



# ΓΛΩΣΣΙΚΗ ΤΕΧΝΟΛΟΓΙΑ

Κάτια Κερμανίδου [kerman@ionio.gr](mailto:kerman@ionio.gr)

Αποσπάσματα κάποιων διαφανειών είναι δανεισμένα από τα  
slides του Speech and Language Processing

ΣΗΜΑΣΙΟΛΟΓΙΑ

ΔΙΑΝΥΣΜΑΤΙΚΕΣ  
ΑΝΑΠΑΡΑΣΤΑΣΕΙΣ

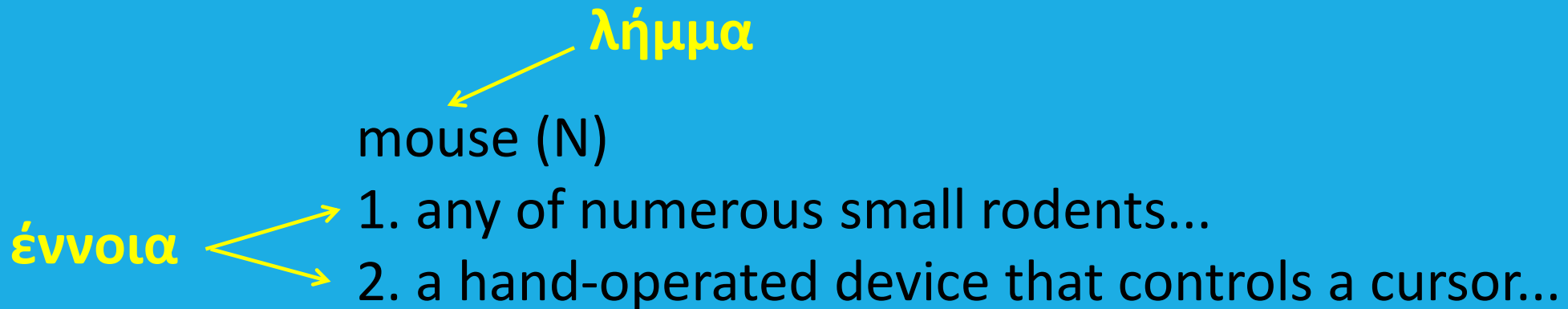
# Οι λέξεις έχουν έννοιες

Οι λέξεις μιας γλώσσας δεν είναι απλά αλφαριθμητικά, ούτε απλά καταχωρήσεις σε ένα λεξικό.

Οι λέξεις ανήκουν σε νοηματικές κατηγορίες (types/categories/classes)

Πχ Όλοι οι άνθρωποι είναι θνητοί.

$\forall x \text{ ANΘΡΩΠΟΣ}(x) \rightarrow \text{ΘNHΤΟΣ}(x)$



# Σχέσεις μεταξύ των εννοιών

## Συνωνυμία (Synonymy)

Οι συνώνυμες λέξεις έχουν ίδιες ή πολύ κοντινές έννοιες σε κάποια ή σε όλα τα συμφραζόμενα

Δύσκολα υπάρχει τέλεια ταύτιση εννοιών (θα υπάρχει διαφοροποίηση στο ύφος)

Αμάξι/αυτοκίνητο    νερό/ύδωρ    θέλω/επιθυμώ

## Αντωνυμία (Antonymy)

Οι αντώνυμες λέξεις έχουν αντίθετες έννοιες

Γρήγορος/αργός    κρύο/ζεστό    πάνω/κάτω

## Μερωνυμία (Meronymy) / Ολωνυμία (Holonymy)

Το χερούλι είναι **μερώνυμο** της πόρτας.

Η πόρτα είναι **ολώνυμο** του χερουλιού.

# Περισσότερες Σχέσεις μεταξύ των Εννοιών

## Σημασιολογική Ομοιότητα (Semantic Similarity)

Λέξεις που μοιράζονται κάποια εννοιολογικά στοιχεία

Αυτοκίνητο/μοτοσικλέτα/ποδήλατο                      άλογο/αγελάδα/πρόβατο

Σε αντίθεση με:            ποδήλατο – ευτυχία - σκουμπρί

## Σημασιολογική Συσχέτιση (Semantic Relatedness)

Λέξεις που συσχετίζονται σημασιολογικά (είναι αναμενόμενο να συνεμφανιστούν σε ένα σενάριο)

καφές/κούπα/ζάχαρη

σερβιτόρος/μενού/φαγητό/επιδόρπιο/σεφ

χειρουργός/νυστέρι/αναισθητικό/νοσοκόμα/κλινική

| word1  | word2      | similarity |
|--------|------------|------------|
| vanish | disappear  | 9.8        |
| behave | obey       | 7.3        |
| belief | impression | 5.95       |
| muscle | bone       | 3.65       |
| modest | flexible   | 0.98       |
| hole   | agreement  | 0.3        |

# Συναίσθημα στην Έννοια (Connotation)

Συχνά οι έννοιες λέξεων έχουν **συναισθηματικό περιεχόμενο**

## Πόλωση (Polarity/Valence)

Θετικό (χαρούμενος, φανταστικά)

Αρνητικό (λυπημένος, χάλια)

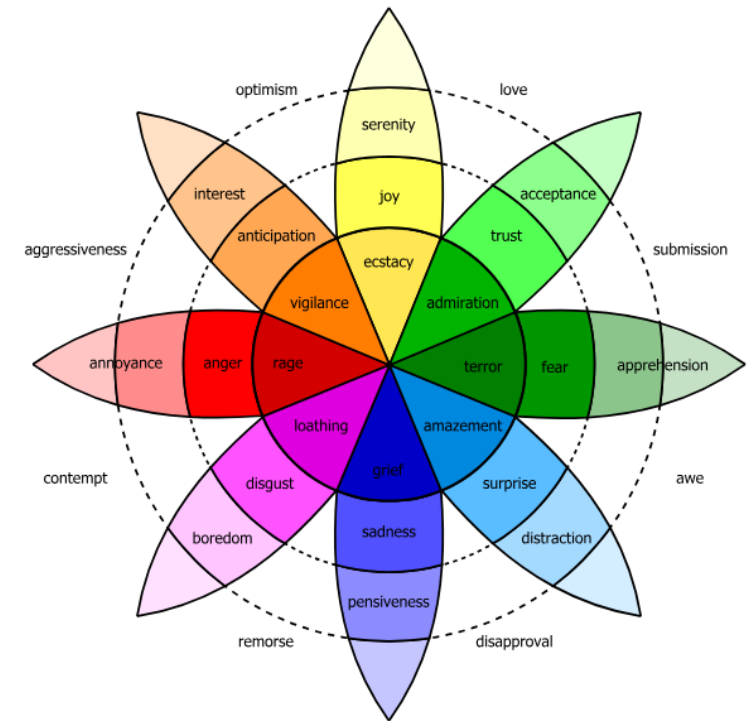
Το πρόσημο του συναισθήματος μπορεί να είναι λιγότερο ευδιάκριτο

απομίμηση/αντίγραφο/αναπαραγωγή

## Ένταση/Επίπεδο Ενέργειας (Arousal)

Υψηλή (εκστασιασμένος, μανιακός)

Χαμηλή (χαλαρός, γαλήνιος, ναρκωμένος)



**Figure 22.2** Plutchik wheel of emotion.

# Πώς αναπαριστώ την έννοια μιας λέξης/φράσης/πρότασης

Με γλώσσες αναπαράστασης νοήματος

- Κατηγορηματική Λογική (Predicate Logic)
  - Πολύ αυστηρή, μαθηματική
  - Δεν έχει μεγάλη περιγραφική ικανότητα ώστε να αναπαραστήσει πληθώρα νοημάτων
- Λογική Υψηλότερης Τάξης (Higher-Order Logic)
- Πιο περιγραφικοί, λιγότερο τυπικοί φορμαλισμοί
  - Σημασιολογικά Δίκτυα, Σενάρια, Πλαίσια, Λογική Φόρμα

# Συνθετική Σημασιολογία / Compositional Semantics

- Συνθετικότητα: Το νόημα ενός συστατικού βρίσκεται αποκλειστικά από το νόημα των υποσυστατικών του
  - π.χ. το νόημα της φράσης «**κρύο νερό**» σχηματίζεται από τον συνδυασμό της έννοιας «**κρύο**» με την έννοια «**νερό**» ή
  - π.χ. το νόημα μιας πρότασης βρίσκεται συνδυάζοντας το νόημα της ονοματικής φράσης και της ρηματικής φράσης που την σχηματίζουν
    - Ο κανόνας συνδυασμού των νοημάτων πρέπει να είναι ανεξάρτητος από το πώς παράγονται τα νοήματα της κάθε φράσης
- Η ανάπτυξη ενός καλού συνθετικού μοντέλου είναι πολύ δύσκολη
- Δεν ισχύει πάντα
  - «Βρέχει καρέκλες»

# Πόροι (Resources) στο Σημασιολογικό Επίπεδο: Σημασιολογικά Λεξικά

- Αναπαράσταση του νοήματος κάθε καταχώρησης μέσω κάποιας γλώσσας αναπαράστασης νοήματος (πχ Κατηγορηματική Λογική)
  - “jumps”: *jumps(x)*
  - “loves”: *loves(y,x)*
  - “Mary”: *mary*
  - “John”: *john*

# Πόροι (Resources) στο Σημασιολογικό Επίπεδο: Ιεραρχία Τύπων (Type Hierarchy) ή Οντολογία

- ❑ Κάθε λέξη έχει τουλάχιστον ένα σημασιολογικό τύπο (κατηγορία, έννοια) που υποδεικνύει τις σημασιολογικές ιδιότητες της λέξης
- ❑ Σε μία ιεραρχία τύπων οι θυγατρικοί τύποι κληρονομούν τις ιδιότητες των μητρικών τύπων



Ποντίκι: ΤΡΩΚΤΙΚΑ ή ΗΛ.ΣΥΣΚΕΥΕΣ

# Πόροι (Resources) στο Σημασιολογικό Επίπεδο: Ιεραρχία Τύπων (Type Hierarchy) ή Οντολογία

- Κάθε τύπος προσδιορίζει συγκεκριμένες σημασιολογικές ιδιότητες
  - ΦΥΣ.ΑΝΤΙΚ: πιάνονται, βλέπονται
  - ΥΛΙΚΑ: δημιουργούνται, καταστρέφονται
  - ΖΩΝΤΑΝΑ: ζουν, πεθαίνουν
  - ΦΥΤΑ: τρώγονται, ανθίζουν
  - ΑΝΘΡΩΠΟΙ: διαβάζουν, τρώνε, πίνουν
  - ΤΡΩΚΤΙΚΑ: τρώνε, τρώγονται?

# Πόροι (Resources) στο Σημασιολογικό Επίπεδο: Γραμματικές Σημασιολογικών Ρόλων (Semantic Roles)

- ❑ Περιλαμβάνουν ένα σύνολο σημασιολογικών ρόλων (semantic roles)
- ❑ Οι σημασιολογικοί ρόλοι μπορούν να αντιστοιχιστούν σε διαφορετικές συντακτικές δομές
  - *Ο Γιάννης έσπασε το παράθυρο με το σφυρί.*
  - *Το σφυρί έσπασε το παράθυρο.*
  - *Το παράθυρο έσπασε.*
- ❑ **Υποκινητής της ενέργειας:** ο Γιάννης
- ❑ **Όργανο** που χρησιμοποιήθηκε: το σφυρί
- ❑ **Αντικείμενο** που δέχτηκε την ενέργεια: το παράθυρο

# Πόροι (Resources) στο Σημασιολογικό Επίπεδο:

## Οι Σημαντικότεροι Σημασιολογικοί Ρόλοι

- ❑ AGENT: ο υποκινήτης μιας ενέργειας
- ❑ THEME: το αντικείμενο που δέχεται μια ενέργεια
- ❑ INSTRUMENT: το αντικείμενο που χρησιμοποιείται σε μια ενέργεια
- ❑ EXPERIENCER: το άτομο που αντιλαμβάνεται κάτι ή βρίσκεται σε μία ψυχολογική κατάσταση
- ❑ BENEFICIARY: το άτομο για το οποίο έγινε μία ενέργεια
- ❑ FROM-LOC: αρχικός τόπος (πχ φεύγω από την Κέρκυρα)
- ❑ TO-LOC: τελικός τόπος (πχ πηγαίνω προς την Πάτρα)
- ❑ AT-LOC: στατικός τόπος (πχ είμαι στο σπίτι)
- ❑ FROM-TIME: αρχικός χρόνος (πχ φεύγω από τις 9πμ)
- ❑ TO-TIME: τελικός χρόνος (πχ θα διαρκέσει μέχρι τις 11μμ)
- ❑ AT-TIME: στατικός χρόνος (πχ η ώρα είναι 1μμ)
- ❑ FROM-POSS: αρχικός κάτοχος
- ❑ TO-POSS: τελικός κάτοχος

# Σημασιολογικοί Ρόλοι: Παραδείγματα

- Ο Γιάννης έσπασε το παράθυρο με το σφυρί.  
AGENT                      THEME                      INSTRUMENT
- Ο αέρας στέγνωσε τα ρούχα.  
INSTRUMENT                      THEME
- Ο Γιάννης κοιμήθηκε.  
EXPERIENCER
- Έδωσα στο Γιάννη το βιβλίο.  
TO-POSS      THEME      (AGENT: εγώ)
- Έδωσα το βιβλίο στο Γιάννη για τη Μαρία.  
THEME      TO-POSS                      BENEFICIARY
- Περπάτησα από εδώ ως το σχολείο.  
FROM-LOC      TO-LOC

# Διανυσματικές Αναπαραστάσεις Λέξεων

## Vector Representations

Κάθε λέξη είναι ένα σημείο/διάνυσμα σε έναν σημασιολογικό χώρο

Σημασιολογικά όμοιες λέξεις βρίσκονται κοντά μεταξύ τους σε αυτόν τον χώρο

**Γιατί διανύσματα;**

Πχ στην αναγνώριση μερών του λόγου αν χρησιμοποιήσω καθαυτές λέξεις στο παράθυρο συμφραζομένων, θα πρέπει το μοντέλο να βρει την ίδια λέξη και στο σώμα εκπαίδευσης και στο σώμα δοκιμής (πχ να ακολουθεί την λέξη εστίασης η λέξη 'cafeteria'.

| Λέξη-3 | Λέξη-2 | Λέξη-1 | Λέξη εστίασης | Λέξη+1    | Λέξη+2  | Λέξη+3 | MTA |
|--------|--------|--------|---------------|-----------|---------|--------|-----|
| --     | --     | --     | The           | cafeteria | remains | closed | DT  |

**Με διανύσματα** η λέξη 'cafeteria' είναι ένα διάνυσμα αριθμών, πχ το [35, 22, 17, ...]. Μπορεί στο σώμα δοκιμής να μην βρω το ίδιο, αλλά κάποιο κοντινό, πχ το [34, 21, 19, ...].

Οπότε μπορώ να κάνω **γενίκευση** σε παρόμοιες λέξεις, ακόμα και αν δεν υπάρχουν στο σώμα εκπαίδευσης!



# Διανυσματικές Αναπαραστάσεις Λέξεων: One-hot vectors

|      | abandon | ability | able | ... | ants | ... | zone |
|------|---------|---------|------|-----|------|-----|------|
| ants | 0       | 0       | 0    | 0   | 1    | 0   | 0    |

- Ο αριθμός των στηλών είναι ο αριθμός των λέξεων στο λεξικό της γλώσσας μου
  - Της τάξης των εκατοντάδων χιλιάδων
- Πώς θα αποδώσω την σημασιολογική ομοιότητα ανάμεσα στο «Seattle motel» και στο «Seattle hotel»;
- τα παρακάτω διανύσματα είναι ορθοκανονικά - (ομοιότητα συνημιτόνου=0)

```
motel = [0 0 0 0 0 0 0 0 0 0 1 0 0 0 0]
hotel = [0 0 0 0 0 0 0 1 0 0 0 0 0 0 0]
```

- Λύση: έμμεση απεικόνιση της ομοιότητας στα διανύσματα, λαμβάνοντας υπόψη το συμφραστικό περιβάλλον που εμφανίζεται η λέξη

# Κατανεμημένα Διανύσματα Λέξεων Distributional word vectors

You shall know a word by the company it keeps" (J. R. Firth, 1957)

government debt problems turning into banking crises as has happened in  
saying that Europe needs unified banking regulation to replace the hodgepodge

↖ These words will represent *banking* ↗

Μια από τις σημαντικότερες ιδέες της σύγχρονης  
στατιστικής ΕΦΓ

# Word Representations μέσω των κειμένων που οι λέξεις εμφανίζονται (1/3)

|       | doc1 | doc2 | doc3 | doc4 | ... |
|-------|------|------|------|------|-----|
| Λέξη1 | 0    | 0    | 0    | 1    |     |
| Λέξη2 | 1    | 0    | 1    | 1    |     |
| Λέξη3 | 1    | 0    | 0    | 0    |     |
| Λέξη4 | 0    | 1    | 0    | 0    |     |
| ...   |      |      |      |      |     |

Term-document incidence matrix

|       | doc1 | doc2 | doc3 | doc4 | ... |
|-------|------|------|------|------|-----|
| Λέξη1 | 0    | 0    | 0    | 3    |     |
| Λέξη2 | 4    | 0    | 17   | 7    |     |
| Λέξη3 | 9    | 0    | 0    | 0    |     |
| Λέξη4 | 0    | 6    | 0    | 0    |     |
| ...   |      |      |      |      |     |

Term-document frequency matrix

# Word Representations μέσω των κειμένων που οι λέξεις εμφανίζονται (2/3)

- tf: η συχνότητα εμφάνισης της λέξης στο κείμενο (term frequency)
- df: ο αριθμός των εγγράφων στα οποία εμφανίζεται η λέξη (document frequency)
- idf:  $\log_{10} (N/df)$ , όπου N ο συνολικός αριθμός εγγράφων (inverse document frequency)
- $tfidf = tf * idf$

|       | doc1 | doc2 | doc3 | doc4 | ... |
|-------|------|------|------|------|-----|
| Λέξη1 | 0    | 0    | 0    | 1.02 |     |
| Λέξη2 | 4.11 | 0    | 1.65 | 2.36 |     |
| Λέξη3 | 2.89 | 0    | 0    | 0    |     |
| Λέξη4 | 0    | 1.08 | 0    | 0    |     |
| ...   |      |      |      |      |     |

Term-document tf-idf matrix

# Word Representations μέσω των κειμένων που οι λέξεις εμφανίζονται

## (3/3): Latent Semantic Indexing

- Singular Value Decomposition
- Ένας πίνακας  $A$  ( $M \times N$ ) μπορεί να διαχωριστεί σε 3 πίνακες ως εξής:

$$A = U \Sigma V^T$$

The diagram illustrates the dimensions of the matrices in the SVD equation  $A = U \Sigma V^T$ . Below the equation, there are three boxes: the first box contains  $M \times M$  and has an arrow pointing to  $U$ ; the second box contains  $M \times N$  and has an arrow pointing to  $\Sigma$ ; the third box contains  $N \times N$  and has an arrow pointing to  $V^T$ .

- Οι στήλες του  $U$  είναι τα ιδιοδιανύσματα του πίνακα  $AA^T$ .
- Οι στήλες του  $V$  είναι τα ιδιοδιανύσματα του πίνακα  $A^T A$ .
- Ο  $\Sigma$  είναι διαγώνιος πίνακας, με στοιχεία της διαγωνίου τα  $\sigma_i = \sqrt{\lambda_i}$ ,  
όπου  $\lambda_1 \dots \lambda_r$ , οι ιδιοτιμές (eigenvalues) του  $AA^T$  (ή του  $A^T A$  – ίδιες είναι)
- $\sigma_i$ : ιδιάζουσες τιμές (singular values)
- Αν στον υπολογισμό του παραπάνω γινομένου κρατήσω μόνο τις  $k$  μεγαλύτερες ιδιάζουσες και μηδενίσω τις υπόλοιπες, τότε υπολογίζω τον  $A_k$ , μια προσέγγιση του  $A$ , τάξης  $k$

# Latent Semantic Indexing: Παράδειγμα

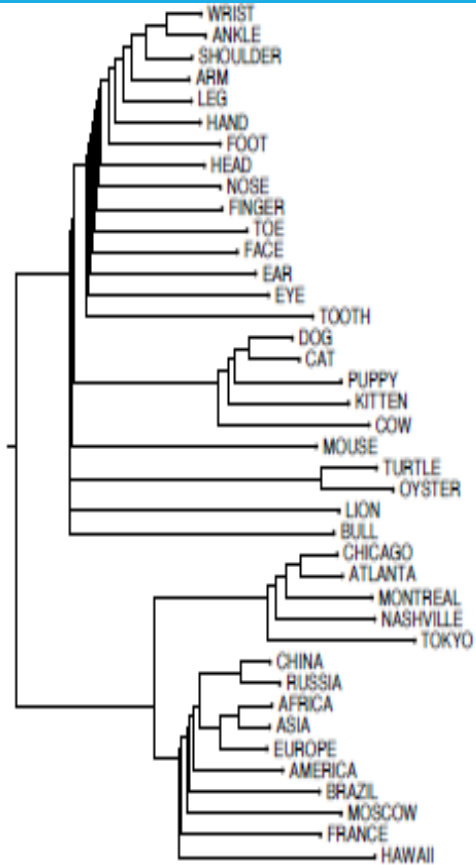
| $C$   | $d_1$ | $d_2$ | $d_3$ | $d_4$ | $d_5$ | $d_6$ |
|-------|-------|-------|-------|-------|-------|-------|
| ship  | 1     | 0     | 1     | 0     | 0     | 0     |
| boat  | 0     | 1     | 0     | 0     | 0     | 0     |
| ocean | 1     | 1     | 0     | 0     | 0     | 0     |
| wood  | 1     | 0     | 0     | 1     | 1     | 0     |
| tree  | 0     | 0     | 0     | 1     | 0     | 1     |

| $C_2$ | $d_1$ | $d_2$ | $d_3$ | $d_4$ | $d_5$ | $d_6$ |
|-------|-------|-------|-------|-------|-------|-------|
| ship  | 0.85  | 0.52  | 0.28  | 0.13  | 0.21  | -0.08 |
| boat  | 0.36  | 0.36  | 0.16  | -0.20 | -0.02 | -0.18 |
| ocean | 1.01  | 0.72  | 0.36  | -0.04 | 0.16  | -0.21 |
| wood  | 0.97  | 0.12  | 0.20  | 1.03  | 0.62  | 0.41  |
| tree  | 0.12  | -0.39 | -0.08 | 0.90  | 0.41  | 0.49  |

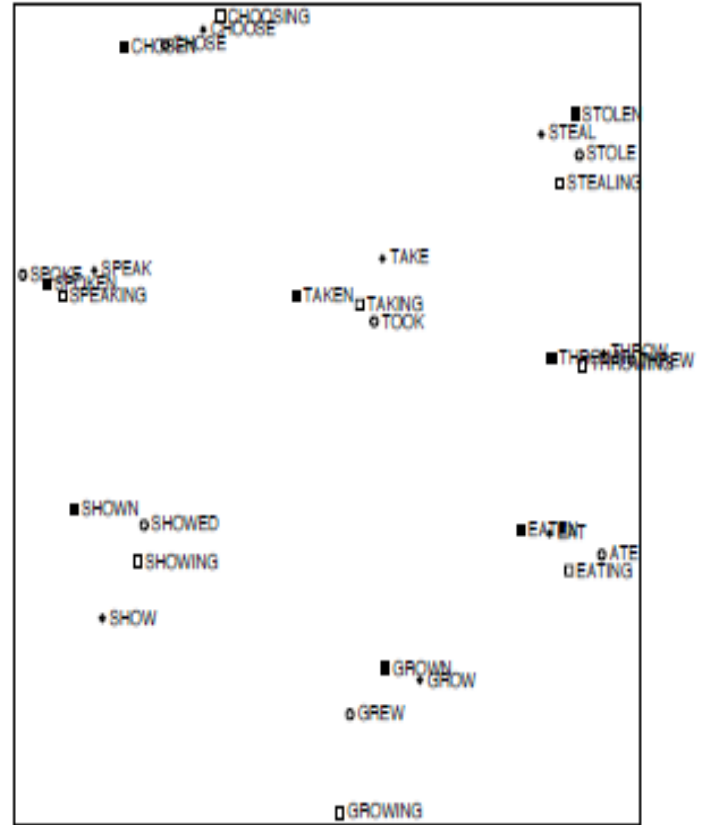
Similarity of  $d_2$  and  $d_3$  in the original space: 0.

Similarity of  $d_2$  and  $d_3$  in the reduced space:  $0.52 * 0.28 + 0.36 * 0.16 + 0.72 * 0.36 + 0.12 * 0.20 + -0.39 * -0.08 \approx 0.52$

# Latent Semantic Indexing: Ενδιαφέροντα Αποτελέσματα



An Improved Model of Semantic Similarity Based on Lexical Co-Occurrence  
Rohde et al. 2005



An Improved Model of Semantic Similarity Based on Lexical Co-Occurrence  
Rohde et al. 2005

# Σημασιολογικά

## ΣΥΝΤΑΚΤΙΚΑ

# Latent Semantic Indexing: Προβλήματα

- Πολύ μεγάλη υπολογιστική πολυπλοκότητα
- Δύσκολα ενσωματώνονται καινούριες λέξεις ή έγγραφα
- Λύση:
  - Αυτόματη μάθηση διανυσμάτων λέξεων χαμηλής διαστατικότητας
  - **Word2vec** (Mikolov et al. 2013)
    - Ρηχά μοντέλα που χρησιμοποιούνται για την παραγωγή word embeddings

# Word Representations μέσω των λέξεων που τις περιβάλλουν σε ένα παράθυρο συμφραζομένων: Window-based co-occurrence matrix

Example corpus:

- I like deep learning.
- I like NLP.
- I enjoy flying.

| counts   | I | like | enjoy | deep | learning | NLP | flying | . |
|----------|---|------|-------|------|----------|-----|--------|---|
| I        | 0 | 2    | 1     | 0    | 0        | 0   | 0      | 0 |
| like     | 2 | 0    | 0     | 1    | 0        | 1   | 0      | 0 |
| enjoy    | 1 | 0    | 0     | 0    | 0        | 0   | 1      | 0 |
| deep     | 0 | 1    | 0     | 0    | 1        | 0   | 0      | 0 |
| learning | 0 | 0    | 0     | 1    | 0        | 0   | 0      | 1 |
| NLP      | 0 | 1    | 0     | 0    | 0        | 0   | 0      | 1 |
| flying   | 0 | 0    | 1     | 0    | 0        | 0   | 0      | 1 |
| .        | 0 | 0    | 0     | 0    | 1        | 1   | 1      | 0 |

το «like» συνεμφανίζεται με το «deep» σε παράθυρο συμφραζομένων  $[-1, +1]$  μια φορά

# Word Representations μέσω των λέξεων που τις περιβάλλουν σε ένα παράθυρο συμφραζομένων: Window-based co-occurrence matrix

- Οι πίνακες αυτοί
  - Αυξάνουν πολύ σε διαστατικότητα καθώς αυξάνεται το λεξικό
  - Απαιτούν μεγάλους πόρους για αποθήκευση
  - Έχουν προβλήματα πολύ αραιών εγγραφών (sparsity)
- Λύση
  - Κρατάμε μόνο την σημαντική πληροφορία μέσω ενός μικρού αριθμού διαστάσεων (25-1000 διαστάσεις)
    - **Word embeddings**
  - Οπότε κάθε λέξη αναπαρίσταται σαν ένα πυκνό διάνυσμα
  - Τα διανύσματα δυο λέξεων που εμφανίζονται σε όμοια περιβάλλοντα είναι όμοια (κοντά το ένα στο άλλο στον διανυσματικό χώρο)

*banking* =

$$\begin{pmatrix} 0.286 \\ 0.792 \\ -0.177 \\ -0.107 \\ 0.109 \\ -0.542 \\ 0.349 \\ 0.271 \end{pmatrix}$$