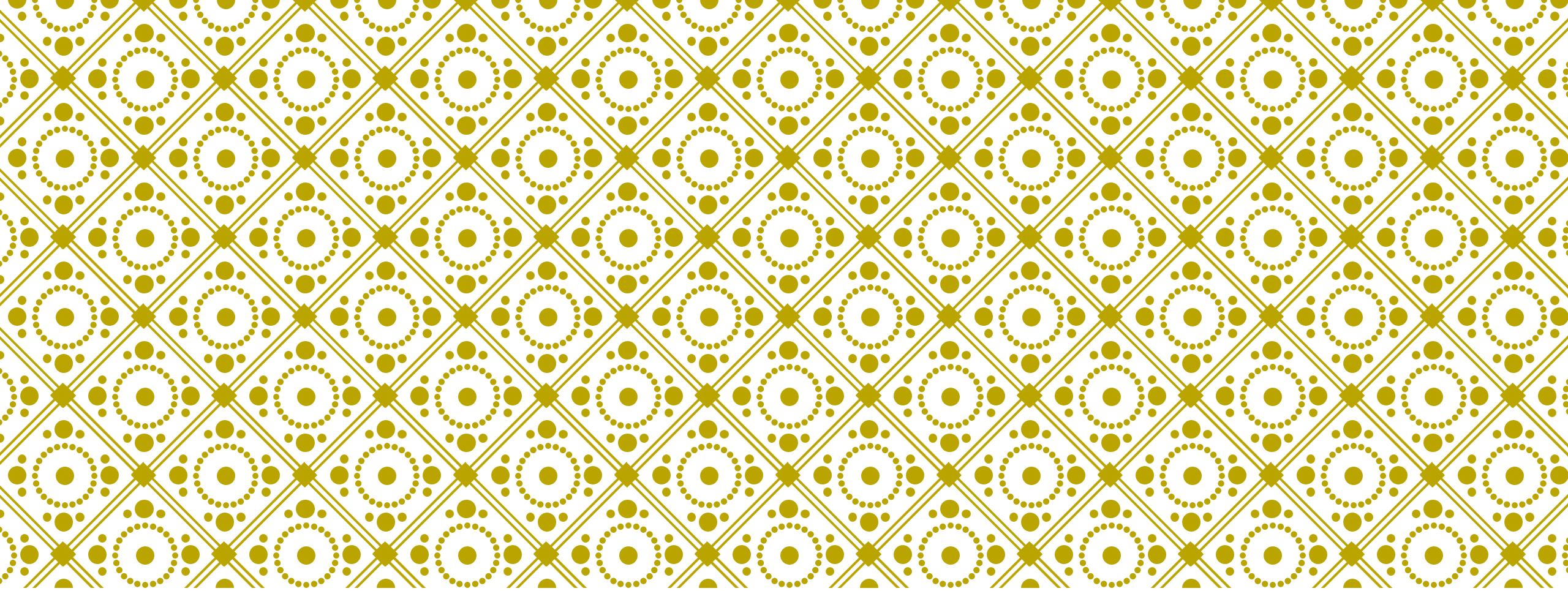


MODULO 1 – FONDAMENTI DI LINGUISTICA DEI CORPORA

Da Freddi, M. *Linguistica dei
corpora*

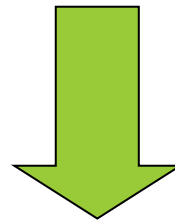


MODULO 1 – FONDAMENTI DI LINGUISTICA DEI CORPORA

1.1. Caratteristiche dei corpora

1.1. LA LINGUISTICA COME DISCIPLINA EMPIRICA

La linguistica è una disciplina empirica



Le sue generalizzazioni traggono il fondamento da dati empirici

I principi teorici rendono conto dei fatti di lingua

1.1. CHOMSKY: LINGUISTICA INTERNA ED ESTERNA

Grammaticalità vs Accettabilità

La prima deriva dall'**osservanza delle regole della grammatica**, nella costruzione di una frase:

Incolori verdi idee dormono furiosamente

La seconda deriva dalla **capacità dei parlanti di attribuire un senso alle frasi e di utilizzarle in contesti appropriati**:

Dovremmo riparare la sedia infelice - Domani ti messaggio

1.1. CHOMSKY: LINGUISTICA INTERNA ED ESTERNA

Grammaticalità vs Probabilità

La frase:

Ho visto un'esile balena

Per quanto sia improbabile nella ordinaria comunicazione linguistica, è perfettamente grammaticale

1.1. CHOMSKY: LINGUISTICA INTERNA ED ESTERNA

A partire dalla pubblicazione delle *Strutture della sintassi* (1957), i dati linguistici sono soprattutto i giudizi di grammaticalità.

Secondo Chomsky. La Linguistica interna riguarda lo studio della *competenza* (Grammatica Generativa), la Linguistica esterna riguarda lo studio dell'*esecuzione*.

1.1. CHOMSKY: LINGUISTICA INTERNA ED ESTERNA

Parallelamente, ed in polemica con la GGT, si sviluppa un approccio nel quale i dati sono il prodotto dell'attività linguistica dei parlanti.

Herdan reinterpreta la dicotomia langue/parole in termini di popolazione statistica/campione statistico

1.1. TIPI DI DATI LINGUISTICI

Dati naturalistici

Aspetto positivo: naturalezza del contesto.

Aspetto negativo: difficoltà di controllo delle variabili pertinenti.

Dati controllati sperimentalmente

Aspetto positivo: astrazione ed idealizzazione.

Aspetto negativo: interferenza dello sperimentatore.

La linguistica computazionale ha bisogno di entrambe le procedure di raccolta dei dati.

1.1. ALCUNE DEFINIZIONI DI CORPORA

Esistono diverse definizioni di corpus:

1. Testo che raccoglie occorrenze di lingua in uso, scelte per caratterizzare uno stato o una varietà linguistica (Sinclair 1991: 171).
2. Una raccolta di testi che si assume essere rappresentativa per una determinata lingua, messa insieme per essere usata ai fini di un'analisi linguistica (Tognini – Bonelli 2001: 2).
3. Una raccolta di esempi di occorrenze di lingua in uso, che consistono di qualsiasi cosa che vada da poche frasi sino a un insieme di testi scritti o registrazioni, che sono stati raccolti per lo studio linguistico. Più recentemente, raccolte di testi memorizzati a cui si accede elettronicamente (Hunston 2002: 2) .
4. Una raccolta di testi o parti di testi su cui si può condurre una qualche analisi linguistica generale. In tempi recenti, si è arrivati a considerare un corpus come un insieme di testi reso disponibile in forma computerizzata per scopi di analisi linguistica (Meyer 2002)
5. Un sacco di testo, memorizzato su un computer (Leech 1992: 106).
6. Una raccolta di parti di lingua selezionate e ordinate secondo espliciti criteri linguistici per essere usate come campioni della lingua (Eagles 1996).

1.1. CORPORA E TRATTI DEFINITORI

Campione estratto da una popolazione più ampia selezionato per condurvi un qualche tipo di analisi linguistica i cui esiti dovrebbero consentirci di inferire qualcosa anche della popolazione da cui il campione è stato tratto , dovrebbero cioè essere generalizzabili (1,2,6)

Scarto esistente tra la concezione attuale di corpus e un'epoca in cui la ricerca linguistica, benché empiricamente fondata e orientata a indagini su esempi di uso naturale, non era ancora supportata dal computer (3,4)

Definizione scherzosa che allude al fatto che i corpora oggi hanno superato i 500 milioni di parole di testo costituendo veri e propri magazzini testuali (5)

1.1. CORPORA E TRATTI DEFINITORI

un **corpus** in linguistica

un insieme di testi che si assume essere rappresentativo dello stato di una lingua, o di una varietà di essa, al fine di ottenere una descrizione complessiva

1.1. I CORPORA

Il *luogo naturale* dei dati linguistici è costituito dai **TESTI**

Una collezione di testi raccolti e organizzati per rispondere alle esigenze dell'analisi linguistica è detta *corpus*.

Esso è un sottoinsieme di tutte le possibili produzioni linguistiche, ossia ne costituisce un *campione*.

1.1. I CORPORA (2)

PROBLEMA: Campionamento da una popolazione infinita o, quantomeno, non delimitabile



Dimensione del campione [possibilità offerte dalla tecnologia-corpora dinamici]



Metodo di campionamento [bilanciamento (campioni stratificati)]

1.1. I CORPORA (3)

La tipologia di un corpus è determinata da:

Generalità [specialistico/generale]

Modalità [lingua scritta/lingua parlata/misto]

Cronologia [sincronico/diacronico]

Lingua [monolingue/multilingue]

1.1 TRATTI DEFINITORI E PROBLEMI

1. Autenticità

2. Rappresentatività e campionamento

3. Informatizzazione e rappresentazione
dei dati linguistici

1.1.1 AUTENTICITÀ (DEI DATI LINGUISTICI)

Dati linguistici sono autentici > di uso reale.

La tecnologia offre grandi quantità di dati linguistici autentici.

MA l'acquisizione di dati orali è complessa

- processi di trascrizione
- permessi per lo sfruttamento delle proprietà intellettuali e i vincoli imposti dalla tutela della privacy.
- Informanti sono informati, le conversazioni sono spontanee?

1.1.2 RAPPRESENTATIVITÀ E CAMPIONAMENTO

In linea teorica, per una ricerca linguistica empiricamente orientata, l'ideale sarebbe poter osservare tutte le occorrenze testuali in una data lingua.

Il campionamento dei testi in corpus è dunque un'operazione necessaria, ma non sufficiente perché la selezione dei testi da includere nel campione deve avvenire secondo dei criteri adeguati alla popolazione che si intende studiare.

1.1.2 RAPPRESENTATIVITÀ E CAMPIONAMENTO (2)

Molti criteri da prendere in considerazione anche in base agli **obiettivi**, ad es.

rappresentazione (a) sincronica o (b) diacronica

(a) In sincronia: studio della parola ingl. *craze* 'mania, moda' > elementi rilevanti?

(1) variazione linguistica

- **diamesica** > lungo l'asse scritto-parlato
 - asse scritto: variazione **diafasica** > variazione per genere e registro (anche nuove forme testualità digitale?)
 - asse orale: variazione **diastratica** > variazione sociale
variazione **diatopica** > geografica

(2) solo testi originali o traduzioni da altre lingue

1.1.2 RAPPRESENTATIVITÀ E CAMPIONAMENTO (3)

(b) In diacronia: studio della variazione nell'uso della parola cioè negli ultimi 40 anni.

Table 9.1 Number of speakers in each age-group, by sex

	1976–1980		2010	
	Male	Female	Male	Female
15–25 – young speakers	6	6	6	6
26–45 – young adult speakers	6	6	6	6
46–65 – adult speakers	6	6	6	6
66–90 – elder speakers	–	–	6	6
Total	18	18	24	24

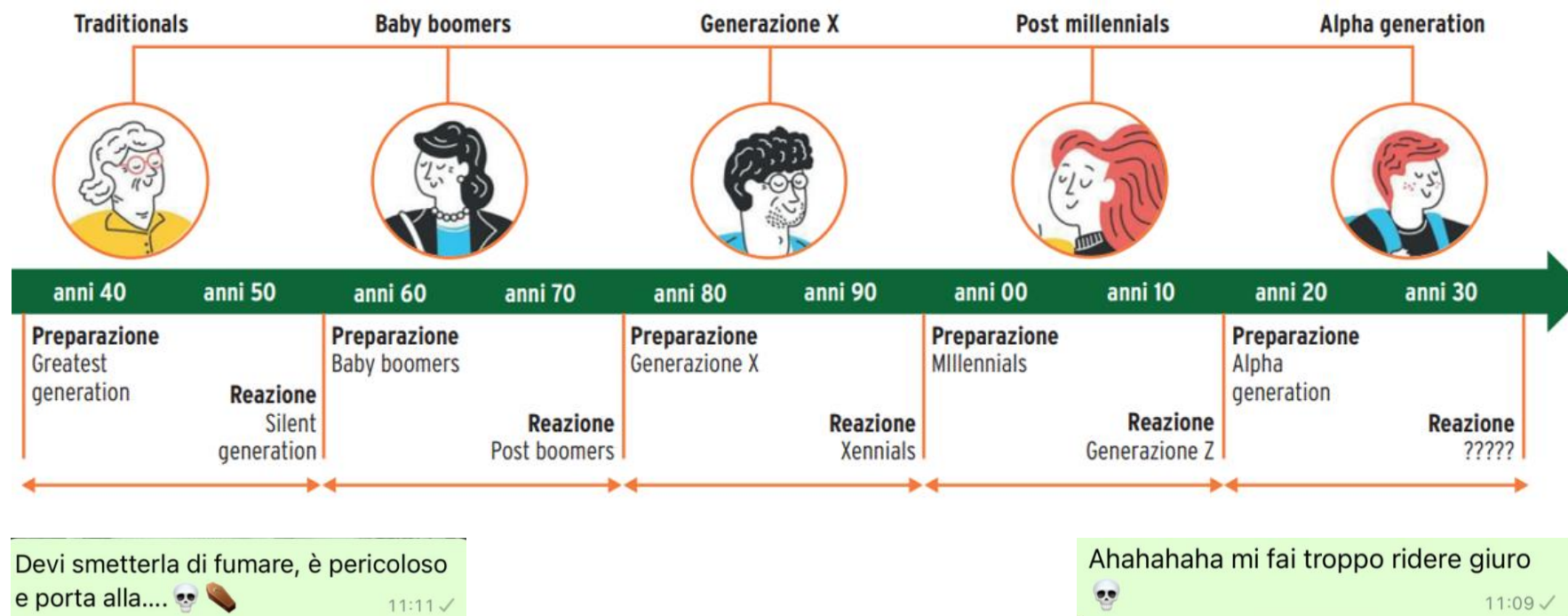
1.1.2 RAPPRESENTATIVITÀ E CAMPIONAMENTO (4)

Il campionamento è fondamentale: tenere in considerazione questi parametri, esterni all'uso linguistico, permette di **fissare le molteplici dimensioni di variabilità** intrinseca alla popolazione di cui il corpus intende fornire una rappresentazione.



la variazione linguistica in termini statistici: esiste una relazione tra uno dei parametri contestuali (*variabile indipendente*) con un certo tipo di comportamento linguistico osservato (*variabile dipendente*)

1.1.2 GENERAZIONI ED EMOJI



1.1.2 RAPPRESENTATIVITÀ E CAMPIONAMENTO

Decisioni rilevanti relative all'**ampiezza del corpus (conteggio parole)**:

- (1) la quantità e proporzione di testi da includere per ogni tipologia individuata,
- (2) la decisione se includere testi interi o porzioni (campionamento casuale).

La costruzione del corpus dipende dagli **obiettivi**. Se obiettivo è:

- (a) offrire uno spaccato della lingua in uso (corpus di riferimento) > corpus molto ampio (molti milioni di parole).
- (b) indagare un fenomeno specifico allora non necessariamente enorme, ma molto rappresentativo in base ai parametri individuati.

1.1.2 RAPPRESENTATIVITÀ E CAMPIONAMENTO

L'utilizzo (o la creazione) di un corpus **non può prescindere** dalla **conoscenza** del modo in cui sono stati affrontati i problemi della **variabilità della lingua** e della necessità di fornirne una **rappresentazione bilanciata** della variabilità della popolazione.

Se usiamo un corpus senza conoscere il modo in cui esso è stato costruito, qualsiasi deduzione fatta risulterebbe un azzardo.

1.1.2 RAPPRESENTATIVITÀ E CAMPIONAMENTO: CORPORA 'MODELLO'

BNC > British National Corpus

corpus generico di riferimento dell'inglese britannico contemporaneo (1970 al 1993).

100 milioni di parole, **90** di inglese scritto e **10** di inglese orale (rapporto di **9:1**).

SCRITTO: variazione diafasica

testi letterari e creativi (25%) e **prosa informativa (75%)** (rapporto di **1:3**). La prosa informativa è equamente divisa tra: scienze applicate, arti, fede e pensiero, commercio e finanza, tempo libero, scienze naturali e pure, scienze sociali, attualità.

90 milioni di parole (**60%** proveniente da libri, **25%** dai periodici, il **5-10%** da forme miste di materiale pubblicato, il **5-10%** da materiale scritto non pubblicato e **5%** di parlato-scritto e parlato-recitato).

ORALE: distinzione tra dati raccolti su **base demografica (età, sesso, occupazione, provenienza geografica)** (4milioni) e dati selezionati sulla **base del contesto più o meno istituzionale** (riunioni di lavoro, riunioni sindacali, lezioni accademiche, telegiornali, incontri ufficiali di governo, sedute parlamentari, telefonate radiofoniche ecc ecc.) (6milioni).

1.1.2 RAPPRESENTATIVITÀ E CAMPIONAMENTO: CORPUS 'MODELLO' (2)

Corpus KIParla: corpus italiano di **lingua orale**

Differenziazione geografica (Torino vs Bologna) **perno** nella costruzione del corpus (specialistico). Situazione sociolinguistica delle due città:

1. compresenza di italiano e dialetto;
2. meta di mobilità interna, così come di flussi migratori esterni.

Collocazione sociale degli individui: i parlanti coinvolti nelle registrazioni sono differenziati primariamente per età, titolo di studio e occupazione (parametri particolarmente significativi).

Tipologia di interazioni: interviste semistrutturate e, in contesto universitario, lezioni ed esami, differenziate in base a parametri situazionali:

- (a) relazione simmetrica/asimmetrica tra i partecipanti,
- (b) presenza/assenza di un argomento predefinito,
- (c) presenza/assenza di norme per la presa di turno, ecc.

1.1.2 CORPUS KIPARLA: LA COSTRUZIONE

Tutti i dati sono registrati a microfono palese.

Le registrazioni sono state trascritte utilizzando il software ELAN, che permette l'allineamento della trascrizione con la traccia audio.

Per le trascrizioni, è stata adottata una versione semplificata del sistema Jefferson, frequentemente utilizzato nell'analisi della conversazione.

<http://kiparla.it>

1.1.2 CORPUS KIPARLA: ESEMPIO ELAN

Esempio di trascrizione
multimodale in ELAN

The screenshot displays the ELAN software interface for a file named 'CHEB.eaf'. The interface includes a menu bar (File, Edit, Annotation, Tier, Type, Search, View, Options, Window, Help) and a toolbar with playback controls. A video window shows a person speaking. A table lists annotations for the 'ling eval P' tier:

Nr	Annotation	Begin Time	End Time	Duration
8	grad + att	00:00:44.880	00:00:46.624	00:00:00.744
9	att + grad	00:00:47.496	00:00:47.896	00:00:00.390
10	att + grad	00:00:52.494	00:00:53.016	00:00:00.522
11	att	00:01:03.162	00:01:03.762	00:00:00.600
12	att	00:01:03.762	00:01:04.200	00:00:00.438
13	grad	00:01:05.352	00:01:05.712	00:00:00.360

Below the table is a playback control bar with a selection range of 00:00:47.496 - 00:00:47.896. The main area shows a video window, an audio waveform, and several annotation tracks:

- trans D (71): we we just redoing those experiments so there's a huge distortion of
- trans P (59):
- genre (4):
- ling eval D (38):
- ling eval P (18): grad + att, att + grad
- paralanguage (9): SDL
- gesture (7): PsUMb1, SPsU
- gaze (7):
- head (7): N, M!
- face (5):

1.1.2 CORPUS KIPARLA: IL SISTEMA JEFFERSON

,	Intonazione ascendente
.	Intonazione discendente
:	Suono prolungato
(.)	Pausa breve
> ciao <	Pronuncia (più) veloce
<ciao>	Pronuncia (più) lenta
[ciao]	Sovrapposizioni tra parlanti
(ciao)	Testo di difficile comprensione (ipotesi del trascrivente)
xxx	Testo non comprensibile
((ride))	Comportamento non verbale
=	Unità unite prosodicamente

1.1.2 CORPUS KIPARLA: LA MODULARITÀ INCREMENTALE

Organizzazione **interna**:

1. moduli **indipendenti** che permettono nel tempo l'aggiunta di nuovi moduli;
2. medesimi **design e metadati**, trascritti da ELAN, e resi disponibili attraverso NoSketch Engine;
3. diverse **dimensioni della variazione linguistica** e possono raccogliere dati da diverse aree geografiche.

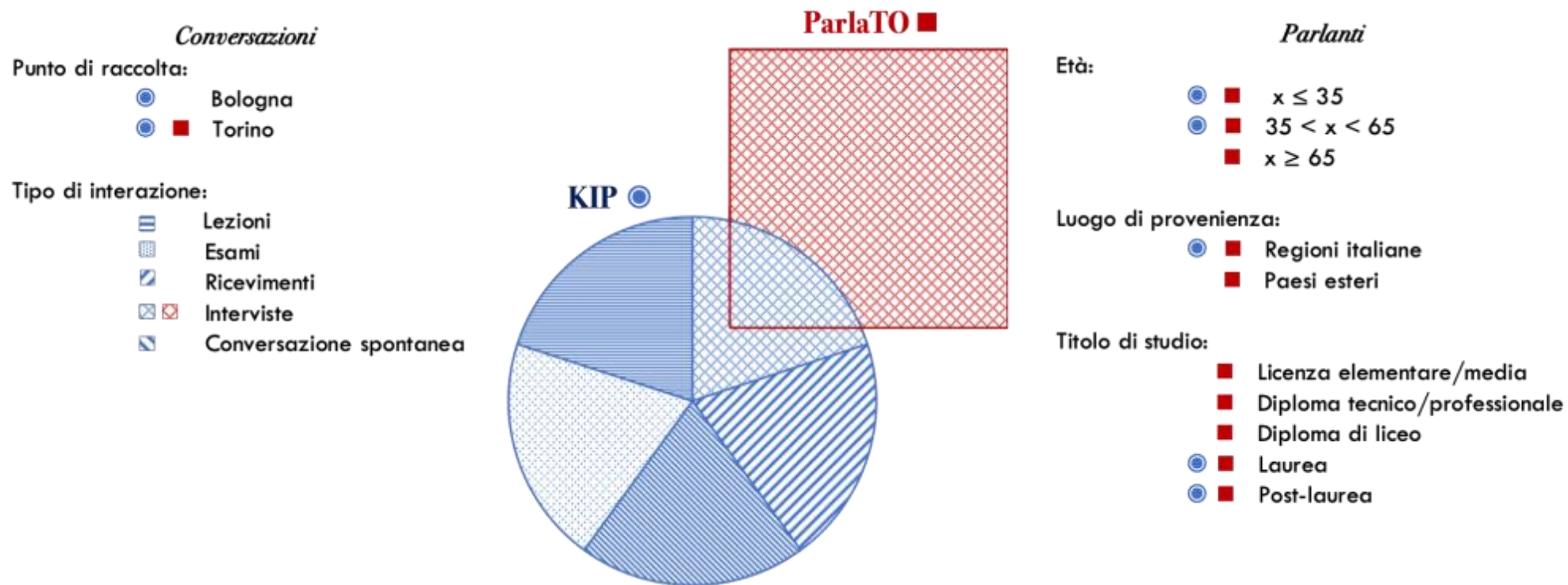
KIParla è un potenziale **corpus monitor**, aperto a integrazioni e aggiornamenti nel tempo.

1.1.2 CORPUS KIPARLA: I MODULI

Ad oggi, il corpus KIParla è costituito da due moduli:

1. Il modulo KIP: registrazioni nelle Università di Torino e Bologna in cinque tipi di situazioni comunicative (lezioni universitarie, 25h : 45m : 12s, esami, 6h : 20m : 22s, ricevimento studenti, 6h : 48m : 19s, interviste semistrutturate a studenti, 14h : 6m : 15s, conversazione libera, 16h : 23m : 33s.
2. Il modulo ParlaTO: conversazioni di più di un centinaio di parlanti con diversa provenienza geografica e diversa collocazione sociale, raccolte a Torino fra il 2018 e il 2020, prevalentemente attraverso interviste individuali e discussioni di gruppo su vari temi (studio, lavoro, attività nel tempo libero o in pensione, ricordi del passato, vita in città, ecc.). Le ore di registrazione sono ripartite in modo pressoché paritario fra parlanti giovani, adulti e anziani.

1.1.2 CORPUS KIPARLA: SINTESI



1.1.3 INFORMATIZZAZIONE E RAPPRESENTAZIONE DEI DATI LINGUISTICI

L'attuale nozione implica una componente elettronica

> problemi di rappresentazione dei dati testuali su un supporto digitale (**codifica informatica dei testi**).

> contenuti determinano analisi interpretative di tipo morfosintattico, fonetico, semantico e pragmatico che lo studioso associa ai dati testuali grezzi per poterli interrogare in maniera più raffinata (**annotazione linguistica per esplorare la struttura linguistica**)

> Importanza di una codifica standardizzata (portabilità dei dati)

1.1.3 INFORMATIZZAZIONE: ANNOTAZIONE E MARKUP

Si arricchiscono dati grezzi (testi) con **metadati**. Normalmente distinzione tra

1. **Mark-up** > codifica di metadati contestuali e oggettivi relativi ai testi da includere nel corpus, come per esempio il titolo, l'autore e l'anno di pubblicazione.
2. **Annotazione (tagging)** > informazioni di tipo interpretativo (anche linguistico) > più soggettive o opinabili.
3. **Etichettatura grammaticale (POS tagging)**, primo livello di annotazione, necessariamente preceduta dalla segmentazione del testo in parole o **token (tokenizzazione)**
4. **Lemmatizzazione**: codifica grazie alla quale si associano varianti morfologiche di una parola e le sue forme flesse come un unico lessema (flessione verbale *do, did, does, doing, done* come forme varianti riconducibili ad un unico lemma: *DO*).

1.1.3 INFORMATIZZAZIONE: ANNOTAZIONE E MARKUP (2)

Diversi livelli di annotazione

1. POS Tagging: etichettatura grammaticale
2. Annotazione sintattica
3. Annotazione semantica
4. Annotazione dei fenomeni di coesione testuale
5. Annotazione pragmatica (ad es. per atti linguistici)
6. Annotazione per categorie di errore

1.1.3 FORMATI E LINGUAGGI DI CODIFICA: XML E DATABASE RELAZIONALI

Diverse modalità con cui le informazioni linguistiche vengono codificate a livello informatico. Obiettivo: formato (idealmente) leggibile e condivisibile dall'intera comunità scientifica (ad es. TEI).

Standard deve essere estensibile e la standardizzazione deve distinguere 3 livelli:

1. Il **formato dei file** con cui i testi sono codificati (.txt, Unicode).
2. La **modalità** con cui avviene l'annotazione (standard .xml).
3. I **contenuti** dell'annotazione (.xml e grammatiche DTD *Document Type Definition* –elenco dei tag, struttura, caratteristiche).

1.1.3 CORPUS ‘MODELLO’: REPUBBLICA

Corpus specialistico e diacronico: testi **giornalistici scritti** > tutti gli articoli pubblicati nel quotidiano “la Repubblica” dal **1985 al 2000**.

Disponibile online, sito Dipartimento di Interpretazione e Traduzione di Forlì (Università Alma Mater Studiorum di Bologna).

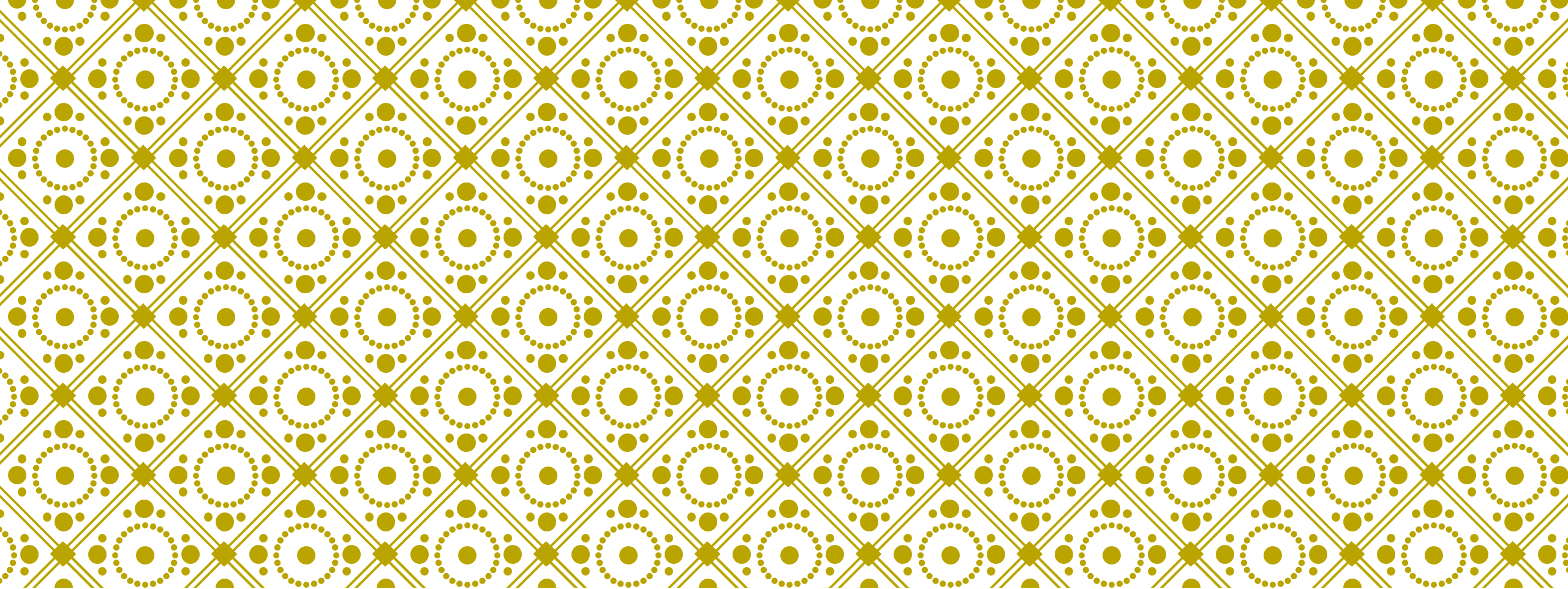
Publicato tramite la piattaforma **NoSketchEngine**. Per accedere:

1. corpora.dipintra.it
2. Cliccare pulsante blu “Public” - per accedere a NoSketchEngine
3. Nel menù a tendina in alto, selezionare “Repubblica”

1.1.3 CORPUS REPUBBLICA: TAGSET

Il **corpus Repubblica** è annotato integralmente per lemma e per parte del discorso. Al link seguente sono riportati tutti i tag utilizzati per l'annotazione per parti del discorso del corpus:

<https://docs.sslmit.unibo.it/doku.php?id=corpora:tagsets:italian>



1. FONDAMENTI DI LINGUISTICA DEI CORPORA

1.2. Creazioni e tipologie di
corpora

1.2. CREAZIONE E TIPOLOGIE DI CORPORA

I criteri che guidano la creazione di corpora rispondono a domande/obiettivi diversi > classificazione in tipologie a scopo orientativo.

1. corpora generici vs corpora specialistici,
2. corpora di parlato vs corpora di scritto,
3. corpora di testi prodotti da parlanti nativi vs non nativi,
4. corpora monolingue vs bilingue,
5. comparabili vs paralleli,
6. annotati vs non annotati.

1.2. CORPORA GREZZI VS ANNOTATI

Distinguendo in **corpora grezzi** e **corpora annotati** si indica la presenza o meno in un corpus di un qualche livello di annotazione linguistica.

BNC ad esempio contiene annotazione POS.

1.2. CORPORA GREZZI VS ANNOTATI (2)

Corpora annotati grammaticalmente (POS) prendono il nome dal modello di grammatica su cui è basata l'annotazione.

Due modelli sintattici :

1. analisi dei costituenti di frase: segmentano la frase in gruppi di parole in relazione logica tra di loro (ad es. Penn Treebank)
2. relazioni di dipendenza: specificano le relazioni gerarchiche tra il verbo e i suoi argomenti (ad es. PDT)

1.2. ANALISI DEI COSTITUENTI DI FRASE: *PENN TREEBANK*

Tag	Description	Example	Tag	Description
CC	coord. conjunction	<i>and, or</i>	RB	adverb
CD	cardinal number	<i>one, two</i>	RBR	adverb, c
DT	determiner	<i>a, the</i>	RBS	adverb, s
EX	existential there	<i>there</i>	RP	particle
FW	foreign word	<i>noire</i>	SYM	symbol
IN	preposition or sub-conjunction	<i>of, in</i>	TO	"to"
JJ	adjective	<i>small</i>	UH	interjection
JJR	adject., comparative	<i>smaller</i>	VB	verb, bas
JJS	adject., superlative	<i>smallest</i>	VBD	verb, pas
LS	list item marker	<i>1, one</i>	VBG	verb, gen
MD	modal	<i>can, could</i>	VBN	verb, pas
NN	noun, singular or mass	<i>dog</i>	VBP	verb, non
NNS	noun, plural	<i>dogs</i>	VBZ	verb, 3sg
NNP	proper noun, sing.	<i>London</i>	WDT	wh-deter
NNPS	proper noun, plural	<i>Azores</i>	WP	wh-pronc
PDT	predeterminer	<i>both, lot of</i>	WP\$	possessiv
POS	possessive ending	<i>'s</i>	WRB	wh-adver
PRP	personal pronoun	<i>he, she</i>		

tk treebank viewer

TREEBANK VIEWER Sandway Fong University of Arizona (dec 2006) (beta version)

Sentence File: /Users/sandway/Desktop/freesearch/tesj1 Prolog Tree File: /Users/sandway/Desktop/freesearch/tesj1 Load

Sentence Count: 49209 Displayed Tree (Sentence): 37975

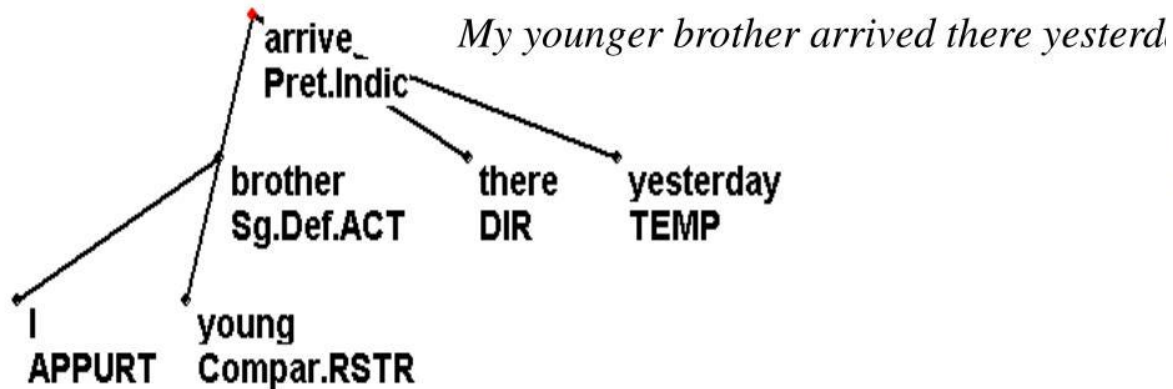
The announcement , made after the close of trading , c
The company closed at \$ 12 a share , down 62.5 cents
Pinnacle West slashed its quarterly dividend to 40 cents
A company spokesman said the decision to eliminate th
He declined to elaborate .
Edward J. Tirello Jr. , an analyst at Shearson Lehman H
Analysts have estimated that Pinnacle West may have to
The latest financial results at the troubled utility and thr
Third-quarter net income slid to \$ 5.1 million , or six o
Utility operations , the only company unit operating in th
In other operations , losses at Merabank totaled \$ 85.7
The latest quarter includes a \$ 42.7 million addition to
As recently as August , the company said it did n't forec
Pinnacle 's SunCor Development Co. , real-estate unit 's
The latest period included a \$ 9 million write-down on
Losses at its Malapai Resources Co. , uranium-mining ur
Losses at El Dorado Investment Co. , the venture-capita
The latest quarter included a \$ 6.6 million write-down
Equitec Financial Group said it will ask as many as 100,
Under the proposal by Equitec , a financially troubled ri
Shares of the new partnership would trade on an excha
Hallwood is a merchant bank whose activities include th
in a statement , Equitec Chairman Richard L. Saalfeld sa
While he did n't describe the partnerships ' financial cor

```

graph TD
    S --> ADVP-TMP
    S --> NP-SBJ
    S --> VP
    ADVP-TMP --> ADVP
    ADVP --> RB
    ADVP --> RB
    ADVP --> IN
    NP-SBJ --> DT
    NP-SBJ --> NN
    NP-SBJ --> NP
    NP --> NNP
    NP --> NNP
    NP --> NNP
    VP --> VBD
    VP --> SBAR
    SBAR --> NONE
    SBAR --> S
    S --> NP-SBJ
    NP-SBJ --> PRP
    NP-SBJ --> VBD
    VBD --> VBD
    VBD --> VBD
  
```

1.2. RELAZIONI DI DIPENDENZA: *PRAGUE DEPENDENCY TREEBANK*

Dependency tree



Linearized form, one-to-one relation:

((I)_{Appurt} (younger)_{Rstr} brother)_{Act} arrive.Pret.Indic (Dir there) (Temp yesterd.

- Pred - Predicate if it depends on the tree root
- Sb - Subject
- Obj - Object
- Adv - Adverbial
- Atv - Complement
- AtvV - Complement, if one governor is present
- Atr - Attribute
- Pnom - Nominal predicate's nominal part, depends on the copula „to be“
- AuxV - Auxiliary verb „to be“
- Coord - Coordination node
- Apos - Apposition node
- AuxR - Reflexive particle, which is neither Obj nor AuxT (passive)
- AuxT - Reflexive particle, lexically bound to the verb

1.2. UN CASO SPECIALE DI ANNOTAZIONE DEL PARLATO

PCFD (Pavia Corpus of Film Dialogue): parlato filmico in due lingue, inglese e italiano.

Rappresentazione del parlato attraverso trascrizione > scelte con rilevanza teorica (ad es. trascrizione ortografica vs prosodica).

Scelta consona al tipo di analisi che si condurrà sul corpus.

Allineamento > corpora paralleli e comparabili, l'individuazione e la marcatura delle corrispondenze tra porzioni di testo equivalenti o confrontabili in due o più lingue.

Unità di allineamento? In PCFD è la battuta.

1.2. CORPORA E VARIAZIONE DIAFASICA

1. corpus di riferimento: testi di tutte le varietà diafasiche, diastratiche, diatopiche e diamesiche, considerando le caratteristiche di una lingua nel suo insieme
2. corpus specialistico: circoscritto ad un singolo genere o dominio

1.2. CORPORA GENERICI DI RIFERIMENTO

Grandi quantità di parole in gamma di testi il più possibile varia e completa (ad es. costruire una grammatica).

Oggi questo tipo di corpus raggiunge le centinaia di milioni di parole, etichettate grammaticalmente, e copre un'enorme quantità di testi scritti e orali.

Al suo interno è possibile individuare ulteriori distinzioni

(a) inglese in BNC vs (b) americano in COCA

oppure

1. corpus statico: fornisce quadro di una lingua attraverso un numero di parole prefissato e raccolte in un arco temporale preciso (ad es. BNC),
2. corpus dinamico: costantemente aggiornato (KIParla, COCA).

1.2. CORPORA GENERICI DI RIFERIMENTO

I corpora generici di riferimento sono usati anche per gli studi di genere (a metà tra la sociolinguistica e l'analisi del discorso) perché mettono in rilievo distribuzioni di frequenza diversamente associati al sesso, all'età o alla condizione sociale del parlante.

Nei corpora di riferimento sono dunque l'ampiezza e la gamma di tipologie testuali che determinano il criterio dominante di selezione dei testi.

1.2. CORPORA SPECIALISTICI

I corpora specialistici sono generalmente più piccoli (in media 1 o 2 milioni di parole).

Tra i più noti

(a) Repubblica, per italiano

(b) il MICASE (Michigan Corpus of Academic Spoken English) per l'inglese accademico orale nella varietà americana (1.8 milioni di parole che corrispondono a 200 ore di parlato trascritto offrendo una rappresentazione della lingua in uso nei diversi contesti accademici dell'Università del Michigan),

(c) il BASE (British Academic Spoken English) di 1.6 milioni di parole (Università di Warwick e Reading).

1.2. CORPORA E VARIAZIONE DIAMESICA

La lingua scritta ha caratterizzato i primi corpora, invece i corpora di solo parlato sono molto più recenti.

Il potenziamento dei mezzi informatici ha contribuito solo parzialmente a velocizzare lo sviluppo dei corpora del parlato: l'interesse primario è di associare, alla trascrizione di testi orali, file audio e video.

CORPUS DI SCRITTO: BROWN esemplifica ogni corpora di scritto prodotto successivamente.

1.2. CORPORA DI PARLATO

Le componenti orali dei corpora di riferimento consentono l'osservazione della variazione diamesica all'interno di una stessa varietà geografica.

CANCODE-> Raccoglie trascrizioni di parlato spontaneo in inglese britannico registrato in Gran Bretagna nelle situazioni più disparate (5 milioni di parole)

Wellington Corpus-> trascrizioni raccolte tra il 1990 e il 1994 (12% discorsi formali o monologhi – 13% discorsi semiformali o monologhi elicitati – 75% conversazioni informali o dialoghi). Gli estratti sono suddivisi in 15 categorie che coprono una vasta gamma di contesti d'uso: monologo o dialogo, pubblico o privato, pianificato o non.

1.2. CORPORA E VARIAZIONE DIACRONICA

1. corpora sincronici: offrono uno spaccato di una lingua in un momento definito,
2. corpora diacronici: contengono testi di periodi diversi in una stessa lingua; adatti per studi di linguistica storica, pragmatica storica, sociolinguistica storica, ecc.

Alcuni esempi:

CORPORA SINCORNICI: l'ICE raccoglie dati linguistici delle diverse varietà nazionali e regionali dell'inglese nel mondo.

CORPORA DIACRONICI: Helsinki Corpus raccoglie testi inglesi di 3 grandi periodi della storia che vanno dal VIII secolo fino all'inizio del XVIII (inglese antico, medio e primo moderno).

I dati sono annotati secondo parametri sociolinguistici > mettere in relazione osservazioni sulla lingua con altre variabili (ad es. sesso, età, status sociale).

1.2. CORPORA DINAMICI O DI MONITORAGGIO

Un'altra categoria di corpora viene utilizzata per lo studio del cambiamento linguistico, ossia i corpora di monitoraggio, distinti perché a loro modo sono corpora dinamici.

1.2. CORPORA E APPRENDIMENTO LINGUISTICO

Raccolta di testi prodotti da apprendenti di una lingua straniera per confrontare l'interlingua degli apprendenti rispetto alla lingua materna dei nativi.

Il più noto corpus inglese L2 è sicuramente l'ICLE (International Corpus of Learner English) che contiene saggi di tipo argomentativo scritti da studenti universitari con 16 lingue madre diverse. I testi contengono metadati come sesso, età e livello di competenza.

Nella sua ultima versione del 2009: 3,7 milioni di parole e 6085 testi. Tutti i testi sono lemmatizzati ed etichettati grammaticalmente.

1.2. APPLICAZIONI DI CORPORA DI APPRENDIMENTO LINGUISTICO

Numerose applicazioni:

1. lo sviluppo di sillabi e materiale per apprendenti,
2. definire sequenze didattiche di argomenti grammaticali,
3. definire l'ordine di presentazione del lessico in base ai diversi obiettivi di apprendimento (ad es. creazione di prove d'esame finali o test d'ingresso).

Inizialmente la realizzazione di questi corpus era limitata all'inglese L2, oggi la gamma di lingue straniere è molto più ampia.

1.2. CORPORA MULTILINGUE

Con la bipartizione tra corpora paralleli e corpora comparabili si entra nella dimensione del confronto interlinguistico.

Si possono scegliere

1. testi tradotti e confrontati con i testi originali
2. testi confrontabili per genere, registro, funzione ma redatti in lingue diverse.

Si tratta però in genere di corpora monodirezionali

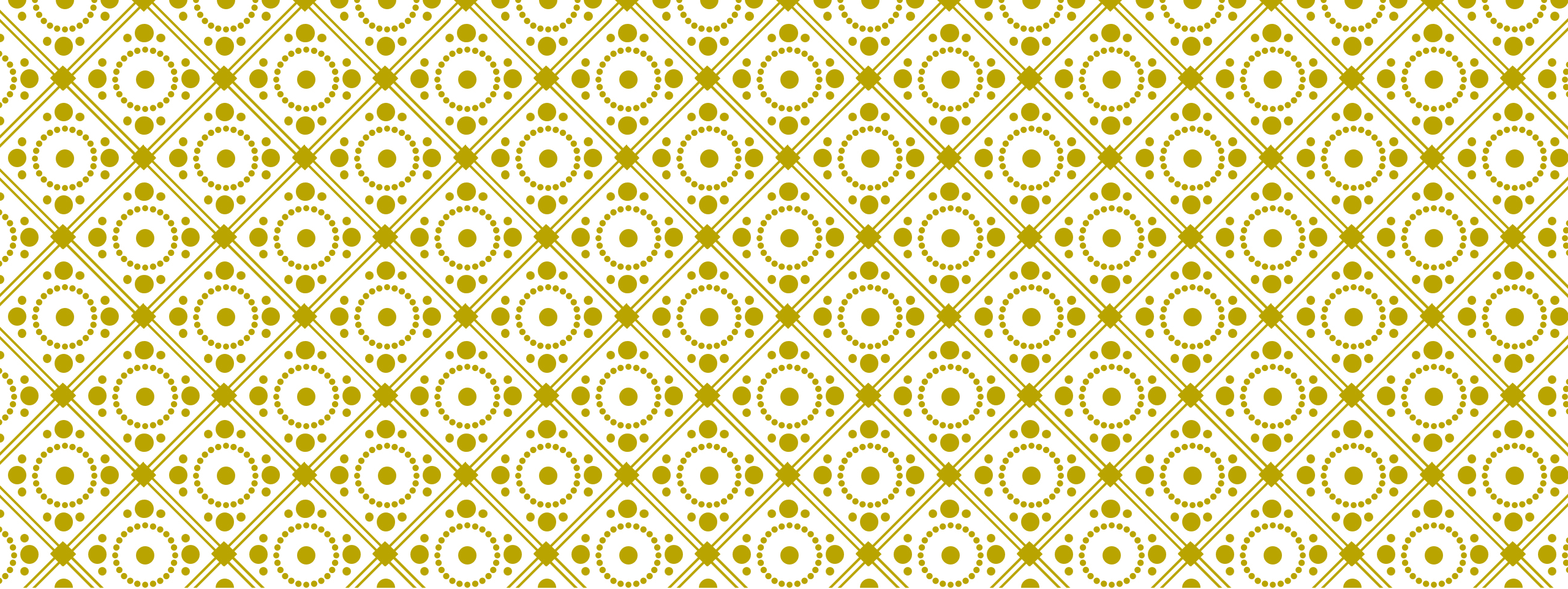
1.2. CORPORA E WWW

Il World Wide Web è oggi la più grande banca dati di testi esistenti facilmente accessibile.

Però

1. non contiene testi selezionati secondo criteri espliciti legati ad una ricerca linguistica
2. l'origine dei documenti non è sempre verificabile
3. aspetti legati alla volatilità dei testi sul web e alla loro qualità.

Alcuni strumenti importanti per la creazione di corpora “fai da te” che attingono ad informazioni dal web (*WebBootCat* e il *WebCorp* che ‘catturano’ dalla rete testi che si inseriscono in un corpus).



1. FONDAMENTI DI LINGUISTICA DEI CORPORA

1.3. Corpora e analisi linguistica

1.3. CORPORA E ANALISI LINGUISTICA

L'analisi si compone di due momenti distinti:

1. procedure di compilazione del corpus (che dipendono dagli obiettivi dell'analisi, dalle ipotesi, dai vincoli esterni) (sezione 1 e 2)
2. Utilizzo di strumenti informatici per l'interrogazione e l'utilizzo del corpus (sezione 3)

1.3. DISTRIBUZIONI DI FREQUENZA E APPROCCIO PROBABILISTICO

Il campionamento degli usi linguistici in corpora informatizzati **quantifica i fenomeni osservati.**

Un esempio: calcolo del numero di volte con cui un fenomeno linguistico si presenta in un campione scelto.

1.3. DISTRIBUZIONI DI FREQUENZA E APPROCCIO PROBABILISTICO

Frequenza delle parole che compongono un corpus (contando il numero di occorrenze di ciascuna parola). In generale: conteggio aritmetico del numero di elementi linguistici (tokens) che appartengono ad ogni classificazione (type).

Tokens: 6

The cat sat on the mat

Types: 5

The = 2

Cat = 1

Sat = 1

On = 1

Mat = 1

Per ogni parola diversa/nuova è indicato il numero di occorrenze in **valore assoluto (frequenza assoluta/raw frequency)**

1.3. DISTRIBUZIONI DI FREQUENZA E APPROCCIO PROBABILISTICO

La frequenza assoluta (ovvero il conteggio EFFETTIVO di occorrenze) è utile quando si usa UN solo corpus/sottocorpus.

MA

Se si devono CONFRONTARE corpora diversi (o segmenti di uno stesso corpus) con grandezze diverse, la frequenza assoluta deve essere **normalizzata**.

1.3. DISTRIBUZIONI DI FREQUENZA E APPROCCIO PROBABILISTICO

Frequenza assoluta vs frequenza relativa (permette di confrontare delle frequenze di parole tra di loro o dati tra campioni diversi)

Frequenza relativa (FR, *relative frequency*) di ciascun tipo:

Numero di occorrenze (di una parola)/numero totale delle parole del corpus

$$2/6 = 0,33 \text{ (eventualmente } * 100) \text{ (33\%)}$$

$$1/6 = 0,16 \text{ (eventualmente } * 100) \text{ (16\%)}$$

1.3. DISTRIBUZIONI DI FREQUENZA E APPROCCIO PROBABILISTICO

Ma cosa succede in corpora di maggiori dimensioni?

$$\begin{array}{l} 6.000.000 \quad \rightarrow \quad 2/6.000.000 \quad = \quad 0,0000003 \quad \rightarrow \\ 0,0000003 * 1.000.000 = 0,33 \text{ per milione} \end{array}$$

Normalizzazione: frequenza riferita a un numero fisso di parole

Il moltiplicatore \rightarrow base comune: 1.000.0000 (la frequenza che la parola avrebbe avuto se il corpus fosse stato composto da un numero totale di parole pari alla base comune)

1.3. DISTRIBUZIONI DI FREQUENZA E APPROCCIO PROBABILISTICO

La grandezza del corpus influisce sul **significato statistico** QUINDI la base comune per la normalizzazione deve essere **comparabile** alla grandezza del corpus.

Se confrontiamo la sezione orale del BNC (10 milioni di parole) e quella scritta (90 milioni di parole) la normalizzazione a 1000 parole è inappropriata.

I risultati ottenuti su basi comuni troppo grandi o troppo piccole sono distorti.

1.3. DISTRIBUZIONI DI FREQUENZA E APPROCCIO PROBABILISTICO

ESEMPIO

confronto tra uso del termine *fucker* in BNC scritto (90 milioni) e orale (10 milioni)

Scritto: 50 volte (frequenza assoluta)

Orale: 25 volte (frequenza assoluta)

Occorrenze **devono** essere messe in relazione con numerosità del campione

$50/90.000.000 = 0,55$ (frequenza relativa)

$25/10.000.000 = 2,5$ (frequenza relativa)

1.3. DISTRIBUZIONI DI FREQUENZA E APPROCCIO PROBABILISTICO

BNC quasi 100.000.000 parole

CONCORDANCE

British National Corpus (BNC)

Get more space [+](#) [↶](#) [?](#) [💬](#) [👤](#)

simple **cat** • 5,275
46.95 per million tokens • 0.0047%

[🔍](#) [📄](#) [☰](#) [↶](#) [👁](#) [📖](#) [✂](#) [≡](#) [≡](#) [📄](#) [📄](#) [⋮](#) [📄](#) **KWIC** [+](#) [👁](#) [☆](#)

Details

Left context KWIC Right context

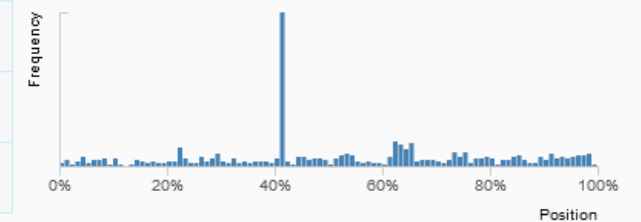
1	<input type="checkbox"/>	Written books a... ine library, interesting friends... and Joshua, the Principal's tabby	cat	. </s><s> ' Pamela Walford wrote: 'The revelation to me was beir
2	<input type="checkbox"/>	Written books a... versity Womens' Club (worth joining even if it is known as Pussy	Cat	Hall by my irreverent 88 year old cousin, a St. Hugh's gal) gave u
3	<input type="checkbox"/>	Written books a... he Scotsman 29 May </s><s> Transport </s><s> Platinum-free"	cats	" developed </s><s> The Japanese car company, Nissan, has ar
4	<input type="checkbox"/>	Written books a... ree catalytic converters. </s><s> Using only palladium, the new"	cats	" would be up to one-third cheaper, Nissan claim. </s><s> Financ
5	<input type="checkbox"/>	Written books a... ats, hunting, and the introduction of alien predators such as rats,	cats	, stoats and mongoose. </s><s> Among those in imminent dange
6	<input type="checkbox"/>	Written books a... gkok's Chatachuk market, with police seizing bab		
7	<input type="checkbox"/>	Written books a... allaby in the Australian outback have been set ba		
8	<input type="checkbox"/>	Written books a...) of the rufous hares survive outside captivity. </s>		
9	<input type="checkbox"/>	Written books a... rows (20) and birds of prey (36). </s><s> Some 6		

RESULT DETAILS

simple **cat**

Number of hits	5,275
Number of hits per million tokens	46.95
Percent of whole corpus	0.004695%
Corpus size (tokens)	112,345,722

Distribution of hits in the corpus



1.3. DISTRIBUZIONI DI FREQUENZA E APPROCCIO PROBABILISTICO

Rango	Tipi di parole	N. occorrenze	Frequenza (%)
1	The	2	33,3%
2	Cat	1	16,7%
3	Mat	1	16,7%
4	On	1	16,7%
5	Sat	1	16,7%

Righe: elenchi di tipi di parole presenti nel corpus

Colonne: numero di occorrenze di ciascun tipo espresse in **valore assoluto** o riferite alla **base comune**

NB:

- elenco dei tipi normalmente in ordine **decrescente** rispetto alla base comune (in alto valori di frequenza più alti, in basso *hapax legomena* – parole che occorrono una sola volta)
- A parità di frequenza relativa, ordinamento alfabetico

WORDLIST

Italian Web 2016 (itTenTen16) 🔍 ⓘ

Get more space + 🔗 ? ! 👤

Domain_it × **word** (2,042,948 items | 4,522,630,925 total frequency)

🔍 ⬇️ 👁️ ⓘ ☆

Word	Frequency ?	Word	Frequency ?	Word	Frequency ?	Word	Frequency ?
1 di	186,127,403 ...	14 della	38,669,608 ...	27 delle	17,333,211 ...	40 questo	10,634,568 ...
2 e	133,863,386 ...	15 i	38,624,004 ...	28 alla	17,267,760 ...	41 nella	10,431,362 ...
3 il	88,897,560 ...	16 non	37,023,409 ...	29 dell'	17,257,980 ...	42 all'	10,076,131 ...
4 la	88,216,145 ...	17 le	36,752,708 ...	30 come	17,218,577 ...	43 essere	8,410,475 ...
5 che	79,077,438 ...	18 una	35,493,650 ...	31 ma	16,270,900 ...	44 su	8,155,406 ...
6 in	70,727,655 ...	19 si	33,448,613 ...	32 anche	16,168,676 ...	45 cui	8,076,029 ...
7 a	65,405,934 ...	20 da	31,114,319 ...	33 o	14,190,742 ...	46 alle	7,970,176 ...
8 per	62,535,452 ...	21 al	25,258,772 ...	34 gli	13,720,854 ...	47 tra	7,880,725 ...
9 un	51,257,588 ...	22 dei	21,707,593 ...	35 ed	11,461,093 ...	48 ci	7,838,432 ...
10 del	50,356,523 ...	23 sono	19,622,661 ...	36 se	11,415,311 ...	49 ai	7,641,081 ...
11 è	49,064,717 ...	24 nel	18,799,876 ...	37 dal	11,141,255 ...	50 degli	7,196,515 ...
12 l'	40,533,641 ...	25 più	17,389,349 ...	38 ad	11,103,852 ...		

5 nuove notifiche

Domain_it x **word** (2,042,948 items | 4,522,630,925 total frequency)

Word	Frequency ?	Word	Frequency ?	Word	Frequency ?	Word	Frequency ?
51 dalla	7,189,517 ...	64 prima	5,666,016 ...	77 può	4,847,507 ...	90 dall'	3,952,074 ...
52 tutti	6,840,701 ...	65 sua	5,591,637 ...	78 nei	4,777,354 ...	91 nelle	3,929,264 ...
53 sul	6,817,722 ...	66 sia	5,562,131 ...	79 c	4,755,955 ...	92 fare	3,920,492 ...
54 solo	6,727,036 ...	67 sempre	5,484,211 ...	80 ogni	4,741,419 ...	93 fatto	3,902,903 ...
55 mi	6,683,237 ...	68 un'	5,457,715 ...	81 poi	4,656,099 ...	94 vita	3,799,984 ...
56 d'	6,631,605 ...	69 tutto	5,366,913 ...	82 quando	4,615,680 ...	95 quanto	3,793,517 ...
57 hanno	6,332,522 ...	70 ho	5,167,723 ...	83 nell'	4,421,250 ...	96 proprio	3,763,946 ...
58 due	6,187,797 ...	71 era	5,151,504 ...	84 senza	4,342,846 ...	97 lavoro	3,759,760 ...
59 questa	6,183,073 ...	72 uno	5,097,732 ...	85 perché	4,326,891 ...	98 altri	3,735,270 ...
60 anni	6,175,291 ...	73 suo	5,096,808 ...	86 così	4,283,942 ...	99 chi	3,727,908 ...
61 stato	6,050,961 ...	74 dopo	5,024,891 ...	87 ancora	4,235,376 ...	100 già	3,585,612 ...
62 loro	6,007,077 ...	75 sulla	4,930,443 ...	88 quello	4,180,207 ...		
63 parte	5,698,504 ...	76 molto	4,880,163 ...	89 tempo	4,154,104 ...		

You are only allowed to access 1,000 items. [Get more](#)

Rows per page: 50 51-100 of 1,012 < < 2 /21 > >

WORDLIST

Italian Web 2016 (itTenTen16)

Domain_it x word (2,042,948 items | 4,522,630,925 total frequency)

	Word	Frequency ?
1,001	cerca	476,783 ...
1,002	voglia	476,768 ...
1,003	consente	475,856 ...
1,004	sugli	475,127 ...
1,005	epoca	473,250 ...
1,006	sostegno	473,036 ...
1,007	documenti	472,992 ...
1,008	pertanto	472,352 ...
1,009	confronto	471,541 ...
1,010	indirizzo	470,859 ...

	Word	Frequency ?
1,011	dobbiamo	470,838 ...
1,012	legno	470,415 ...

You are only allowed to access 1,000 items. [Get more](#)

Rows per page: 50 1,001–1,012 of 1,012 < > 21 / 21 > >|

1.3. DISTRIBUZIONI DI FREQUENZA E APPROCCIO PROBABILISTICO

A seconda dello scopo, le parole grammaticali che compaiono in cima si possono escludere dal computo compilando una **stoplist** in modo da vedere subito le parole lessicali più frequenti (anche in base al contenuto del corpus → Sketch Engine).

Vediamo un esempio

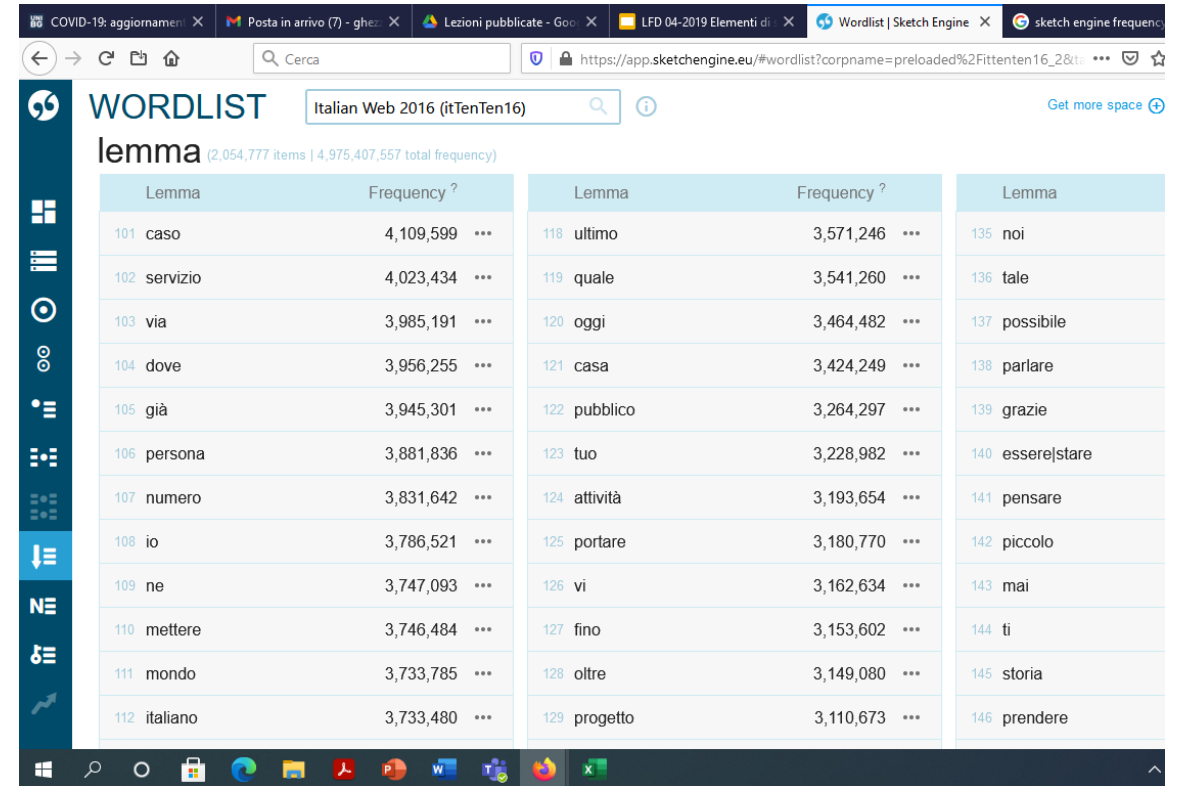
1.3. DISTRIBUZIONI DI FREQUENZA E APPROCCIO PROBABILISTICO

L'analisi basata sulla frequenza d'uso di espressioni (ad es. lessicografia o didattica delle lingue) incentrata su fasce di frequenza (*frequency bands*)

1. fascia alta: poche parole a frequenza massima.
2. fascia media: a partire dalla prima coppia di parole con la stessa frequenza.
3. fascia bassa: le parole a frequenza bassa e gli hapax.

1.3. DISTRIBUZIONI DI FREQUENZA E APPROCCIO PROBABILISTICO

Liste di frequenza lemmatizzate in cui vengono ridotte le forme flesse di una parola ad un unico lemma.



The screenshot shows the Sketch Engine Wordlist interface. The browser address bar displays the URL: https://app.sketchengine.eu/#wordlist?corpname=preloaded%2Fitten16_2&t.... The page title is "WORDLIST" and the selected corpus is "Italian Web 2016 (itTenTen16)". The main content area shows a list of lemmas with their frequencies, sorted in descending order. The table is divided into three columns, with the first column containing items 101 through 112, the second column containing items 118 through 129, and the third column containing items 135 through 146. Each row lists a lemma and its frequency, with a small icon indicating that the frequency is truncated.

Lemma	Frequency ?	Lemma	Frequency ?	Lemma
101 caso	4,109,599 ...	118 ultimo	3,571,246 ...	135 noi
102 servizio	4,023,434 ...	119 quale	3,541,260 ...	136 tale
103 via	3,985,191 ...	120 oggi	3,464,482 ...	137 possibile
104 dove	3,956,255 ...	121 casa	3,424,249 ...	138 parlare
105 già	3,945,301 ...	122 pubblico	3,264,297 ...	139 grazie
106 persona	3,881,836 ...	123 tuo	3,228,982 ...	140 essere stare
107 numero	3,831,642 ...	124 attività	3,193,654 ...	141 pensare
108 io	3,786,521 ...	125 portare	3,180,770 ...	142 piccolo
109 ne	3,747,093 ...	126 vi	3,162,634 ...	143 mai
110 mettere	3,746,484 ...	127 fino	3,153,602 ...	144 ti
111 mondo	3,733,785 ...	128 oltre	3,149,080 ...	145 storia
112 italiano	3,733,480 ...	129 progetto	3,110,673 ...	146 prendere

1.3. DISTRIBUZIONI DI FREQUENZA E APPROCCIO PROBABILISTICO

Liste di frequenza per qualsiasi altro aspetto linguistico (se è stato annotato!) ad es. per POS (parte del discorso).

The screenshot shows a web interface with three tabs: BASIC, ADVANCED, and ABOUT. The ADVANCED tab is selected. On the left, there is a 'find ?' section with a dropdown menu for 'lemmas' containing: nouns, verbs, adjectives, adverbs, pronouns, conjunctions, and prepositions. A secondary dropdown menu is open, showing options: all, starting with, ending with, containing, matching regex, and from this list. On the right, there is a section for 'Exclude these words:' with a checked checkbox and a text area for pasting a list. Below this, there is an unchecked checkbox for 'Include nonwords?'. A checked checkbox for 'A = a?' is also present. Underneath, there are two input fields for 'Frequency min?' (set to 5) and 'Frequency max?' (set to 0). At the bottom, there is a 'result format' section with two radio buttons: 'simple list?' (selected) and 'display as?'. A red circular 'GO' button is located in the bottom right corner of the interface.

1.3. DISTRIBUZIONI DI FREQUENZA: EQUAZIONE DI ZIPF

George Zipf è stato tra i primi a lavorare sulle distribuzioni di frequenza, a partire dall'*Ulisse* di James Joyce. In particolare osserva il rapporto tra

1. la frequenza di una parola
2. la sua posizione in un ordinamento decrescente di frequenze.

Più in generale parleremo dell'equilibrio tra parole nuove, **tipi**, e le loro ripetizioni, **token**.

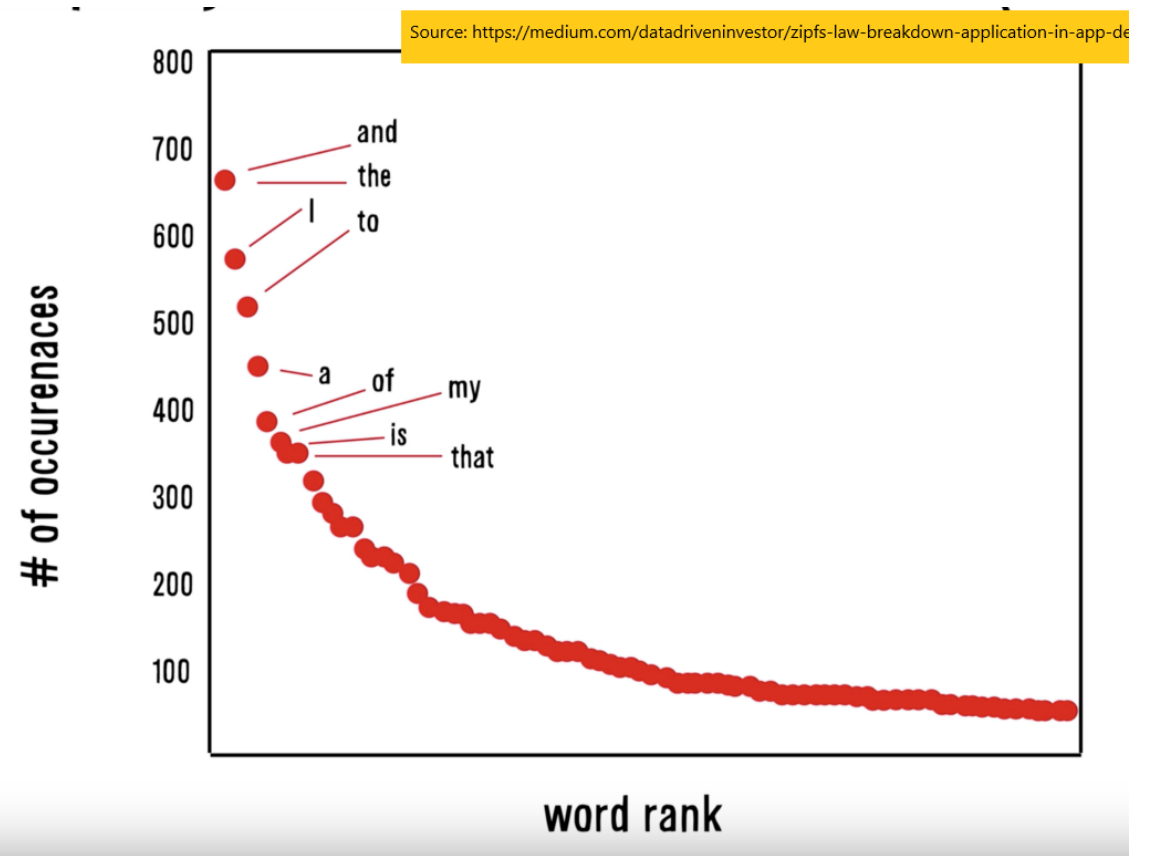
$$r \times f = C$$

r è il rango/la posizione, f la frequenza, C il loro prodotto.

Con questa equazione Zipf intuisce l'equilibrio caratteristico del vocabolario tra **novità** (tipi) e **ripetizione** (token) (Vocabulary balance)

1.3. DISTRIBUZIONI DI FREQUENZE E ZIPF

Frequenza e rango in *Romeo and Juliet*



1.3. DISTRIBUZIONI DI FREQUENZE

Fr. Assoluta (token)	Numero di tipi che hanno una certa frequenza assoluta	Conteggio dei tipi	Tipi (%)	Conteggio dei token	Token (%)
1	85	85	75	85	45
2	15	100	88	115	61
3	4	104	92	127	67
4	1	105	93	131	69
5	3	108	95	146	77
6	1	109	96	152	80
8	2	111	98	168	89
10	1	112	99	178	94
11	1	113	100	189	100

1.3. DISTRIBUZIONI DI FREQUENZE: INDICE DI GUIROT

Rapporto tra **tipi** e **token** > **più è alta la frequenza di ogni singola occorrenza** più lentamente cresce un vocabolario.

Sinclair (2004) > implicazioni rilevanti di questo rapporto per apprendenti di L2/LS che incontrano molte parole nuove con scarsa possibilità di ripetizione.

> Rapporto tra tipi e token (***Type Token Ratio – TTR***, anche **indice di Guirot**) come una possibile misura della varietà lessicale di un testo.

Risultato di questo rapporto = 1 > ogni parola usata è una parola diversa.

> Più TTR si avvicina ad 1 più ampio è il vocabolario usato.

MA sappiamo che all'aumentare dei token rallenta la crescita del vocabolario e quindi la TTR appare poco significativa perché sensibile alla numerosità del campione.

1.3. DISTRIBUZIONI DI FREQUENZE: INDICE DI GUIROT (2)

TTR standardizzata: per esplorare la variazione stilistica rispetto alla ricchezza lessicale o ampiezza del vocabolario (utile quando si comparano scrittori o generi diversi).

Diversi modi per calcolarla, spesso dipendenti dal software.

Ad es. corpus specialistico con capitoli introduttivi di 10 manuali di linguistica (250.000 parole) > TTR standardizzate con pochissima variabilità (tipico dei linguaggi settoriali).

Aitchinson	Akmajian	Brown	Crystal	Lyons
42,20	39,10	36,38	40,67	37,21
Radford	Robins	Wallwark	Widdowson	Yule
37,33	41,07	42,96	37,12	40,28

1.3. DISTRIBUZIONI DI FREQUENZE

1. I dati vanno analizzati in termini di **significatività** > interpretazione probabilistica delle lingue > pattern d'uso più o meno frequente entro la naturale variabilità osservabile in campioni di lingua).
2. approccio probabilistico allo studio del linguaggio che riconosce fenomeni di gradienza e di non categoricità (anche per i giudizi di grammaticalità).
3. Enorme impatto sulla ricerca empirica: ad es. lo studio del mutamento linguistico (**linguistica storica**), della produzione e comprensione delle lingue (**psicolinguistica**), dell'acquisizione della propria lingua, dell'apprendimento delle lingue seconde, della descrizione grammaticale e della lessicografia.

1.3. CONFRONTO TRA CORPORA: LE PAROLE CHIAVE

Parola chiave anche *key-words* - **Mike Scott**: caratterizzano un corpus e ne rappresentano la deviazione rispetto a una norma presa come riferimento.

Quindi, dati due corpora, interessa verificare se

1. la distanza tra di essi, misurata in differenza tra le frequenze osservate, è **significativa**;
2. rappresentano campioni **casuali** della stessa popolazione o se le differenze osservate **ci dicono qualcosa** della naturale variazione del linguaggio.

1.3. LETTURA DI CONCORDANZE E COLLOCATI

Un ulteriore strumento rilevante nella linguistica dei corpora è relativo a

concordanze: elenco di tutte le occorrenze di una **parola (nodo)** nell'ambiente che la circonda (**cotesto - orizzonte**). Normalmente associata alla modalità di visualizzazione KWIC (*Key Word In Context*)

Il nodo viene allineato al centro a distanza fissa dalla porzione di testo che lo precede e che lo segue.

	<input type="checkbox"/> Details	Left context	KWIC	Right context	
1	<input type="checkbox"/> cinqueterre.it	è di Volastra, come ricorda la lapide sulla facciata. </s><s> La	pianta	è a tre navate, mentre la facciata è impreziosita da un rosone	
2	<input type="checkbox"/> grandespirito.i...	con il viso rivolto a ovest, e chiede aiuto. </s><s> Parla con le	piante	, ed esse rispondono. </s><s> Ascolta con attenzione le voci d	
3	<input type="checkbox"/> unito.it	l'inizio dei Settanta ricostruirono il cinema americano di sana	pianta	, unendo cinefilia e senso del botteghino e quasi tutti provenier	
4	<input type="checkbox"/> cottodeste.it	ue prestigiosi edifici, su cui crescono più di 1.000 esemplari di	piante	, si integrano alla perfezione nel verde che li circonda. </s><s>	
5	<input type="checkbox"/> tibursuperbum.i...	:ui si è sviluppato il nucleo abitato. </s><s> Oggi si presenta a	pianta	rettangolare, protetto da due torri quadrate e caratterizzato da	
6	<input type="checkbox"/> pinocchio.it	per lo sviluppo ed il coordinamento delle attivit ... </s><s> Le	piante	per clima mediterraneo sono state presentate come una scelta	
7	<input type="checkbox"/> pinocchio.it	entate come una scelta ideale per i giardini di oggi ... </s><s>	Piante	preistoriche nel giardino di domani è la proposta del secondo v	

1.3. LETTURA DI CONCORDANZE E COLLOCATI

L'estensione del contesto, di solito misurata in caratteri, può variare a seconda del tipo di osservazione.

La concordanza può essere osservata in modi diversi. La concordanza può essere ordinata in vari modi, secondo l'ordine di occorrenza nel corpus, alfabeticamente a destra o a sinistra, ecc.

1.3. LETTURA DI CONCORDANZE E COLLOCATI

Importante sviluppo è la *teoria della collocazione* (Sinclair).
Obiettivo: mostrare che il **significato** di una parola è in parte derivabile dal **suo contesto d'uso**

1. contesto verbale immediatamente precedente e successivo (in orizzontale, espressione dell'asse sintagmatico).
2. contesto più astratto derivante dalla ripetizione di tale contesto in un elenco di concordanze (in verticale, asse paradigmatico).

Queste due dimensioni possono rispondere di usi e restrizioni situazionali (contesto).

1.3. LETTURA DI CONCORDANZE E COLLOCATI

Contesto e *cotesto* possono servire a disambiguare sensi di parole molto comuni e polisemiche.

Il corpus, letto tramite concordanze, diventa il modo di scardinare l'idea radicata secondo cui l'unità di senso è la singola parola.

Nuova concezione del significato: ***meaning shift unit*** o **MSU (unità di passaggio di senso)** estensione della nozione di collocazione a comprendere la co-selezione di più elementi concomitanti e la ripetizione lungo l'asse delle scelte possibili.

1.3. LETTURA DI CONCORDANZE E COLLOCATI

Ruolo di primo piano nell'individuazione dei significati è giocato dalla frequenza con cui certe combinazioni occorrono in un corpus.

La **ripetitività** di una scelta fa sì che **l'uso** (*la parole*) tenda al **sistema** (*langue*).

La frequenza di un fenomeno non è casuale e anche se non è condizione sufficiente all'individuazione di sensi di una parola è una condizione necessaria.

WORD SKETCH

Italian parliamentary debates (ParlaMint 2.1)

social as adjective 580x

modifiers of "social"
green green social
via via social e
proprio proprio social
molto molto social

nouns modified by "social"
network sui social network
card social card
media sui social media
compact un social compact
APE l' APE social , la pensione
housing il social housing
canale i canali social
impact social impact investing
maggiore che gestiscono i maggiori social network , che
Ape Ape social

prepositional phrases with nouns
"social" per
"social" a
"social" da

verbs before "social"
denominare denominato social
determinare determinati social
chiamare chiamate social
definire definita molto social

"social" and/or ...
chiuso social , subito chiuso



1.3. LETTURA DI CONCORDANZE E COLLOCATI

Ricorrenze **lessico-grammaticali significative**

- (1) fanno emergere pattern strutturali e semantici
- (2) evidenziano la natura **fraseologica** del linguaggio.

La lettura delle concordanze, in corpora molto ampi,

- a) mostra come la ricorrenza di pochi elementi è sufficiente a disambiguare i significati più frequenti.
- b) segnala inoltre come significati meno frequenti siano spesso associati a restrizioni di tipo situazionale o dipendenti dal contesto.

1.3. LETTURA DI CONCORDANZE E COLLOCATI

Frequenza di alcune combinazioni gioca ruolo di primo piano nell'individuazione di significati.

Nel vocabolario sono molte poche le scelte libere non condizionate da fattori contestuali e cotestuali.

Collocazione > teoria del significato analizzabile in queste componenti:

1. Collocazione («collocation»): rapporto con altre parole (lessicali), combinazione (o co-occorrenza) di due o più parole che tendono a presentarsi insieme (ad es. *bandire un concorso, amara sorpresa*)
2. Colligazione («colligation»): rapporto con categorie di parole funzionali, funzioni grammaticali, le classi o parti del discorso a cui appartengono i collocati
3. preferenza semantica («semantic preference»): rapporto con parole appartenenti ad uno stesso campo semantico, insieme a cui sono riconducibili le diverse lessicalizzazioni dei collocati
4. prosodia semantica («semantic prosod»): rapporto con parole che esprimono intensione o atteggiamento di chi parla/scrive, aura di significato di cui si colora una parola in ragione dei suoi collocati

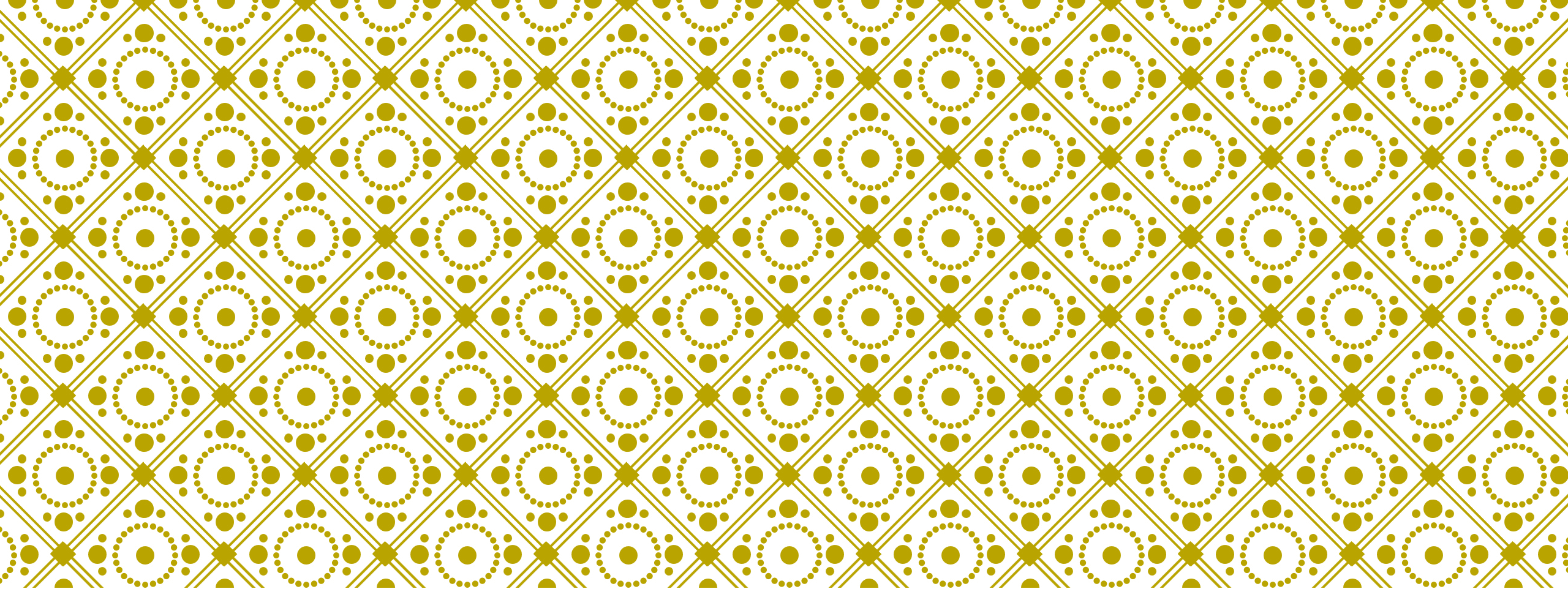
1.3 LETTURA DI CONCORDANZE E COLLOCATI

T-score	Mutual information
Last	Saturday
At	Wednesday
the	tomorrow
That	Friday
one	Tuesday
Tomorrow	last
Friday	Monday
All	Sunday
sunday	previous

1.3 LETTURA DI CONCORDANZE E COLLOCATI

unstoppable, torrent, ferocious, fury, wrath, mayhem, fierce, creativity, devastating, deadly, entrepreneurial, artillery, terror, awesome, inner, shot, tremendous, storm, hell, wave, powerful, creative, potential, flood, debut, upon, power, talent, destruction, evil, onto, massive, violence, superb, force, nuclear, weapon, true, dog, full, war, attack, energy, against, his, their, new, your, city, another, its, low

Tab. 2 I primi 50 collocati di *unleash* (*Mutual Information*, posizione +1 e +2).



1. FONDAMENTI DI LINGUISTICA DEI CORPORA

1.4 Fare analisi sui corpora

1.4 ESEGUIRE UNA QUERY

Query: analisi/ricerca tramite motore di ricerca (di un corpus o di un database). Letteralmente, una “richiesta” fatta al motore di ricerca di cercare (nel corpus o nel database) tutti i dati che rispettano i **vincoli di ricerca** inseriti (una specifica sequenza di caratteri, o anche una combinazione di più parametri, ad es. tutte le occorrenze di “lo” etichettato come articolo).

Come fare: inserire del testo in uno dei **campi di ricerca** (una o anche più parole) > i risultati dipendono dal campo utilizzato:

- **Simple query:** cerca la sequenza di caratteri ovunque nei dati
- **Lemma:** permette di cercare tutte le forme di un lemma
- **Word form:** cerca una specifica forma di parola
- **Character:** cerca una sequenza di caratteri, anche all’interno della parola
- **CQL:** permette di fare ricerche complesse, combinando più vincoli, tramite il **Corpus Query Language (CQL)**

Query **baby, killer** 129 (0.34 per million)

Page of 7 [Next](#) | [Last](#)

#38001580 DETROIT , CAPITALE DELL' AUTO È NELLE MANI DEI **BABY KILLER** . NEW YORK Si muore giovani nelle strade
#55632800 voluto dare il suo nome di battesimo . Il **baby killer** è già stato ascoltato dal magistrato ,
#79366822 significativo titolo : Dai muschilli ai **baby killer** . I minorenni denunciati sono passati dai
#94511219 anni , che dodici anni fa fu uno dei primi **baby killer** . Andò nel cortile del tribunale con due
#96197270 quarto film . Dal suo esordio con me , in Cry **Baby Killer** , è nata una bella amicizia . Quando abbi
#101560740 , Nicholson aveva avuto una parte in Cry **Baby Killer** , uno psycho-film che aveva per protagoni
#102624613 Catania . Conosciuto anche come Nuccio ' u **baby killer** per aver accompagnato il padre Francesco
#112056865 agghiaccianti dati sulla delinquenza minorile , sui **baby killer** . A tutti i ministri che incontravo chiedevo
#124808444 bloccato dai carabinieri Sulle mani del **baby killer** tracce di polvere da sparo . Secondo gli
#124808885 La sua cattura non è stata semplice . Il **baby killer** era stato bloccato con altre persone sospet
#124875519 di forza per il clan che ha prodotto un **baby killer** . Un potere del quale il ragazzino Michele
#125093838 del clan rivale . NAPOLI , SCARCERATO IL **BABY KILLER** . NAPOLI I killer siete voi , dei giornali
#125094457 Ed eccola , la verità di Michele Papi , **baby killer** per quattro giorni . Accanto a lui le due
#125095384 mattina era ritenuto un pericolosissimo **baby killer** , lascia subito spazio a una reazione viole
#125095848 Sbordone per spiegare la scarcerazione del **baby killer** . Il fermo è stato convalidato ma il giudic
#125096230 del gruppo Uno . Poi affronta il caso del **baby killer** : Alla luce della mia esperienza da investi
#125202128 Papi , ritenuto sino a sabato mattina il **baby killer** della strage di Casoria . A due giorni
#126791905 stesso Geri in stato di choc . " PER NOI È UN **BABY KILLER** DEVE RITORNARE IN CARCERE " . NAPOLI II
#126791934 sufficientemente gravi per ritenerlo il **baby killer** della strage di Casoria nella quale cadde
#128388655 . Congelata l' ipotesi di scaricare sui **baby killer** la tragedia della delinquenza minorile

Page of 7 [Next](#) | [Last](#)

Un esempio di query: “baby killer”

1.4 CONCORDANZE E KEYWORD-IN-CONTEXT (KWIC)

I risultati di una query vengono visualizzati in una lista, di solito nella **visualizzazione Keyword-in-context** (abbreviato: **KWIC**).

Nella visualizzazione KWIC, le **concordanze** (le righe dei risultati, “concordanti”, perché riproducono la parola o la sequenza di parole cercata) sono incolonnate in modo da avere la sequenza di testo cercata (la **keyword**) al centro, contornata da porzioni di testo immediatamente precedenti (**contesto sinistro**) e seguenti (**contesto destro**).

Query **baby, killer** 129 (0.34 per million)

Page of 7 [Next](#) | [Last](#)

#38001580	DETROIT , CAPITALE DELL' AUTO È NELLE MANI DEI BABY KILLER . NEW YORK Si muore giovani nelle strade
#55632800	voluta dare il suo nome di battesimo . Il baby killer è già stato ascoltato dal magistrato ,
#79366822	significativo titolo : Dai muschilli ai baby killer . I minorenni denunciati sono passati dai
#94511219	anni , che dodici anni fa fu uno dei primi baby killer . Andò nel cortile del tribunale con due
#96197270	quarto film . Dal suo esordio con me , in Cry Baby Killer , è nata una bella amicizia . Quando abbiamo
#101560740	, Nicholson aveva avuto una parte in Cry Baby Killer , uno psycho-film che aveva per protagonisti
#102624613	Catania . Conosciuto anche come Nuccio ' u baby killer per aver accompagnato il padre Francesco

Query **baby, killer** 129 (0.34 per million) → num. risultati

Page 1 of 7 Go [Next](#) | [Last](#) concordanza

#38001580	DETROIT , CAPITALE DELL' AUTO È NELLE MANI DEI	BABY KILLER	NEW YORK Si muore giovani nelle strade
#55632800	voluto dare il suo nome di battesimo. Il	baby killer	è già stato ascoltato dal magistrato,
#79366822	significativo titolo : Dai muschilli ai	baby killer	. I minorenni denunciati sono passati dai
#94511219	anni , che dodici anni fa fu uno dei primi	baby killer	. Andò nel cortile del tribunale con due
#96197270	quarto film . Dal suo esordio con me , in Cry	Baby Killer	, è nata una bella amicizia . Quando abbiamo
#101560740	, Nicholson aveva avuto una parte in Cry	Baby Killer	, uno psycho-film che aveva per protagonisti
#102624613	Catania . Conosciuto anche come Nuccio ' u	baby killer	per aver accompagnato il padre Francesco
#112056865	ragghiaccianti dati sulla delinquenza minorile , sui	baby killer	. A tutti i ministri che incontravo chiedevo
#124808444	bloccato dai carabinieri Sulle mani del	baby killer	tracce di polvere da sparo . Secondo gli
#124808885	La sua cattura non è stata semplice . Il	baby killer	era stato bloccato con altre persone sospettate
#124875519	di forza per il clan che ha prodotto un	baby killer	. Un potere del quale il ragazzino Michele
#125093838	del clan rivale . NAPOLI , SCARCERATO IL	BABY KILLER	. NAPOLI I killer siete voi , dei giornali
#125094457	Ed eccola , la verità di Michele Papi ,	baby killer	per quattro giorni . Accanto a lui le due
#125095384	mattina era ritenuto un pericolosissimo	baby killer	, lascia subito spazio a una reazione violenta
#125095848	Sbordone per spiegare la scarcerazione del	baby killer	. Il fermo è stato convalidato ma il giudice
#125096230	del gruppo Uno . Poi affronta il caso del	baby killer	: Alla luce della mia esperienza da investigatore
#125202128	Papi , ritenuto sino a sabato mattina il	baby killer	della strage di Casoria . A due giorni

contesto sinistro keyword contesto destro

Concordanze e Keyword-in-context (KWIC)

1.4 CONTESTO / CO-TESTO

Attenzione! Quello che qui viene chiamato **contesto (sinistro o destro)** dovrebbe essere definito **co-testo** (più specifico e appropriato).

Contesto: l'insieme degli elementi testuali compresenti con la porzione di messaggio che stiamo analizzando.

co-tèsto (o cotèsto) s. m. [comp. di *co-1* e *testo3*]. – In linguistica testuale, l'insieme degli elementi intrinsecamente testuali (detti anche *intra-testuali*), come per es. le frasi, gli elementi costitutivi di esse, ecc., le cui relazioni compongono un testo; in questo senso il *co-testo* si contrappone al *contesto*, che si riferisce agli elementi *extra-testuali*, cioè non facenti parte del testo, ma che ne influenzano la produzione e la ricezione, come per es. la situazione comunicativa.

Treccani

<http://www.treccani.it/vocabolario/co-testo/>

1.4. METADATI IN (NO) SKETCH ENGINE

Cliccando sul codice identificativo al margine sinistro della concordanza (il numero in blu), si apre un box giallo con i **metadati** relativi al testo (in basso nella finestra):

- Autore del testo
- Genere testuale
- Sezione del quotidiano
- Anno
- Titolo dell'articolo
- Conteggio parole

```
article.id      80486
article.author  P S
article.gen     commento
article.top     cronaca
article.year    1987
article.title   <subtitle> Ha ucciso per vendicare suo
article.wordcount 233
```


article.id 80486
article.author P S
article.gen commento
article.top cronaca
article.year 1987
article.title <subtitle> Ha ucciso per vendicare suo padre e suo fratello. </subtitle>
article.wordcount 233

Cliccando sul codice identificativo in blu (#12...) si apre un box in con i metadati.

Cliccando sulla concordanza, compare un box con una porzione più ampia del testo.

#94511219 anni , che dodici anni fa fu uno dei primi baby killer . Andò nel cortile del tribunale con due
#96197270 quarto film . Dal suo esordio con me , in Cry Baby Killer , è nata una bella amicizia . Quando abbiamo
#101560740 , Nicholson aveva avuto una parte in Cry Baby Killer , uno psycho-film che aveva per protagonisti
#102624613 Catania . Conosciuto anche come Nuccio ' u baby killer per aver accompagnato il padre Francesco
#112056865 agghiaccianti dati sulla delinquenza minorile , sui baby killer . A tutti i ministri che incontravo chiedevo
#124808444 bloccato dai carabinieri Sulle mani del baby killer tracce di polvere da sparo . Secondo gli
#124808885 La sua cattura non è stata semplice . Il baby killer era stato bloccato con altre persone sospettate
#124875519 di forza per il clan che ha prodotto un baby killer . Un potere del quale il ragazzino Michele
#125093838 del clan rivale . NAPOLI , SCARCARATO I BABY KILLER . NAPOLI I killer siete voi , dei giornali
#125094457 Ed eccola , la verità di Michele Papi , baby killer per quattro giorni . Accanto a lui le due
#125095384 mattina era ritenuto un pericolosissimo baby killer , lascia subito spazio a una reazione violenta
#125095848 Sbordone per spiegare la scarcerazione del baby killer . Il fermo è stato convalidato ma il giudice
#125096230 del gruppo Uno . Poi affronta il caso del baby killer : Alla luce della mia esperienza da investigatore

Metadati e “contesto” (= co-testo immediato)

1.4 RICERCA DI COLLOCATI

Le **collocazioni** sono

- ◆ sequenze di **tokens** (i.e. parole) che co-occorrono in un corpus.
- ◆ combinazioni di n elementi (**n-grams**) che mostrano un grado di **solidarietà semantica**, ovvero la combinazione risulta sensibilmente più frequente e lessicalmente appropriata rispetto ad altre combinazioni possibili tra uno degli elementi della collocazione e altri possibili sostituti (es. *condurre un'indagine* vs. *fare un'indagine*).

Esistono diversi tipi e gradi di solidarietà semantica (collocazioni, locuzioni, polirematiche etc.).

Collocation candidates ?

Attribute: lemma In the range from: -1 to: 1

Minimum frequency in corpus: 5

Minimum frequency in given range: 3

Show functions: T-score MI MI3 log likelihood min. sensitivity logDice

Sort by: T-score MI MI3 log likelihood min. sensitivity logDice

Make candidate list

1.4 RICERCA DI COLLOCATI

Per individuare le possibili collocazioni di un elemento

- (1) si esegue prima una query normale con un'unità di riferimento;
- (2) si filtrano i risultati tramite una **query “Collocations”** (voce in basso nel menù di sinistra di NoSketchEngine).

La query per le collocations richiede di indicare:

- il **tipo di unità** co-occorrenti con la keyword (**attribute**)
- il **range** in cui cercare unità ricorrenti (ovvero, quante parole prima o dopo la keyword occorre scansionare)

102

Collocation candidates ?

Attribute: lemma in the range from: -1 to: 1

Minimum frequency in corpus: 5

Minimum frequency in given range: 3

Show functions: T-score, MI, MI3, log likelihood, min. sensitivity, logDice

Sort by: T-score, MI, MI3, log likelihood, min. sensitivity, logDice

Make candidate list

1.4 LOG-DICE

Log-dice: misura statistica che individua collocati in termini di 'tipicalità' (in Sketch Engine nella sezione WordSketch, in noSketch Engine nella sezione Collocations)

Basato sulla frequenza del nodo (parola) e del collocato e sulla frequenza della collocazione (nodo + collocato). Non dipende dalla grandezza del corpus e quindi può essere usato con corpora di diversa grandezza.

bedroom

noun 921,752x



modifiers of "bedroom"	nouns modified by "bedroom"	verbs with "bedroom" as object	verbs with "bedroom" as subject
master ... the master bedroom	apartment ... bedroom apartment	furnish ...	decorate ... bedroom decorating ideas
double ... double bedroom	furniture ... bedroom furniture	decorate ...	overlook ... bedroom overlooking
spacious ... spacious bedrooms	villa ... bedroom villa	boast ...	sleep ... bedrooms sleeping
spare ... a spare bedroom	suite ... bedroom suite	carpet ... carpeted bedrooms	redecorate ...
en-suite ... en-suite bedrooms	bathroom ... bedroom , bathroom	comprise ...	feature ... bedroom features a
upstairs ... upstairs bedroom	condo ... bedroom condo	feature ...	boast ... bedroom boasts a
twin ... twin bedroom	door ... the bedroom door	size ... sized bedrooms	terrace ... bedroom terraced house
guest ... guest bedroom	closet ... bedroom closet	condition ... air conditioned bedrooms	share ... bedrooms share a
		equip ... bedrooms are equipped	pillow ... bedroom pillows

1.4 MI-SCORE E T-SCORE

MI (Mutual Information)-score: la misura con cui le parole **ricorrono contemporaneamente** rispetto al numero di volte in cui appaiono separatamente.

Fortemente influenzato dalla frequenza, le parole a bassa frequenza tendono a raggiungere un punteggio MI alto che può essere fuorviante (Sketch Engine consente di impostare un limite di frequenza in modo che le parole a bassa frequenza possano essere escluse dal calcolo).

T-Score: la certezza con cui si può sostenere che esiste un'associazione tra le parole, ovvero la loro co-occorrenza non è casuale. Il valore è influenzato dalla frequenza dell'intera collocazione, motivo per cui combinazioni di parole molto frequenti tendono a raggiungere un punteggio T elevato nonostante non siano collocazioni significative.

Nella maggior parte dei casi, il punteggio T è più affidabile o più utile del punteggio MI.

1.4 COLLOCATI CANDIDATI

La maschera di ricerca per le collocations usa la denominazione “**Collocation candidates**”, perché la lista risultante può solo indicare la presenza di elementi co-occorrenti con frequenza significativa, ma non può - da sola - individuare reali rapporti di solidarietà semantica, che pertengono all’analisi dell’osservatore.

Qui si osservi *sire* (lemmatizzazione errata di *sitter* come forma di un verbo inesistente), *Bells*, *Jane* e *Achtung* (si tratta di nomi propri: *Baby Bells*, *Baby Jane*, *Achtung Baby*).

Concordance
Word list
Corpus info
My jobs



Home
User guide

Save
< concordance
Sample
Filter

Overlaps
1st hit in doc

Frequency
Node tags
Node forms

ConcDesc
Visualize



Collocation candidates

Page [Next >](#)

	Frequency	T-score	MI	logDice
P N sire	507	22.516	16.351	11.990
P N doc	125	11.178	12.493	9.457
P N boomer	71	8.426	16.730	9.385
P N Bells	63	7.937	15.993	9.198
P N gang	90	9.485	12.360	9.115
P N doll	39	6.245	16.101	8.525
P N boom	119	10.901	10.527	8.294
P N SITTER	31	5.568	16.298	8.202
P N pensionare	45	6.705	10.944	7.951
P N M	38	6.162	11.236	7.913
P N Bell	36	5.998	11.313	7.885
P N killer	120	10.943	9.952	7.858
P N Pretty	23	4.795	13.311	7.686
P N pensionato	73	8.535	9.813	7.583
P N Jane	32	5.653	10.524	7.489
P N Achtung	18	4.243	15.325	7.415
P N bye	18	4.242	13.731	7.374
P N Doll	19	4.358	12.126	7.325
P N my	21	4.581	11.317	7.313
P N Fae	13	3.606	16.301	6.959
P N Pozzi	18	4.240	10.552	6.957
P N pretty	13	3.605	14.827	6.945
P N parking	13	3.605	14.426	6.938
P N rapinatore	25	4.994	9.710	6.925

1.4 QUERY SU PIÙ LIVELLI

La ricerca di collocati può essere fatta anche tramite la maschera di partenza, usando il **Corpus Query Language**.

Ad es., per individuare tutte le combinazioni di baby seguito da un nome, possiamo combinare le query:

- **[word = "baby"]**
- **[tag = "NOUN"]**

Corpus: Repubblica

Simple query:

[Query types](#) [Context](#) [Text types](#) ?

Query type simple lemma phrase word character

Lemma:

Phrase:

Word form:

Character:

CQL:

[Tagset summary](#)

leggerli si eviterebbero molti guai . Il **baby boom** degli anni 1960 - 64 , ad esempio , ha
di sei mesi del servizio militare , del **baby boom** e , soprattutto , della mutata propensione
svaniranno gli effetti dei decenni del **baby boom** , ma nel frattempo gli anziani con disponibilità
fare bambine piccole in pose osè , in **baby doll** scomposti , in calze velate e gesti invitanti
Stato questo è il credo dei figli del " **baby boom** " Oggi chi è nato tra il 1946 e il 1964
minante in America . L' avvento dei " **baby boomers** " è ormai completo . La definizione si
a fascia più visibile e più mobile dei " **baby boomers** " : i consumatori più prodighi , i più
. Ci sono alcune caratteristiche dei " **baby boomers** " e degli " yuppies " che rendono difficili
lei primi anni 30 . Ma a quei tempi i " **baby boomers** " non erano ancora nati . CINQUE ERGASTOLI
Piper Laurie . Disney a parte , ecco il **baby puffo** di Hanna e Barbera e Le ultime avventure
clienti . Così facemmo dei corsi per i **baby animatori** e pensammo che era bene cominciare aiutandoli
bisogna dire che quelle piume , quei **baby doll** forestali , quei lamè , quei fazzolettoni
. Quei 73 milioni di figli prodotti dal **baby boom** negli anni che corrono fra la bomba di
i avesse osato descrivere la fine del " **baby boom** " come è avvenuta sarebbe stato credibile

1.4. ESPRESSIONI REGOLARI E CQL

Alcune informazioni pratiche per individuare:

- ❖ **Espressioni regolari**, ovvero ‘schemi’ che corrispondono ad un qualche tipo di sequenza nel testo
- ❖ **Corpus query language (CQL)**: linguaggio per costruire ricerche complesse utilizzando :
 - Espressioni regolari
 - Attributi e valori

CORPUS QUERY LANGUAGE (CQL)

Il **Corpus Query Language (CQL)** permette di elaborare queries molto più complesse della ricerca semplice, combinando più elementi in sequenza e più livelli di codifica/annotazione dei dati.

Alcuni esempi:

- ◆ **[word= “lo” & tag= “ART”]**
Individua tutte le occorrenze di *lo* come articolo
- ◆ **[word = “baby”] [tag = “NOUN”]**
Individua tutte le occorrenze di *baby* seguito da un nome
- ◆ **[lemma= “condurre”] [] [lemma= “indagine”]**
Individua tutte le occorrenze del lemma *condurre* (in qualsiasi sua forma) seguito, dopo un'altra parola qualsiasi, dal lemma *indagine*

1.4 NOTA TIPOGRAFICA

Nelle slide che seguono le espressioni regolari sono scritte tra barrette oblique (//) per distinguerle del testo normale.

In generale non è comunque necessario usarle tra barrette laterali.

1.4 ESPRESSIONI REGOLARI (REGULAR EXPRESSIONS = REGEX)

Schemi che corrispondono ad un qualche tipo di sequenza nel testo.
Possono essere composti da:

- Caratteri o stringhe di testo
- Caratteri speciali
- Gruppi

Es. “trova una corrispondenza con una stringa che inizia con la lettera S e finisce con -ane”

1.4 DELIMITARE LE REGEX

Caratteri speciali per indicare l'inizio e la fine

- `/^man/` => qualsiasi sequenza che inizia con "man":
man, manned, manning...
- `/man$/` => qualsiasi sequenza che termina con "man":
doberman, policeman...
- `/^man$/` => qualsiasi sequenza che contenga solo "man"

1.4 GRUPPI DI CARATTERI E SCELTE

/[wh]ood/

- corrisponde a *wood* o *hood*
- [...] significa una scelta di caratteri

/[^wh]ood/

- Corrisponde a *mood*, *food*, ma non a *wood* o *hood*
- /[^...]/ significa qualsiasi carattere con l'eccezione di quelli tra parentesi

1.4 INTERVALLO

Alcuni gruppi di caratteri possono essere espressi in termini di intervalli:

`/[a-z]/`

- Qualsiasi carattere alfabetico minuscolo

`/[0-9]/`

- Qualsiasi numero tra 0 e 9

`/[a-zA-Z]/`

- Qualsiasi carattere alfabetico maiuscolo o minuscolo

1.4 DISGIUNZIONI E JOLLY

/ba./

- Corrisponde a *bat*, *bad*, ...
- /./ significa “ogni singolo carattere alfanumerico”

/gupp(y | ies)/

- *guppy* OR *guppies*
- /(*x* | *y*)/ significa “o X o Y”
- Importante usare le parentesi!

1.4 QUANTIFICATORI (I)

/colou?r/

- Corrisponde a *color* o *colour*

/govern(ment)?/

- Corrisponde a *govern* o *government*

/?/ significa zero o uno dei caratteri o dei gruppi di caratteri precedenti

1.4 QUANTIFICATORI (II)

`/ba+/`

- Corrisponde a *ba, baa, baaa...*

`/(inkiss)+/`

- Corrisponde a *inkiss, inkiss inkiss*
- (notare lo spazio bianco nella regex)

`/+/` significa “uno o più del carattere o del gruppo di caratteri precedente”

1.4 QUANTIFICATORI (III)

/ba/i>*

- Corrisponde a *b, ba, baa, baaa*
- */*/i> significa “zero o più del carattere o del gruppo di caratteri precedente”*

/(ba){1,3}/i>

- Corrisponde a *ba, ba ba* or *ba ba ba*
- *{n, m}* significa “tra n e m del carattere o del gruppo di caratteri precedente”

/(ba){2}/i>

- Corrisponde a *ba ba*
- *{n}* significa “esattamente n del carattere o del gruppo di caratteri precedente”

1.4 CQL SINTASSI (I)

Ricerche CQL consistono in espressioni regolari rispetto ad attributi (parole, lemmi o tag)

Regex rispetto a parole:

```
[word="it"] [word="resulted"] [word="that"]
```

- Corrisponde solo a "it resulted that"

Regex rispetto a parole con caratteri speciali:

```
[word="it"] [word="result.*"] [word="that"]
```

- Corrisponde a *it resulted/results* that

Regex rispetto ad un lemma:

```
[word="it"] [lemma="result"] [word="that"]
```

- Corrisponde a qualsiasi forma di *result* (regex sul lemma)

1.4 CQL SINTASSI (2)

É possibile combinare query con parola, lemma e tag:

Limiti rispetto a parola e tag:

```
[word="it"] [lemma="result" & tag="V.*]
```

Corrisponde a *it* seguito da una variante morfologica del lemma *result* il cui tag inizia con V (i.e. un verbo)

1.4 CQL SINTASSI (3)

Le parentesi quadrate vuote significano “qualsiasi corrispondenza”

L'uso di quantificatori complessi per corrispondenza rispetto ad intervalli:

`[word="confus.*" & tag="V.*"] []{0,2} [word="by"]`

- “verbo che inizia con *confus* taggato come verbo, seguito dalla parola *by*, con parole inframmezzate in numero da 0 a 2”
- *confused by (the problem)*
- *confused John by (saying that)*
- *confused John Smith by (saying that)*