came from SPP, which got them from the web interface. And since there are thousands of word pairs in these datasets, and I don't know of a way to extract values from the web interface non-manually (nor do I know exactly what corpora/settings they used for the web-interface model, which I would need to retrain my own model identically), I don't have those numbers at the moment. Probably I need to just train a new LSA model on corpora I have access to, with reasonable parameters. For now I'll talk about what I have.]

   Consistent with my previous experience using those datasets, the results are a bit all over the place. But GloVeWG (trained on Wiki-Gigaword) seems in general to be winning, with the exception of WS353. In synonymy, GloVeWG is winning as well, except for ESL-50.

   Ideally, we would like to be able to compare relative performance on these datasets with relative performance predicting priming measures, to try to better understand the relationships of these different measures to one another. I'm not sure off the top of my head what we can conclude yet, but I don't think it's hopeless, and I think it's something people would want to know.

2 Additional tasks I'm already planning on

Things I am already planning to do but haven't managed to do yet (I'll probably do these while I await responses ... unless you tell me not to):

- Neighbor rank. You'll see in the Previous Work section that one study found that forward neighbor rank (rank of the target among neighbors of the prime) yields better prediction of priming results. I've written the functions to compute neighbor rank, but since it takes a long time to search the whole space and rank everything, I need to make a slight change to where this fits in the whole analysis infrastructure, so I only have to run it once. (I just ran out of time to do this.) Once I have it, we can see whether neighbor rank indeed performs better.
- Check priming averages within each relation. This seems like a good thing to know in general.
- Run ANOVA on VSM values to see whether we replicate the main effect of priming found in other studies. Because the different relations have different numbers of items, we can't/shouldn't directly test for a main effect of relation type, but we could get around this by iteratively subsetting the data within relation types, if desired.
- See what the classification accuracy of the models is (deciding which is related prime and which is unrelated), for comparison with Lapesa and Evert 2013.
- Train an LSA model that we can actually use to get new values, without using web interface. While I'm at it, train on a corpus identical or comparable to the one w2v was trained on, for more controlled comparison.
- Same with GloVe—train on comparable/same corpus.
- Possibly add other variables into the regression predicting raw RTs. Are certain models doing better on that because they are encoding information about frequency etc? How much variance do other variables explain over and above the model predictions?

3 Results: Correlation with priming magnitude

3.1 LDT-200

	LSAdiff	w2vuk1	w2vbig	w2vukfull	gloveWG100	gloveTW100
FullSPP	rho: 0.148 p: 0.0 r: 0.147 p: 0.0 t: 27.17 n: 3240	rho: 0.136 p: 0.0 r: 0.135 p: 0.0 t: 28.273 n: 3240	rho: 0.142 p: 0.0 r: 0.144 p: 0.0 t: 28.911 n: 3240	rho: 0.139 p: 0.0 r: 0.139 p: 0.0 t: 28.634 n: 3240	rho: 0.153 p: 0.0 r: 0.157 p: 0.0 t: 29.154 n: 3240	rho: 0.126 p: 0.0 r: 0.121 p: 0.0 t: 26.504 n: 3240

3.2 LDT-1200

	LSAdiff	w2vuk1	w2vbig	w2vukfull	gloveWG100	gloveTW100
FullSPP	rho: 0.091 p: 0.0 r: 0.1 p: 0.0 t: 20.774 n: 3240	rho: 0.1 p: 0.0 r: 0.093 p: 0.0 t: 21.739 n: 3240	rho: 0.117 p: 0.0 r: 0.109 p: 0.0 t: 22.581 n: 3240	rho: 0.11 p: 0.0 r: 0.103 p: 0.0 t: 22.291 n: 3240	rho: 0.11 p: 0.0 r: 0.11 p: 0.0 t: 22.344 n: 3240	rho: 0.088 p: 0.0 r: 0.095 p: 0.0 t: 20.976 n: 3240

3.3 naming-200

	LSAdiff	w2vuk1	w2vbig	w2vukfull	gloveWG100	gloveTW100
FullSPP	rho: 0.084	rho: 0.025	rho: 0.032	rho: 0.031	rho: 0.151	rho: 0.151
	p: 0.0	p: 0.147	p: 0.067	p: 0.078	p: 0.0	p: 0.0
	r: 0.091	r: 0.031	r: 0.032	r: 0.033	r: 0.149	r: 0.157
	p: 0.0	p: 0.08	p: 0.071	p: 0.058	p: 0.0	p: 0.0
	t: 11.64	t: 9.816	t: 9.943	t: 9.973	t: 14.172	t: 14.448
	n: 3234	n: 3234	n: 3234	n: 3234	n: 3234	n: 3234

3.4 naming-1200

	LSAdiff	w2vuk1	w2vbig	w2vukfull	gloveWG100	gloveTW100
FullSPP	rho: 0.047	rho: 0.076	rho: 0.075	rho: 0.077	rho: 0.063	rho: 0.06
	p: 0.007	p: 0.0	p: 0.0	p: 0.0	p: 0.0	p: 0.001
	r: 0.049	r: 0.077	r: 0.079	r: 0.077	r: 0.064	r: 0.071
	p: 0.006	p: 0.0	p: 0.0	p: 0.0	p: 0.0	p: 0.0
	t: 10.421	t: 12.014	t: 12.16	t: 12.088	t: 11.541	t: 11.493
	n: 3234	n: 3234	n: 3234	n: 3234	n: 3234	n: 3234

4 Results: Correlation with raw RT

4.1 LDT-200

	LSA	w2vuk1	w2vbig	w2vukfull	gloveWG100	gloveTW100
FullSPP	rho: -0.265	rho: -0.152	rho: -0.157	rho: -0.158	rho: -0.284	rho: -0.285
	p: 0.0	p: 0.0	p: 0.0	p: 0.0	p: 0.0	p: 0.0
	r: -0.222	r: -0.144	r: -0.145	r: -0.146	r: -0.275	r: -0.279
	p: 0.0	p: 0.0	p: 0.0	p: 0.0	p: 0.0	p: 0.0
	t: -11.015	t: -6.856	t: -6.112	t: -6.938	t: -11.648	t: -9.351
	n: 6539	n: 6539	n: 6539	n: 6539	n: 6539	n: 6539

4.2 LDT-1200

	LSA	w2vuk1	w2vbig	w2vukfull	gloveWG100	gloveTW100
FullSPP	rho: -0.197	rho: -0.118	rho: -0.127	rho: -0.125	rho: -0.22	rho: -0.219
	p: 0.0	p: 0.0	p: 0.0	p: 0.0	p: 0.0	p: 0.0
	r: -0.18	r: -0.114	r: -0.121	r: -0.119	r: -0.221	r: -0.225
	p: 0.0	p: 0.0	p: 0.0	p: 0.0	p: 0.0	p: 0.0
	t: -8.856	t: -5.389	t: -5.141	t: -5.609	t: -9.279	t: -7.479
	n: 6539	n: 6539	n: 6539	n: 6539	n: 6539	n: 6539

4.3 naming-200

	LSA	w2vuk1	w2vbig	w2vukfull	gloveWG100	gloveTW100
FullSPP	rho: -0.155	rho: -0.055	rho: -0.064	rho: -0.061	rho: -0.162	rho: -0.185
	p: 0.0	p: 0.0	p: 0.0	p: 0.0	p: 0.0	p: 0.0
	r: -0.129	r: -0.057	r: -0.062	r: -0.061	r: -0.159	r: -0.186
	p: 0.0	p: 0.0	p: 0.0	p: 0.0	p: 0.0	p: 0.0
	t: -7.319	t: -3.37	t: -3.364	t: -3.573	t: -7.763	t: -7.504
	n: 6530	n: 6530	n: 6530	n: 6530	n: 6530	n: 6530

4.4 naming-1200

	LSA	w2vuk1	w2vbig	w2vukfull	gloveWG100	gloveTW100
FullSPP	rho: -0.133	rho: -0.087	rho: -0.097	rho: -0.092	rho: -0.129	rho: -0.138
	p: 0.0	p: 0.0	p: 0.0	p: 0.0	p: 0.0	p: 0.0
	r: -0.114	r: -0.085	r: -0.092	r: -0.087	r: -0.13	r: -0.141
	p: 0.0	p: 0.0	p: 0.0	p: 0.0	p: 0.0	p: 0.0
	t: -5.812	t: -4.358	t: -4.284	t: -4.436	t: -5.608	t: -4.885
	n: 6530	n: 6530	n: 6530	n: 6530	n: 6530	n: 6530

5 Results: Relation-specific correlations (LDT-200-RawRT)

	LSA	w2vuk1	w2vbig	w2vukfull	gloveWG100	gloveTW100
synonym	rho: -0.168	rho: 0.066	rho: 0.074	rho: 0.074	rho: -0.088	rho: -0.141
	p: 0.0	p: 0.078	p: 0.048	p: 0.049	p: 0.019	p: 0.0
	r: -0.177	r: 0.044	r: 0.062	r: 0.055	r: -0.097	r: -0.136
	p: 0.0	p: 0.243	p: 0.1	p: 0.144	p: 0.009	p: 0.0
	t: -4.791	t: 1.17	t: 1.647	t: 1.465	t: -2.611	t: -3.656
	n: 712	n: 712	n: 712	n: 712	n: 712	n: 712

	LSA	w2vuk1	w2vbig	w2vukfull	gloveWG100	gloveTW100
antonym	rho: -0.152	rho: -0.07	rho: -0.071	rho: -0.07	rho: -0.213	rho: -0.258
	p: 0.018	p: 0.279	p: 0.271	p: 0.278	p: 0.001	p: 0.0
	r: -0.164	r: -0.084	r: -0.053	r: -0.098	r: -0.268	r: -0.298
	p: 0.011	p: 0.194	p: 0.411	p: 0.132	p: 0.0	p: 0.0
	t: -2.575	t: -1.306	t: -0.825	t: -1.516	t: -4.307	t: -4.83
	n: 240	n: 240	n: 240	n: 240	n: 240	n: 240

	LSA	w2vuk1	w2vbig	w2vukfull	gloveWG100	gloveTW100
fpa	rho: -0.136	rho: -0.014	rho: -0.01	rho: -0.005	rho: -0.15	rho: -0.195
	p: 0.008	p: 0.785	p: 0.848	p: 0.915	p: 0.003	p: 0.0
	r: -0.135	r: -0.008	r: -0.005	r: 0.002	r: -0.144	r: -0.151
	p: 0.008	p: 0.878	p: 0.919	p: 0.97	p: 0.005	p: 0.003
	t: -2.664	t: -0.153	t: -0.102	t: 0.038	t: -2.837	t: -2.991
	n: 383	n: 383	n: 383	n: 383	n: 383	n: 383

	LSA	w2vuk1	w2vbig	w2vukfull	gloveWG100	gloveTW100
script	rho: -0.029	rho: -0.03	rho: -0.008	rho: -0.012	rho: -0.12	rho: -0.165
	p: 0.703	p: 0.685	p: 0.916	p: 0.872	p: 0.107	p: 0.027
	r: -0.026	r: -0.028	r: 0.009	r: -0.004	r: -0.111	r: -0.208
	p: 0.733	p: 0.705	p: 0.908	p: 0.959	p: 0.138	p: 0.005
	t: -0.343	t: -0.38	t: 0.116	t: -0.051	t: -1.493	t: -2.857
	n: 181	n: 181	n: 181	n: 181	n: 181	n: 181

	LSA	w2vuk1	w2vbig	w2vukfull	gloveWG100	gloveTW100
unclassified	rho: -0.103	rho: 0.035	rho: 0.032	rho: 0.025	rho: -0.148	rho: -0.178
	p: 0.007	p: 0.357	p: 0.396	p: 0.507	p: 0.0	p: 0.0
	r: -0.109	r: 0.041	r: 0.045	r: 0.04	r: -0.157	r: -0.186
	p: 0.004	p: 0.285	p: 0.236	p: 0.293	p: 0.0	p: 0.0
	t: -2.877	t: 1.072	t: 1.187	t: 1.053	t: -4.154	t: -4.939
	n: 684	n: 684	n: 684	n: 684	n: 684	n: 684

	LSA	w2vuk1	w2vbig	w2vukfull	gloveWG100	gloveTW100
category	rho: -0.148	rho: -0.071	rho: -0.05	rho: -0.055	rho: -0.109	rho: -0.2
	p: 0.009	p: 0.217	p: 0.386	p: 0.336	p: 0.056	p: 0.0
	r: -0.142	r: -0.095	r: -0.063	r: -0.073	r: -0.115	r: -0.211
	p: 0.013	p: 0.097	p: 0.273	p: 0.2	p: 0.044	p: 0.0
	t: -2.508	t: -1.668	t: -1.099	t: -1.287	t: -2.027	t: -3.775
	n: 308	n: 308	n: 308	n: 308	n: 308	n: 308

	LSA	w2vuk1	w2vbig	w2vukfull	gloveWG100	gloveTW100
supraordinate	rho: -0.137	rho: -0.041	rho: -0.064	rho: -0.027	rho: -0.201	rho: -0.197
	p: 0.054	p: 0.562	p: 0.372	p: 0.707	p: 0.005	p: 0.006
	r: -0.137	r: -0.043	r: -0.097	r: -0.034	r: -0.204	r: -0.192
	p: 0.054	p: 0.55	p: 0.176	p: 0.637	p: 0.004	p: 0.007
	t: -1.944	t: -0.6	t: -1.363	t: -0.474	t: -2.922	t: -2.747
	n: 198	n: 198	n: 198	n: 198	n: 198	n: 198

	LSA	w2vuk1	w2vbig	w2vukfull	gloveWG100	gloveTW100
instrument	rho: -0.284	rho: -0.338	rho: -0.337	rho: -0.337	rho: -0.269	rho: -0.206
	p: 0.065	p: 0.027	p: 0.027	p: 0.027	p: 0.081	p: 0.186
	r: -0.267	r: -0.267	r: -0.236	r: -0.307	r: -0.336	r: -0.222
	p: 0.084	p: 0.084	p: 0.128	p: 0.045	p: 0.028	p: 0.153
	t: -1.795	t: -1.794	t: -1.571	t: -2.09	t: -2.312	t: -1.475
	n: 43	n: 43	n: 43	n: 43	n: 43	n: 43

	LSA	w2vuk1	w2vbig	w2vukfull	gloveWG100	gloveTW100
functional property	rho: -0.156	rho: -0.138	rho: -0.172	rho: -0.153	rho: -0.115	rho: -0.189
	p: 0.118	p: 0.169	p: 0.086	p: 0.128	p: 0.253	p: 0.058
	r: -0.167	r: -0.083	r: -0.129	r: -0.107	r: -0.115	r: -0.172
	p: 0.094	p: 0.411	p: 0.197	p: 0.286	p: 0.252	p: 0.085
	t: -1.699	t: -0.829	t: -1.306	t: -1.079	t: -1.158	t: -1.749
	n: 101	n: 101	n: 101	n: 101	n: 101	n: 101

	LSA	w2vuk1	w2vbig	w2vukfull	gloveWG100	gloveTW100
bpa	rho: -0.088	rho: 0.032	rho: -0.025	rho: 0.064	rho: -0.055	rho: -0.043
	p: 0.205	p: 0.645	p: 0.722	p: 0.357	p: 0.428	p: 0.532
	r: -0.062	r: 0.065	r: 0.003	r: 0.086	r: -0.034	r: -0.054
	p: 0.371	p: 0.344	p: 0.963	p: 0.212	p: 0.627	p: 0.437
	t: -0.899	t: 0.951	t: 0.046	t: 1.256	t: -0.488	t: -0.781
	n: 211	n: 211	n: 211	n: 211	n: 211	n: 211

	LSA	w2vuk1	w2vbig	w2vukfull	gloveWG100	gloveTW100
perceptual property	rho: -0.28	rho: -0.1	rho: -0.123	rho: -0.132	rho: -0.335	rho: -0.231
	p: 0.001	p: 0.238	p: 0.147	p: 0.12	p: 0.0	p: 0.006
	r: -0.219	r: -0.104	r: -0.109	r: -0.137	r: -0.344	r: -0.256
	p: 0.009	p: 0.223	p: 0.201	p: 0.107	p: 0.0	p: 0.002
	t: -2.645	t: -1.228	t: -1.29	t: -1.627	t: -4.313	t: -3.128
	n: 140	n: 140	n: 140	n: 140	n: 140	n: 140

	LSA	w2vuk1	w2vbig	w2vukfull	gloveWG100	gloveTW100
action	rho: -0.709	rho: -0.588	rho: -0.604	rho: -0.582	rho: -0.527	rho: -0.566
	p: 0.007	p: 0.035	p: 0.029	p: 0.037	p: 0.064	p: 0.044
	r: -0.686	r: -0.543	r: -0.556	r: -0.6	r: -0.446	r: -0.533
	p: 0.01	p: 0.055	p: 0.049	p: 0.03	p: 0.126	p: 0.061
	t: -3.264	t: -2.24	t: -2.315	t: -2.597	t: -1.728	t: -2.182
	n: 13	n: 13	n: 13	n: 13	n: 13	n: 13

	LSA	w2vuk1	w2vbig	w2vukfull	gloveWG100	gloveTW100
norel	rho: -0.173	rho: -0.0	rho: -0.007	rho: -0.01	rho: -0.235	rho: -0.218
	p: 0.0	p: 0.986	p: 0.697	p: 0.58	p: 0.0	p: 0.0
	r: -0.118	r: -0.002	r: -0.003	r: -0.011	r: -0.239	r: -0.218
	p: 0.0	p: 0.912	p: 0.868	p: 0.537	p: 0.0	p: 0.0
	t: -6.802	t: -0.11	t: -0.166	t: -0.618	t: -14.05	t: -12.746
	n: 3270	n: 3270	n: 3270	n: 3270	n: 3270	n: 3270

6 Results: Performance on similarity and synonymy datasets

	w2vuk1	w2vbig	w2vukfull	gloveWG100	gloveTW100
MC-30	rho: 0.605 p: 0.0 r: 0.626 p: 0.0 n: 30	rho: 0.532 p: 0.002 r: 0.584 p: 0.001 n: 30	rho: 0.612 p: 0.0 r: 0.68 p: 0.0 n: 30	rho: 0.629 p: 0.0 r: 0.627 p: 0.0 n: 30	rho: 0.607 p: 0.0 r: 0.62 p: 0.0 n: 30
MEN-full	rho: 0.676 p: 0.0 r: 0.669 p: 0.0 n: 3000	rho: 0.691 p: 0.0 r: 0.676 p: 0.0 n: 3000	rho: 0.689 p: 0.0 r: 0.68 p: 0.0 n: 3000	rho: 0.697 p: 0.0 r: 0.693 p: 0.0 n: 3000	rho: 0.573 p: 0.0 r: 0.573 p: 0.0 n: 3000
RG-65	rho: 0.605 p: 0.0 r: 0.598 p: 0.0 n: 65	rho: 0.622 p: 0.0 r: 0.627 p: 0.0 n: 65	rho: 0.64 p: 0.0 r: 0.643 p: 0.0 n: 65	rho: 0.682 p: 0.0 r: 0.691 p: 0.0 n: 65	rho: 0.683 p: 0.0 r: 0.675 p: 0.0 n: 65
SimLex-999	rho: 0.295 p: 0.0 r: 0.284 p: 0.0 n: 998	rho: 0.31 p: 0.0 r: 0.296 p: 0.0 n: 998	rho: 0.313 p: 0.0 r: 0.303 p: 0.0 n: 998	rho: 0.324 p: 0.0 r: 0.296 p: 0.0 n: 998	rho: 0.131 p: 0.0 r: 0.12 p: 0.0 n: 998
ws353	rho: 0.617 p: 0.0 r: 0.635 p: 0.0 n: 334	rho: 0.614 p: 0.0 r: 0.628 p: 0.0 n: 334	rho: 0.619 p: 0.0 r: 0.636 p: 0.0 n: 334	rho: 0.546 p: 0.0 r: 0.528 p: 0.0 n: 334	rho: 0.526 p: 0.0 r: 0.521 p: 0.0 n: 334

	w2vuk1	w2vbig	w2vukfull	gloveWG100	gloveTW100
ESL-50-single-word	52.08 48	50.0 48	58.33 48	50.0 48	31.25 48
RD-300-single-word	57.14 63	58.73 63	60.32 63	65.08 63	49.21 63
TOEFL-80-single-word	66.67 69	66.67 69	66.67 69	79.71 69	52.17 69

7 Previous Work (more detail)

There are several studies indicating that VSMs can simulate priming. Four of these I would qualify as modeling-oriented, with success being measured by ability to produce a significant effect in the correct directions/conditions as determined by the human experimental results being modeled.

- Three studies run ANOVAs on VSM-derived values for word pairs from Hodgson 1991, and claim successful simulation when they find a main effect of prime type and no effect of relation type (which is apparently what Hodgson found – equivalent priming in reading times across relations, for six relation types among 144 total prime-target pairs).
- Another, Jones et al 2006, compares values derived from three different VSM types and compares against a suite of different priming studies, and checks whether the VSM values pattern with the means of the various conditions in these studies, and whether the differences are significant (looks like they used t-tests for pairwise comparisons).
- In the human experiments being modeled, LDT, naming latency, and reading time measures are all represented. Jones et al say that they prefer naming latencies, but as far as I can tell they have a mix of LDT and naming.

- I didn't collect information on latency of target presentation for all the modeled studies, but good chance they aren't all the same either.

One study, Lapesa and Evert 2013, aims to examine which VSM features most affect performance on two tasks: 1) prime vs. unrelated classification, and 2) prediction of RT. They test a suite of VSMs, systematically varying all of the different features/parameters of interest.

- The classification task, as far as I can guess, just involves getting the relatedness (e.g. cosine similarity) between target and each of prime/unrelated, and choosing the closer one as the prime.
- Prediction of RT is done by Pearson correlation.
- They find that as an alternative to cosine similarity, rank of the target among neighbors of the prime has greater benefit for performance on these tasks.
- They evaluate against results from three priming studies (Feretti et al 2001, McRae et al 2004, Hare et al. 2009), within which LDT, naming latency, and animacy/concreteness decision time are all represented.