Generating Semantic Representations From Simple Word Co-occurrence Statistics

John A. Bullinaria

School of Computer Science
The University of Birmingham
Birmingham B15 2TT
UK

j.a.bullinaria@cs.bham.ac.uk

http://www.cs.bham.ac.uk/~jxb

Plan of Today's Talk

Extracting Semantic Representations from Word Co-occurrence Statistics: A Computational Study

John A. Bullinaria & Joseph P. Levy

Behavior Research Methods (2007), **39**, 510-526

Semantic Categorization Using Simple Word Co-occurrence Statistics

John A. Bullinaria

Proceedings paper presenting results on the Workshop Challenge Tasks

Some Ideas for Going Beyond Simple Word Co-occurrence Statistics

John A. Bullinaria

Introduction

The basic idea here is very simple:

word co-occurrence statistics from large text corpora

⇒ certain aspects of word meaning / lexical semantics

This leaves many important questions, such as:

1. **Which word co-occurrence statistics are best?**
2. **Does it depend on which aspects of word meaning we require?**
3. **What are the limitations of this idea?**
4. **Do we need to go beyond simple word co-occurrence statistics?**
5. **If so, what exactly do we need to do?**
6. **Does any of this tell us anything about human language acquisition?**
7. **…**

Simple Word Co-occurrence Statistics

target word

Take a large corpus of language usage and count the word co-occurrences.

context words

For each target word ***t*** we can count how many times each context word ***c*** appears within a window of a certain type and size around it, and thus compute a vector of conditional probabilities ***p(c|t)***.

These result in the basic vector space that we hope will constitute a useful representation of lexical semantics.

How Important Are The Details?

Early studies indicated that getting the details right was crucial.

Bullinaria & Levy, *Behavior Research Methods*, 2007 considered:

Varying the context window type

Varying the context window size

Varying the vector dimensionality

Varying the corpus size

Varying the corpus quality

Different semantic tasks

Different distance metrics

Different vector components (other than conditional probabilities)

I'll now summarize the key results obtained using an 89.6M word BNC corpus

Four Different Tasks

TOEFL (Test of English as a Foreign Language) – (Landauer & Dumais, 1997)

Pick which of four given words is closest to the target word – implemented using semantic distance comparisons. [80 target words]

Distance Comparison – (Bullinaria & Levy, 2007)

Tests larger scale structure of the semantic space by comparing distances to semantically related words against those for ten random control words. [200]

Semantic Categorization – (Patel, Bullinaria & Levy, 1997)

Compares distances between target words and their correct semantic category centers against distances to the centers of other categories. [530]

Syntactic Categorization – (Levy, Bullinaria & Patel, 1998)

Compares distances between target words and their correct syntactic category centers against distances to the centers of other categories. [1200]

Six Different Distance Metrics

Euclidean

$$d(t_1, t_2) = \left(\sum_c \left| p(c \mid t_1) - p(c \mid t_2) \right|^2 \right)^{1/2}$$

City Block

$$d(t_1, t_2) = \sum_c \left| p(c \mid t_1) - p(c \mid t_2) \right|$$

Cosine

$$d(t_1, t_2) = 1 - \frac{\left(\sum_c p(c \mid t_1).p(c \mid t_2) \right)}{\left(\sum_c p(c \mid t_1).p(c \mid t_2) \right)^{1/2} \left(\sum_c p(c \mid t_2).p(c \mid t_2) \right)^{1/2}}$$

Hellinger

$$d(t_1, t_2) = \sum_c \left(p(c \mid t_1)^{1/2} - p(c \mid t_2)^{1/2} \right)^2$$

Bhattacharya

$$d(t_1, t_2) = -\ln \sum_c \left(p(c \mid t_1) \right)^{1/2} \left(p(c \mid t_2) \right)^{1/2}$$

Kullback-Leibler

$$d(t_1, t_2) = \sum_c p(c \mid t_1) \log \left(\frac{p(c \mid t_1)}{p(c \mid t_2)} \right)$$

Four Different Vector Components

Raw Conditional Probabilities (P)

$$p(c \mid t)$$

Ratios of Conditional Probabilities (R)

$$r(c,t) = \frac{p(c \mid t)}{p(c)}$$

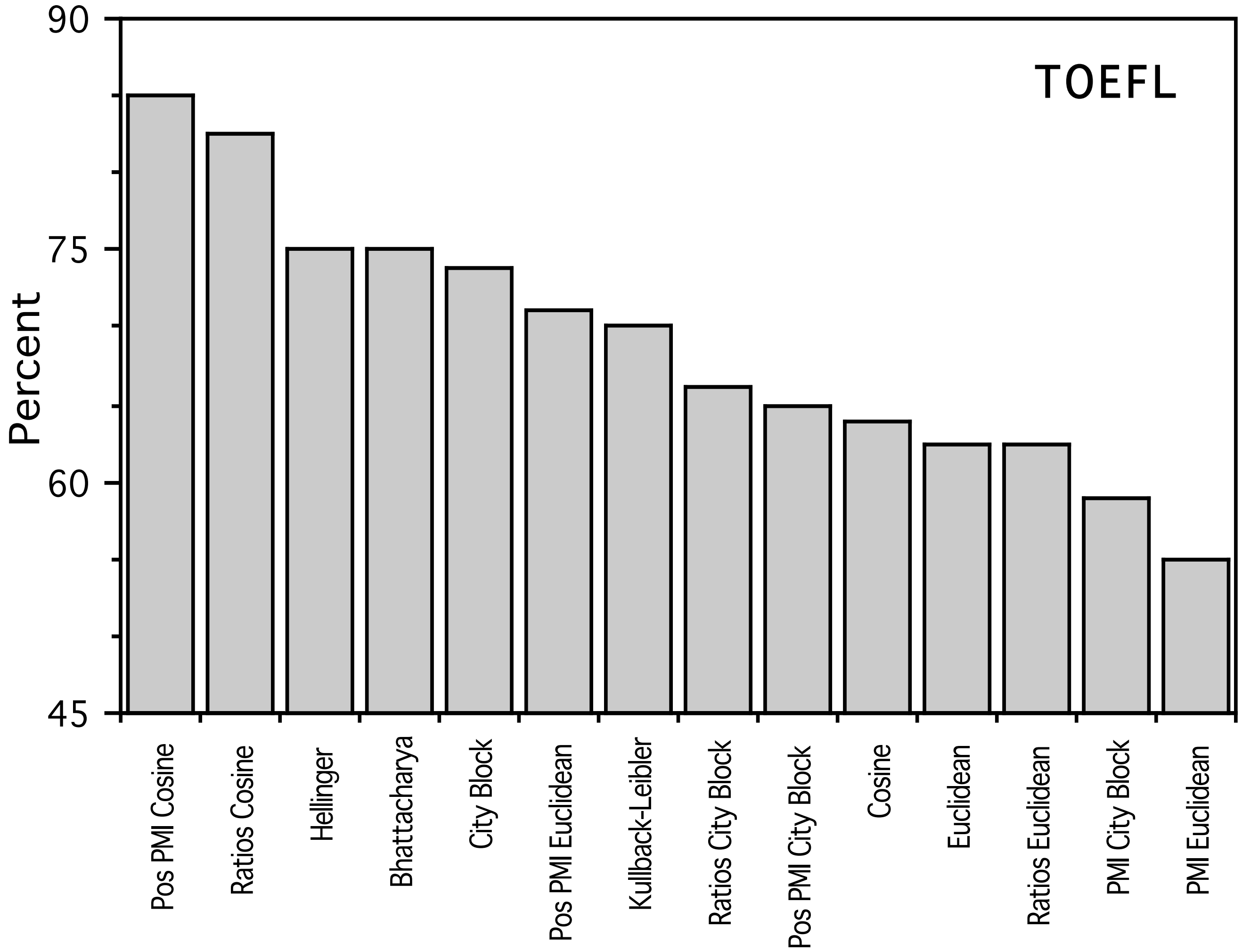
Pointwise Mutual Information (PMI)

$$i(c,t) = \log \frac{p(c \mid t)}{p(c)}$$

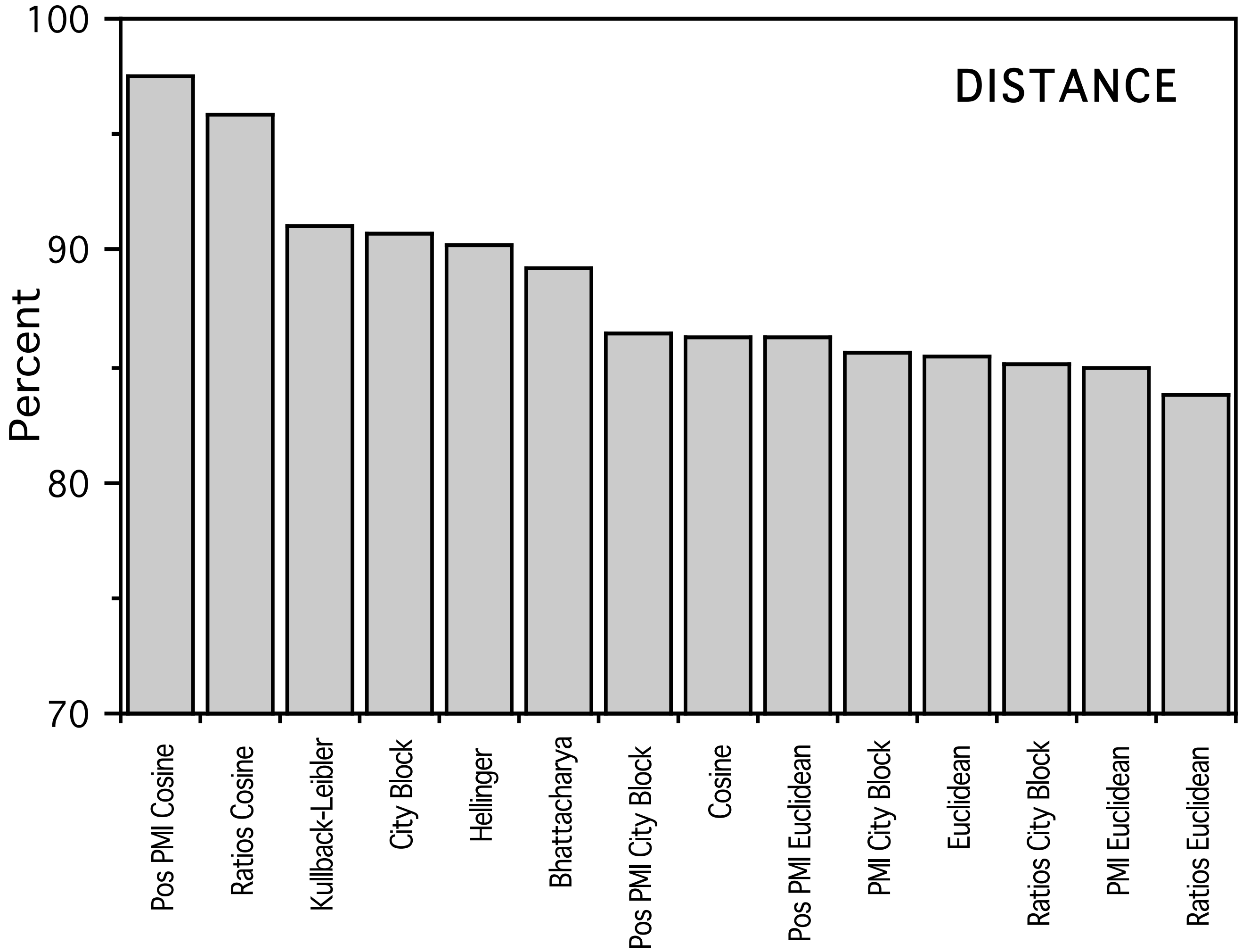
Positive Pointwise Mutual Information (PPMI)

$$i_{+}(c,t) = \begin{cases} 0 & \text{if} \quad i(c,t) \leq 0 \\ i(c,t) & \text{if} \quad i(c,t) > 0 \end{cases}$$

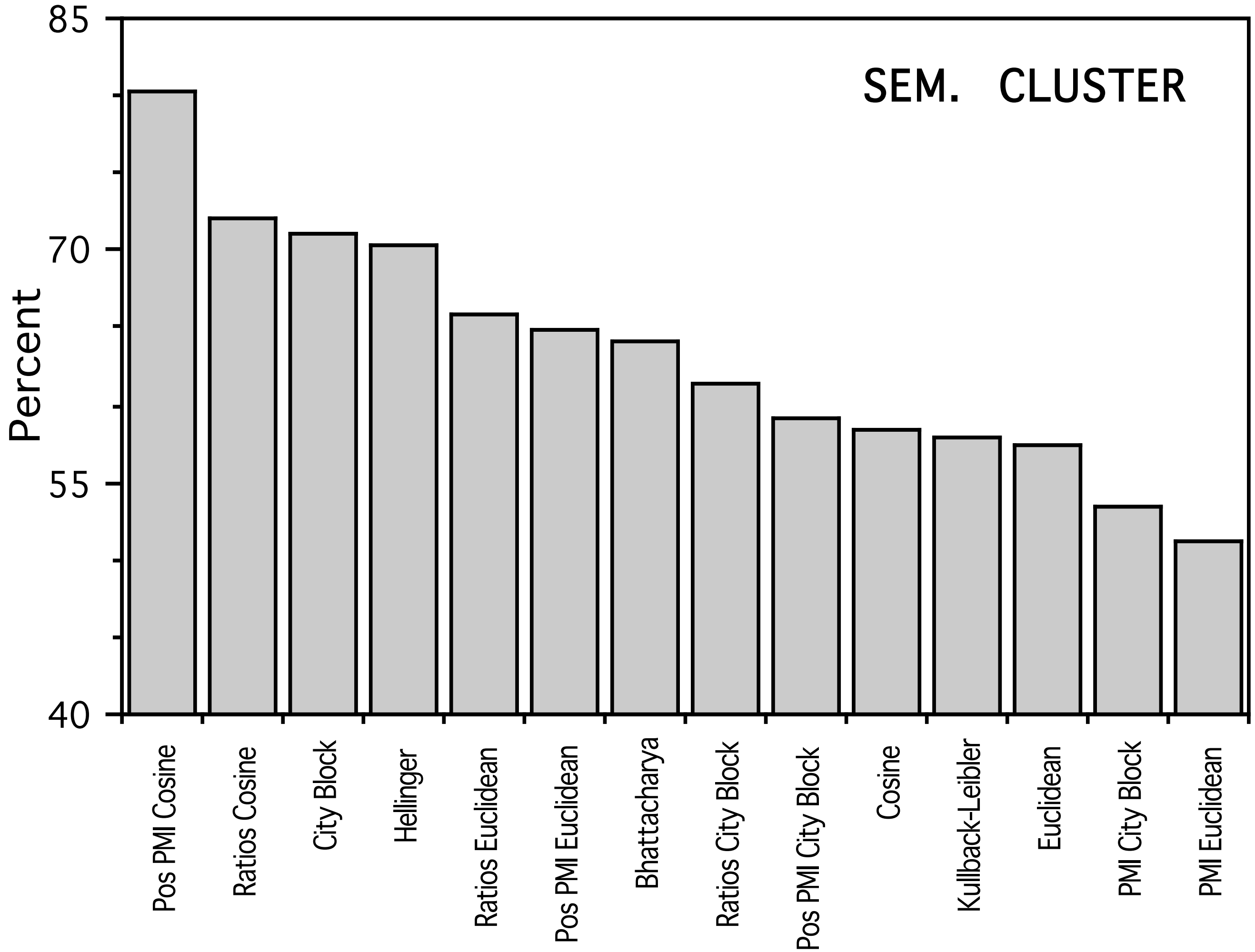
Best Results Across Component Types and Distance Metric



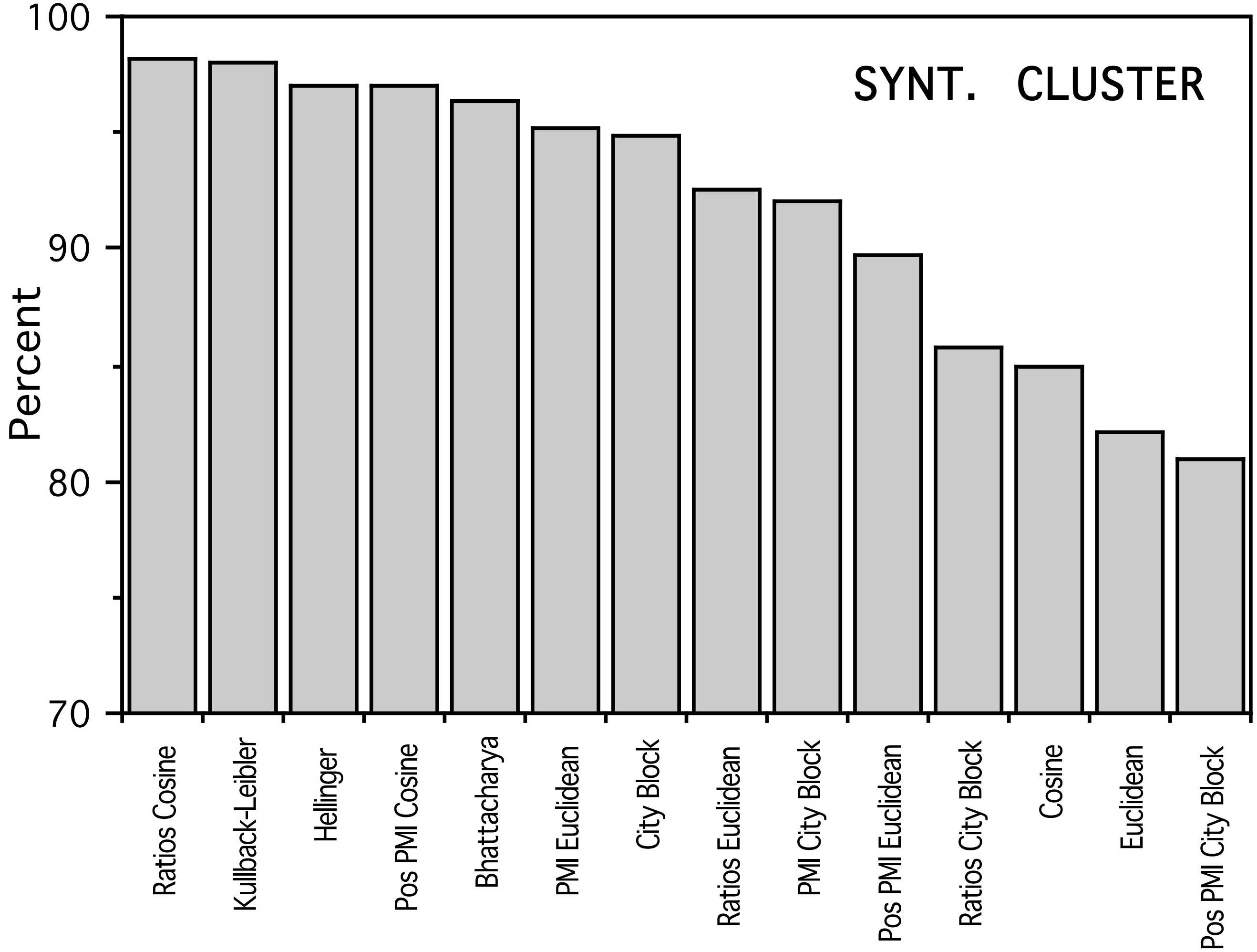
Best Results Across Component Types and Distance Metric



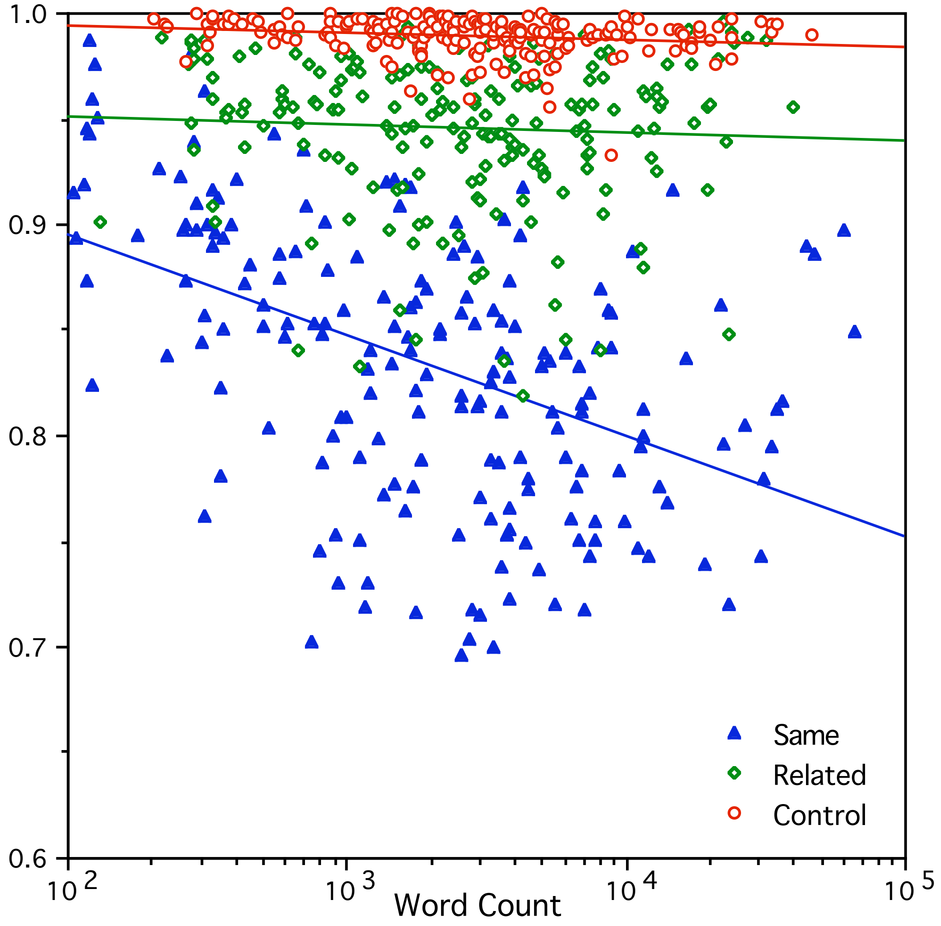
Best Results Across Component Types and Distance Metric

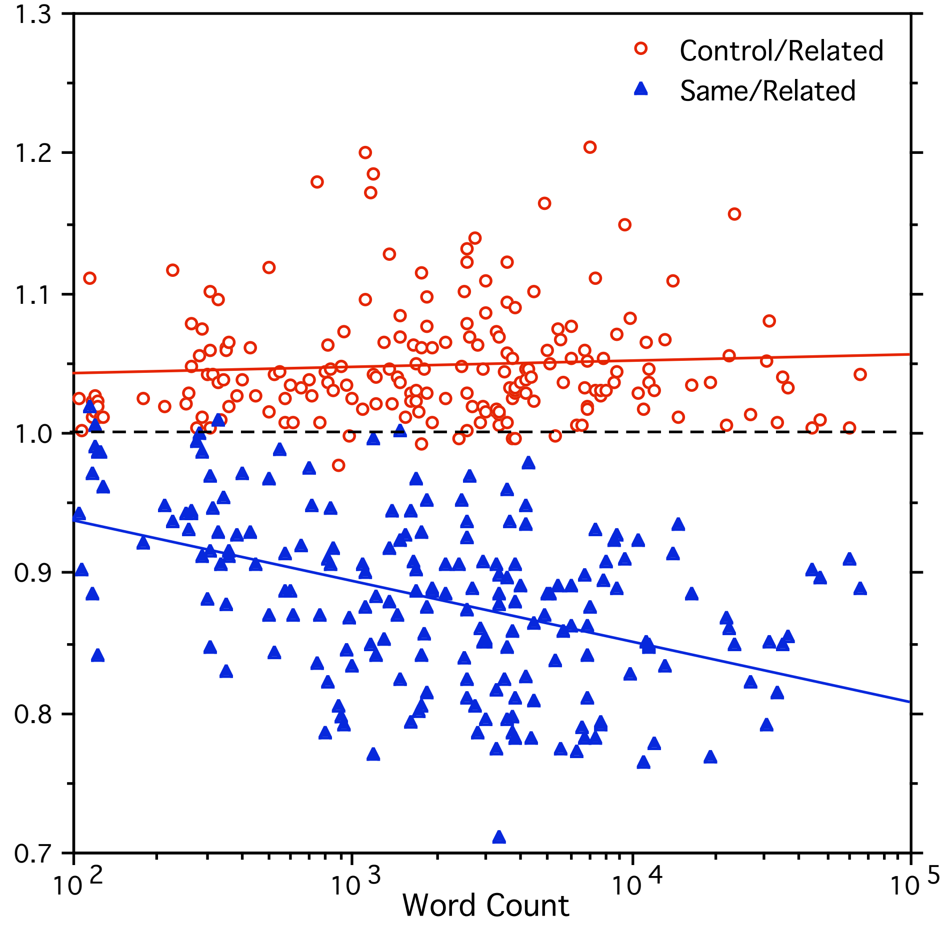


Best Results Across Component Types and Distance Metric

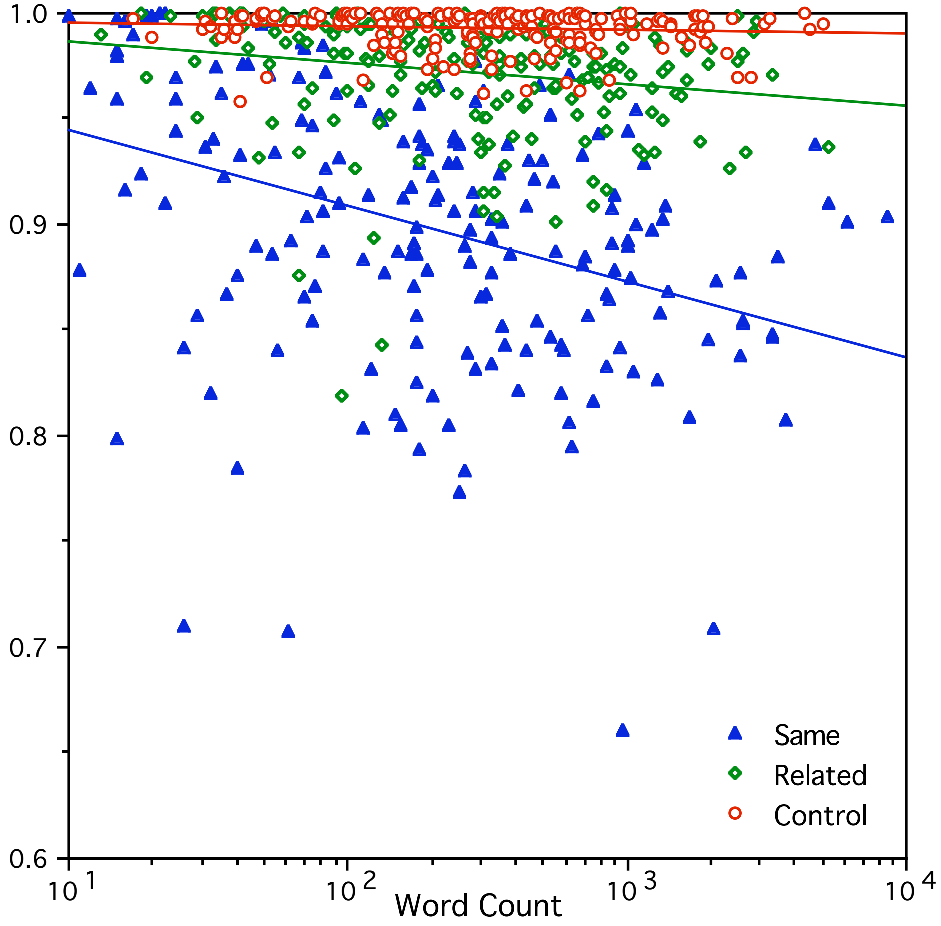


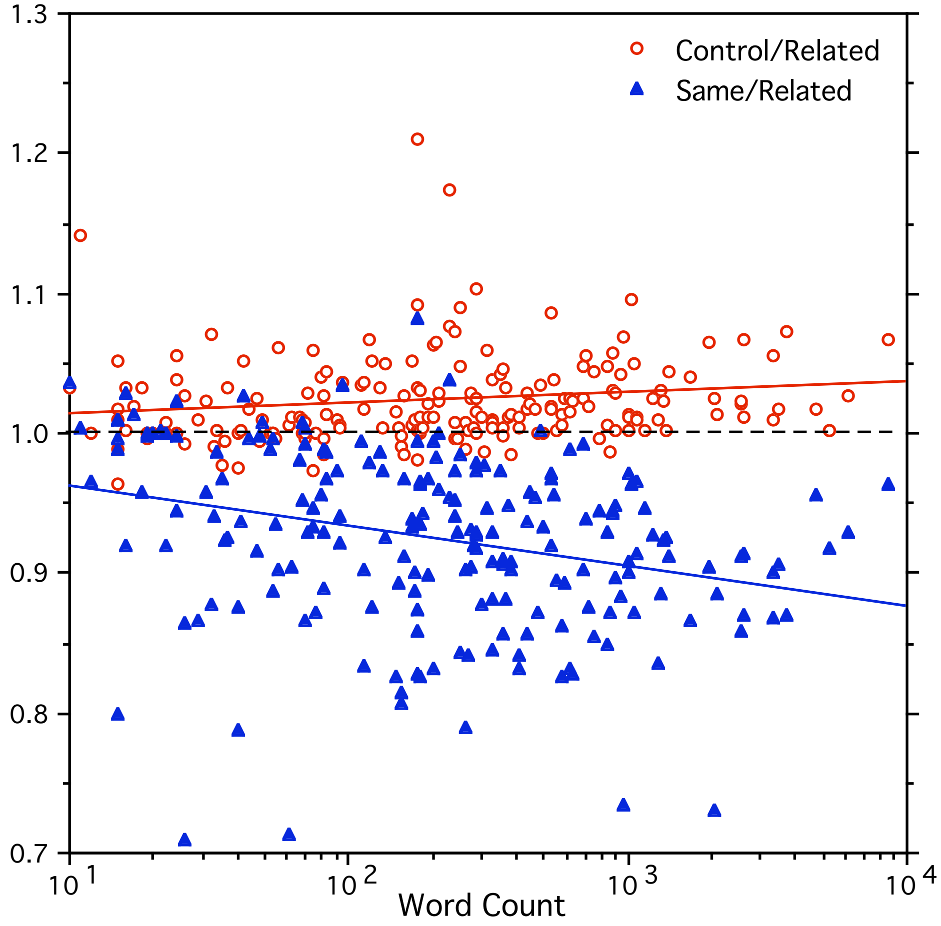
Statistical Reliability – PPMI Cos – Halves BNC Corpus (44.8M)



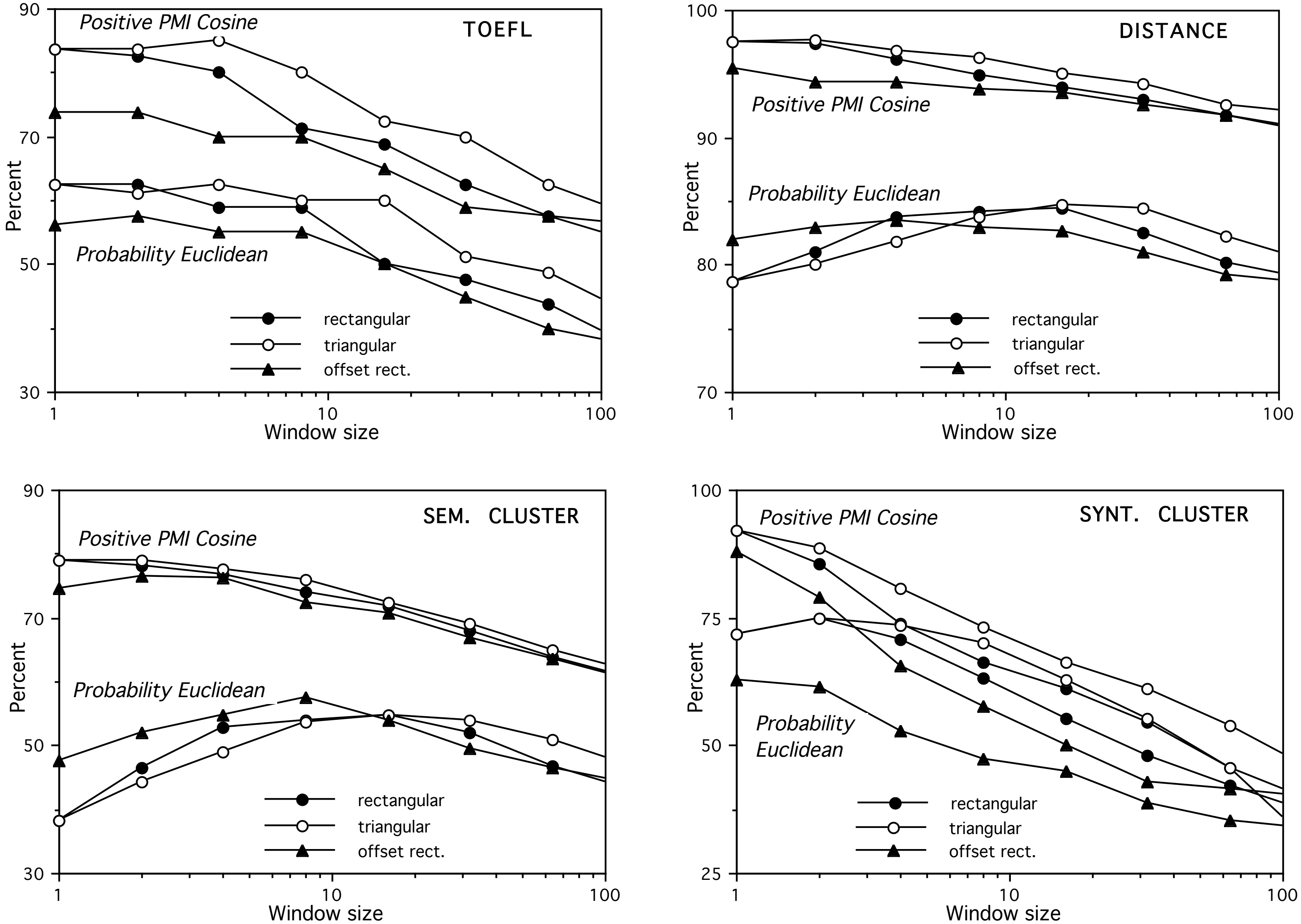


Statistical Reliability – PPMI Cos – Smaller Corpora (4.6M)

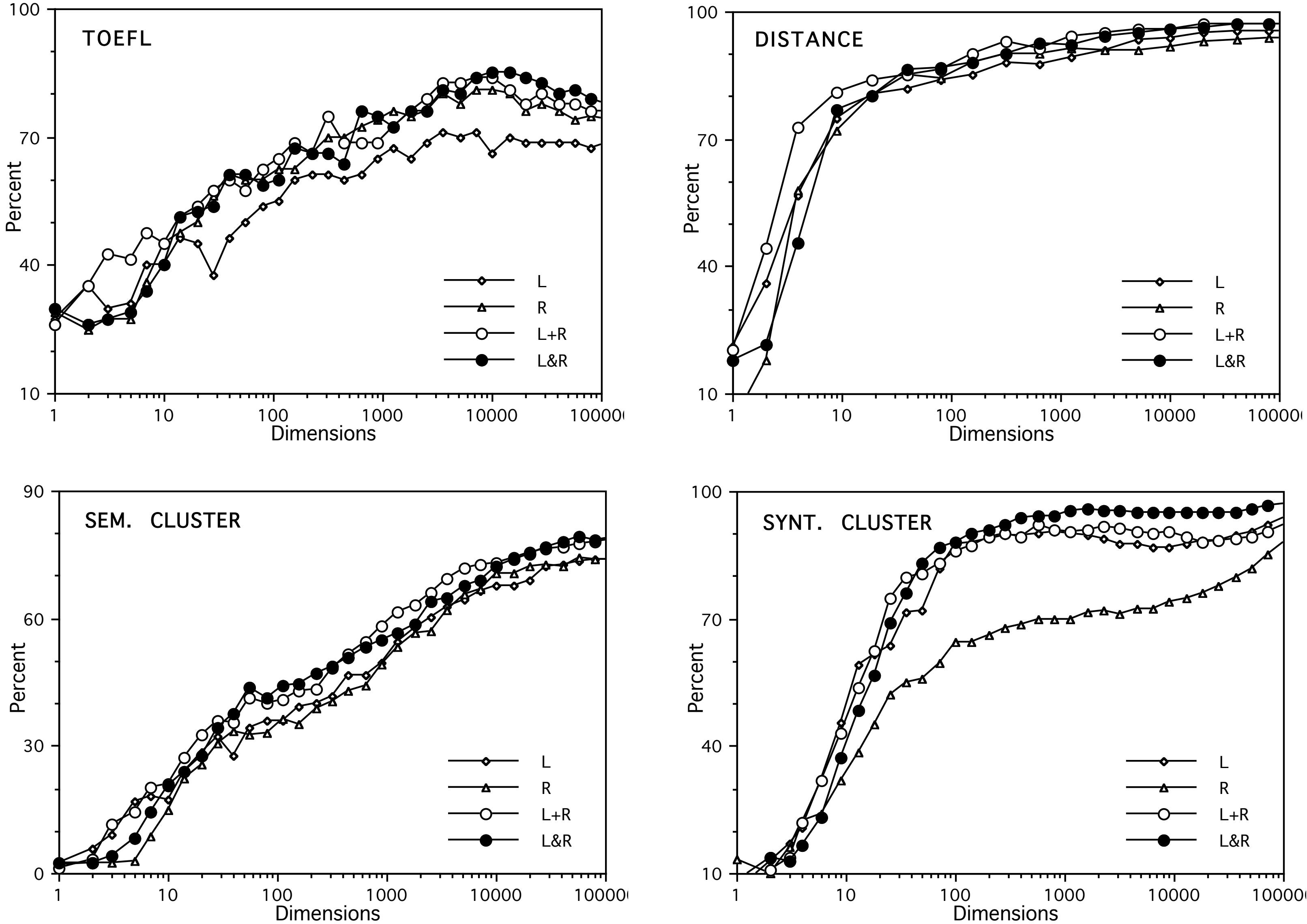




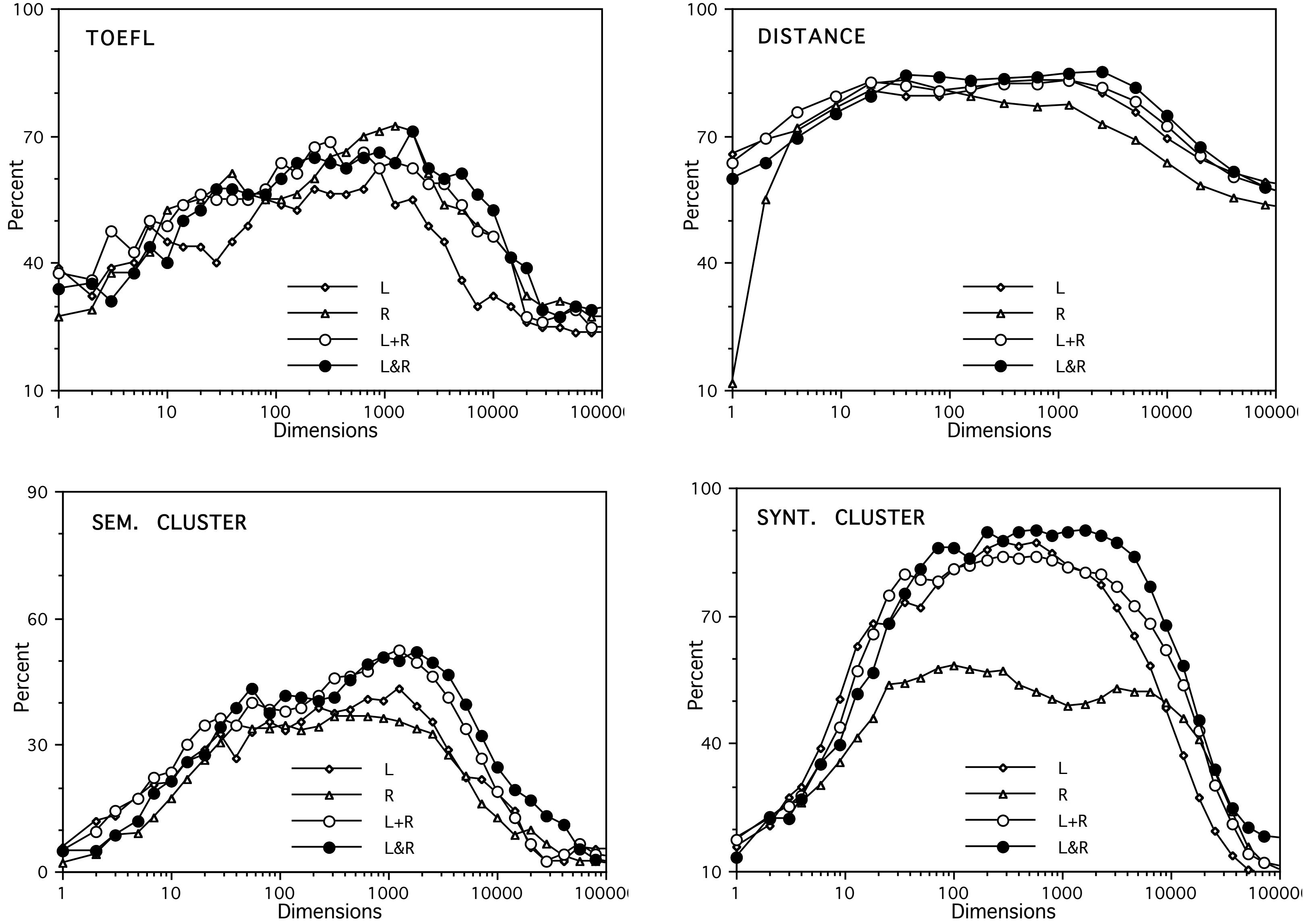
Effect of Window Size and Shape – PPMI Cos, Prob Eucl



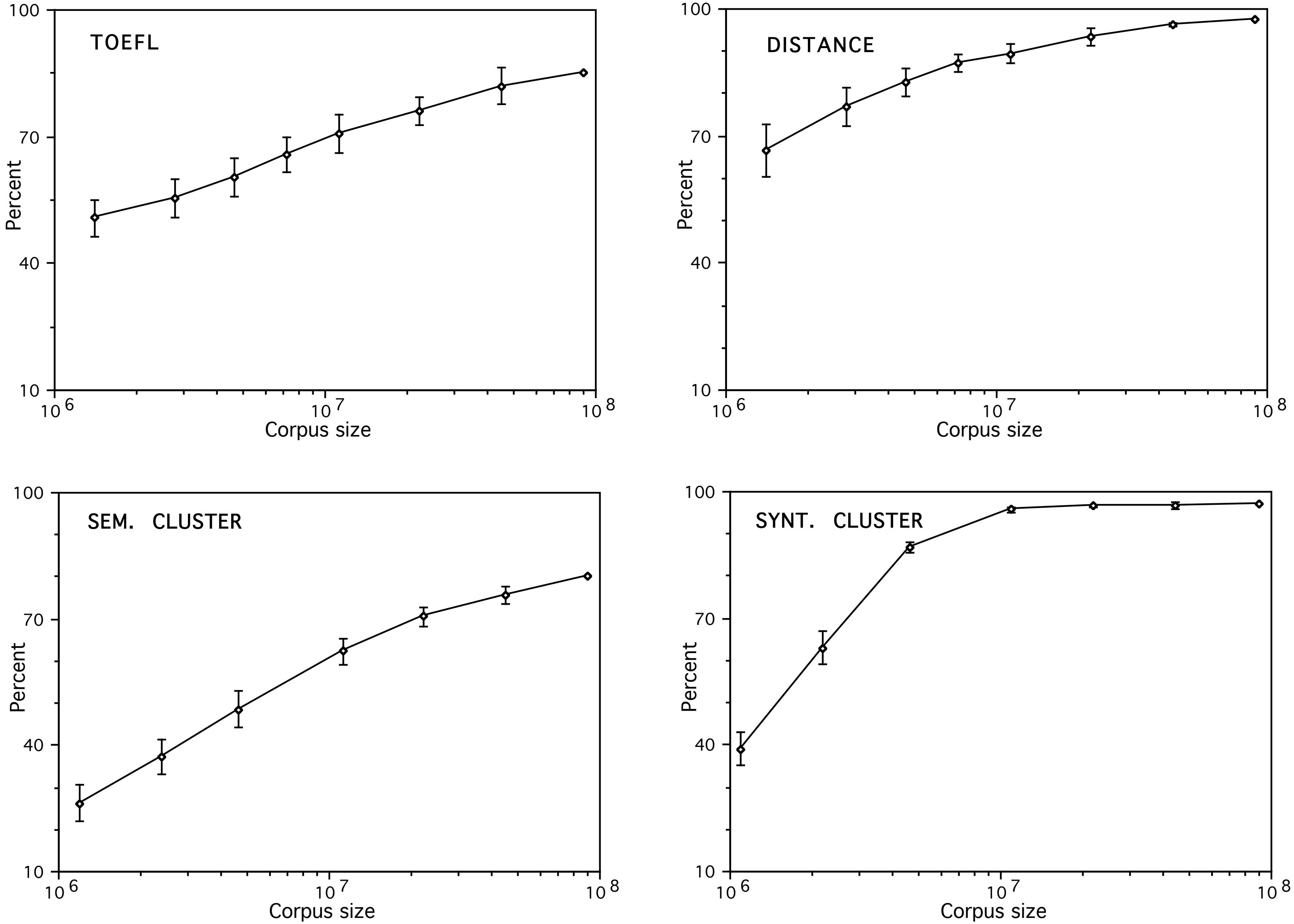
Effect of Window Type and Vector Dimensionality – PPMI Cos



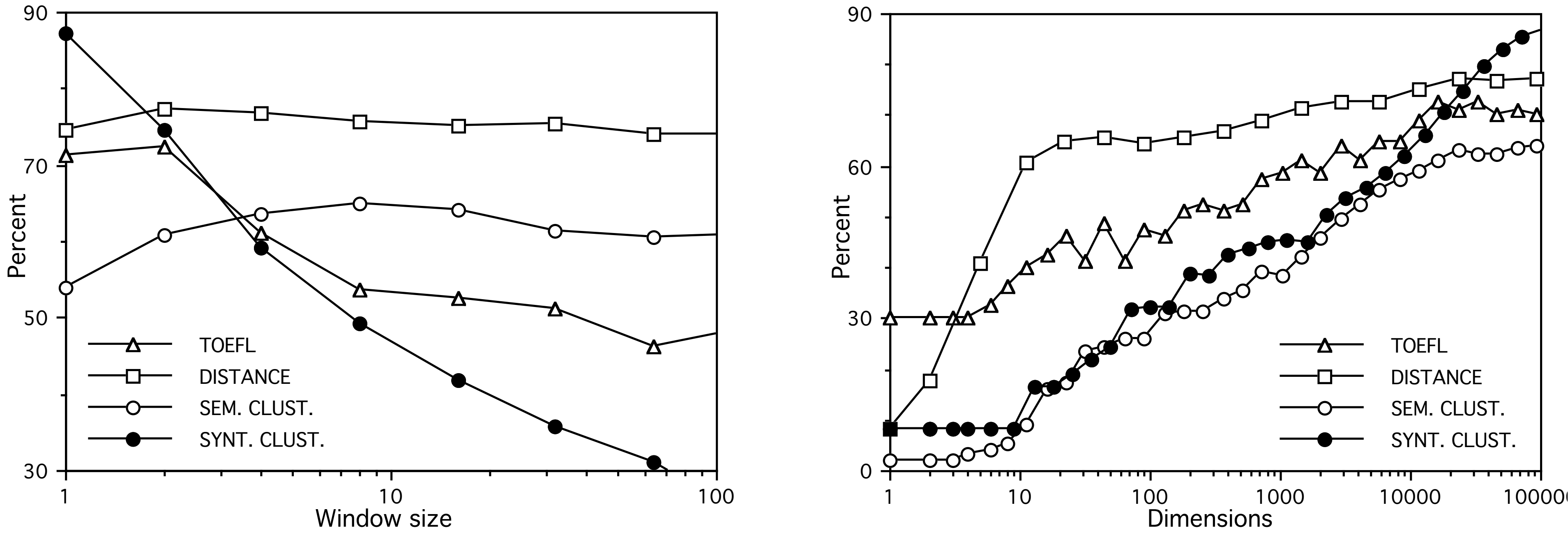
Effect of Window Type and Vector Dimensionality – PPMI Eucl



Effect of Corpus Size – PPMI Cos



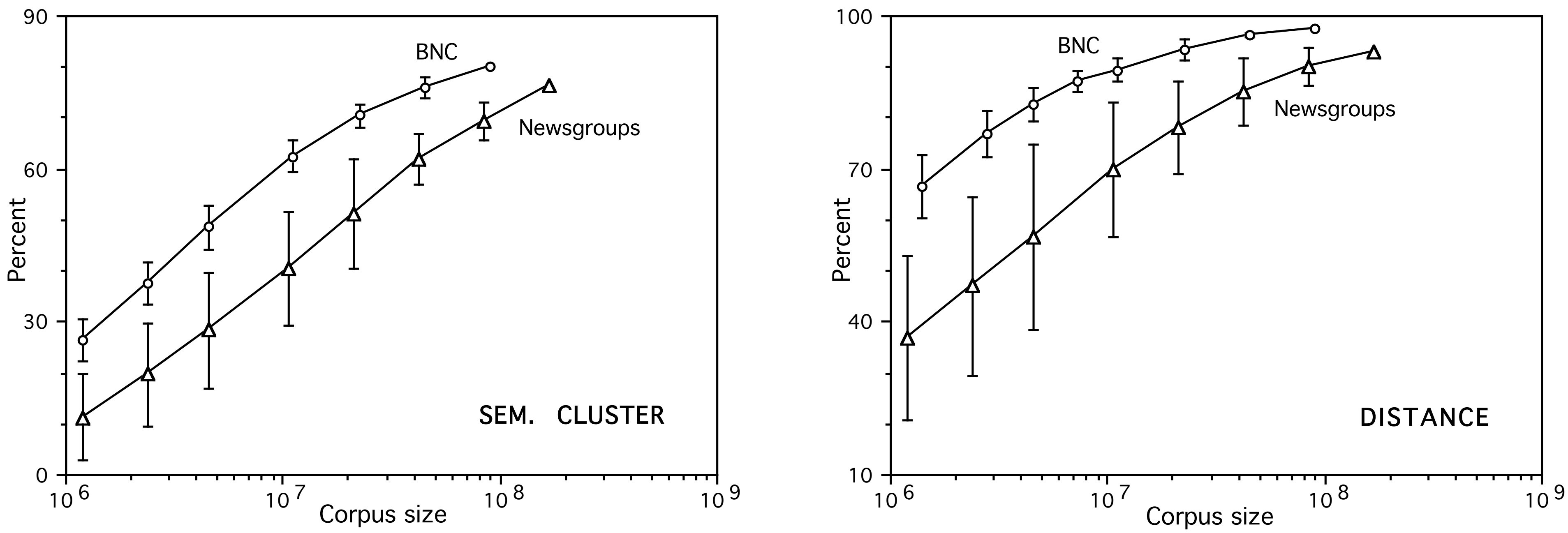
Performance For Smaller Corpus – 4.6M – PPMI Cos



For smaller corpora, statistical reliability issues arise and the performance falls

In these cases the optimal window size may be larger to compensate

Effect of Corpus Quality – PPMI Cos



Ceiling performance with respect to corpus size has not yet been reached

Corpus quality is also crucial – just increasing the size is not enough!

General Conclusions So Far?

Drawing general conclusions from such a small sample is dangerous, but it looks like the best semantic representations arise from:

Vectors of Positive Pointwise Mutual Information

Using the standard Cosine distance measure

Very small windows, just one context word each side of the target

As many vector components as possible

The biggest and highest quality corpus available

The obvious way to proceed now is to:

Find a bigger and better corpus

Test the semantic vectors on more tasks

Understand the limitations of the approach

The Lexical Semantics Workshop Challenge

The ukWaC corpus – 1984.4M words derived from web-pages

- ~20 times the size of the BNC corpus
- 1M words with a frequency of five or more

Categorization tasks

- 44 concrete nouns – 6 hand-labelled semantic categories
- 45 verbs – 9 hand-labelled semantic categories

CLUTO Clustering Toolkit

- Direct k-way clustering algorithm
- Default parameter settings

Does PPMI Cosine still give good results with the same optimal parameters?

Are the limitation of the approach clearer with the bigger corpus and new tasks?

Measures of Clustering Quality

Two measures of clustering quality are built into CLUTO - both compare the clusters against hand-crafted class labels:

Entropy

$$E = \sum_{r=1}^{k} \frac{n_r}{n} E_r \qquad , \qquad E_r = -\frac{1}{\log q} \sum_{i=1}^{q} \frac{n_r^i}{n_r} \log \frac{n_r^i}{n_r}$$

Purity

$$P = \sum_{r=1}^{k} \frac{n_r}{n} P_r \qquad , \qquad P_r = \frac{1}{n_r} \max_i \left(n_r^i \right)$$

for clustering of n words, with r labelling k clusters, and i labelling q classes.

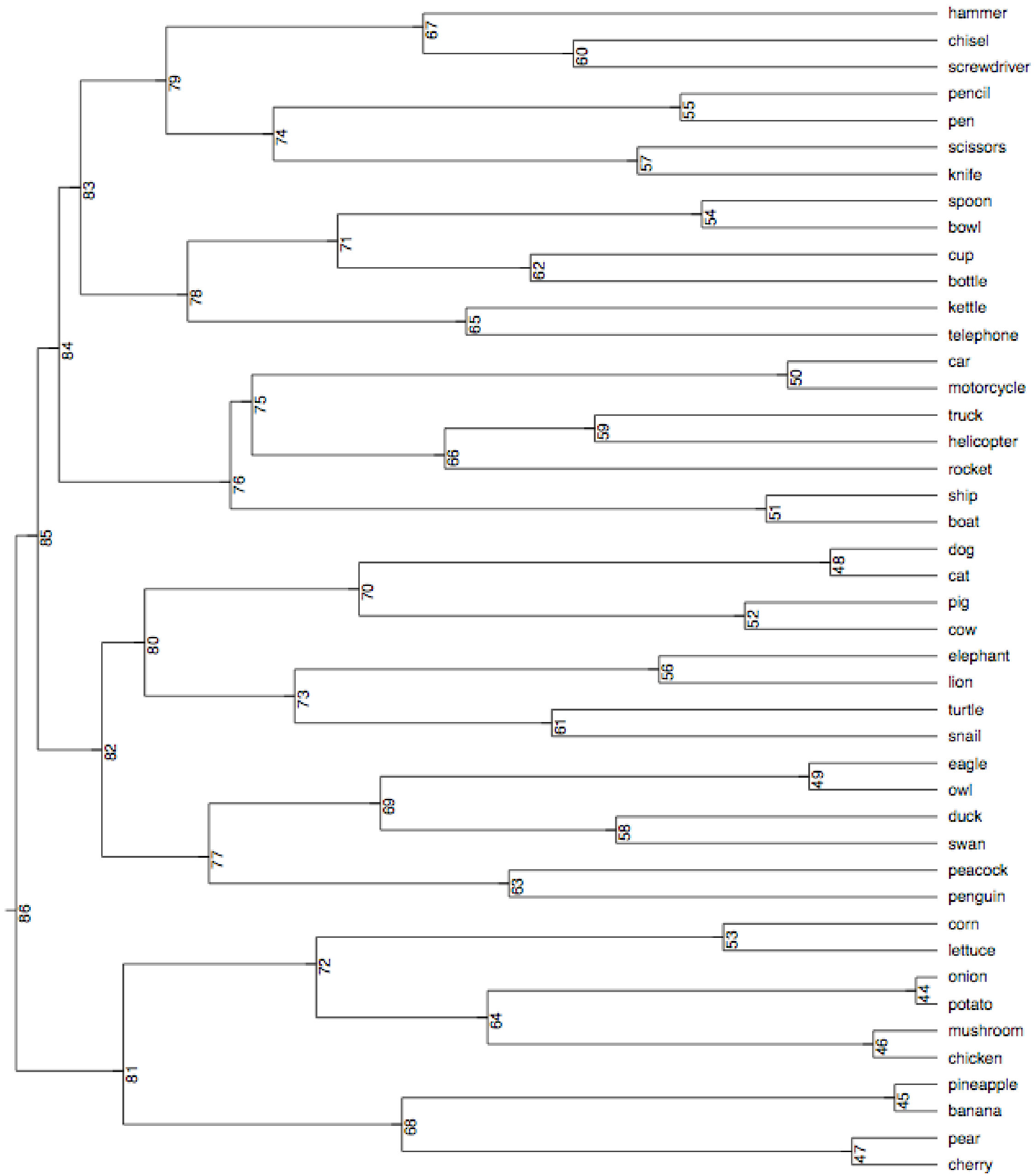
Both range from 0 to 1. Perfect clusters have entropy 0 and purity 1.

Concrete Noun Clustering

Six clusters:

Purity
= 0.886

Entropy
= 0.120



Comments on the Concrete Noun Clustering

Good clustering is obtained, right down to individual word pairs.

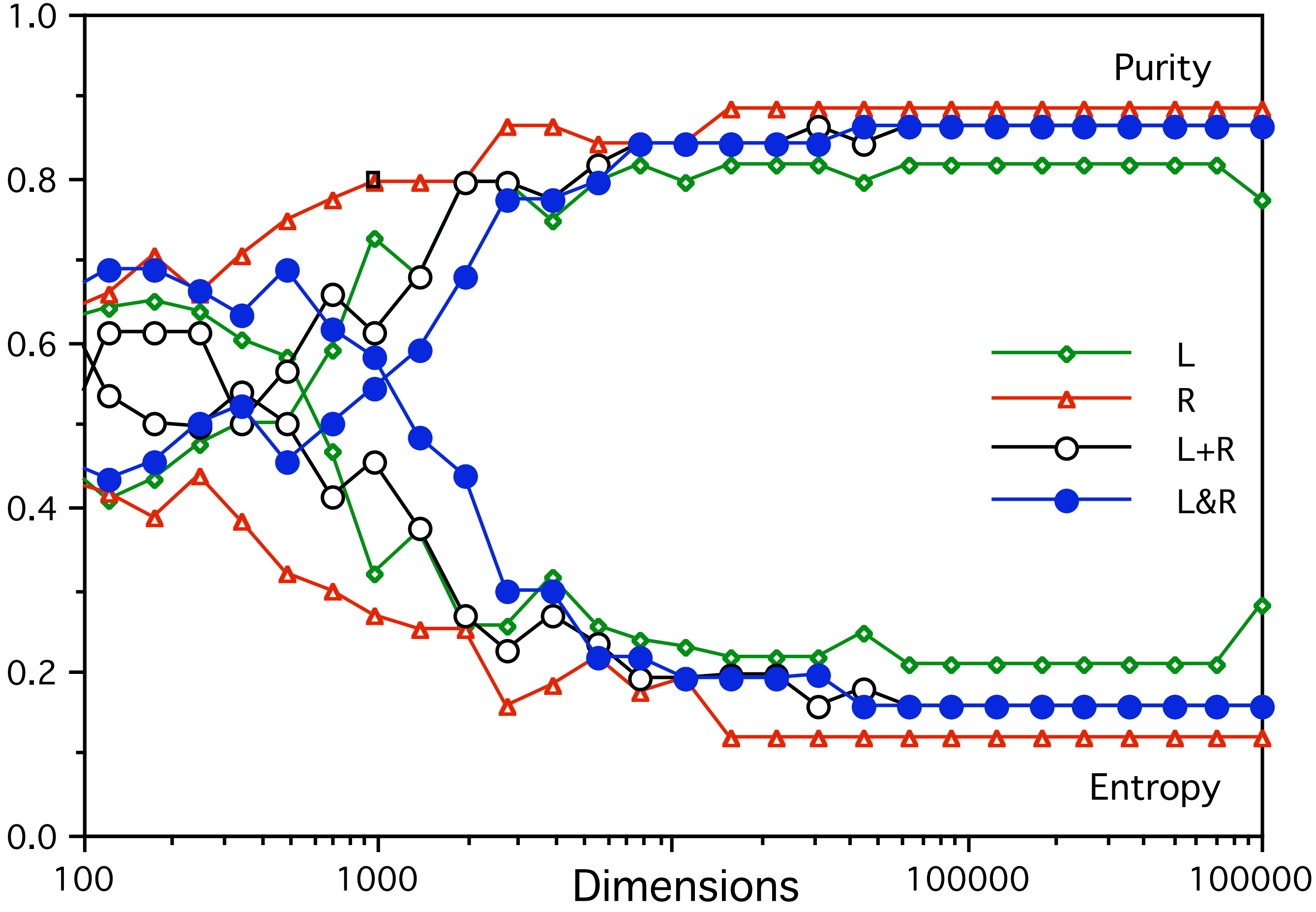
One understandable “mistake” – ‘chicken’ in a ‘foodstuffs’ cluster rather than in the ‘animal’ cluster.

The six main clusters do not line up with the handcrafted clusters – ‘fruit’ and ‘vegetable’ clusters are combined, and the ‘tools’ split. This is responsible for the poor purity and entropy scores. And asking for different numbers of clusters doesn’t help.

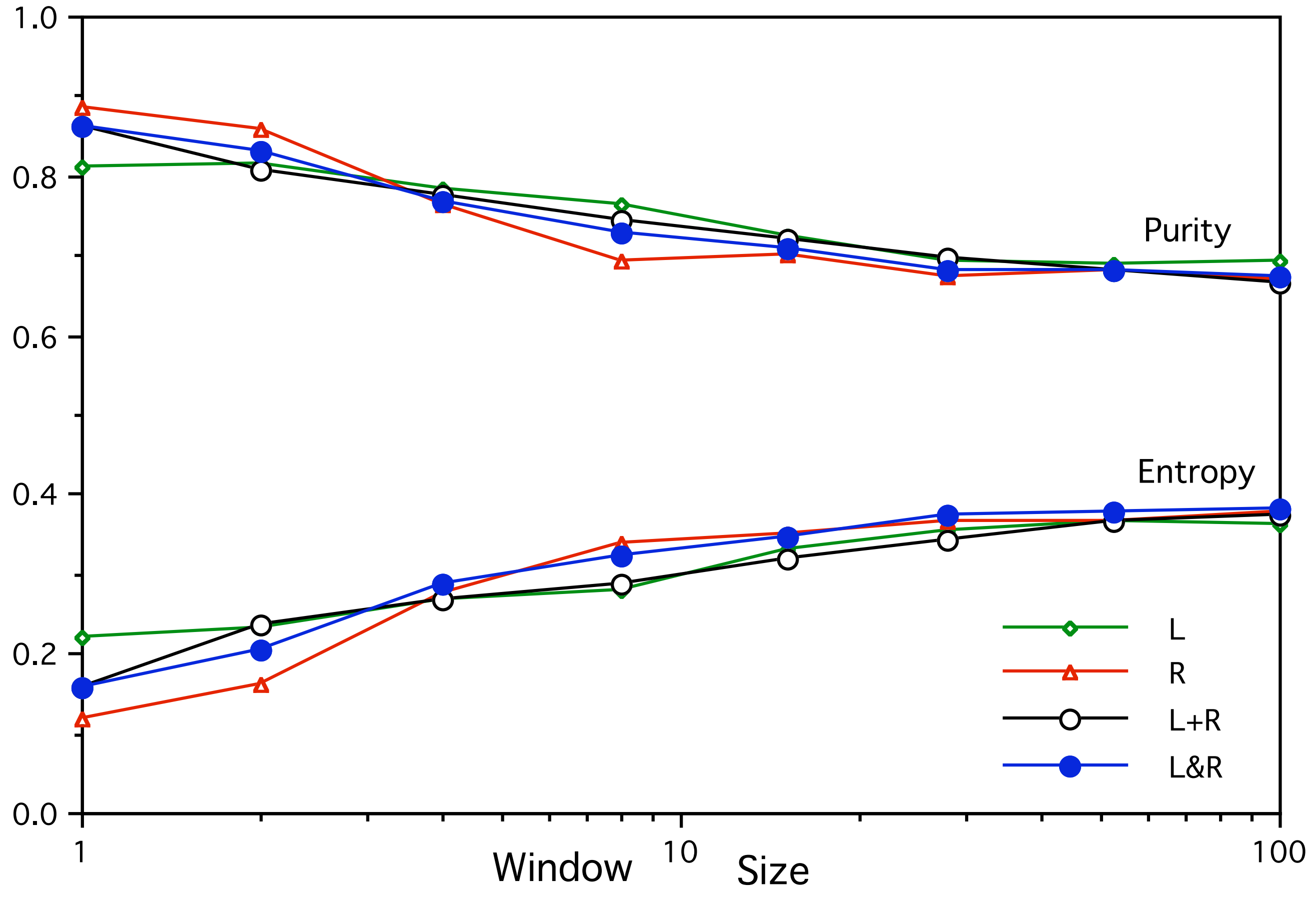
Nevertheless, it is worth asking if we get the same dependence on Vector Dimensionality, Window Size and Corpus Size as in the earlier study?

Then, what about Verbs and other tasks such as TOEFL?

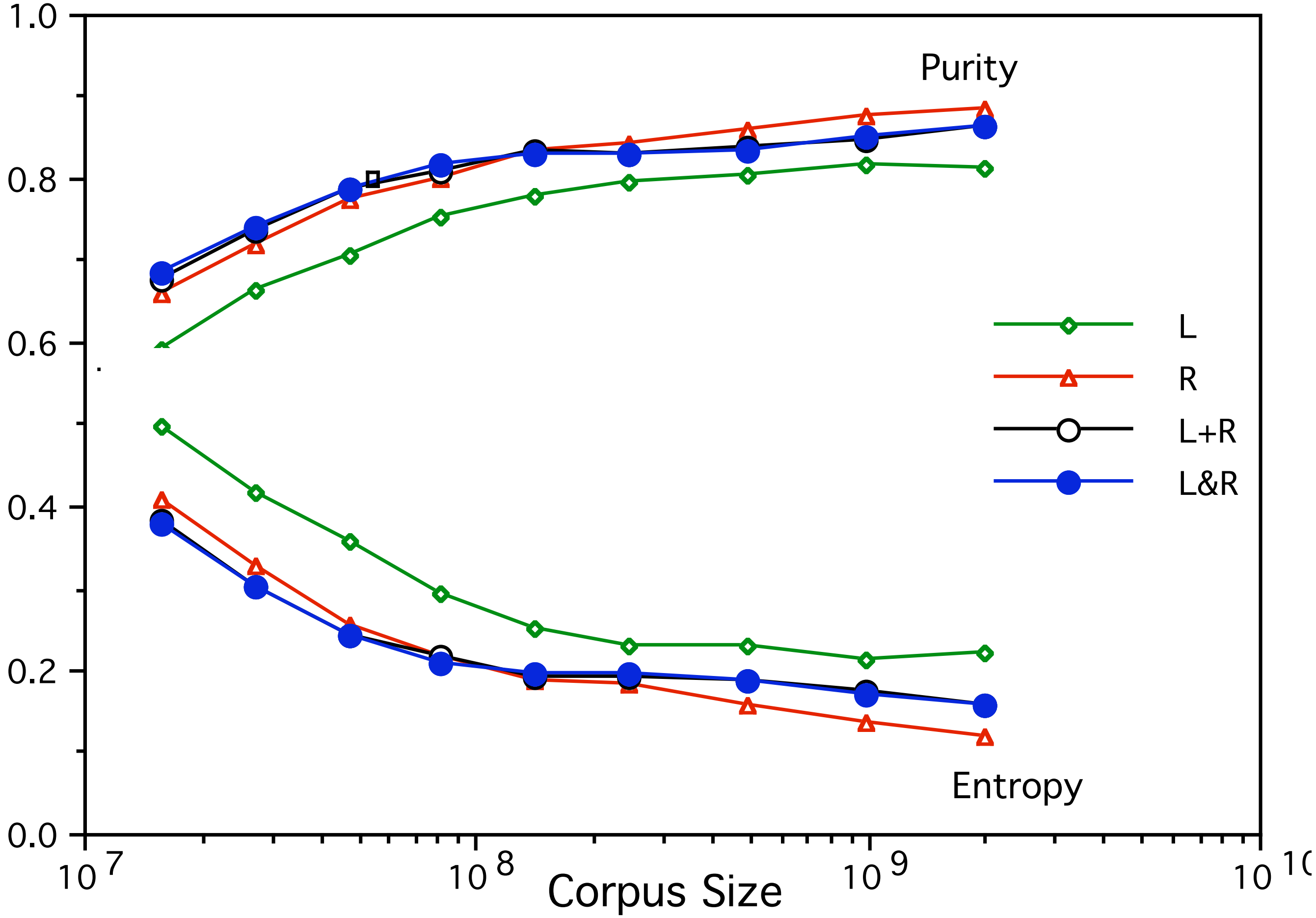
Effect of Vector Dimensionality



Effect of Window Size



Effect of Corpus Size

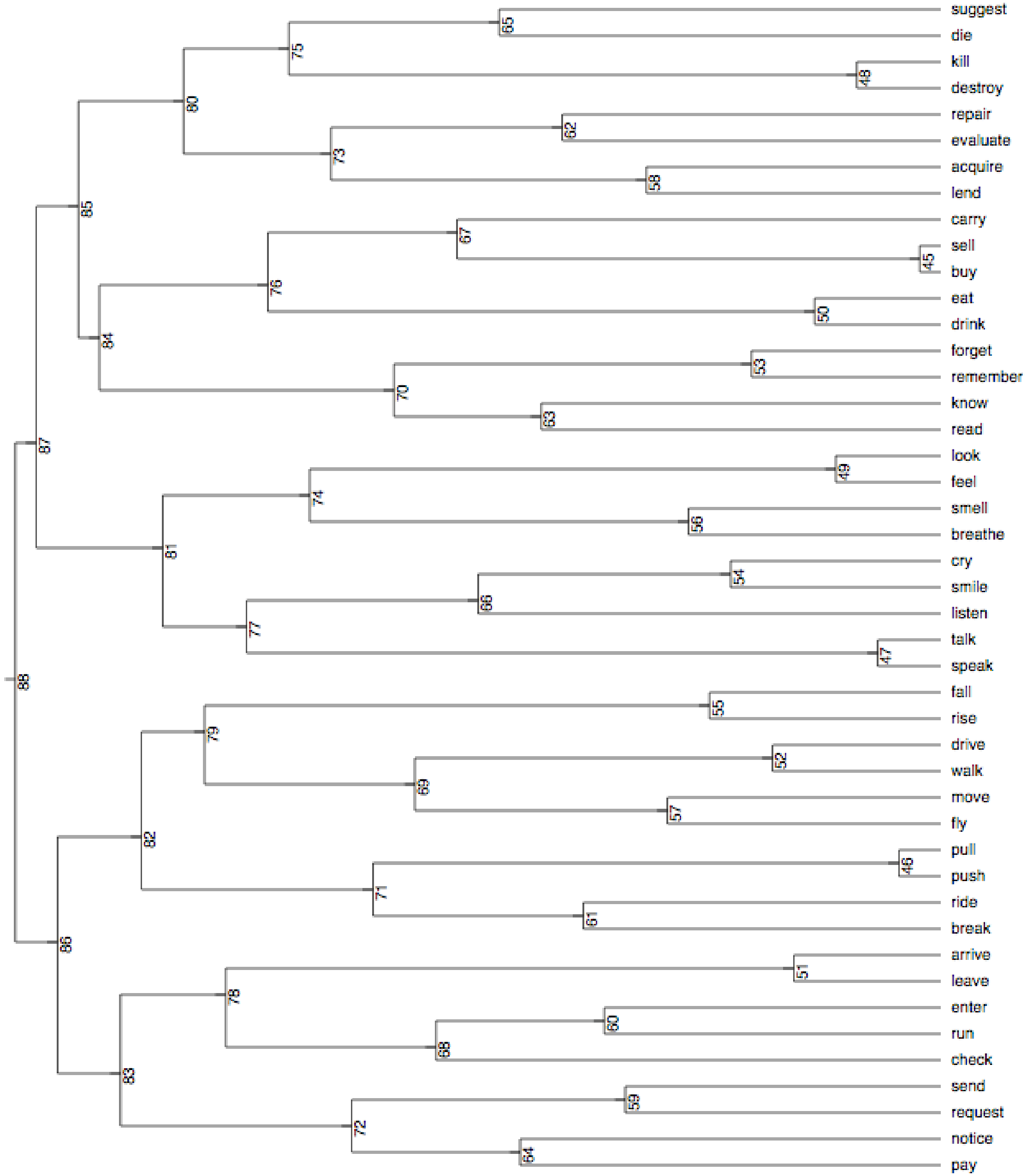


Verb Clustering

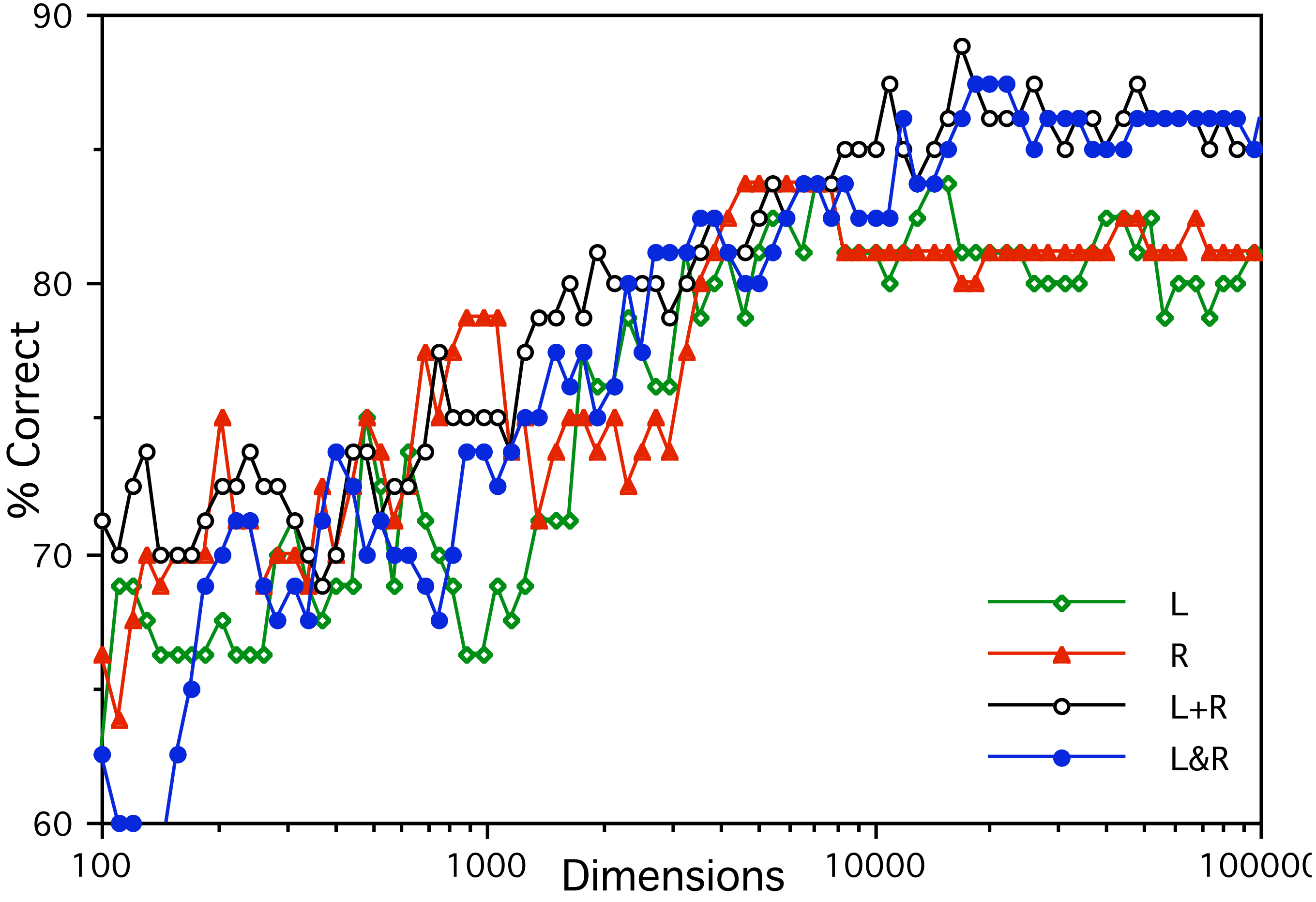
Five clusters:

Purity
= 0.644

Entropy
= 0.527



ukWaC TOEFL Performance



General Conclusions So Far?

Drawing general conclusions from such a small sample is dangerous, but it seems that vectors of simple word co-occurrence statistics lead to good semantic representations for concrete nouns, but not for verbs.

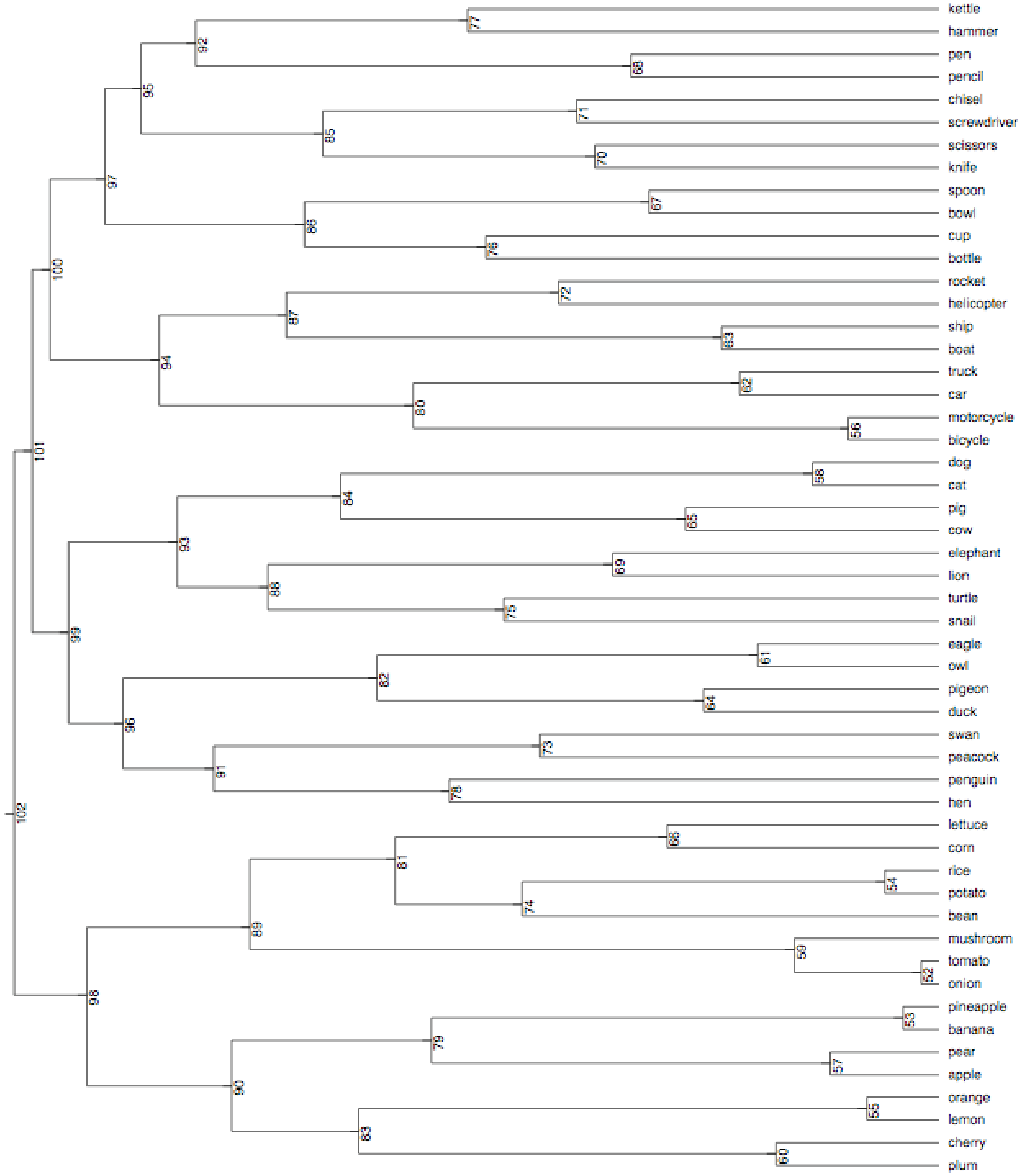
One current technical problem is that small word sets lead to sparse clusters, but larger word sets are difficult to manage computationally.

There are also more fundamental problems with the merging of vectors for different word meanings and different valid dimensions of semantics.

Perhaps, before dealing with these issues, we should first improve the current small concrete noun set by removing outliers and increasing/evening the class sizes, and optimise the semantic vector generation process on that?

Varying the noun set: 'chicken' → 'hen', adding 'pork' and 'beef', ...

Extended Concrete Noun Set Clustering



Six clusters:

Purity = 1.000

Entropy = 0.000

Making Further Progress?

There are clearly fundamental limitations to the simple co-occurrence statistics approach for generating semantic representations

But there remain many potential avenues for future work:

- Machine learning to split merged representations?
- Discriminant analysis for different aspects of semantics?
- Totally different co-occurrence statistics?
- Other ideas from other speakers?

There is certainly much scope for future progress in this field…

That's all for today!