Laboratoire de Sciences Cognitives et Psycholinguistique
Ecole Normale Supérieure, Paris

A Corpus-based Evaluation Method For Distributional Semantic Models

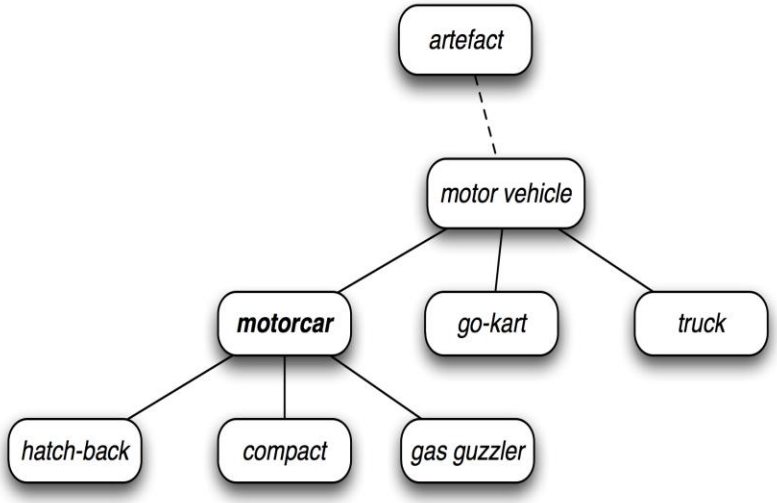
Abdellah Fourtassi & Emmanuel Dupoux

ACL- SRW, Sofia 2013

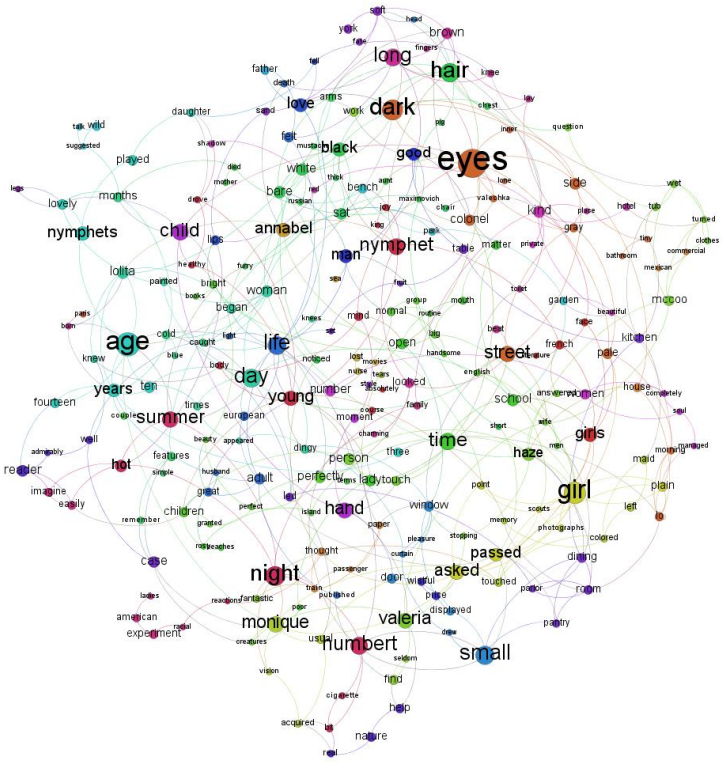
Word Similarity

Two classes of algorithms to measure word similarity:

Thesaurus-based
(WordNet)



Distributional
(LSA, LDA,..)



Evaluation Methods for DSMs

Extrinsic Methods

Quantitatively : Compare to human jugement (Benchmarks : Word association norms (Nelson et al, 1996), TOEFL synonym test (Laudauer, 1997)…)

Qualitatively : Nearest neighbers (researcher intuition)

HYPOTHESIS	STUDY	**CLASS**	ENGINE
EXPERIMENT	**TEST**	MARX	FUEL
SCIENTIFIC	STUDYING	ECONOMIC	ENGINES
OBSERVATIONS	HOMEWORK	CAPITALISM	STEAM
SCIENTISTS	NEED	CAPITALIST	GASOLINE
EXPERIMENTS	**CLASS**	SOCIALIST	AIR
SCIENTIST	MATH	SOCIETY	**POWER**
EXPERIMENTAL	TRY	SYSTEM	COMBUSTION
TEST	TEACHER	**POWER**	DIESEL
METHOD	WRITE	RULING	EXHAUST

Griffiths et al. (2007)

What if the researcher does not speak the language?

Evaluation Methods for DSMs

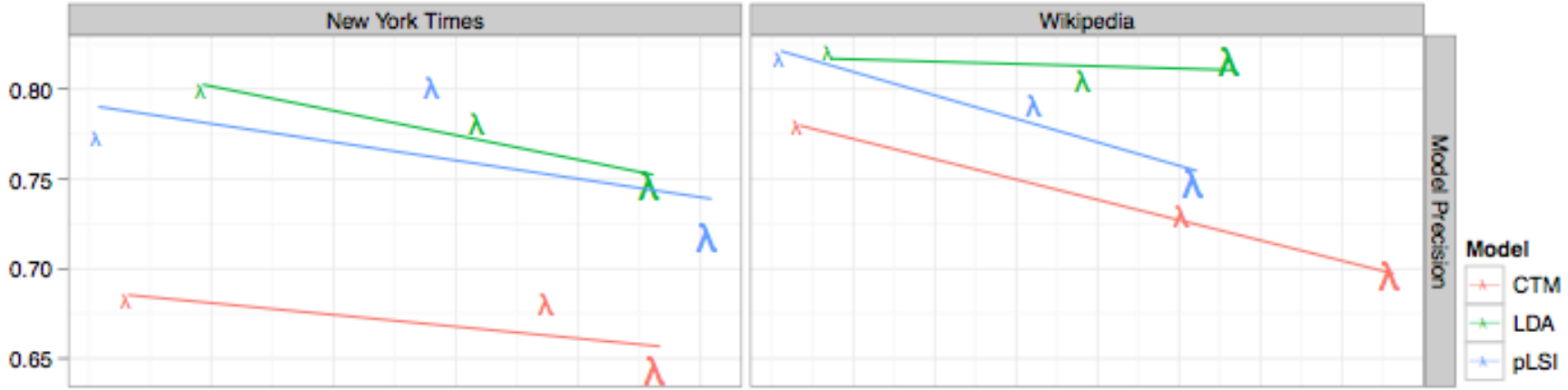
Intrinsic Methods

Held-out Likelihood

→ Possible only for proabilistic models (pLSA, LDA,..)

→ Costly computationally

→ Do not predict human jugement !! (Chang, 2009)



Evaluation Method

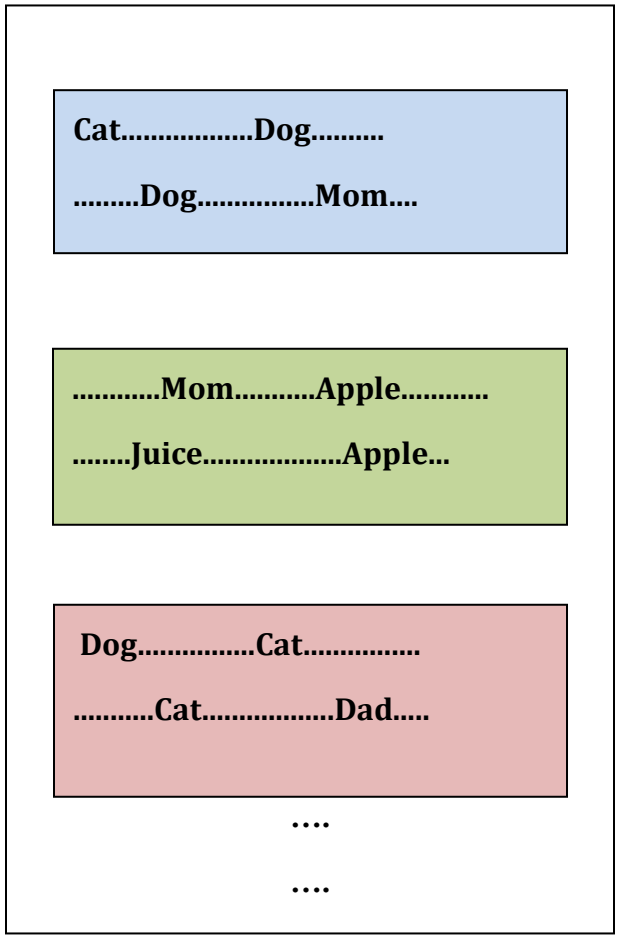
Our Method

→ Intrinsic : no external resources or mastery of the language required

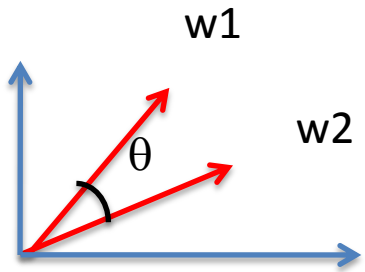
→ Predicts human jugement

→ Easy to implement

Latent Semantic Analysis



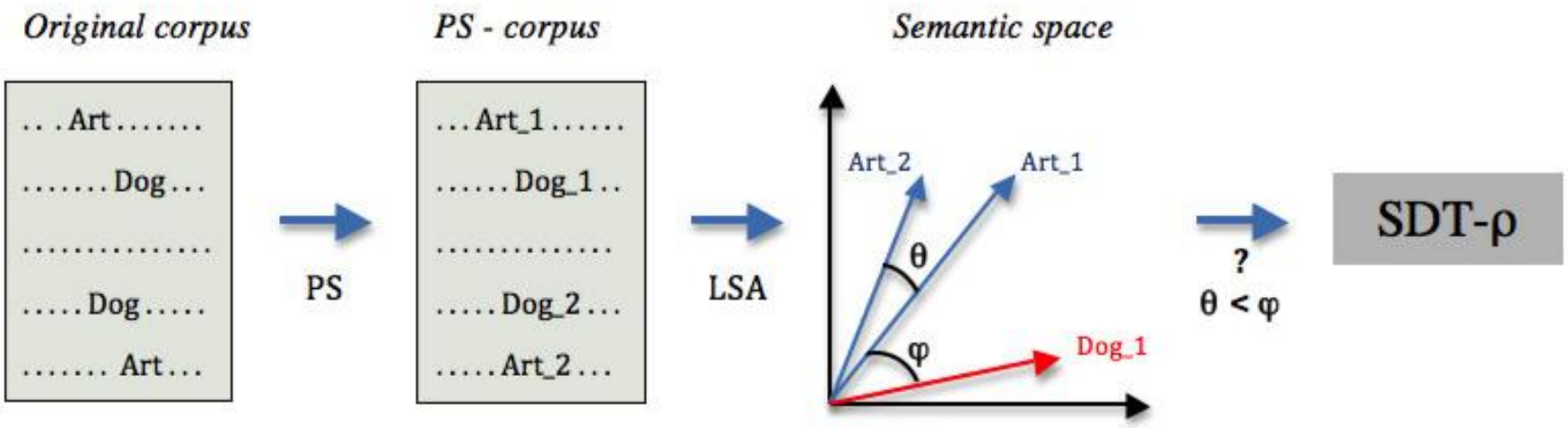
	Context 1	Context 2	Context 3	...	
Cat	1	0	2	...	
Mom	1	1	0	...	
Dog	2	0	1	...	
Juice	0	1	0	...	
Apple	0	2	0	...	
Dad	0	0	1	...	
	...	...	...	...	

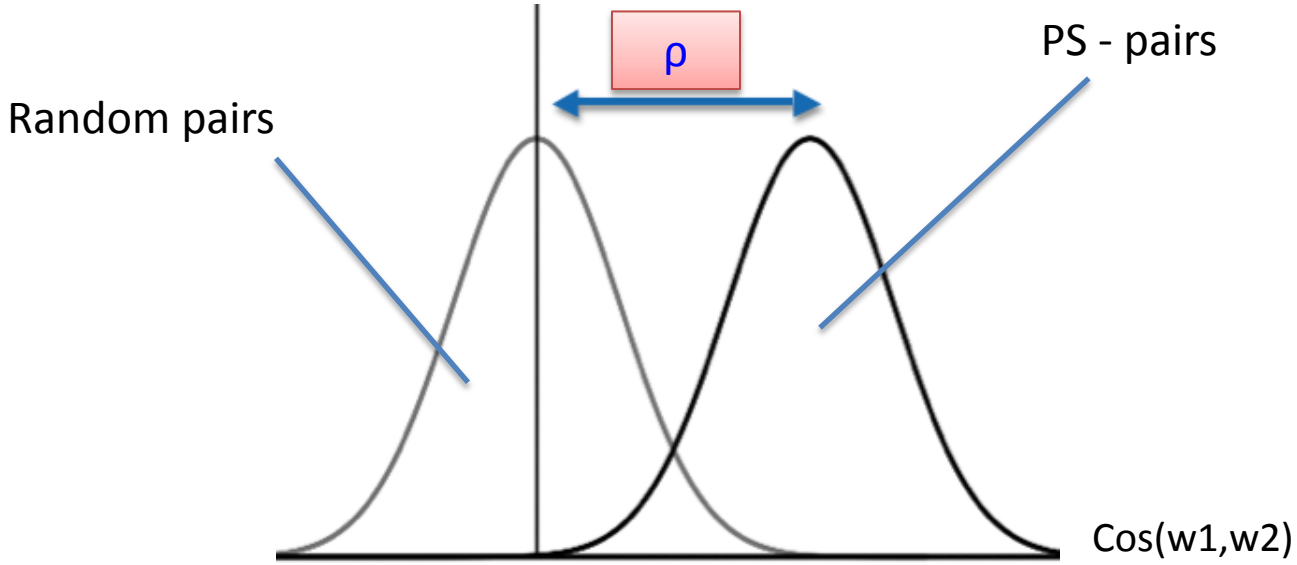


$$\cos(w_1, w_2) = \frac{w_1^T w_2}{\|w_1\|\|w_2\|}$$

Landauer and Dumais (1996)

SDT-ρ


Original corpus
PS - corpus
Semantic space
. . . Art . . . . . .
. . . . . Dog . . .
. . . . . .
. . . . . Dog . . . . .
. . . . . Art . . .
PS
. . . Art_1 . . . . . .
. . . . . Dog_1 . .
. . . . . .
. . . . . Dog_2 . . .
. . . . . Art_2 . . .
LSA
Art_2
Art_1
θ
φ
Dog_1
?
θ < φ
SDT-ρ


ρ
PS - pairs
Random pairs
Cos(w1,w2)

Evaluation Method

SDT-ρ

→ Intrinsic : no external resources or mastery of the language required

→ Predicts human jugement

→ Easy to implement

Experiment I

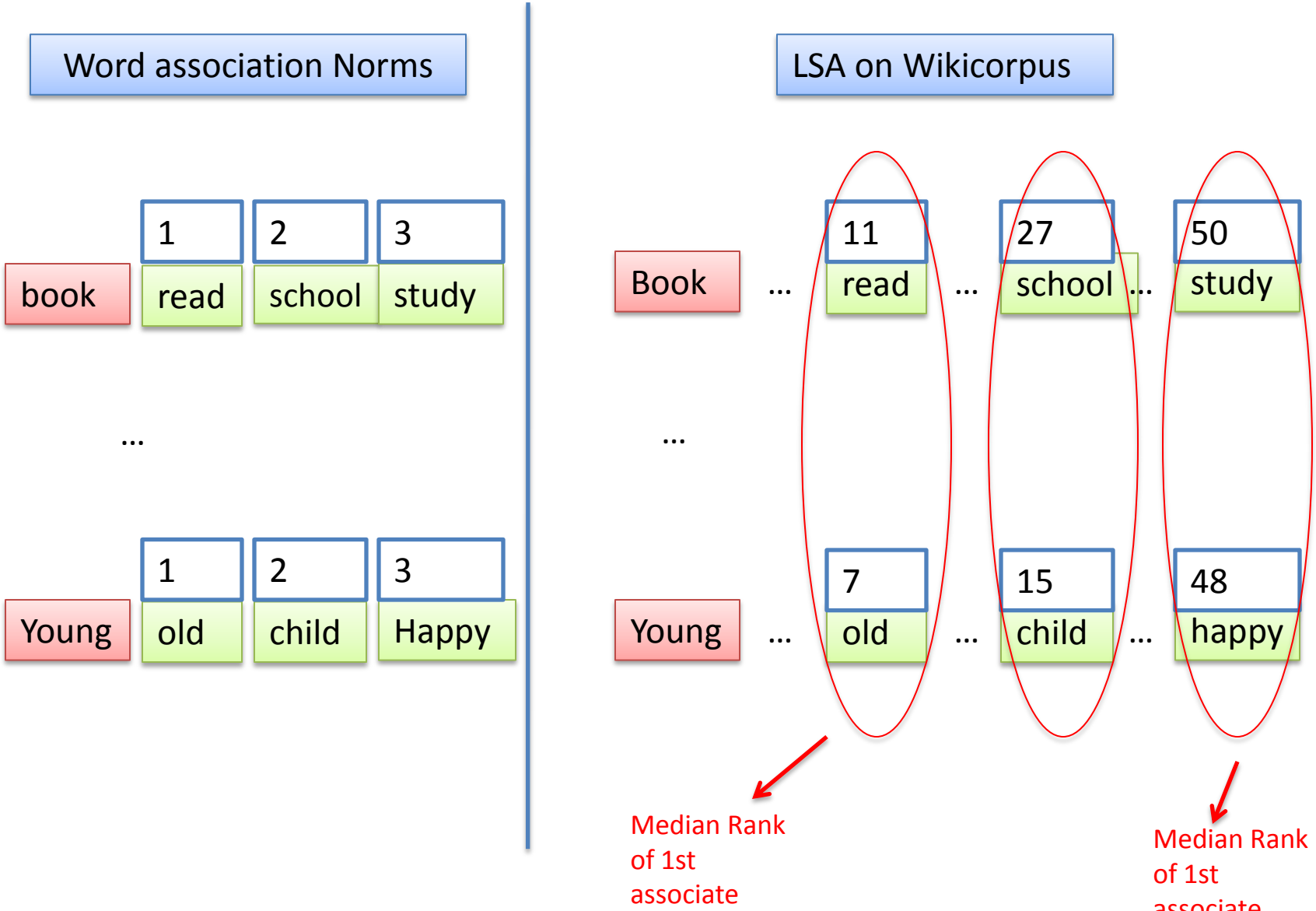
Word association Norms

	1	2	3
book	read	school	study

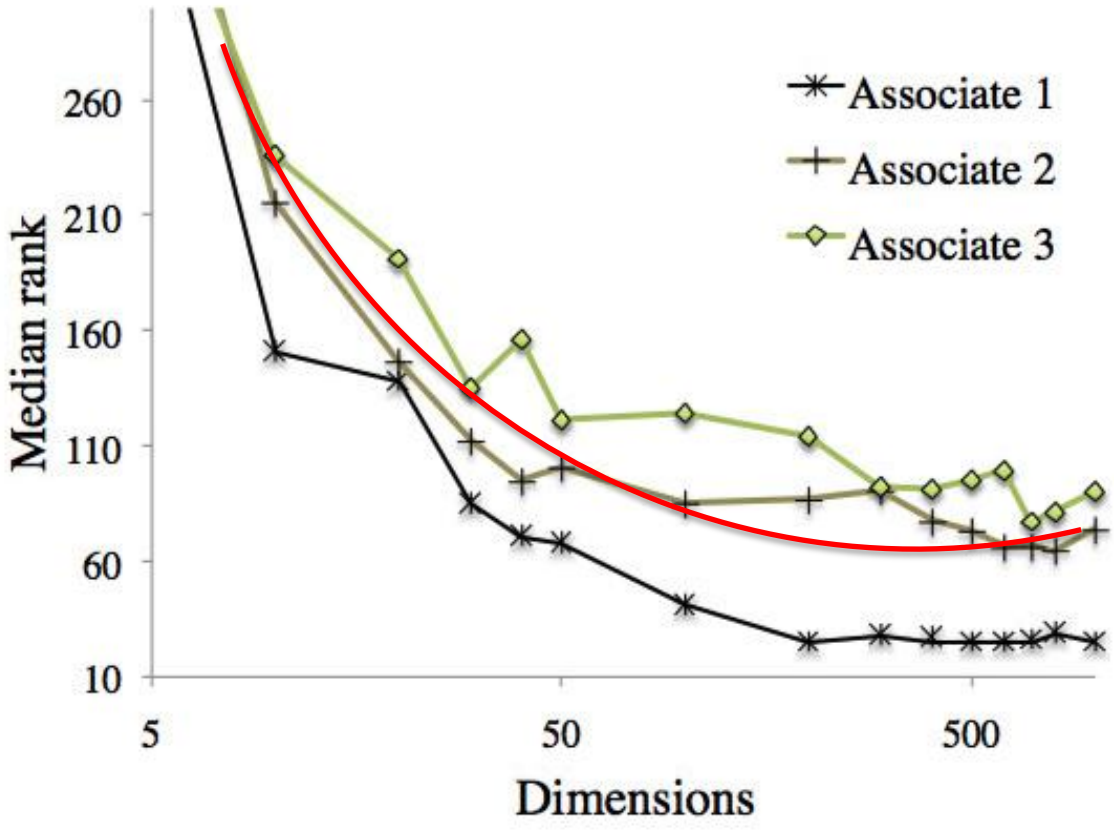
…

	1	2	3
Young	old	child	Happy

Experiment I



Experiment I



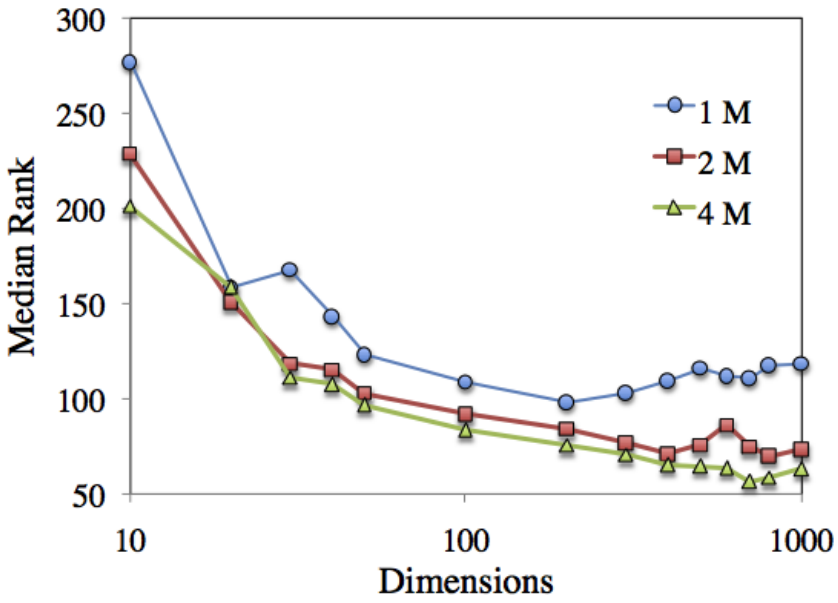
Median Rank

↕ ?

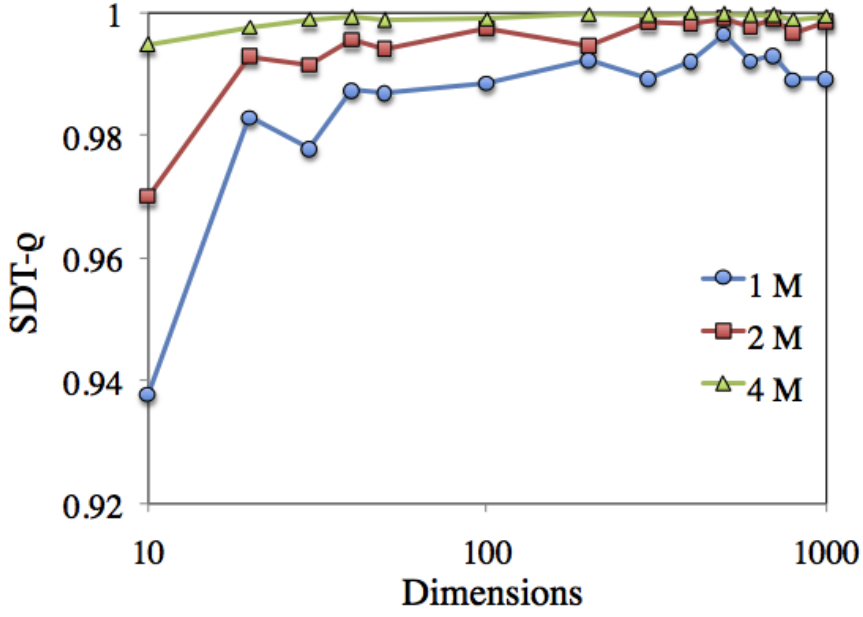
SDT-ρ

Results I

Correlation :
Dimensions and Corpus size



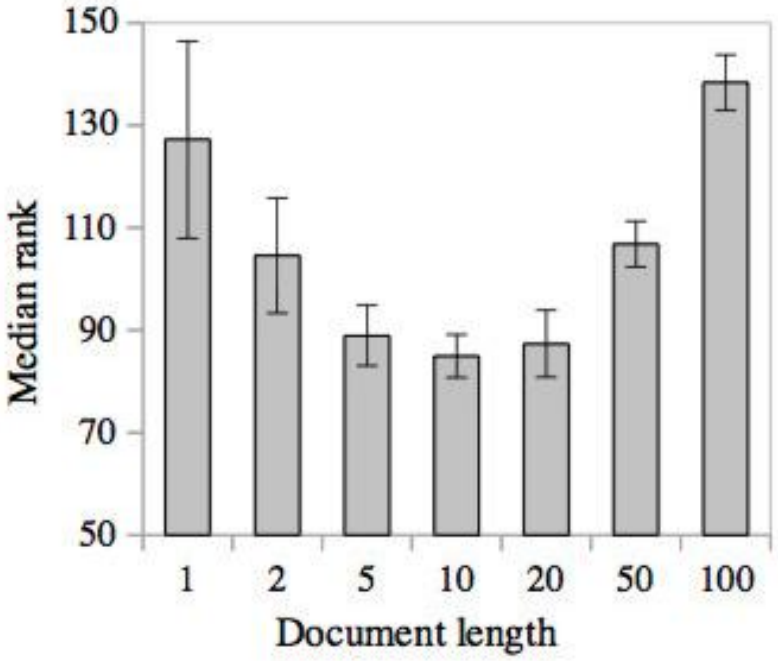
Lower is better



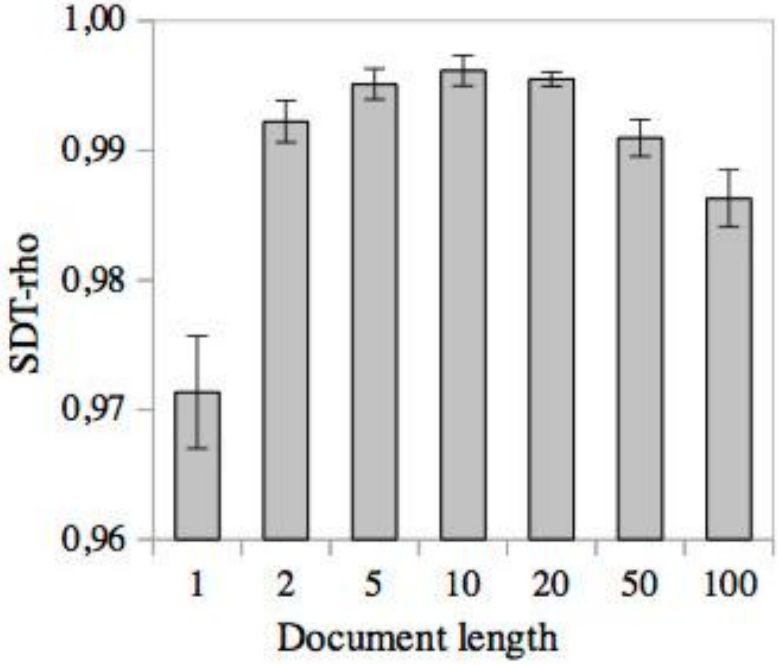
Higher is better

Results I

Correlation :
Document length



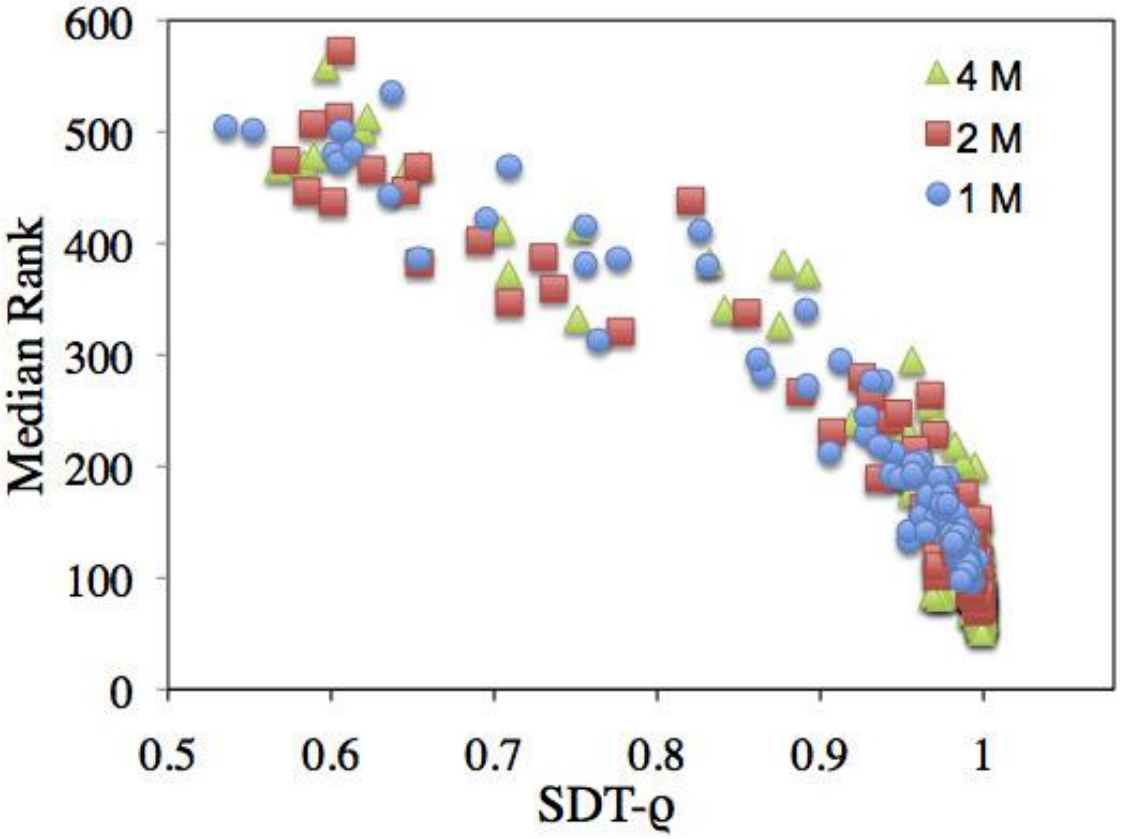
Lower is better



Higher is better

Results I

Correllation :
Overall



MIC=0.677
p<0.0001

Experiment II

TOEFL Synonym test (Landauer & Dumais, 1997)

Costly
→beautiful
→expensive
→popular
→complicated

Prolific
→productive
→serious
→capable
→promising

.....

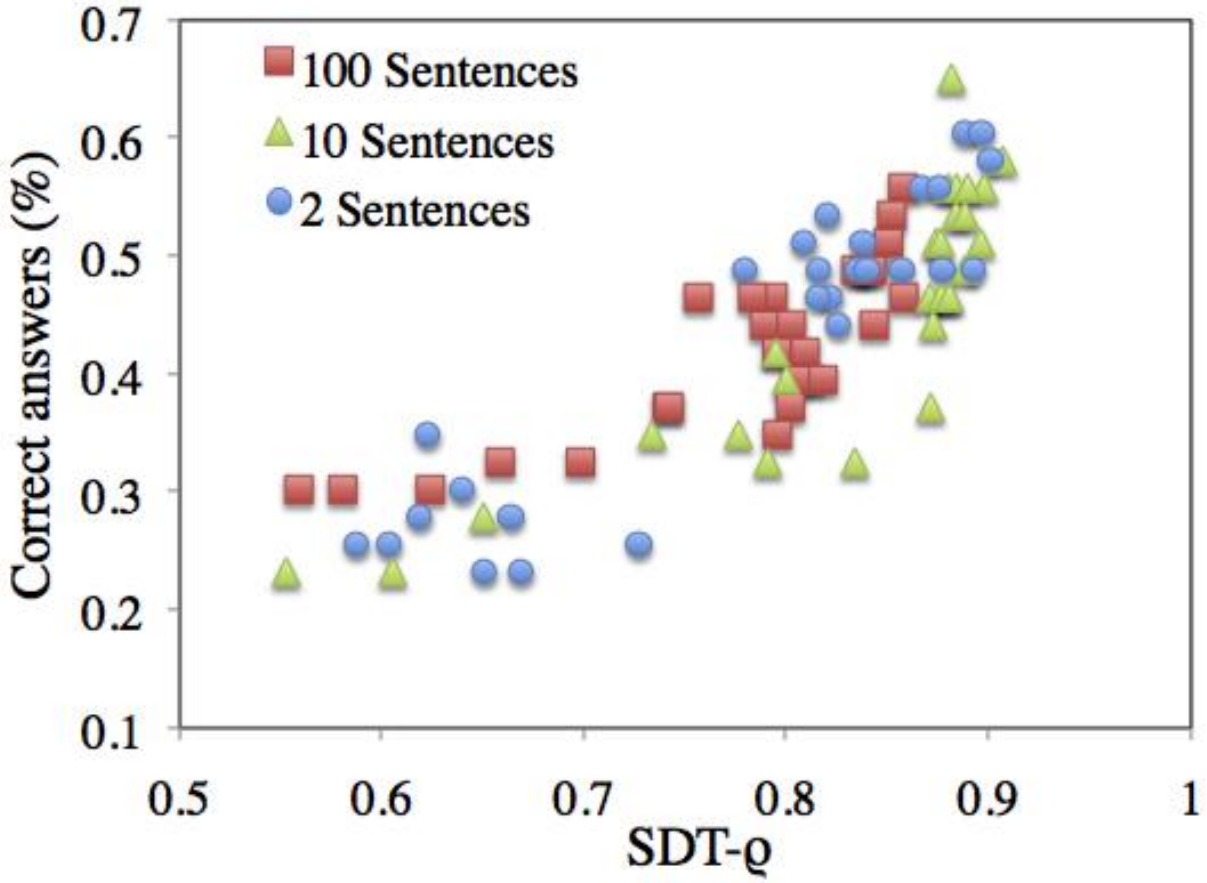
% of Correct Answers

?

SDT-ρ

Results II

Correllation :
Overall



MIC=0.675
p<0.0001

Evaluation Method

SDT-ρ

→ Intrinsic : no external resources or mastery of the language required

→ Predict human jugement

→ Easy to implement

How to use in practice

→ Correlation computed using the ovelap with the benchmarks (Norms, TOEFL)

→ What set of words to use as Pseudo-Synonyms in computing the SDT-ρ?

- Random words?
- Set size?
- Word frequency?

Correlation with Median Rank

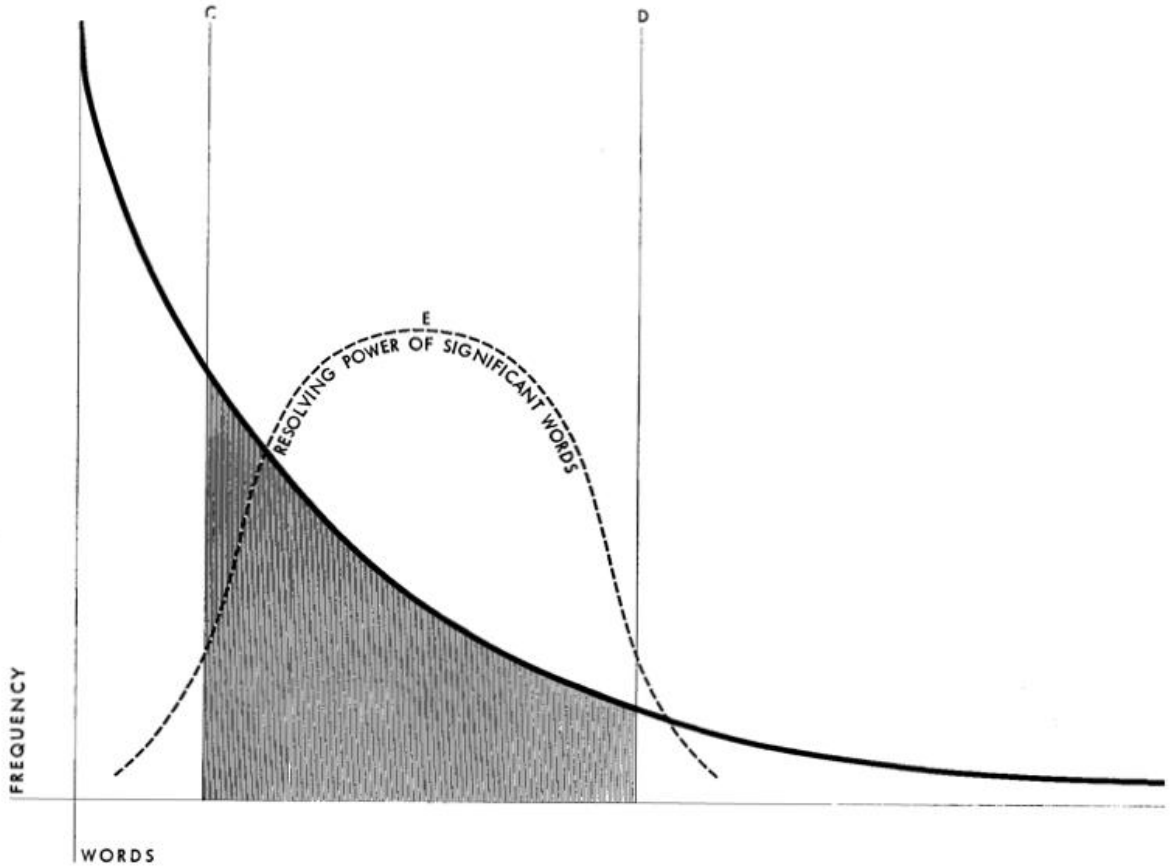
Freq. x	$1 < x < 40$			$40 < x < 400$			$x > 400$			**All**	**Overlap**
Size	100	500	1000	100	500	1000	100	500	1000	∼ 4 M	1093
MIC	0.311	0.219	0.549*	0.549*	**0.717***	**0.717***	0.311	0.205	0.419	0.549*	**0.717***

* : $p < 0.05$

A small set of random mid-frequency words give the highest correlation with human jugement

How to use in practice

Mid-frequency words have high disriminating
(« resolving ») power



Luhn (1958)

Evaluation Method

SDT-ρ

→ Intrinsic : no external resources or mastery of the language required

→ Predict human jugement

→ Easy to implement

Conclusion

→ An evaluation method to be used as a proxy when no human generated data is available

→ This method could be used to set the parameters of the semantic models(semantic dimension, doc length, corpus size,..)

→ Gives a global and rough measure of the quality of the semantic structure

→ **Future work** : a word level variant of the method will, a priori, enable us to assess some fine grained properties (distribution properties of absract/concrete words or linguistic categories,…)

Questions

Thanks for your attention!

Questions?