

Università Ca' Foscari di Venezia

Linguistica Informatica Mod. 1

Anno Accademico 2010 - 2011

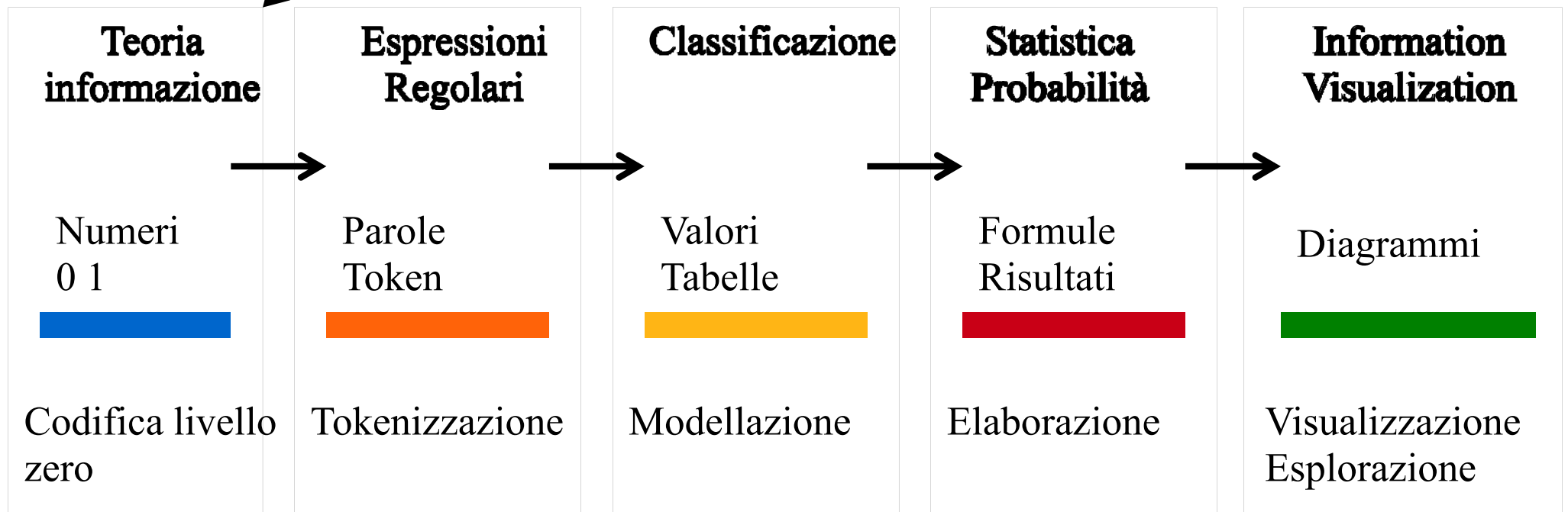


Esplorazioni e visualizzazioni

Rocco Tripodi
rocco@unive.it

Schema

Input Text



Data visualization 1

Le visualizzazioni servono ad offrire informazioni in formato grafico
Comunicare idee complesse con chiarezza precisione e efficienza
Trasformando aspetti quantitativi delle variabili e dei risultati in figure.

Peirce classifica i diagrammi come *icone* e li definisce come segni che riproducono le relazioni tra le parti. Le *icone* sono segni che si caratterizzano per il tipo di legame presunto con il referente, come insieme alle immagini (segni che sono simili all'oggetto per alcuni caratteri) e alle metafore (segni per i quali viene generato un parallelismo più generico con i loro oggetti)

Data visualization 2

Acquisire: fonti

Preparare: tabelle e valori

Filtrare: separare i dati

Processare: analisi (statistica)

Rappresentare: scegliere la forma

Raffinare: dare risalto ai risultati

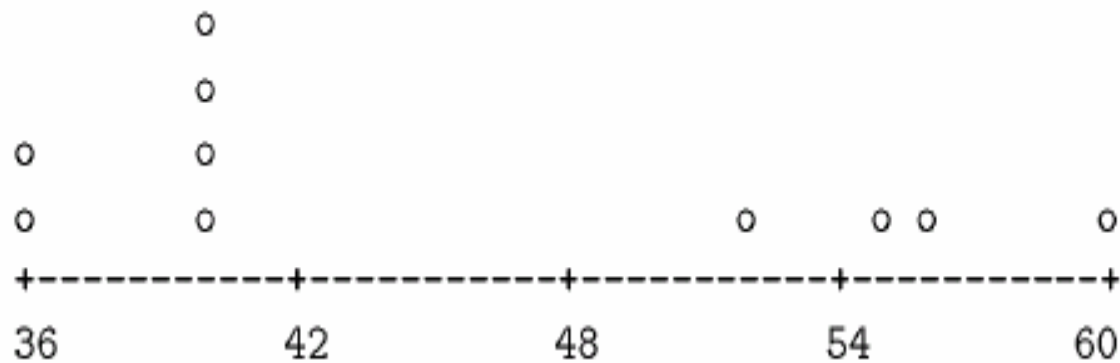
Interazione: fornire strumenti di interazione

Tipi di visualizzazioni 1

Scatter unidimensionale

Utile se il carattere è continuo e il numero di osservazioni non è molto elevato altrimenti si deve ricorrere a strumenti come lo zoom o il filtraggio

{40; 36; 52; 36; 60; 55; 56; 40; 40; 40}



Tipi di visualizzazioni 2

Areogramma

è un tipo di rappresentazione grafica in cui diverse percentuali dei risultati di un'indagine statistica sono visualizzate da aree proporzionali

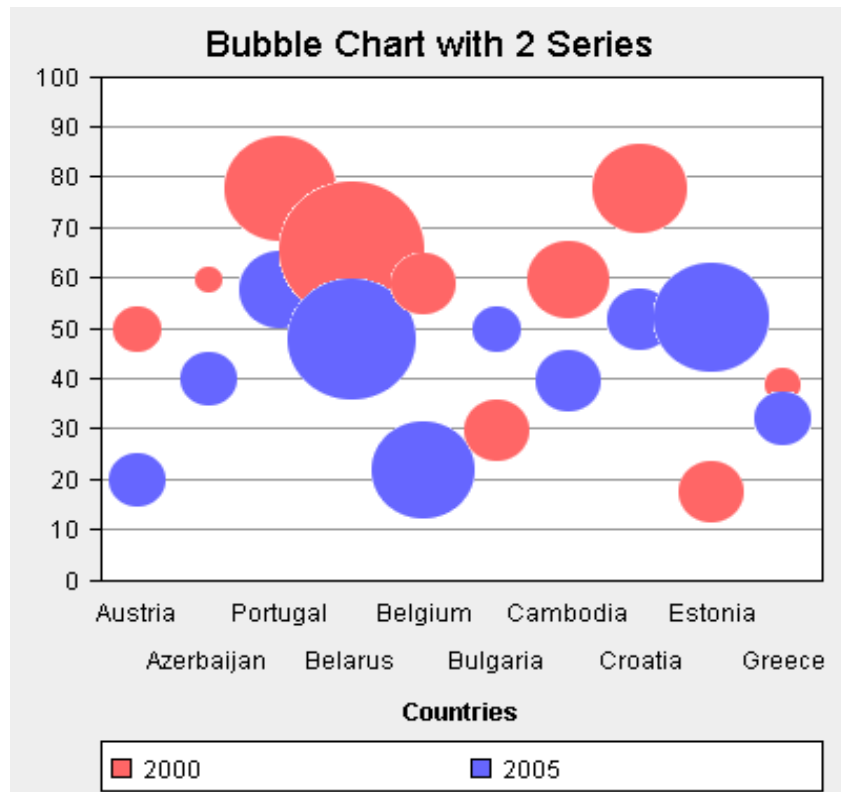
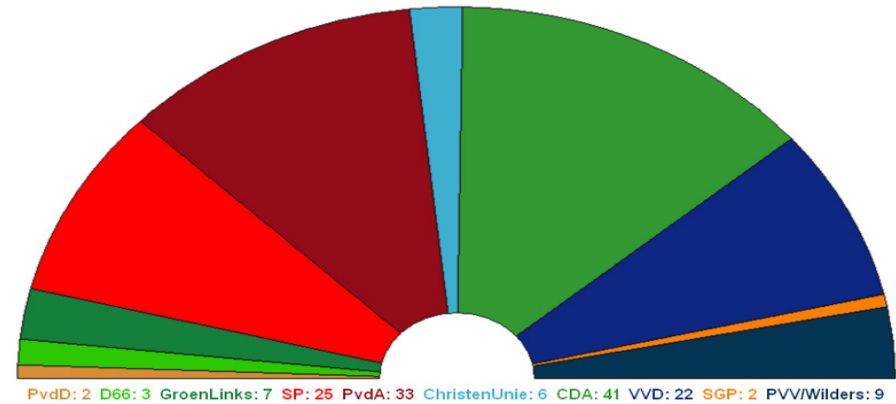
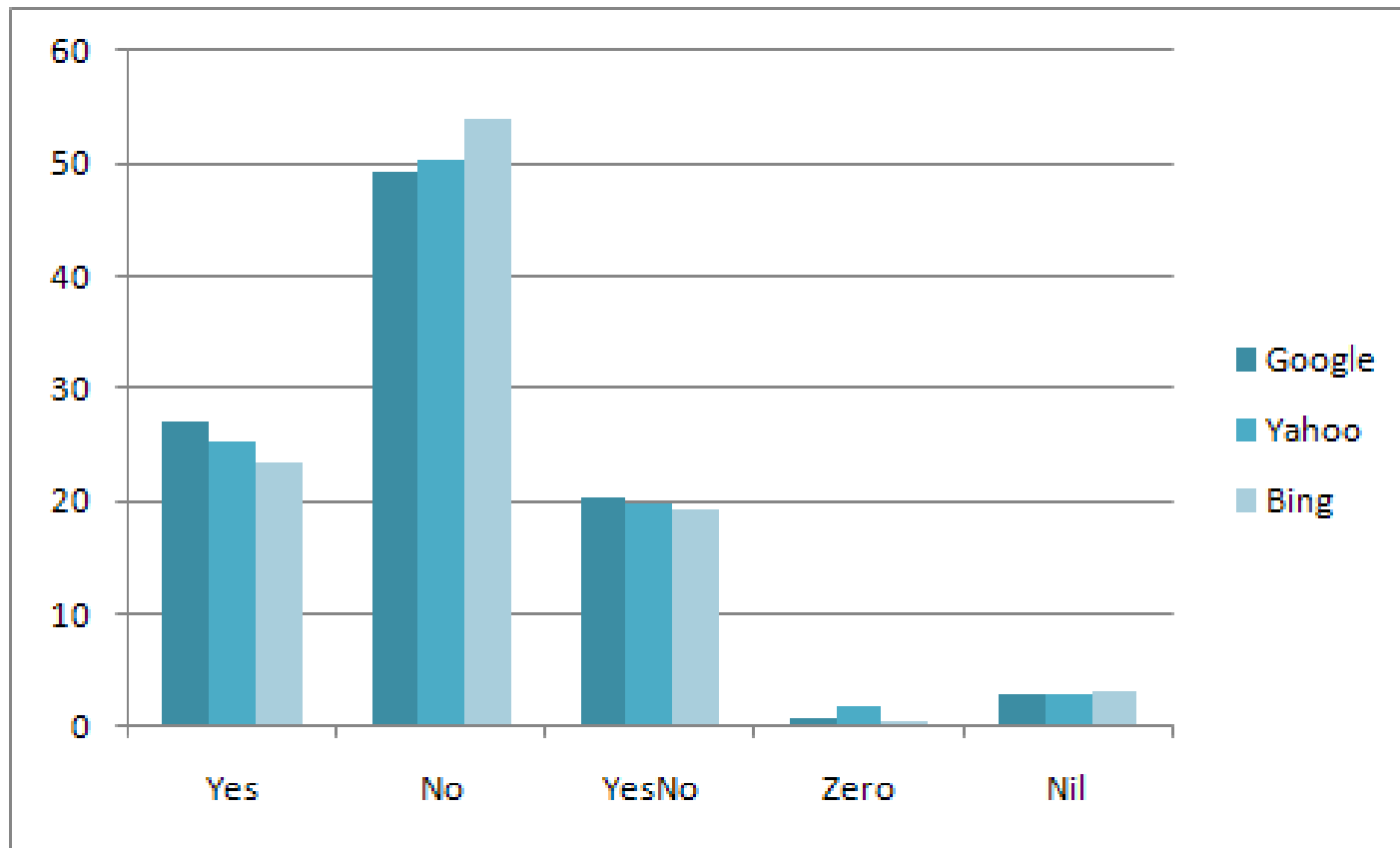


Grafico a bolle
Simile allo scatter

Tipi di visualizzazioni 3

Diagrammi a barre

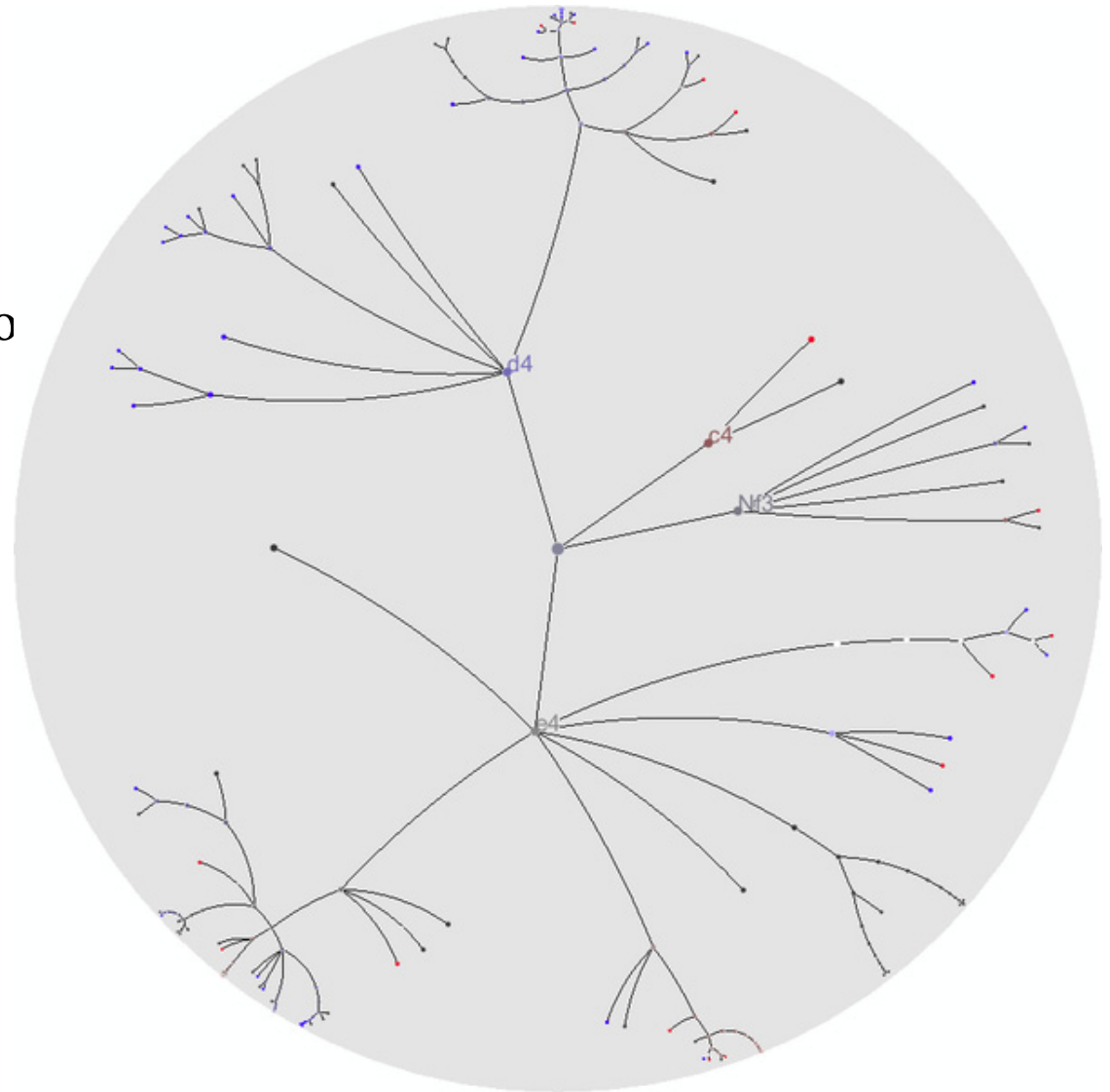
Metodo di visualizzazione in cui le frequenze sono disposte su un asse e le categorie su l'altro



Tipi di visualizzazione 4

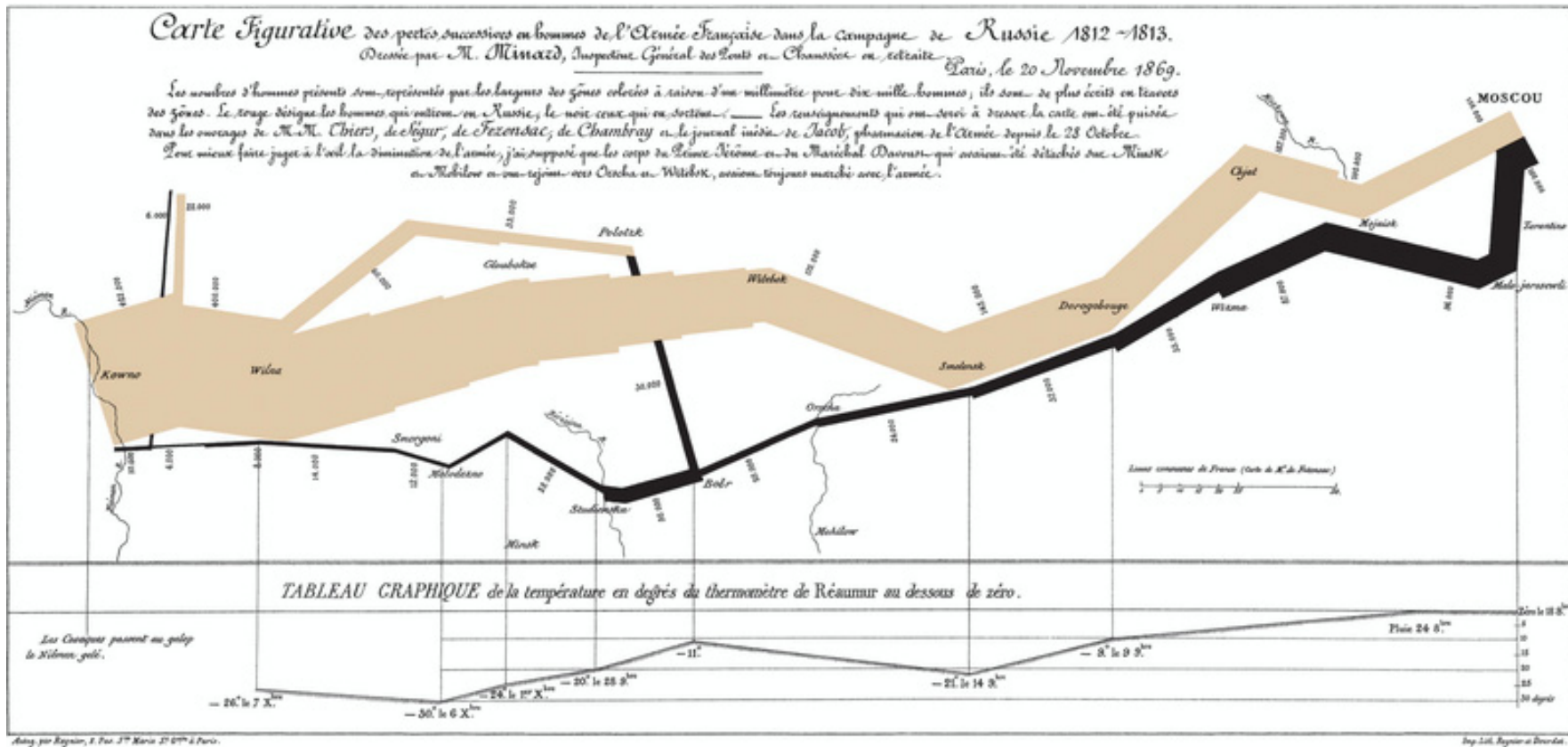
Hyperbolic tree

Utile per rappresentare rapporti gerarchici e le relazioni tra i costituenti del concetto rappresentato



Infografica

I diagrammi acquisiscono una forma prettamente narrativa



Esplorare il testo

Rappresentare l'informazione linguistica nei testi

Osservare il comportamento delle espressioni nei loro contesti d'uso

Dimensione lineare del testo

successione delle forme (x_1, x_2, \dots, x_n)

Dimensione verticale del testo:

informazioni ricavate dalla struttura linguistica (etichette sintattiche, tematiche, semantiche, ecc.). Classificazione delle forme lessicali concrete in categorie astratte. $(x_1 = p, q, r)$

L'annotazione consente di esplorare e analizzare quantitativamente il testo

Metodi di esplorazione

Quantitativi

Calcolare la rilevanza di un fenomeno in base alla probabilità con cui compare in un testo (indici di associazione)

Qualitativi

Individuare elementi che descrivano un particolare uso, ricorrendo al contesto (concordanze)

Separare le informazioni dal rumore

Esplorazione quantitativa

Indici di associazione

Elenco dei termini associati ad una determinata parola e tracciamento del numero di volte che l'associazione viene ripetuta.

Token 1	Associazione1	Quantità	Tipo
Token 1	Associazione2	Quantità	Tipo

Esplorazione qualitativa 1

Concordanze

Lista delle occorrenze di una parola. Ogni entrata della lista presenta il contesto in cui la parola compare nel testo

Utilità

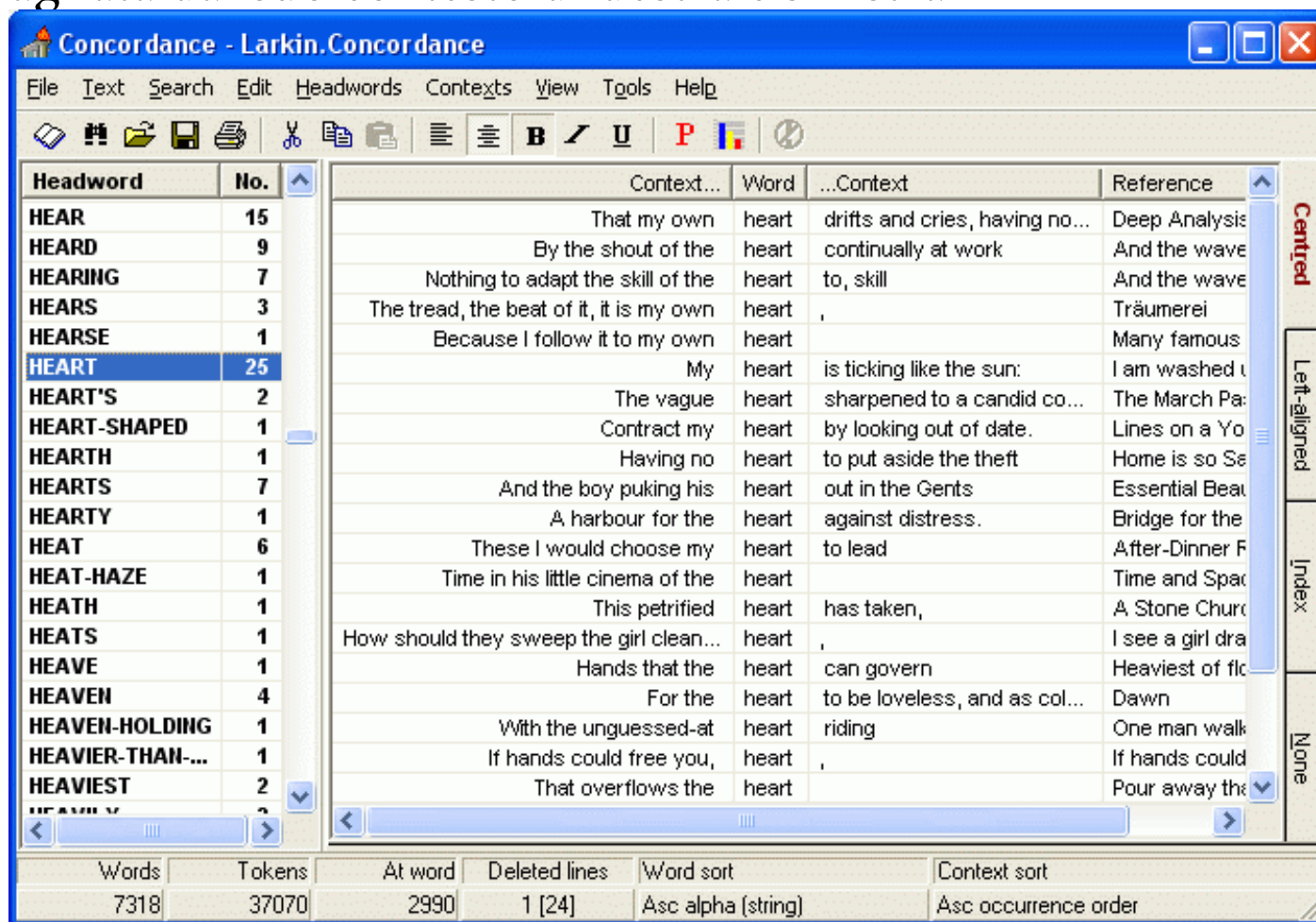
Si evidenzia come vengono usati gli elementi testuali
Indicano come sono legate determinate parole chiave
Individuazione degli omografi aventi sensi diversi
Tracciano la caratterizzazione di un personaggio

Esplorazione qualitativa 2

Key Word In Context

Viene ricercata una parola chiave

Si ottengono tante righe quante sono le occorrenze della parola chiave. La parola chiave è accompagnata dal suo contesto di destra e sinistra



The screenshot shows the Larkin Concordance software interface. The window title is "Concordance - Larkin.Concordance". The menu bar includes File, Text, Search, Edit, Headwords, Contexts, View, Tools, and Help. The toolbar contains various icons for file operations and editing. The main window is divided into several sections:

- Headword List:** A list of words and their frequencies. The word "HEART" is highlighted with a frequency of 25.
- Context Table:** A table showing the context of each occurrence of the selected word. The columns are "Context...", "Word", "...Context", and "Reference".
- Alignment Options:** A vertical sidebar on the right with buttons for "Centred", "Left-aligned", "Index", and "None".
- Status Bar:** At the bottom, it displays statistics: Words (7318), Tokens (37070), At word (2990), Deleted lines (1 [24]), Word sort (Asc alpha (string)), and Context sort (Asc occurrence order).

Headword	No.	Context...	Word	...Context	Reference
HEAR	15	That my own	heart	drifts and cries, having no...	Deep Analysis
HEARD	9	By the shout of the	heart	continually at work	And the wave
HEARING	7	Nothing to adapt the skill of the	heart	to, skill	And the wave
HEARS	3	The tread, the beat of it, it is my own	heart	,	Träumerei
HEARSE	1	Because I follow it to my own	heart		Many famous
HEART	25	My	heart	is ticking like the sun:	I am washed u
HEART'S	2	The vague	heart	sharpened to a candid co...	The March Pa
HEART-SHAPED	1	Contract my	heart	by looking out of date.	Lines on a Yo
HEARTH	1	Having no	heart	to put aside the theft	Home is so Se
HEARTS	7	And the boy puking his	heart	out in the Gents	Essential Bea
HEARTY	1	A harbour for the	heart	against distress.	Bridge for the
HEAT	6	These I would choose my	heart	to lead	After-Dinner F
HEAT-HAZE	1	Time in his little cinema of the	heart		Time and Spa
HEATH	1	This petrified	heart	has taken,	A Stone Churc
HEATS	1	How should they sweep the girl clean...	heart	,	I see a girl dra
HEAVE	1	Hands that the	heart	can govern	Heaviest of flc
HEAVEN	4	For the	heart	to be loveless, and as col...	Dawn
HEAVEN-HOLDING	1	With the unguessed-at	heart	riding	One man walk
HEAVIER-THAN...	1	If hands could free you,	heart	,	If hands could
HEAVIEST	2	That overflows the	heart		Pour away the

Esplorazione qualitativa 3

È possibile determinare la lunghezza del contesto che si vuole visualizzare sia con numero un numero fisso di tokens che con un delimitatore

Inverted index: elenco delle parole con associata la frequenza e la posizione nel testo (riga).

Ordine di presentazione: segue in genere l'ordine di apparizione nel testo ma può essere previsto presentare in base all'ordine alfabetico dei contesti.

Sortare: ordinare alfabeticamente un indice rovesciando le parole (rimario)

Esplorazione qualitativa 4

Programmi di concordanze scaricabili

[Concordance](#)

[TACT](#)

[WordSmith](#)

[Monoconc](#)

Un programma per le concordanze può essere facilmente scritto con Perl e le espressioni regolari

Tutti i corpus moderni sono dotati di un software che ne consente l'esplorazione tramite concordanze

Collocazioni

Definizione

Co-occorrenza privilegiata, associazione abituale di una parola con un'altra all'interno di una frase.

Possono essere generate linguaggi settoriali

Sistema operativo

Possono essere generate da espressioni idiomatiche

Tagliare la corda

Costruzioni a verbo supporto

Dare manforte, prendere posto

Sono escluse le combinazioni lessicali che sfruttano semplicemente le regole combinatorie morfo – sintattiche

Proprietà delle collocazioni

Elevata convenzionale

termini tecnici, uso abituale

Ridotta composizionalità semantica

Il significato generale non è dedotto componendo i significati dei costituenti

Forte rigidità strutturale (strutture pre-confezionate)

Possono ricorrere tramite costruzioni specifiche

A notte fonda

A notte profonda

Le parole si selezionano a vicenda e funzionano come una parola unica

Misurare le collocazioni 1

Problema

Vaghezza della nozione di collocazione

Soluzione

Trasformare la nozione in un indice misurabile. Affinché si possa valutare la forza di un legame

Se due o più parole in un testo ricorrono insieme è presumibile che si ripetano in maniera statisticamente rilevante

Misure

Frequenza assoluta di bigrammi, trigrammi, ecc di un testo

Calcolando solo la frequenza assoluta si trascurano le volte in cui le parole compaiono da sole o accompagnate da altre parole

Misurare le collocazioni 2

Escludere le collocazioni formate da semplici regole combinatorie

Es: un determinante ricorre molte più volte nel testo rispetto alle volte che ricorre accoppiato ad una determinata parola

Mutua informazione (MI)

Confrontare la probabilità di incontrare un bigramma, con le probabilità dei suoi costituenti considerati come mutuamente indipendenti

$$MI(v_1, v_2) = \log [p(v_1, v_2) / p(v_1) * p(v_2)]$$

$p(v_1, v_2)$: si calcola il rapporto tra la frequenza assoluta del bigramma e il numero di bigrammi tipo nel corpus

Misurare le collocazioni 3

Problema

La Mutua Informazione è estremamente sensibile agli eventi rari. I bigrammi formati da apax avranno un MI molto alta.

Questo perché la MI privilegia i casi isolati di collocazione e così facendo riesce ad eliminare le false collocazioni ma diventa sproporzionata nei casi isolati.

Soluzione (parziale)

Stabilire una soglia di frequenza al di sotto della quale le collocazioni non vengono calcolate. Questa soluzione però riduce la quantità di collocazioni individuali

Bigrammi astratti

Estendere il concetto di collocazione a gruppi formati da più di due unità

Ricorrendo alla struttura sintattica

Es: Verbo + Frase nominale

Dare un contributo

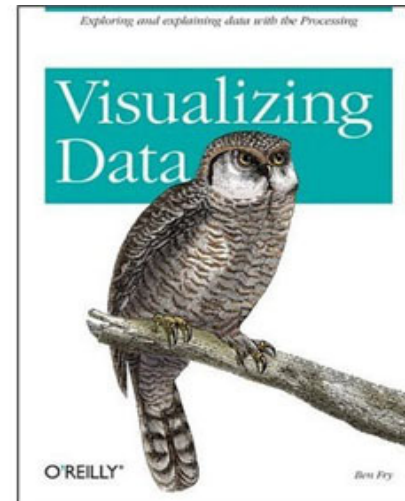
Dare un importante contributo

Dare un significativo contributo

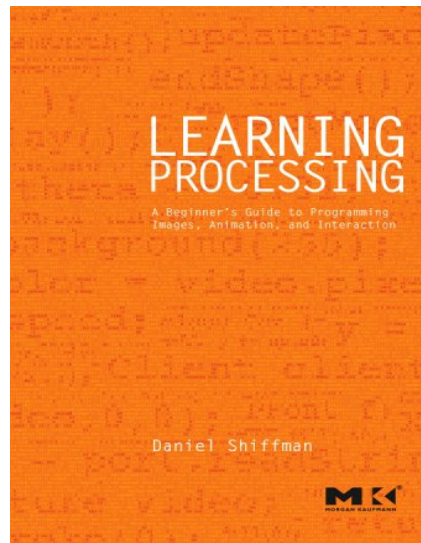
Libri consigliati



Learn Prolog Now
[On - Line](#)



Visualizing Data
Ben Fry
(BAS)



Learning Processing
Daniel Shiffman
[Link](#)