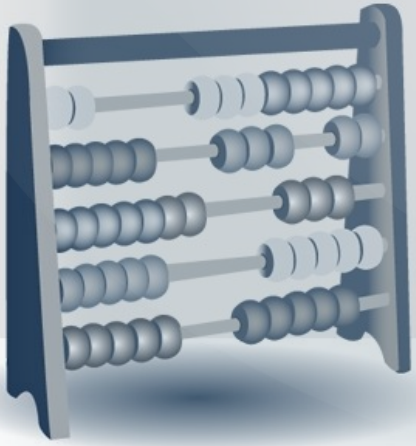
Iperspazi del mediterraneo

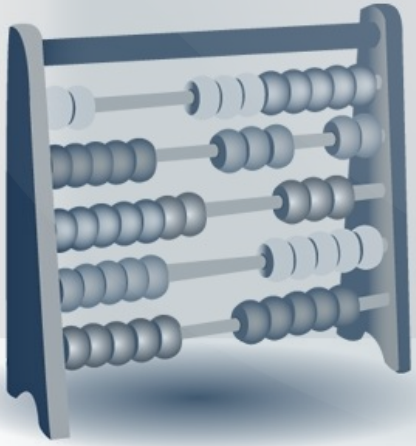
Nuvole di parole greche, latine e arabe

Federico Boschetti – Ouafae Nahli – Angelo Mario Del Grosso

Istituto di Linguistica Computazionale del CNR di Pisa

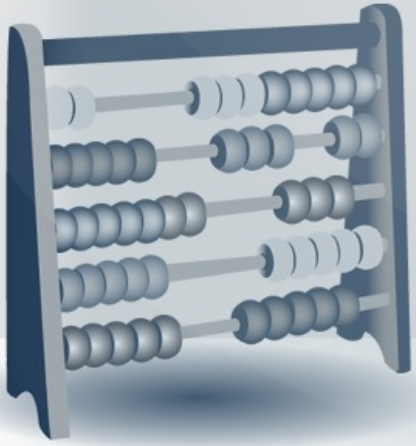


- Che cos'è il significato di una parola?
- C'è un senso delle parole nascosto dentro alle parole stesse?
  - C'è qualcosa di meglio delle traduzioni per capire il senso delle parole di una cultura diversa dalla nostra?
- Come si fa a capire il senso di una lingua che non è più parlata da nessuno?



Facciamo un esperimento ...

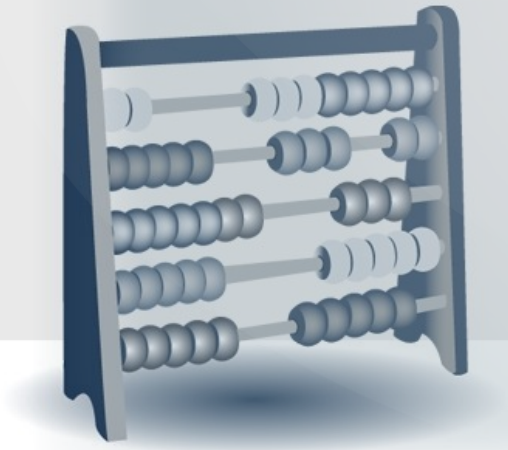
- Dividetevi in due gruppi
- Senza guardare cosa fa l'altro gruppo disponete sulla vostra lavagna le parole latine che secondo voi sono in relazione con la parola AMOR
- Fate in modo che le parole simili fra di loro siano vicini e quelle dissimili siano lontane
- Confrontate i risultati



Problema...

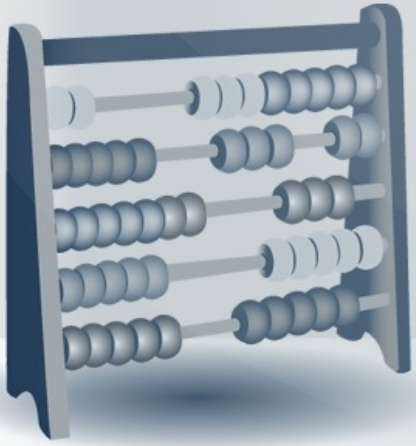
- Come facciamo a valutare i risultati dei due gruppi?

- E' possibile evitare la soggettività quando si ragiona sul significato delle parole?



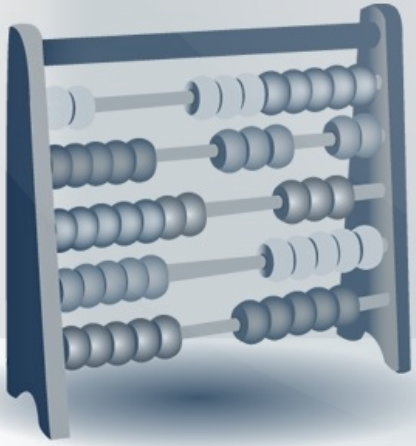
Una risposta...

- Si può ridurre la soggettività se ci basiamo su metodi quantitativi per misurare ciò che i testi stessi contengono
- I dati prodotti tramite l'applicazione di un metodo rigoroso sono ripetibili, falsificabili e quindi analizzabili da tutti
- L'interpretazione dei risultati non potrà mai escludere la soggettività di chi li interpreta



Metodi tradizionali “Spiegare Omero con Omero”

- La filologia fin dalla sua nascita cerca di spiegare il senso delle parole di un corpus mettendole in relazione con i contesti dove si trovano



Corpus (pl. Corpora): collezione organica di testi

musisque deoque

un archivio digitale di poesia latina

home

ricerca
- semplice
- avanzata

indice
- alfabetico
- cronologico

metrica
- metri
- opere

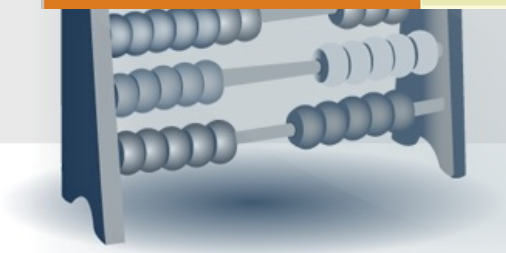
epigraphica

collaboratori

1220 luoghi trovati per la chiave: **amor** NUOVA RICERCA SALVA I RISULTATI

1 2 3 4 5 6 7 8 9 10 11 12 13

PLAVT. Amph. 893	Atque illi dudum meus amor negotium
PLAVT. Aul. 750	Nos fecisse amoris caussa. Nimi' uilest uinum atque amor,
PLAVT. Bacch. 21	Cupidon tecum saeuit anne Amor?
PLAVT. Bacch. 115	Amor, Voluptas, Venu', Venustas, Gaudium,
PLAVT. Cas. 221	Nam ubi amor condimentum inierit, quoiuis placitura <escam> credo;
PLAVT. Cas. 222	Neque salsum neque suaue esse potest quicquam, ubi amor non admiscetur:
PLAVT. Cas. 802	At ego amo. At ego hercle nihili facio. Tibi amor pro cibost,
PLAVT. Cist. 69	Namque ecastor Amor et melle et felle est fecundissumus;
PLAVT. Cist. 72	Perfidiosus est Amor. Ergo in me peculatum facit.



Greek | Arabic | Composite Search

Word Index

(81 of 83) 5

- χρή - 1
- χρόνος - 1
- χωρέω - 1
- χωριστός - 1
- ψυχή - 50

(81 of 83) 5

SEARCH GREEK

Word A	Word B	Word C
lemma	form	form
ψυχή	Search for...	Search for...
Every PoS	Every PoS	Every PoS
Operator:	OR	
Search	Save Parameters	Clear Parameters

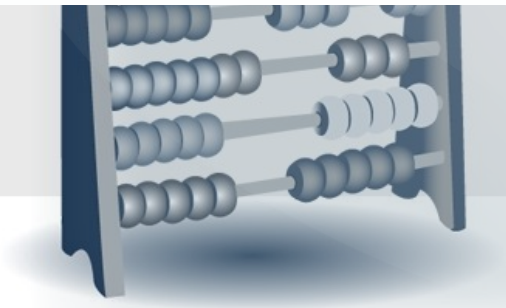
IV 7, 8(3).22 πρότερον ἄρα καὶ νοῦς καὶ ψυχὴ φύσεως.

III, p. 51.5 فالعقل والنفس قبل الطبيعة،

IV 7, 8(3).22-23 Οὐκ ἄρα οὕτως¶ ψυχὴ ὡς πνεῦμα οὐδ' ὡς σῶμα.

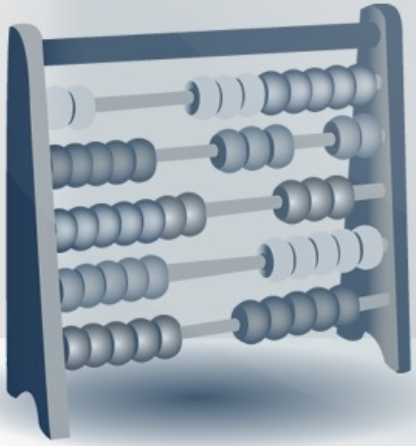
III, p. 52.8 فليست النفس إذاً غريزيٌ ولا بجرم البتة بروح.

Εἰ δὴ ταῦτα ὀρθῶς λέγεται, λύοιντο ἂν ἤδη αἱ ἀπορίαι



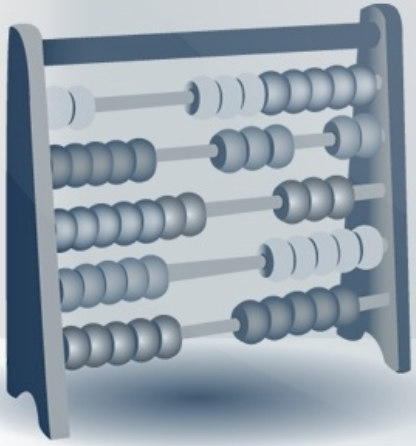
L’ipotesi Distribuzionale

- Due parole sono tanto più semanticamente simili quanto più tendono a ricorrere in contesti linguistici simili [Lenci 2008]
  - Se osserviamo due parole che occorrono sempre nello stesso contesto è ragionevole assumere che indicano “cose simili” [Sahlgren 2001]



Ipotesi Distribuzionale

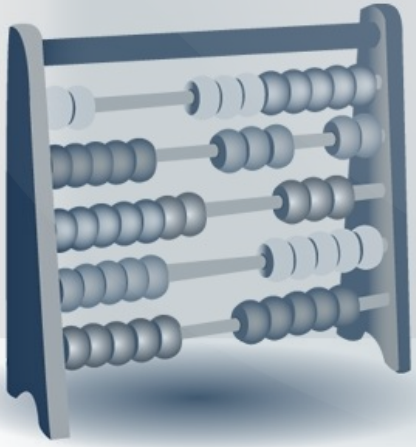
- Parole legate tra loro non devono necessariamente co-occorrere
- E’ sufficiente che tali parole co-occorrano in contesti simili, cioè contenenti un certo numero di parole identiche



co-occorrenza: numero di volte in cui due o più parole sono contemporaneamente presenti all'interno degli stessi contesti

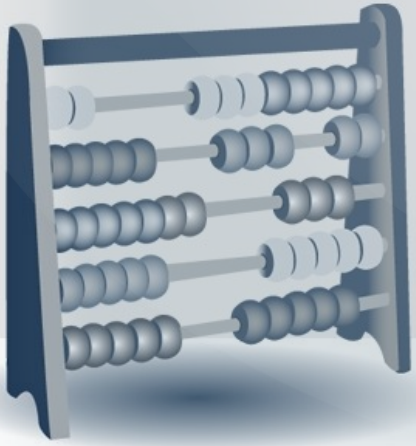
A cosa si applica l'ipotesi distribuzionale

- Selezione di sinonimi
- Ragionamento analogico
- Giudizi di similarità semantica
- Acquisizione lessicale
- ...



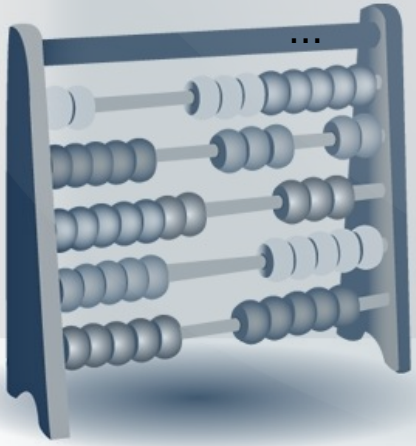
Come si costruisce la tabella delle co-occorrenze

- Le RIGHE rappresentano ciascun termine del corpus
- Le COLONNE rappresentano i termini che si trovano all'interno di una finestra di *n* parole di contesto a destra e a sinistra del termine oggetto di studio



Esempio...

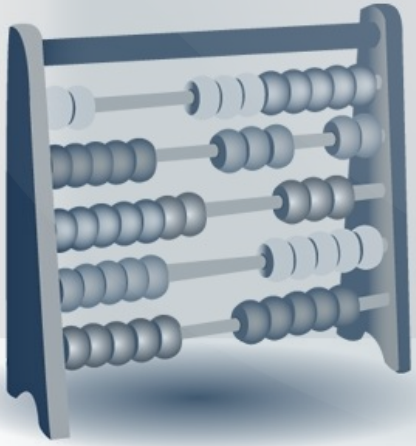
	.	amor	amicus	...	gallus	gaudium	...
...	.	...	...	...	...	...	...
amor	.	3	25	...	0	18	...
...	.	...	...	...	...	...	...
gaudium	.	21	18		2	0	...
...	.	...	...	...	...	...	...
venus	.	18	9	...	0	19	...
venustas	.	16	5	...	0	19	...
voluptas	.	14	4	...	2	21	...
...	.	...	...	...	...	...	...



Il Modello dello spazio delle parole

Analogia con lo spazio geometrico

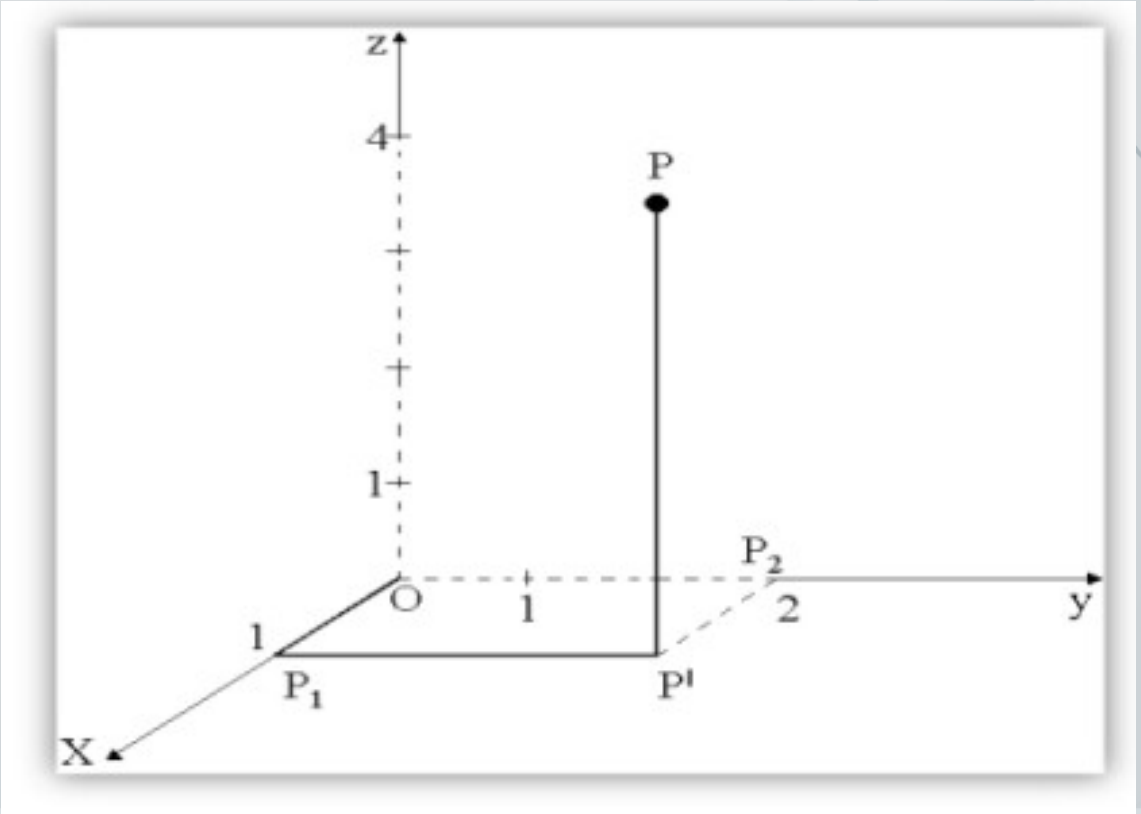
Ciascun punto dello spazio geometrico è definito da un insieme di numeri che rappresentano le sue coordinate rispetto agli assi cartesiani...

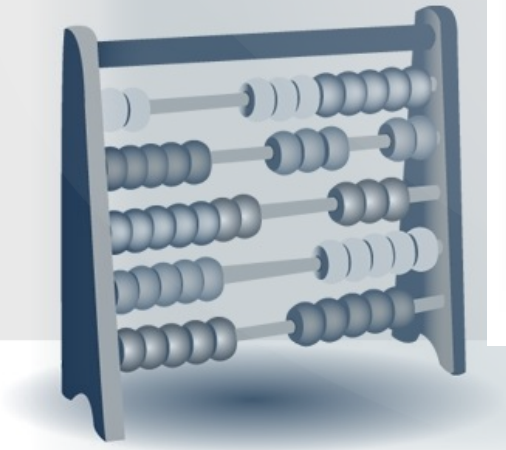


Dimensione dello spazio:
N numeri di coordinate rispetto a N assi di riferimento

Il Modello dello spazio delle parole

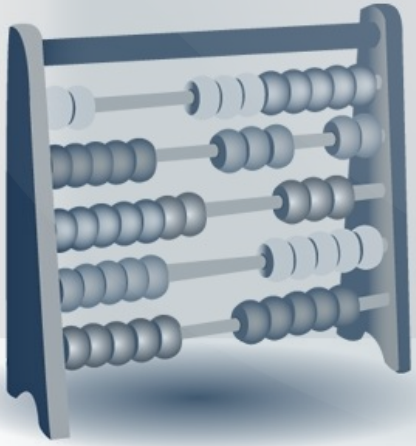
Analogia con lo spazio geometrico





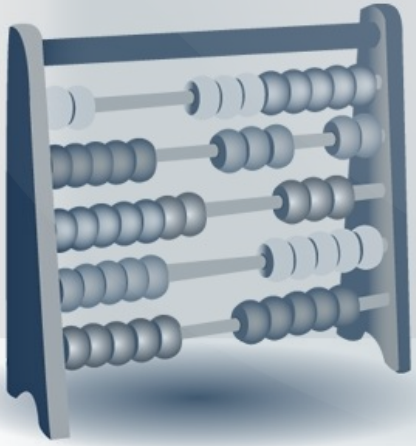
Il Modello dello spazio delle parole

Per analogia il significato di una parola è rappresentabile attraverso la sua posizione in uno spazio le cui coordinate sono determinate dai contesti linguistici in cui la parola può ricorrere

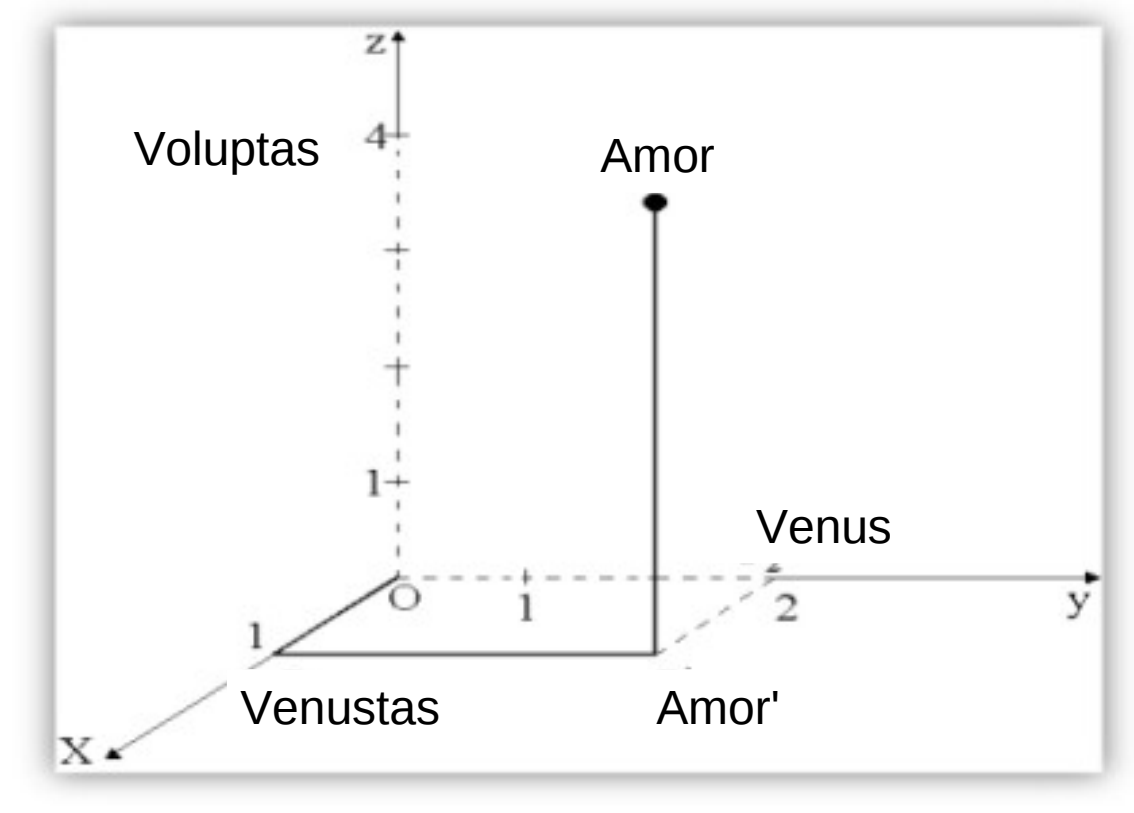


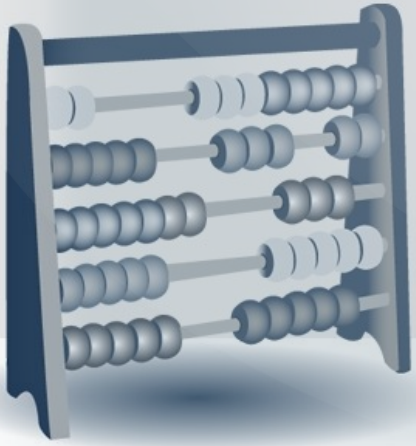
Misura della distanza semantica

Per determinare la posizione di due punti nello spazio delle parole e misurarne la distanza è necessario comparare i loro vettori (*context vectors*) rispetto a tutte le dimensioni che li costituiscono



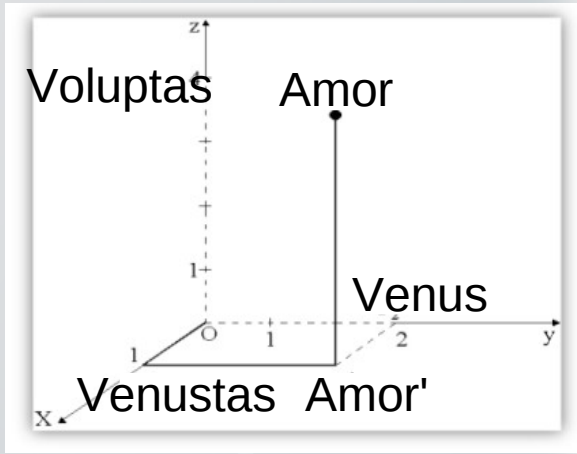
Componenti di Amor nello spazio delle parole

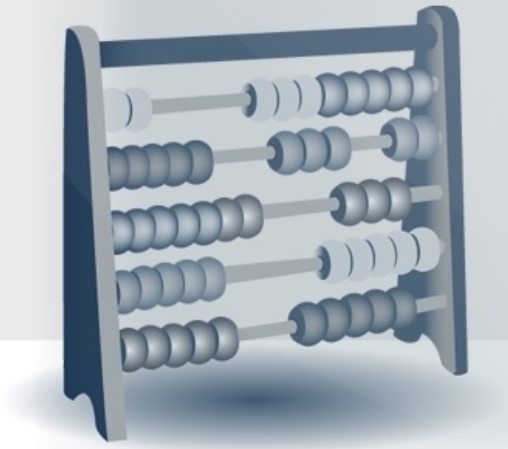




Componenti di Amor nello spazio delle parole

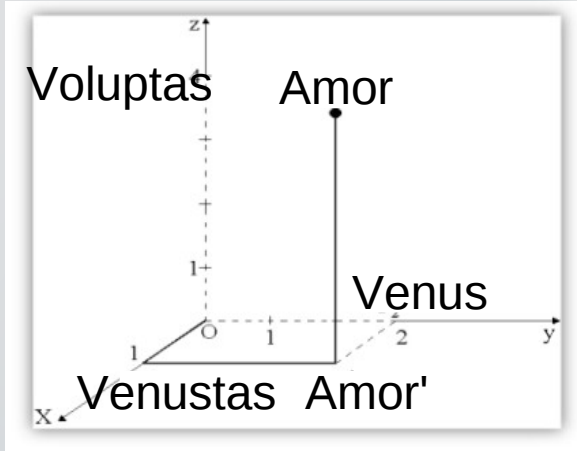
Ricordiamo che ciascuna colonna della matrice di co-occorrenze ci dice quante volte nella collezione dei testi la parola considerata occorre vicino alla parola rappresentata da quella colonna

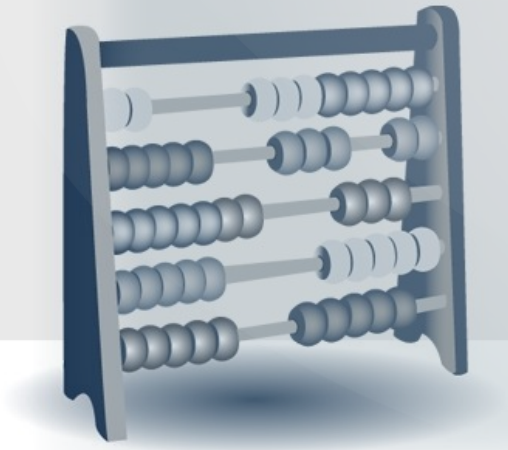




Componenti di Amor nello spazio delle parole

I numeri che indicano le occorrenze possono essere considerati le coordinate del punto in uno spazio ad n dimensioni

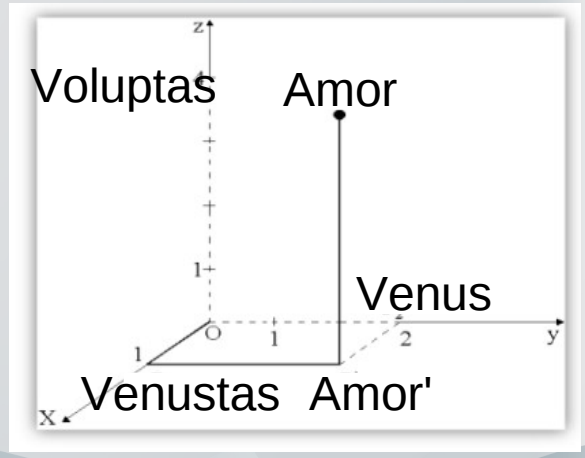


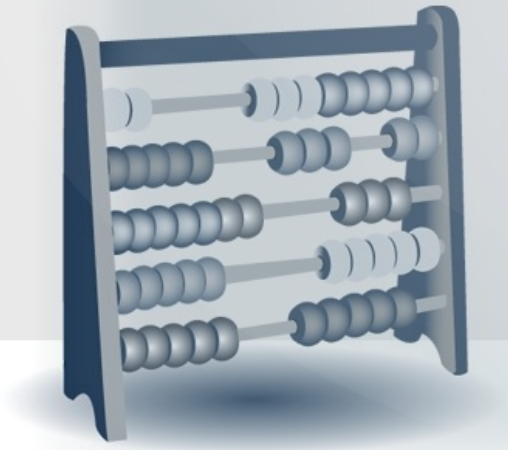


Componenti di Amor nello spazio delle parole

Ad esempio, dato un termine da studiare, se avessimo solo tre parole di contesto, potremmo rappresentare il nostro termine come un punto in uno spazio tridimensionale

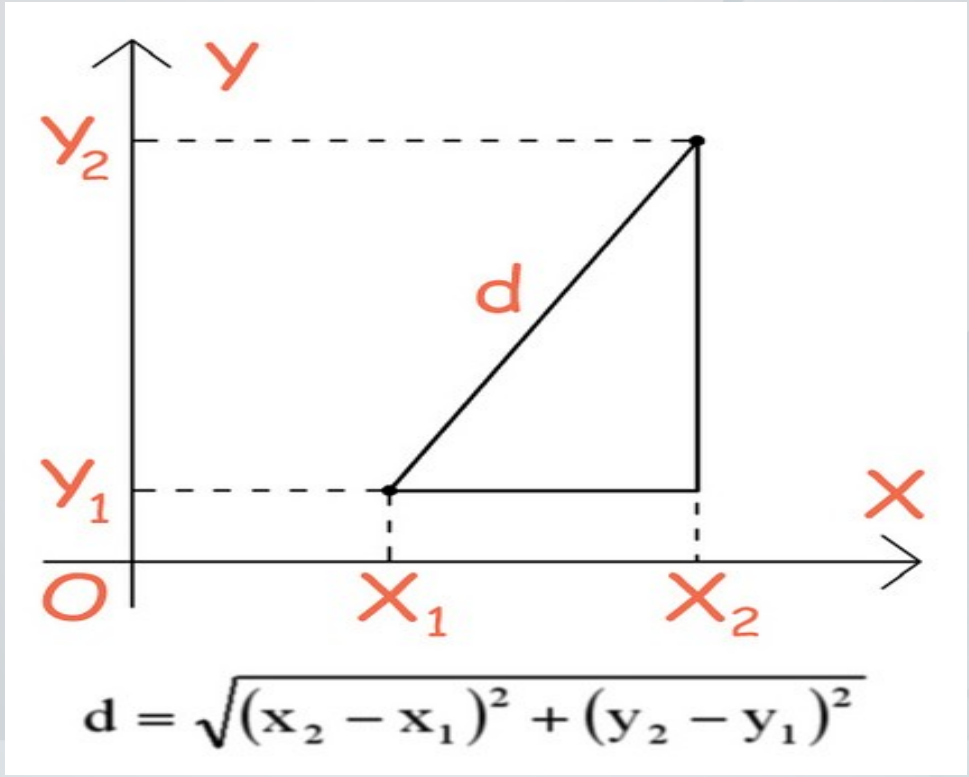
Le sue coordinate sarebbero determinate dal numero di co-occorrenze effettive con ciascuno dei tre termini che si trovano vicine nei vari contesti

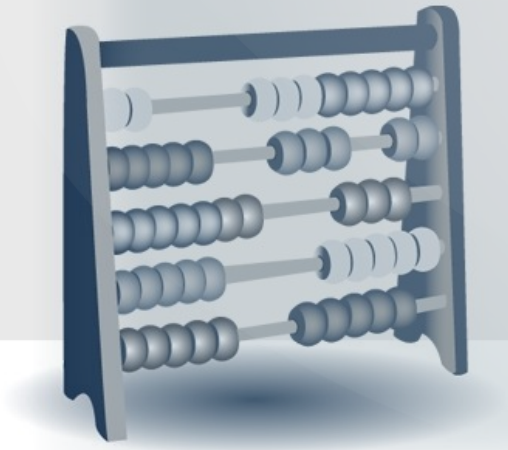




Misura della distanza di due punti nello spazio

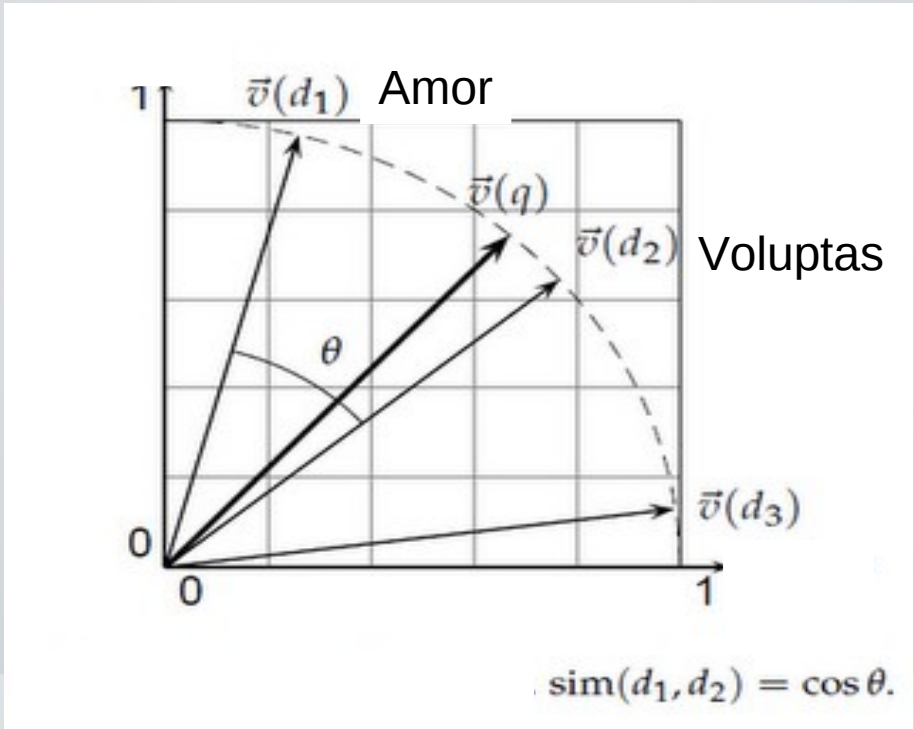
La distanza euclidea di due punti nello spazio si misura applicando il teorema di Pitagora

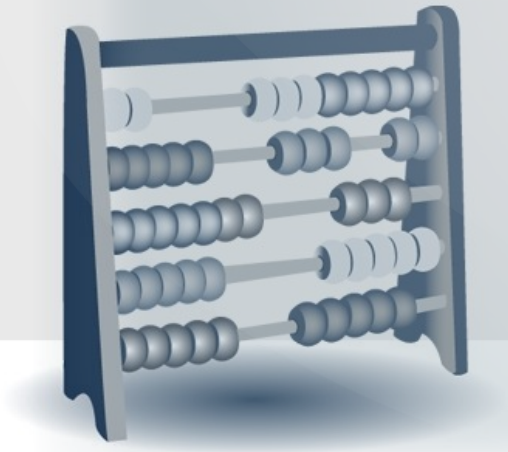




Misura della distanza di due punti nello spazio

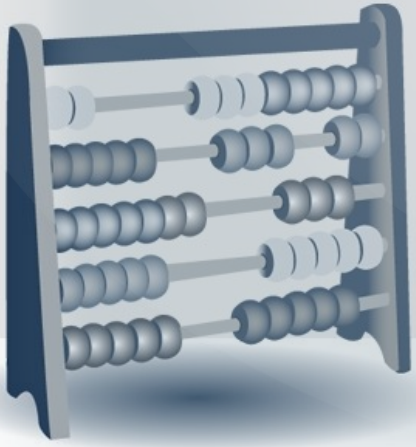
La similitudine fra due parole può essere calcolata applicando il teorema del coseno: l'indice di somiglianza di una parola con un'altra è dato dal coseno dell'angolo compreso fra i due vettori che rappresentano le parole nello spazio semantico





Efficacia del modello spaziale

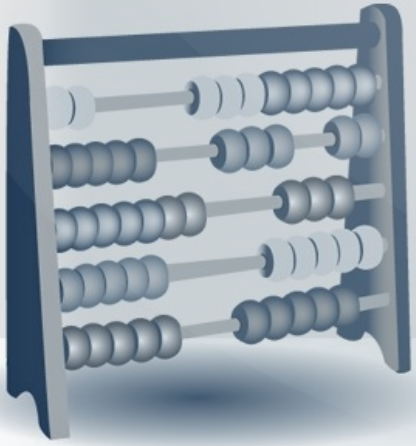
- Consente di trattare in modo computazionale aspetti semantici e definire in termini matematici concetti di similarità tra parole
- Costituisce un approccio descrittivo al modello semantico operando solo su dati concreti e disponibili



Problemi del modello spaziale

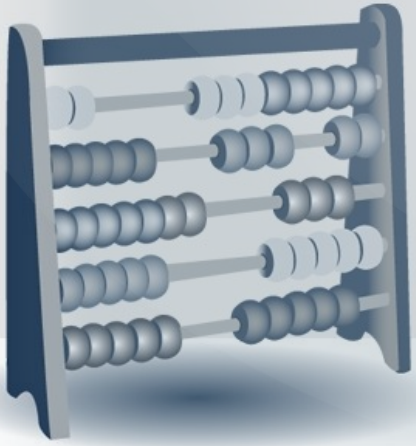
- Le applicazioni reali sono costituite da vastissime quantità di dati e quindi la dimensione dei vettori di contesto è abitualmente molto elevata

- La tabella delle co-occorrenze è molto sparsa (cioè è piena di zeri) in quanto gli incontri pertinenti sono molto meno numerosi degli incontri di parole non pertinenti. Ad esempio Amor non co-occorre mai con *scandula* (assicella), con *tegula* (tegola) e una grandissima quantita di altre parole con cui l'amore non ha nulla a che fare



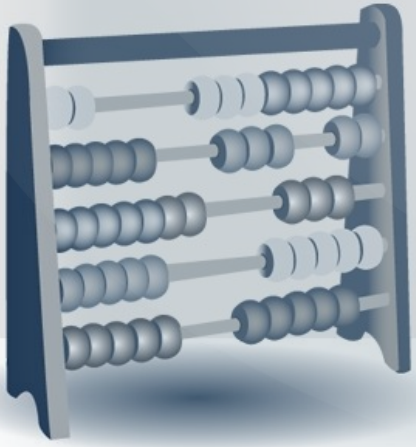
Sparsità della tabella delle co-occorrenze

- La maggior parte delle parole occorrono solo in gruppi di contesti molto limitati
  - In una matrice di co-occorrenza tipicamente più del 99% delle entrate sono pari a zero



Alla ricerca del significato nascosto

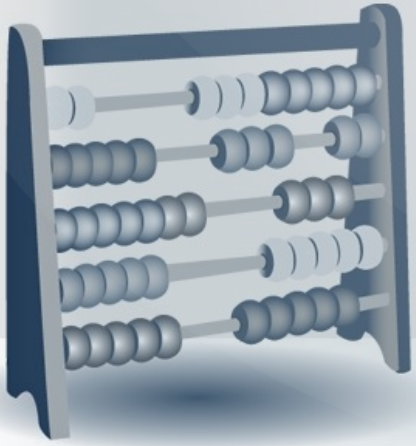
- Con tecniche statistiche è possibile ridurre il numero di dimensioni della tabella delle co-occorrenze eliminando le informazioni poco importanti (“rumore”) agendo come una specie di filtro sulle informazioni importanti. In questo modo le relazioni semantiche latenti vengono alla luce. Per questa ragione tale tipo di analisi è detta **Latent Semantic Analysis (LSA)**



LSA / SVD

- La Singular Value Decomposition (SVD) è una tecnica statistica di riduzione della dimensione dei vettori contestuali

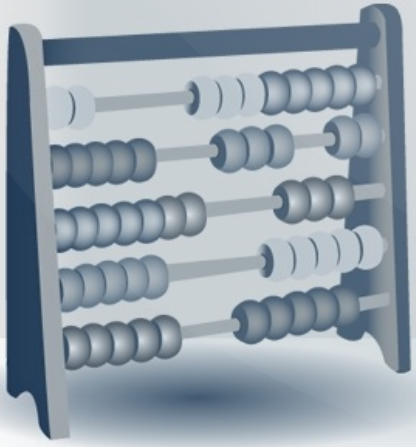
- La matrice di co-occorrenza viene trasformata, così, in una matrice con meno colonne, ma siccome la matrice originale è molto sparsa, cioè contiene molti zeri, la nuova matrice è molto più densa e quindi, pur essendo più piccola, preserva gran parte dell'informazione originale



Random Indexing

- Esiste un sistema molto efficiente per evitare la costruzione della matrice di co-occorrenza e poi sprecare risorse computazionali per ridurla

- Il Random Indexing è un modello di spazio di parole incrementale basato sull'idea di accumulare i vettori contestuali in base alla presenza delle parole nei contesti ove occorrono



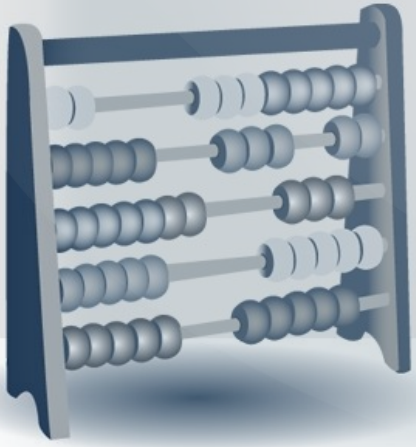
- Questa tecnica non prevede una fase di riduzione della dimensione della matrice successiva alla lettura dei dati

Semantic Vectors

- Esistono strumenti informatici che permettono di generare spazi di parole a partire da una collezione di testi qualsiasi

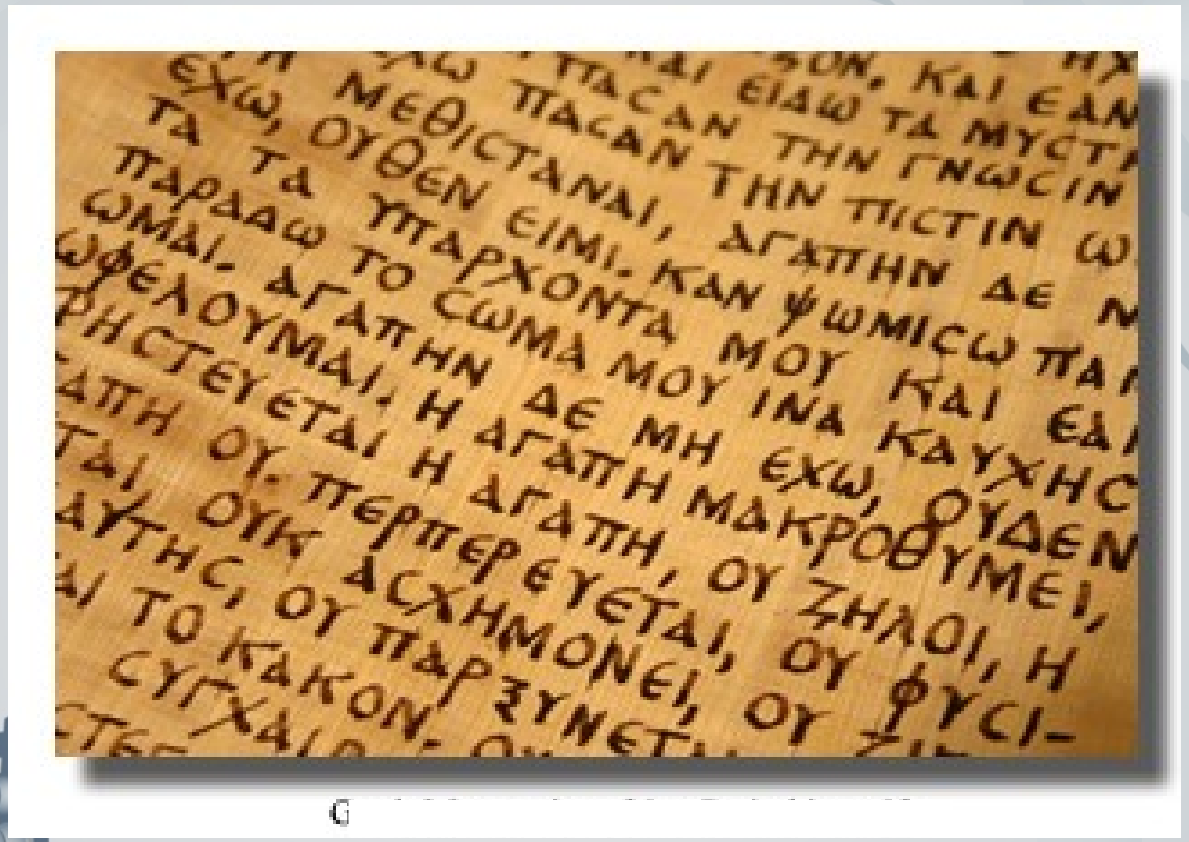
http://code.google.com/p/semanticvectors/

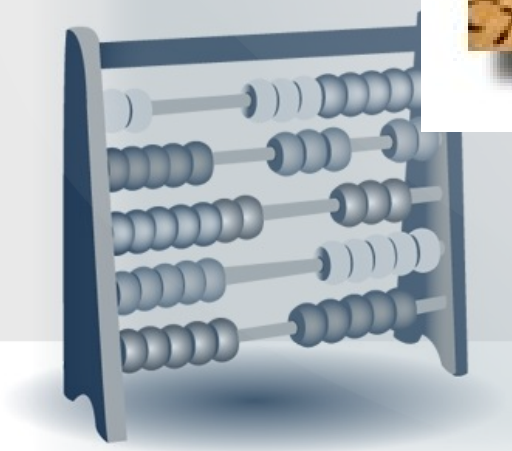
Anche voi potete fare degli esperimenti con queste applicazioni.
Per esempio, potreste usare i vostri dati per la tesina di maturità!
Per discutere i vostri risultati potete scrivere a noi!!



federico.boschetti@ilc.cnr.it
ouafae.nahli@ilc.cnr.it
angelo.delgrosso@ilc.cnr.it

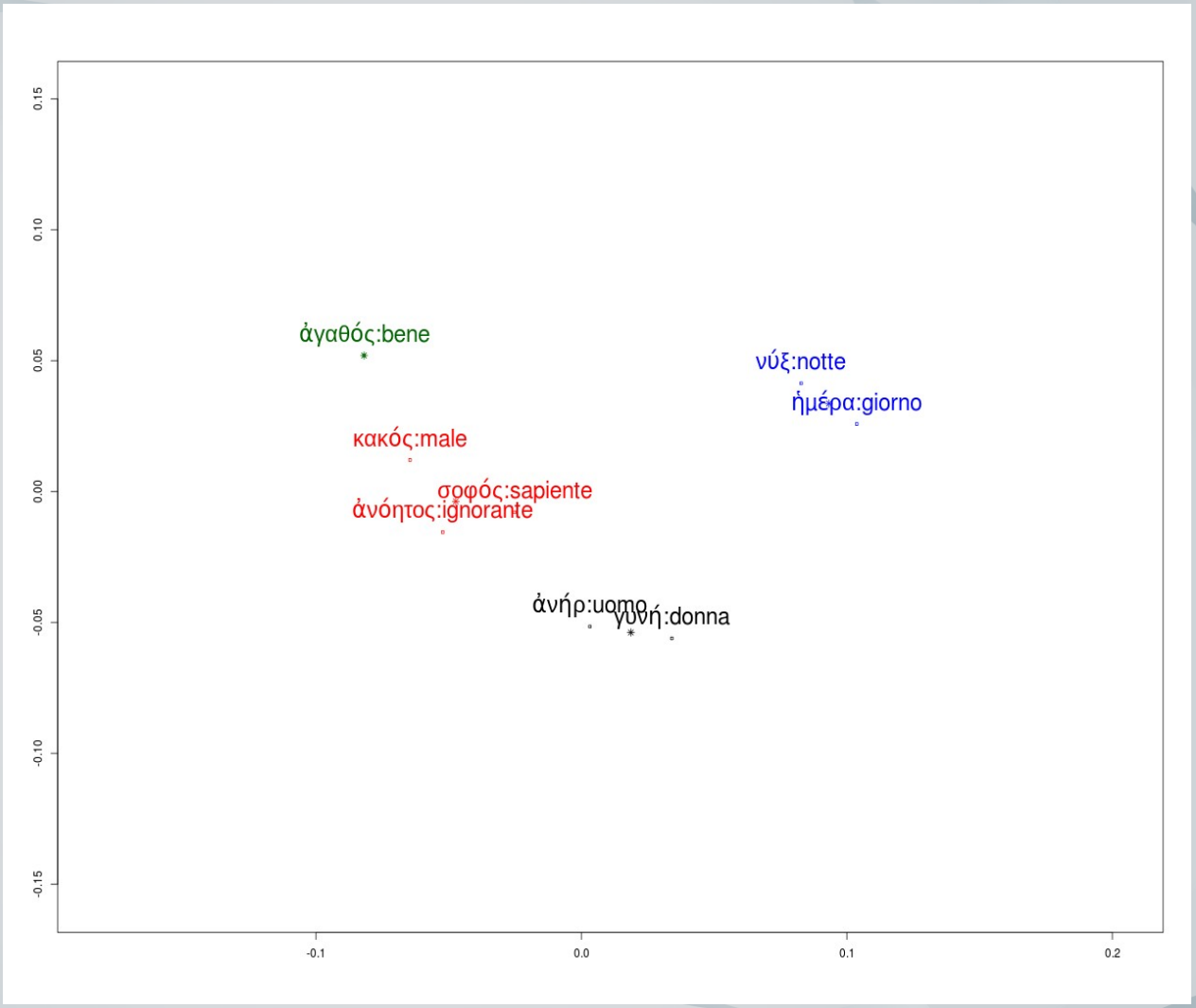
Spazi di parole greche

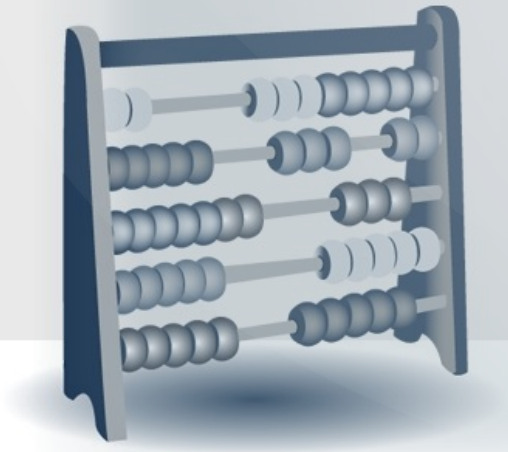




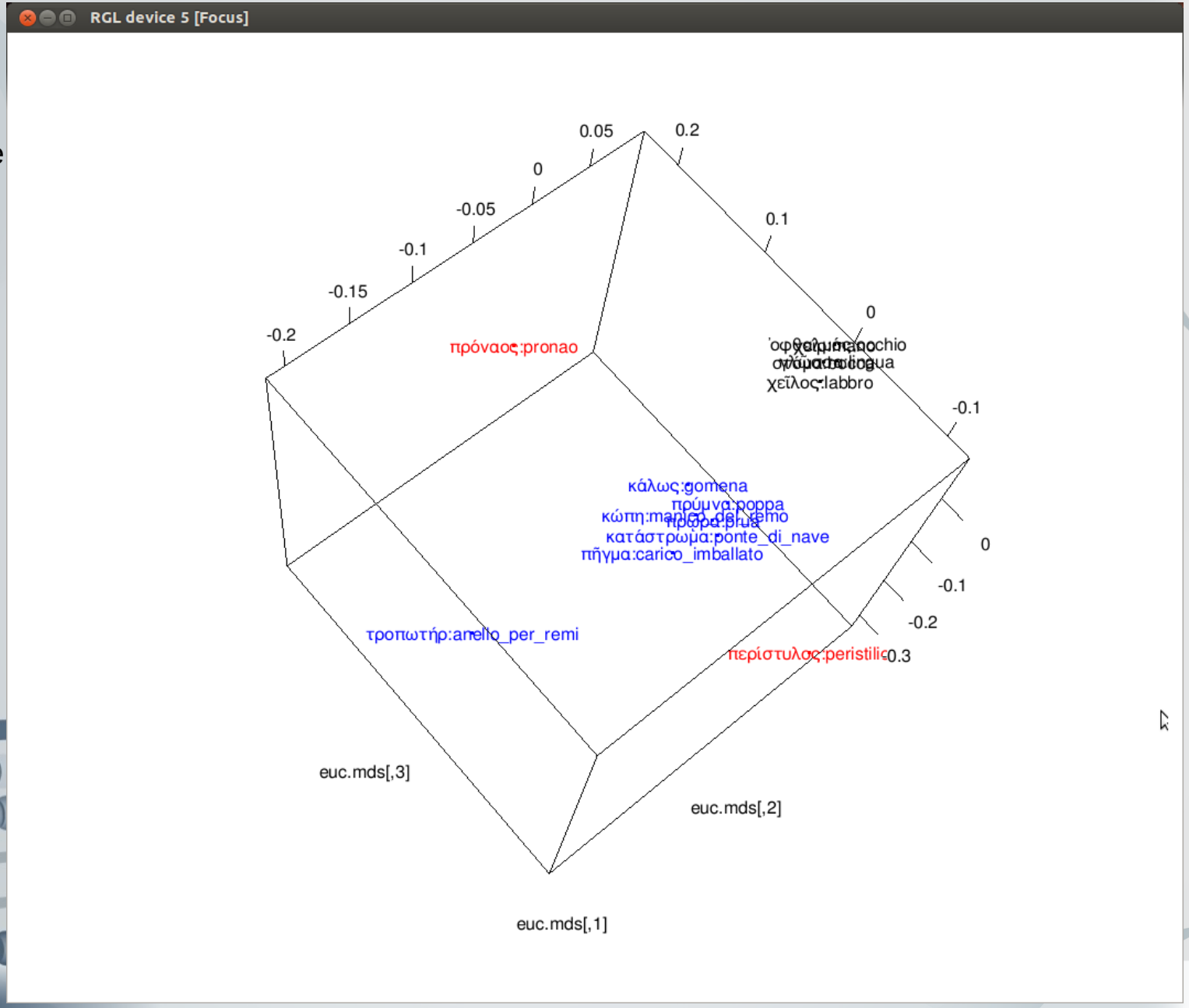
Corpus: collezione di testi digitalizzati
della letteratura greca dall'VIII sec. a.C. al XV sec. d.C.

I contrari tendono a raggrupparsi in modo molto evidente, anche se a volte uno dei due contrari può essere attratto da un termine con cui co-occorre molto spesso, come “male” e “ignorante”

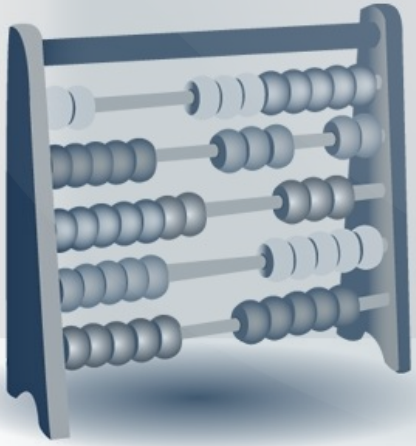


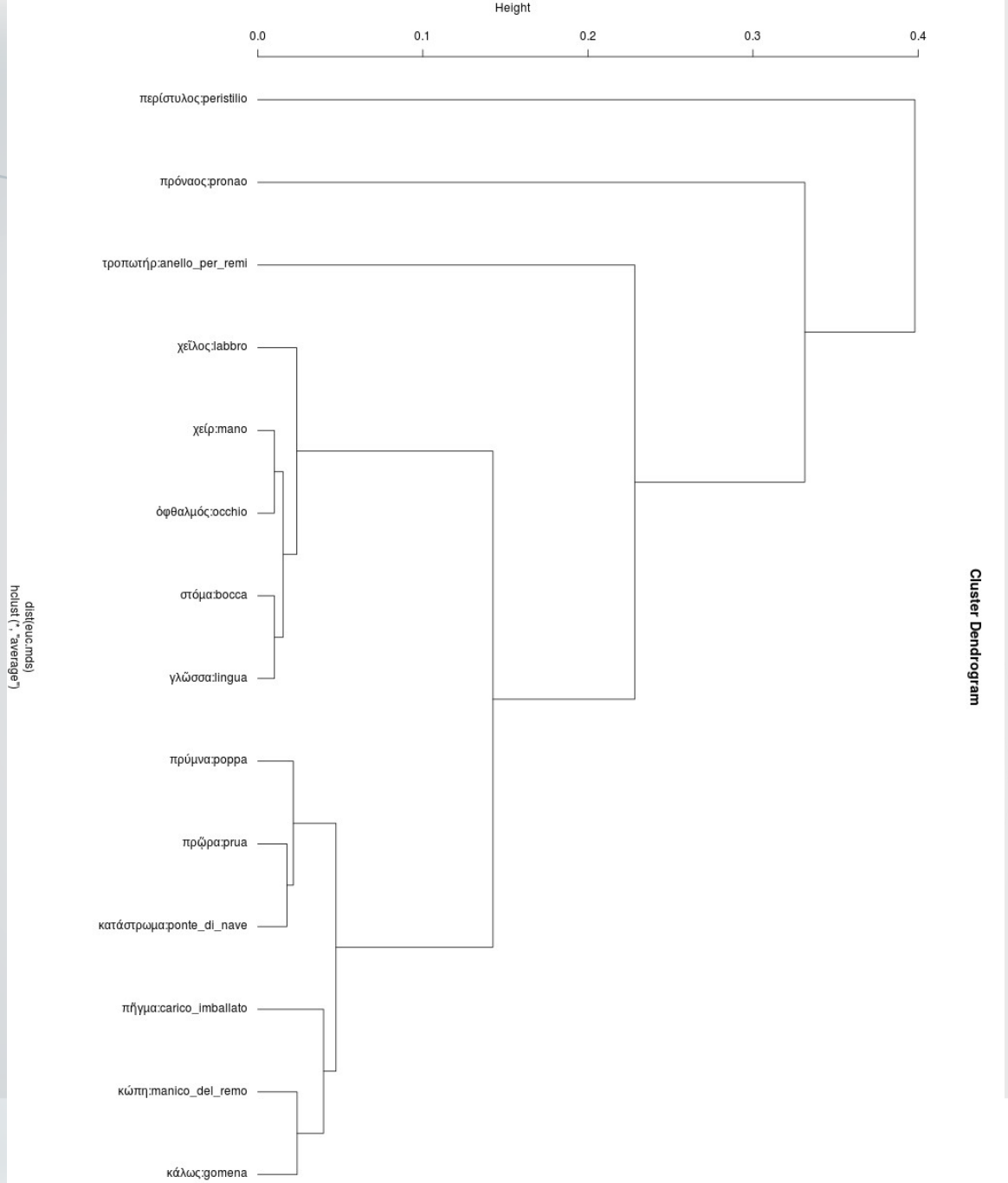


Gli spazi di parole sono in grado di raggruppare in modo molto evidente la relazione della parte con il tutto e delle parti fra di loro.
Un raggruppamento creato in modo automatico e distinto con un colore si chiama *cluster*

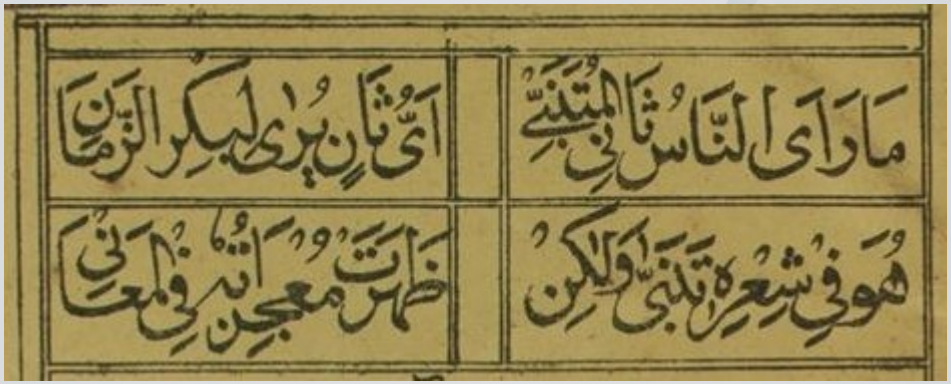


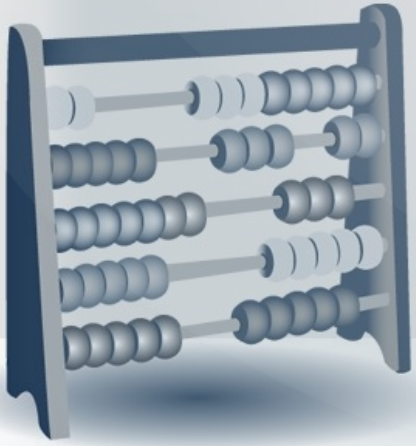
Questo tipo di grafico, detto *dendrogramma perché ha la forma di un albero,* rappresenta *gli stessi dati del grafico precedente ma evidenzia le distanze reciproche fra i termini e le eventuali relazioni gerarchiche fra di essi*



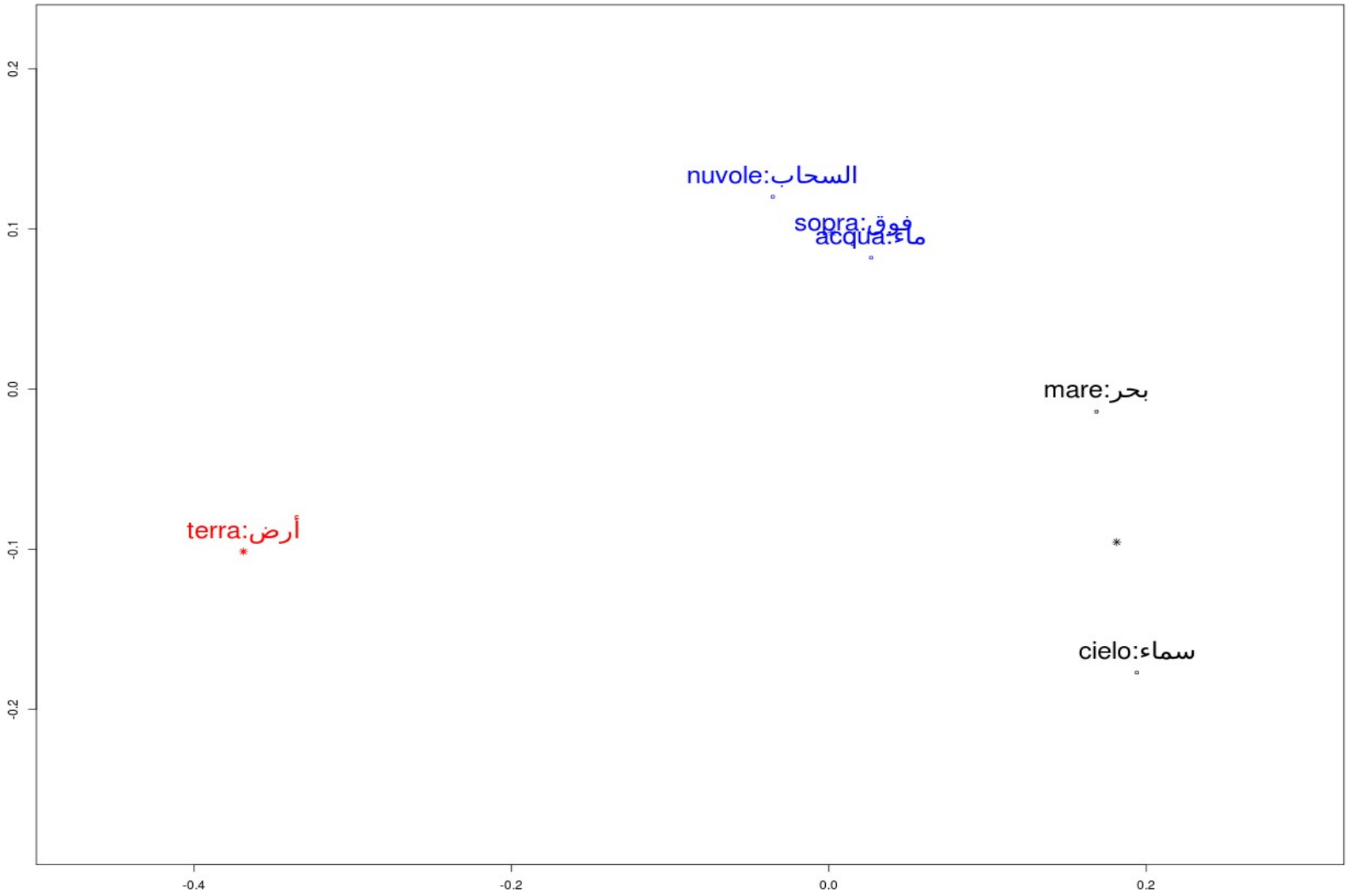


Spazi di parole arabe

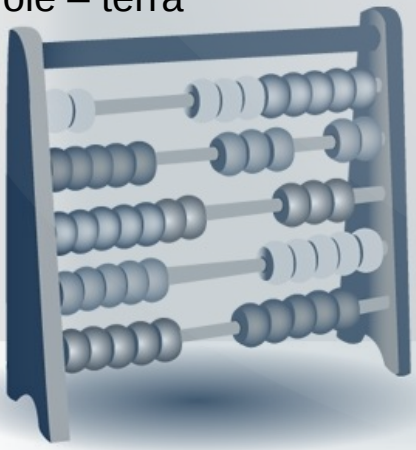


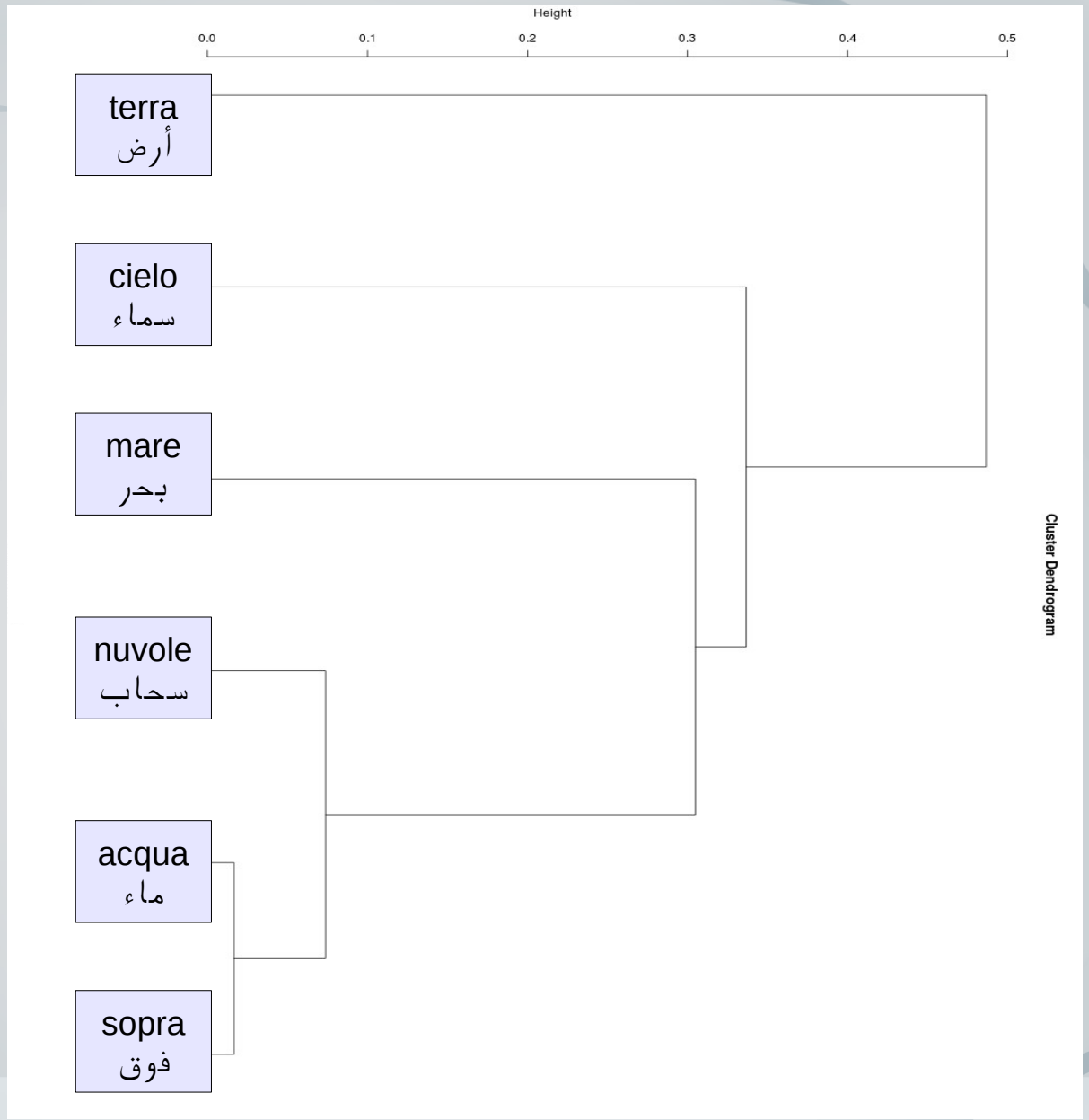


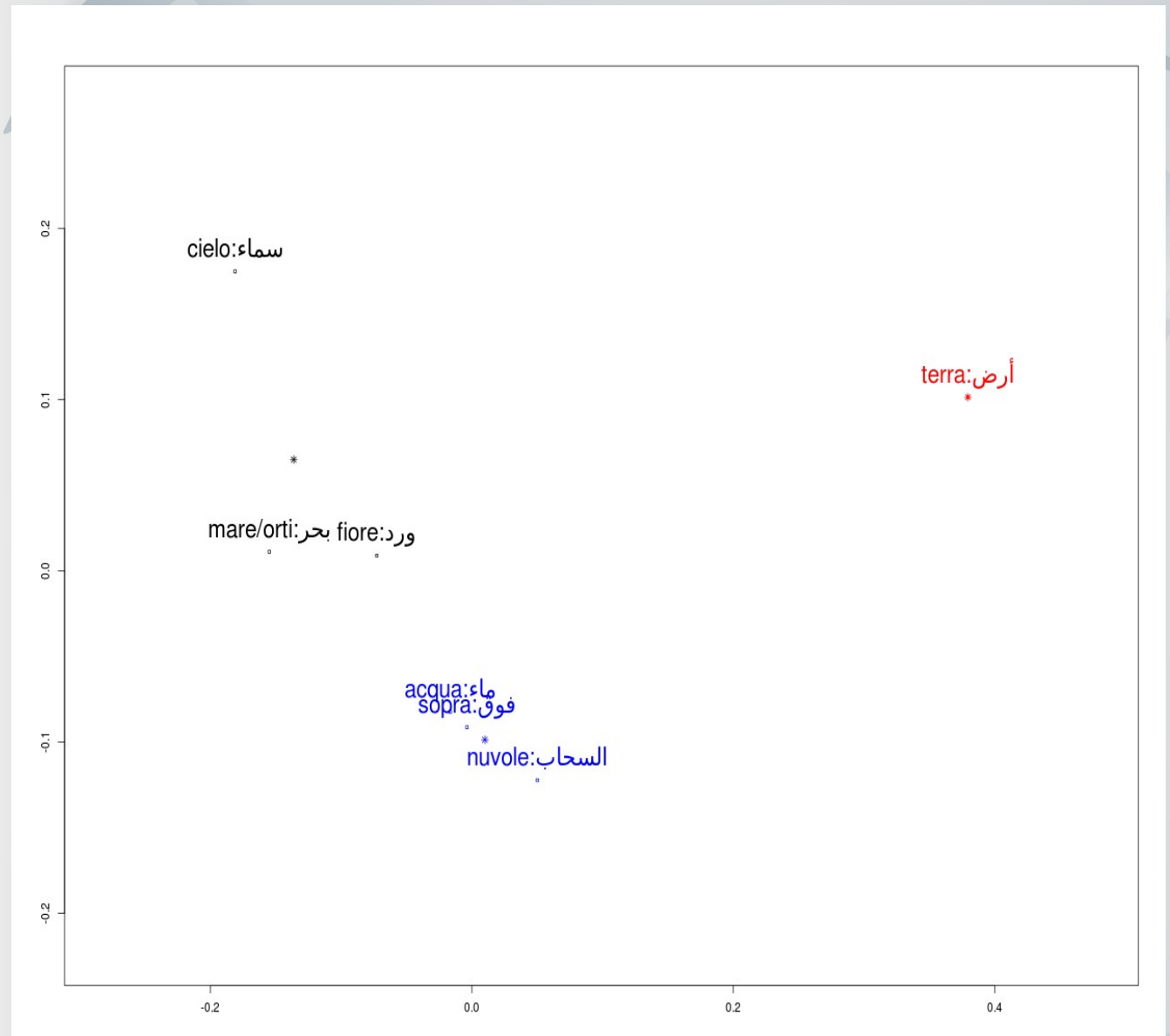
Corpus: *Lisan al-Arab*, enciclopedia in lingua araba del XIV sec.


nuvole:السحاب
sopra:فوق
acqua:ماء
mare:بحر
terra:أرض
cielo:سماء
0.2
0.1
0.0
-0.1
-0.2
-0.4
-0.2
0.0
0.2

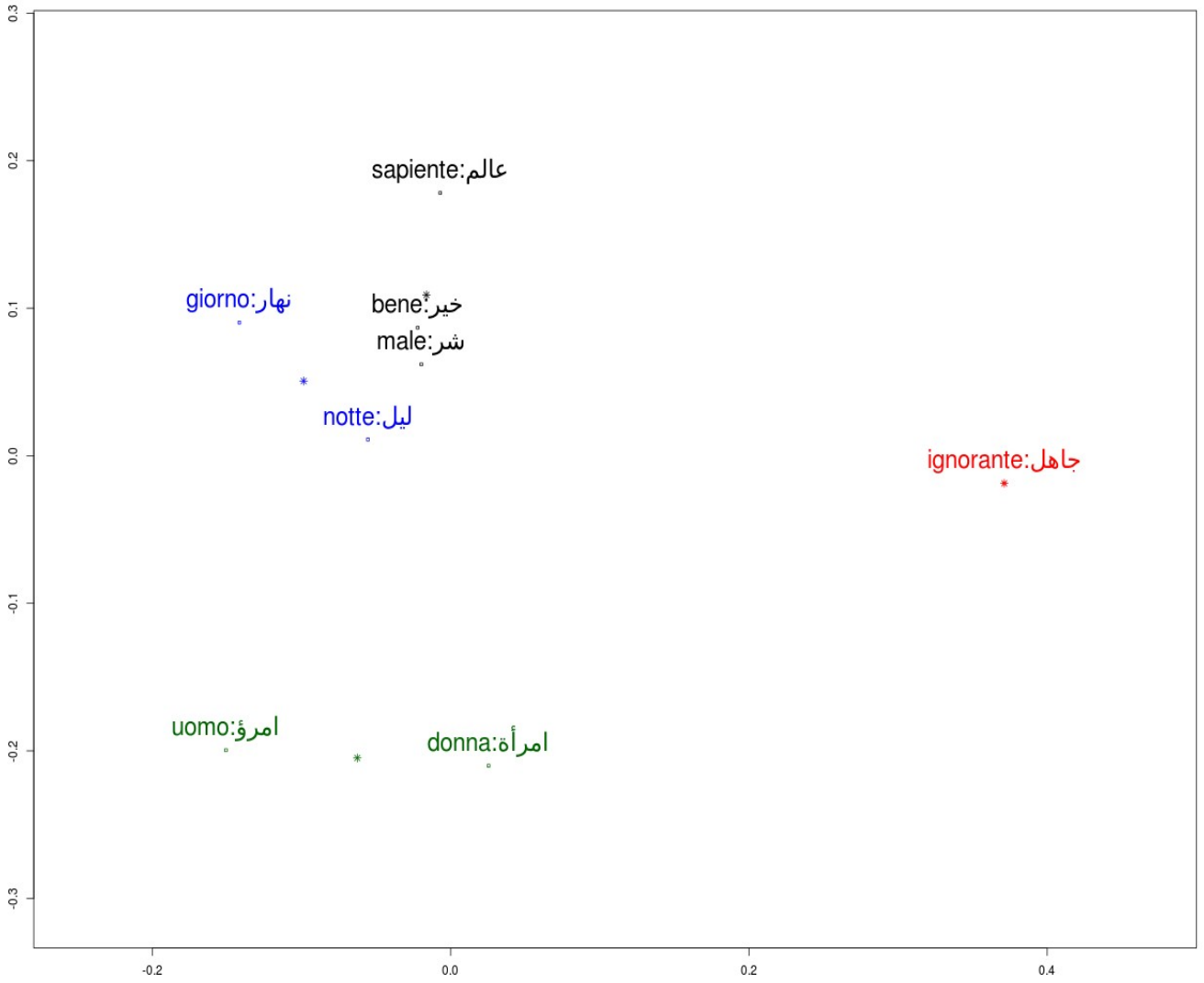
Sia lo scatterplot della diapositiva precedente sia il dendrogramma di questa diapositiva mostrano la (non sorprendente!) relazione fra “nuvole” ed “acqua”. L'associazione non è dovuta solo al legame *in absentia* di nuvole – [pioggia] e [pioggia] – acqua, ma anche dall'associazione *in praesentia* di nuvole – mare, che è molto più forte dell'associazione nuvole – terra





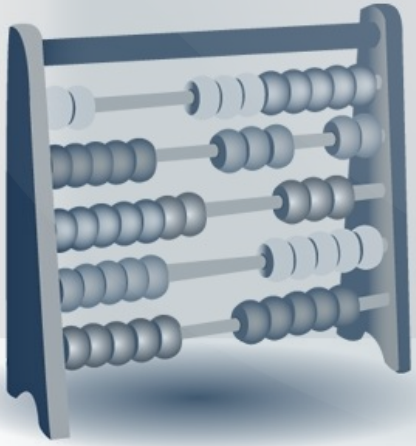


Ci si aspetterebbe che la parola “fiore” venga attratta dalla parola “terra” e invece la troviamo nel cluster di “mare” e “cielo”. Questo si spiega bene se teniamo conto che la parola araba *bHr*, nel nostro corpus, non è vocalizzata e quindi può indicare sia *baHr* , “mare”, sia *biHar,* “orti” oppure “prati”.

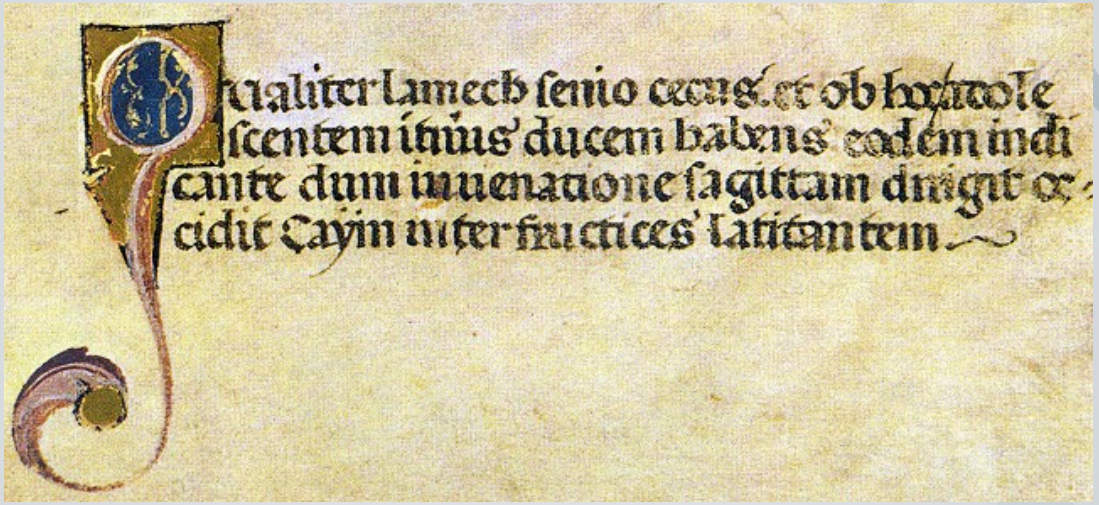


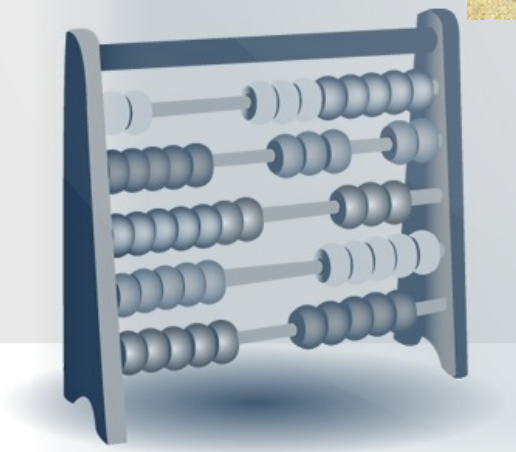
La parola “sapiente” è attratta dalle parole “bene” e “male” perché il sapiente è a conoscenza dei due concetti e sa fare la differenza tra i due. Si noti che in arabo la parole che indica “ignorante” si usa anche per esprimere il concetto di “insipiente”, “stolto”, cioè una persona che non è a conoscenza di niente soprattutto per sua scelta

Le parole “pace”, “città” ed “Allah” sono strettamente legate fra di loro in uno stesso cluster.
Le parole “uomo” e “donna” formano un secondo raggruppamento.
Infine, le relazioni parentali “padre”, “coppia di fratelli” e “gemello” formano un terzo raggruppamento a sé stante.



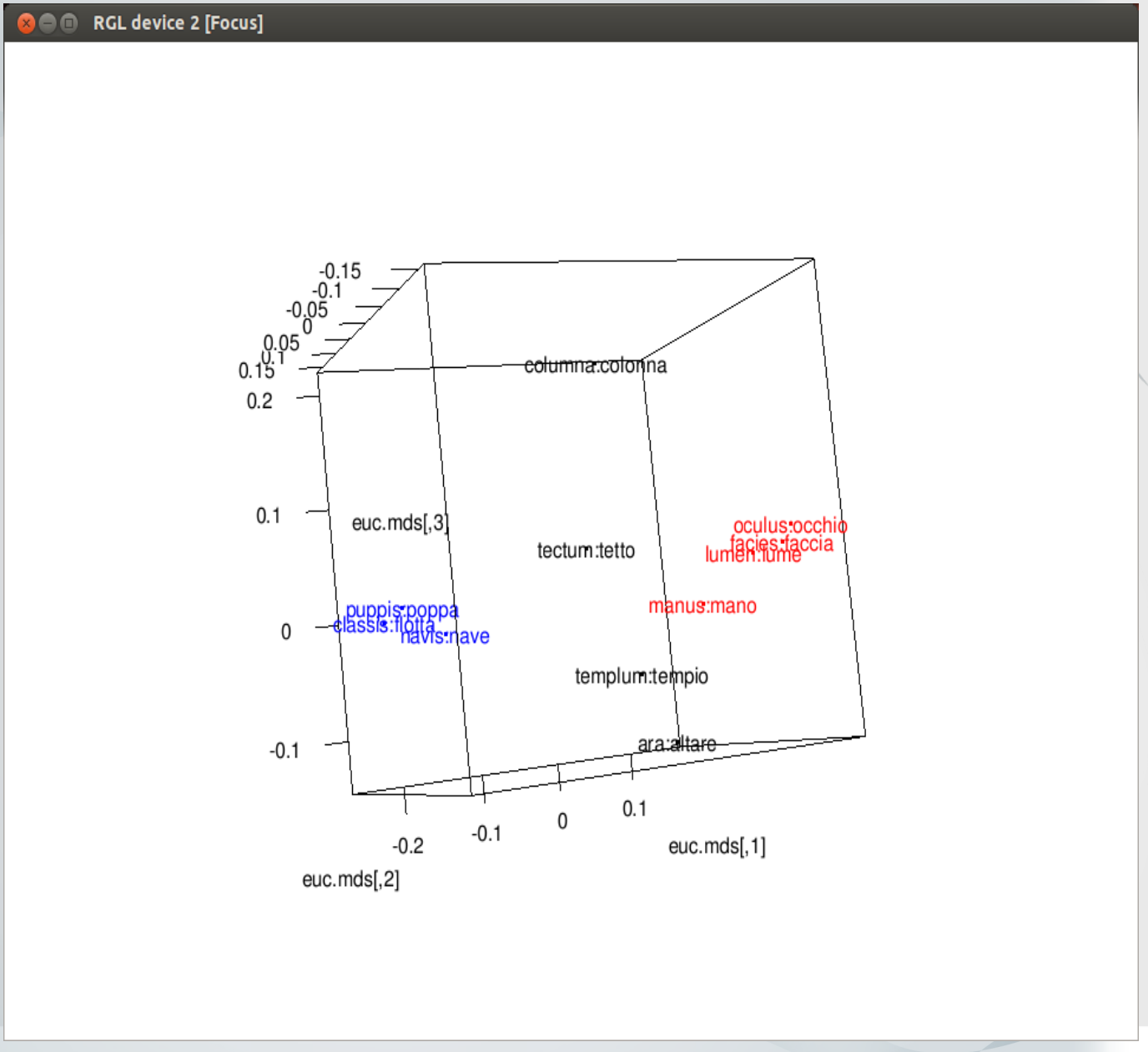
Spazi di parole latine

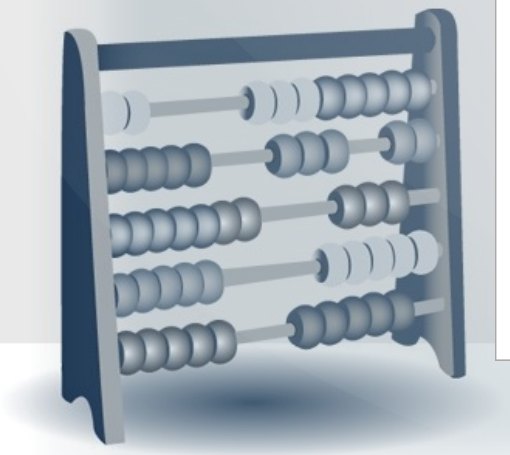




Corpus: corpus di testi poetici latini dell'Umanesimo e del Rinascimento

… e ora commentate voi questo grafico!





Bibliografia

- Papadimitriou, C. H., Raghavan, P., Tamaki, H. & Vempala S. (1998): Latent semantic indexing: A probabilistic analysis. In Proc. 17th ACM Symp. on the Principles of Database Systems, p.159-168, 1998.

- Kanerva, P., Kristoferson, J. & Holst, A. (2000): Random Indexing of Text Samples for Latent Semantic Analysis. In Gleitman, L.R. and Josh, A.K. (Eds.): Proceedings of the 22nd Annual Conference of the Cognitive Science Society, p. 1036. Mahwah, New Jersey: Erlbaum, 2000.

- Karlgren, J. & Sahlgren, M. (2001): From Words to Understanding. In Uesaka, Y., Kanerva, P. & Asoh, H. (Eds.): Foundations of Real-World Intelligence, pp. 294-308, Stanford: CSLI Publications.

- Sahlgren, M. & Karlgren, J. (2005): Automatic Bilingual Lexicon Acquisition Using Random Indexing of Parallel Corpora. Journal of Natural Language Engineering, Special Issue on Parallel Texts, 11(3) September 2005.

- Lenci, A. (2008), "Distributional semantics in linguistic and cognitive research", in A. Lenci (a cura di), From context to meaning: distributional models of the lexicon in linguistics and cognitive science, numero speciale dell'Italian Journal of Linguistics, XX/1:1-31