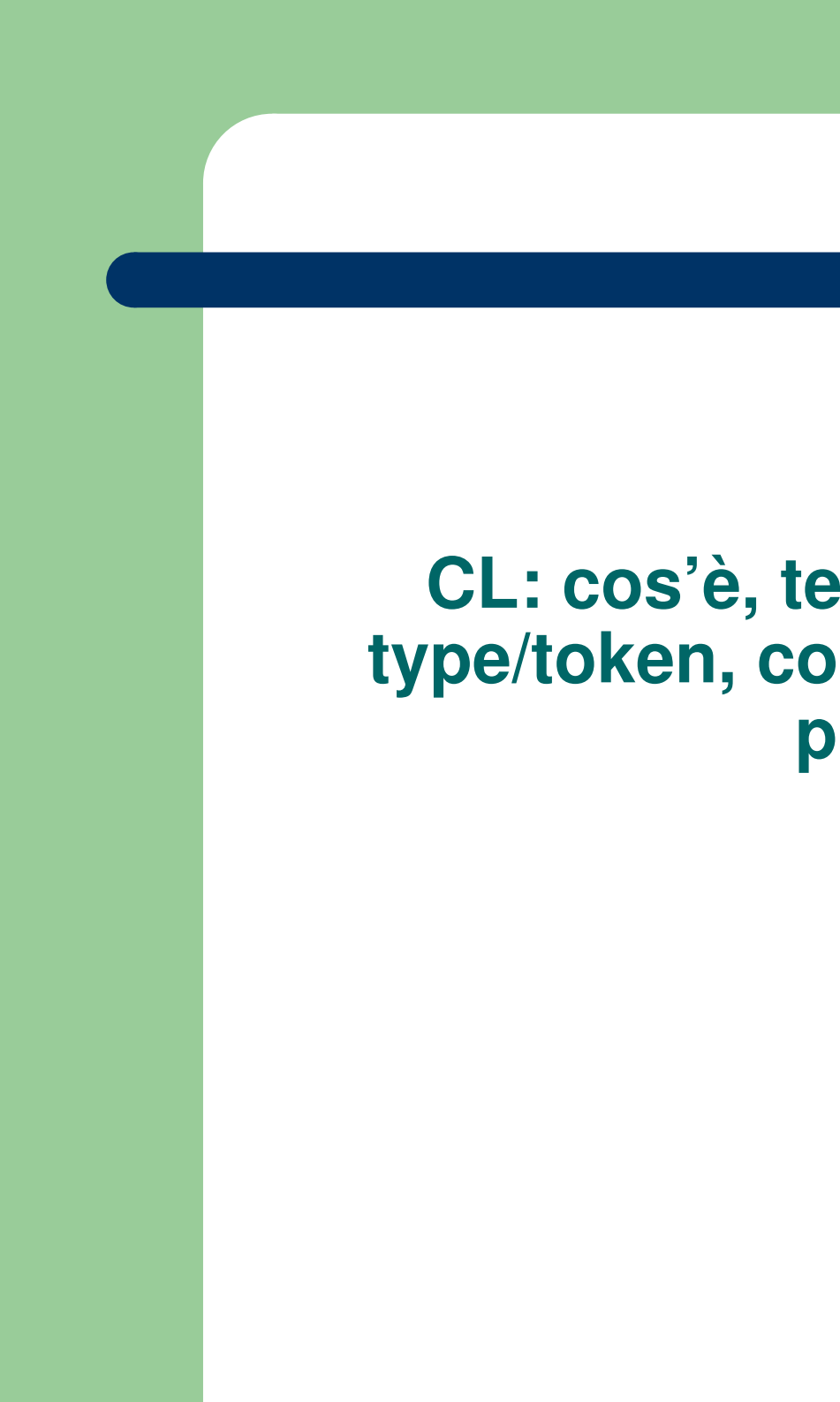


Linguistica dei corpora: una introduzione

Luigi Talamo (talamo.luigi@gmail.com)

A green L-shaped decorative element is in the top-left corner. A thick dark blue horizontal bar is positioned below it, starting from the left edge and extending across the top of the slide.

**CL: cos'è, teoria vs. metodologia,
type/token, collocazione, frequenza e
produttività**

Cos'è

La linguistica dei corpora (*Corpus Linguistics*: CL) è lo studio del linguaggio così come lo si trova espresso in 'campioni di lingua' (corpora).

Seguendo qualche suggestione naturalistica, lo possiamo paragonare all'esame dei campioni che viene effettuato nelle scienze naturali e della vita, come i campioni di sangue per la medicina e i carotaggi in geologia.

Teoria o metodologia?

E' la domanda con cui si apre Gries 2009, sul quale questa prima sezione è largamente basata.

Tra i linguisti dei corpora, gli approcci le risposte sono differenti:

- taluni, come Geoffrey Leech, considerano la CL una vera e propria 'filosofia del linguaggio';
- altri, probabilmente la maggioranza, trattano la CL come un 'semplice' strumento di indagine e di analisi.

CL come teoria

Se consideriamo la CL come teoria, dovremmo poter elaborare una spiegazione al linguaggio (e di conseguenza una grammatica) guidati (*driven*) dai dati ‘grezzi’ del campione linguistico.

E' l'oggetto della *corpus-driven linguistics*:

- uno dei principali obiettivi del trattamento automatico del linguaggio è quello di insegnare alle macchine il linguaggio naturale somministrando loro grosse quantità di dati linguistici;
- discipline più teoriche come la semantica distribuzionale cercano di catturare i significati delle parole dai contesti in cui si trovano: “You shall know a word by the company it keeps” (Joseph Rupert Firth)
- “to derive linguistic categories systematically from the recurrent patterns and the frequency distributions that emerge from language in context” (?:87, citato in ?:59)

CL come metodologia di indagine

Considerare la CL come ‘semplice’ metodologia di indagine non vuol dire affatto ridurre la sua portata a ‘mero strumento’; tuttavia, se la CL non è teoria di per sé, a quale sistema teorico la possiamo riferire?

- alcune discipline linguistiche semplicemente non si *curano* di avere o di esplicitare basi teoriche: la maggior parte dei dizionari contemporanei è basata su corpora linguistici, così come molte opere di linguistica applicata e didattica delle lingue;
- la CL è ormai lo strumento di base della linguistica cognitiva e di stampo funzionalista, cioè di una linguistica basata sull’uso effettivo che i parlanti fanno della lingua (*usage-based cognitive-linguistic theories*). In questo senso, la CL si oppone anche teoricamente alla linguistica di stampo generativista, che è tradizionalmente basata sull’analisi del proprio (= del linguista) linguaggio.

CL e teorie costruttiviste

L'uso della lingua da parte dei parlanti è uno dei pilastri fondamentali di una linguistica molto *à la page*, ovvero della linguistica di impianto costruttivista (*constructional grammar*). Esistono almeno una mezza dozzina di teorie costruttiviste: la Construction Grammar (Goldberg 2013), la Radical Construction Grammar (Croft 2001), la Construction Morphology (Booij 2010), ...

Un altro pilastro fondamentale di queste teorie è ovviamente la costruzione, identificata come un'unità fondamentale di forma e significato, che include in sé le tradizionali categorie linguistiche di morfema, parola, sintagma, ecc.

- Cagliari;
- cagliaritano;
- acchiappa-titolo;
- macchina da scrivere.

Da un punto di vista teorico la CL ha come unità fondamentale la collocazione.

Type, token e collocazione

Espressione = morfema, parola, composto, parola complessa, frase -> in CL 'type'

Occorrenze = tutte le volte in cui trovo una data espressione su un corpus -> in CL 'token'

Collocazione = insieme dei contesti di un corpus in cui si trova una data espressione, cioè insieme delle occorrenze

Esempio: 188 token del type 'rebbe' sul corpus ItWac (circa 1 miliardo di token)

The screenshot shows the NoSketch Engine interface. At the top, the search term 'rebbe' is entered in the search bar, with a magnifying glass icon and a 'Send feedback' link. Below the search bar, the query results are displayed: 'Query **rebbe** 188 (0.10 per million)'. The results are paginated, showing 'Page 1 of 10' with 'Go', 'Next', and 'Last' buttons. The main area displays a list of concordance lines, each starting with a line number (e.g., #3989292, #39963595) followed by the context and the word 'rebbe' in red. The left sidebar contains navigation links: 'Concordance', 'Word list', 'Corpus info', 'My jobs', 'Home', 'User guide', 'Save as subcorpus', 'View options', 'KWIC', 'Sentence', 'Sort', and 'Left'.

NoSketch Engine

rebbe 🔍 Send feedback

Query **rebbe** 188 (0.10 per million)

Page 1 of 10 Go Next Last

#3989292 premendola bene sul bordo con le dita , poi con i **rebbe** di una forchetta sfioracchiate la superficie

#39963595 e , appoggiando uno gnocco per volta sui **rebbe** (ben puliti ed infarinati) schiacciarlo

#49332708 si possono decorare semplicemente con i **rebbe** della forchetta . 3 1/3 tazze di farina

#49332863 leggero fare un motivo decorativo con i **rebbe** della forchetta . Informare in forno già

#59316299 , come può mangiare il risotto tenendo i **rebbe** della forchetta in giù ? POSIZIONE PER

#86137550 strato di pasta , bucherellare la pasta con i **rebbe** della forchetta , per permettere la fuoriuscita

#102310863 salsiccia (punzecchiata qua e là con i **rebbe** della forchetta) , unire poi le castagne

#114997201 , le patate solitamente in 20 minuti . I **rebbe** di una forchetta devono entrare nelle carote

#120686074 si possono decorare semplicemente con i **rebbe** della forchetta . 3 1/3 tazze di farina

#120686229 leggero fare un motivo decorativo con i **rebbe** della forchetta . Informare in forno già

#123724413 della pasta con il tuorlo , rigatela con i **rebbe** di una forchetta e bucatela con la punta

#124824537 quantità della verdura e sentendole con i **rebbe** della forchetta . Man mano che le verdure

#133268303 , le patate solitamente in 20 minuti . I **rebbe** di una forchetta devono entrare nelle carote

#134147135 stessi vogliamo spaventare . Venite ! Con **rebbe** e con forche e con torce e raganelle ,

#134147174 lancino urli ! Coro delle guardie Venite con **rebbe** e con forche , come il diavolo delle loro

Concordance
Word list
Corpus info
My jobs
Home
User guide
Save as subcorpus
View options
KWIC
Sentence
Sort
Left

Collocazioni di *rebbio*

Cosa scopriamo dalle collocazioni di *rebbio* viste sopra?
Abbastanza!

La collocazione di una espressione ci svela le sue caratteristiche grammaticali (formali) e il suo significato, proprio come la costruzione è la rappresentazione della forma e del significato di una espressione.

Collocazioni di *rebbio*

Cosa scopriamo dalle collocazioni di *rebbio* viste sopra?
Abbastanza!

- è un nome, compare, tranne che in due casi, solo al plurale e spesso nel caso strumentale;

La collocazione di una espressione ci svela le sue caratteristiche grammaticali (formali) e il suo significato, proprio come la costruzione è la rappresentazione della forma e del significato di una espressione.

Collocazioni di *rebbio*

Cosa scopriamo dalle collocazioni di *rebbio* viste sopra?
Abbastanza!

- è un nome, compare, tranne che in due casi, solo al plurale e spesso nel caso strumentale;
- è spesso modificato da ‘una forchetta’;

La collocazione di una espressione ci svela le sue caratteristiche grammaticali (formali) e il suo significato, proprio come la costruzione è la rappresentazione della forma e del significato di una espressione.

Collocazioni di *rebbio*

Cosa scopriamo dalle collocazioni di *rebbio* viste sopra?
Abbastanza!

- è un nome, compare, tranne che in due casi, solo al plurale e spesso nel caso strumentale;
- è spesso modificato da ‘una forchetta’;
- è retto da verbi che hanno a che fare con il campo semantico della cucina: ‘bucherellare’, ‘punzecchiare’, ...;

La collocazione di una espressione ci svela le sue caratteristiche grammaticali (formali) e il suo significato, proprio come la costruzione è la rappresentazione della forma e del significato di una espressione.

Collocazioni di *rebbio*

Cosa scopriamo dalle collocazioni di *rebbio* viste sopra?
Abbastanza!

- è un nome, compare, tranne che in due casi, solo al plurale e spesso nel caso strumentale;
- è spesso modificato da ‘una forchetta’;
- è retto da verbi che hanno a che fare con il campo semantico della cucina: ‘bucherellare’, ‘punzecchiare’, ...;
- più in generale, si trova vicino a nomi e verbi che denotano questo campo semantico: melanzana, spennellare, ...;

La collocazione di una espressione ci svela le sue caratteristiche grammaticali (formali) e il suo significato, proprio come la costruzione è la rappresentazione della forma e del significato di una espressione.

Frequenza

Nella sua accezione più semplice, la frequenza è la somma aritmetica delle collocazioni, cioè: l'espressione X si trova Y nel corpus Z.

Frequenza

Nella sua accezione più semplice, la frequenza è la somma aritmetica delle collocazioni, cioè: l'espressione X si trova Y nel corpus Z.

- auto di piazza: ?

Frequenza

Nella sua accezione più semplice, la frequenza è la somma aritmetica delle collocazioni, cioè: l'espressione X si trova Y nel corpus Z.

- auto di piazza: ?
- non la conosco: tre espressioni distinte. Auto che si prende in piazza? A noleggio?

Frequenza

Nella sua accezione più semplice, la frequenza è la somma aritmetica delle collocazioni, cioè: l'espressione X si trova Y nel corpus Z.

- auto di piazza: ?
- non la conosco: tre espressioni distinte. Auto che si prende in piazza? A noleggio?



```
REPUBBLICA3-2> "auto" "di" "piazza";  
46175412: In ogni caso l' arrivo del Comandante in capo al Quartier generale egiziano , subito dopo lo scoppio della guerra , avvenne su una traballante  
<auto di piazza> ...  
57016013: Chiamano un taxi , tornano in albergo , prendono le mitragliette , con la stessa <auto di piazza> si fanno accompagnare a Pigalle .  
181374124: Ma la libert? cromatica per le " <auto di piazza> " ? finita : la prossima tinta che dovranno avere i taxi italiani su tutto il territorio na  
ionale sar? il bianco .  
225794830: E pure sferraglianti <auto di piazza> guidate da tassisti provati e sbuffanti nella canicolare serata romana .
```

Frequenza

Nella sua accezione più semplice, la frequenza è la somma aritmetica delle collocazioni, cioè: l'espressione X si trova Y nel corpus Z.

- auto di piazza: ?
- non la conosco: tre espressioni distinte. Auto che si prende in piazza? A noleggio?



```
REPUBBLICA3-2> "auto" "di" "piazza";  
46175412: In ogni caso l' arrivo del Comandante in capo al Quartier generale egiziano , subito dopo lo scoppio della guerra , avvenne su una traballante  
<auto di piazza> ...  
57016013: Chiamano un taxi , tornano in albergo , prendono le mitragliette , con la stessa <auto di piazza> si fanno accompagnare a Pigalle .  
181374124: Ma la libert? cromatica per le " <auto di piazza> " ? finita : la prossima tinta che dovranno avere i taxi italiani su tutto il territorio na  
ionale sar? il bianco .  
225794830: E pure sferraglianti <auto di piazza> guidate da tassisti provati e sbuffanti nella canicolare serata romana .
```

- ora la conosco: una unica espressione ('parola') con tre token in ItWac

Frequenza come *entrenchment*

Il concetto di frequenza è nuovamente un concetto che ha un corrispettivo funzionale nella linguistica cognitiva.

Più alta è la frequenza, maggiore è la possibilità che una data costruzione sia immagazzinata (radicata) nel lessico mentale dei parlanti.

- se l'espressione è immagazzinata nel lessico, non la devo scomporre: <auto di piazza>;
- se non è immagazzinata, la devo analizzare 'al volo': <auto> <di> <piazza>.



CL al lavoro: qualche applicazione

Annotazioni: lemma e PoS

Negli esempi fatti fino ad ora abbiamo interrogato il nostro corpus direttamente in base alla parola. Un corpus linguistico può offrire però molto di più, potendo essere annotato a tutti i livelli di analisi linguistica. Per ora vediamo le due principali annotazioni, quelle presenti 'di default' (o quasi...) su ogni corpus:

Annotazioni: lemma e PoS

Negli esempi fatti fino ad ora abbiamo interrogato il nostro corpus direttamente in base alla parola. Un corpus linguistico può offrire però molto di più, potendo essere annotato a tutti i livelli di analisi linguistica. Per ora vediamo le due principali annotazioni, quelle presenti 'di default' (o quasi...) su ogni corpus:

- lessicografica, indicando cioè il lemma o 'forma di citazione'; ad es., i token 'rosolerà', 'rosolino', 'rosolavano' sono forme dello stesso lemma ROSOLARE;

Annotazioni: lemma e PoS

Negli esempi fatti fino ad ora abbiamo interrogato il nostro corpus direttamente in base alla parola. Un corpus linguistico può offrire però molto di più, potendo essere annotato a tutti i livelli di analisi linguistica. Per ora vediamo le due principali annotazioni, quelle presenti 'di default' (o quasi...) su ogni corpus:

- lessicografica, indicando cioè il lemma o 'forma di citazione'; ad es., i token 'rosolerà', 'rosolino', 'rosolavano' sono forme dello stesso lemma ROSOLARE;
- sintattica. In base ad un insieme di etichette finito e specifico per ogni corpus (*tagset*), ciascun token è annotato per classe di parola (categoria lessicale, parte del discorso: *Parts of Speech*: PoS) ed eventuali altre caratteristiche morfo-sintattiche; ad., il verbo 'rosolino' è annotato nel *tagset* di ItWac come verbo finito: VER:fin

Annotazioni: il corpus come tabella

Possiamo immaginare un corpus linguistico come una gigantesca tabella, le cui linee corrispondono ai token e le colonne alle annotazioni.

WORD	PoS	LEMMA
La	ART	la
nebbia	NOUN	nebbia
agli	ARTPRE	al
irti	ADJ	irto
colli	NOUN	colle collo
piovigginando	VER:geru	piovigginare
sale	NOUN	sala sale
,	PUN	,

La prima colonna di solito coincide con WORD, ma è di fatto una convenzione.

Trova l'errore!

Annotazioni: come utilizzarle

Le annotazioni possono essere combinate nella nostra ricerca, al fine di ottenere dei risultati più raffinati. Concettualmente, questo equivale a scegliere le righe di una tabella filtrandole attraverso una o più colonne.

Ad es., ‘Dammi tutte le righe in cui compare un nome’, ‘Dammi tutte le righe in cui NON compare un verbo’, ‘Dammi tutte le righe in cui compare un nome al plurale in -i’.

Una operazione che può essere fatta semplicemente attraverso un foglio di calcolo come Excel o Numbers: tuttavia, questo non è né computazionalmente ottimale né umanamente efficace, soprattutto se ci troviamo di fronte a corpora con milioni di token (= milioni di righe!) e dozzine di annotazioni (= dozzine di colonne).

Inoltre, richieste del tipo ‘Dammi tutte le righe in cui nome compare insieme ad un aggettivo’ sono difficili da soddisfare con un programma che ragiona solo per colonne...

Maschere di ricerca

Molti corpora dispongono di una interfaccia grafica, con una comoda maschera di ricerca dove effettuare le proprie query. Utilizzeremo come software di ricerca NoSketch Engine, la cui maschera di ricerca consente di ricercare sia espressioni formate da un solo token (simple) che da più di un token (phrase).

Purtroppo ci sono dei limiti:

- è possibile cercare solo per alcuni tipi di annotazione: token e lemma;
- non è possibile combinare più di un tipo di annotazione per lo stesso token. Ad es., non è possibile una ricerca come 'token = sveglia e PoS = aggettivo';
- è possibile combinare due token solo selezionando il secondo token come lemma. Ad es., è possibile combinare il token 'cipolla' solo con il lemma 'rinvenire', ma non, ad esempio, con una parte del discorso come 'verbo'.

Corpus Query Language

Il modo più efficace per interrogare un corpus è attraverso uno speciale linguaggio, il Corpus Query Language (CQL), originariamente creato all'inizio degli anni '90 per uno dei primi software di interrogazione, il Corpus Query Processor (CQP). CQL è 'parlato' da diversi software di interrogazione e utilizzato da molti corpora: è il linguaggio con cui si può interrogare Sketch Engine, un software commerciale che include 500 corpora pre-caricati per più di 90 lingue. Essendo NoSketch Engine l'implementazione *open source* di Sketch Engine, CQL 'funziona' anche su questo software!

Corpus Query Language: l'unità di base

L'unità di base di CQL è la seguente:

```
[attributo="valore"]
```

dove ciascun token è racchiuso tra due parentesi quadre e 'attributo' sta per qualsiasi tipo di annotazione: word, PoS, lemma.

Per cercare espressioni formate da più token, ad esempio 'la cipolla rinviene', ripeteremo dunque questa unità di base:

```
[attributo="valore"] [attributo="valore"] [...]
```

Corpus Query Language: le espressioni regolari

CQL può utilizzare le cosiddette ‘espressioni regolari’ ovvero simboli speciali che possono essere utilizzati per cercare delle sequenze di caratteri piuttosto che caratteri singoli.

Alternanza di caratteri nella parola. Ad esempio, vogliamo cercare le forme femminili e maschili di ‘gatto’:

```
[word="gatt[ao]"]
```

Qualsiasi carattere nella parola. Ad esempio, vogliamo cercare in quali modi è flesso il nome ‘dito’:

```
[word="dit."]
```

Qualsiasi sequenza di caratteri (alla fine o all’inizio della parola).

Ad esempio, vogliamo cercare tutte le parole in ‘poli’:

```
[word=".*poli$"]
```

oppure che iniziano con ‘tele’:

```
[word="^tele.*"]
```

Corpus Query Language: combinare le annotazioni

Possiamo infine combinare le annotazioni all'interno dello stesso token. Ad esempio, possiamo cercare tutte le occorrenze di 'sveglia' come aggettivo:

```
[lemma="svegliò" & tag="ADJ"]
```

dove ciascun token è racchiuso tra due parentesi quadre e 'attributo' sta per qualsiasi tipo di annotazione: word, PoS, lemma.

Oltre il lessico

Oltre al lessico, che è il primo, tradizionale dominio di utilizzo della CL (ad es., i progetti lessicografici avviati dal De Mauro dagli anni settanta già utilizzavano dei corpora), la CL trova impiego in moltissimi altri campi legati alle scienze linguistiche:

Oltre il lessico

Oltre al lessico, che è il primo, tradizionale dominio di utilizzo della CL (ad es., i progetti lessicografici avviati dal De Mauro dagli anni settanta già utilizzavano dei corpora), la CL trova impiego in moltissimi altri campi legati alle scienze linguistiche:

- abbiamo già menzionato sopra dizionari e usi didattici della CL;

Oltre il lessico

Oltre al lessico, che è il primo, tradizionale dominio di utilizzo della CL (ad es., i progetti lessicografici avviati dal De Mauro dagli anni settanta già utilizzavano dei corpora), la CL trova impiego in moltissimi altri campi legati alle scienze linguistiche:

- abbiamo già menzionato sopra dizionari e usi didattici della CL;
- calcolare la produttività di una costruzione morfologica;

Oltre il lessico

Oltre al lessico, che è il primo, tradizionale dominio di utilizzo della CL (ad es., i progetti lessicografici avviati dal De Mauro dagli anni settanta già utilizzavano dei corpora), la CL trova impiego in moltissimi altri campi legati alle scienze linguistiche:

- abbiamo già menzionato sopra dizionari e usi didattici della CL;
- calcolare la produttività di una costruzione morfologica;
- predire quali scelte sintattiche fanno i parlanti di una lingua (e perché): ad es., il congiuntivo italiano è veramente morto?

Oltre il lessico

Oltre al lessico, che è il primo, tradizionale dominio di utilizzo della CL (ad es., i progetti lessicografici avviati dal De Mauro dagli anni settanta già utilizzavano dei corpora), la CL trova impiego in moltissimi altri campi legati alle scienze linguistiche:

- abbiamo già menzionato sopra dizionari e usi didattici della CL;
- calcolare la produttività di una costruzione morfologica;
- predire quali scelte sintattiche fanno i parlanti di una lingua (e perché): ad es., il congiuntivo italiano è veramente morto?
- identificare il reale utilizzo di due parole all'apparenza tra loro sinonimiche: es. di sopra, quando utilizziamo *auto di piazza* al posto di 'taxi'?

Costruzioni morfologiche: frequenza e produttività

Facciamo ora il caso delle costruzioni morfologiche di tipo derivazionale, come le prefissazioni e le suffissazioni. Ad es., quanto e come i parlanti di lingua italiana utilizzano

- il suffisso *-ame*? legname, pietrame, bambiname, berlusconame, grillame
- il prefissoide *tele-*? televendita, telecomando, telepresentatore
- il suffissoide *-poli*? tangentopoli, vallettopoli, guerciopoli
- il primo membro del composto *acchiappa-*?
acchiappa-macchie, acchiappa-titoli

ovvero: quanto e come è produttiva una data costruzione?

Costruzioni morfologiche: type/token

Definire la frequenza e la produttività di una costruzione morfologica è un compito leggermente diverso dal definire gli stessi valori per una parola come *rebbi* o *auto di piazza*.

Abbiamo detto prima che in CL una parola equivale ad un type, di cui troviamo un certo numero di token in un corpus.

Nelle costruzioni morfologiche ci troviamo di fronte a una regola - la costruzione, appunto - che crea:

- un certo numero di type diversi;
- ciascuno di questi type mostra un certo numero di token.

Tre tipi di produttività morfologica

Per quanto riguarda le costruzioni morfologiche, Baayen 2009 distingue tre tipi di produttività:

1. produttività realizzata;
2. produttività in espansione;
3. produttività potenziale.

Produttività realizzata

La produttività realizzata è il tipo più semplice di produttività e coincide di fatto con la frequenza dei types di una determinata costruzione morfologica.

Volendo fare un'analogia con l'economia, il primo tipo di produttività è simile alla fetta che ha una compagnia detiene sul mercato.

Se la produttività realizzata è alta, la costruzione morfologica avrà una grossa quota consolidata nel 'mercato' delle derivazioni morfologiche.

■ Come si calcola: Numero di types (costruzione)

Produttività in espansione

Nel nostro paragone con l'economia di mercato, il secondo tipo di produttività misura quanto la costruzione morfologica si sta espandendo, anche a danno di altre derivazioni morfologiche. E' inoltre interessante notare che una costruzione può avere una scarsa produttività realizzata, ma un'alta produttività in espansione -> è il caso dei nuovi affissi

- Come si calcola. $P = \frac{\text{numero di hapax legomena (costruzione)}}{\text{numero di hapax legomena (corpus)}}$

Produttività potenziale

Il terzo tipo di produttività misura quanto una costruzione morfologica è in grado di occupare una fetta di mercato; una azienda può essere anche in espansione, ma se il mercato è ormai saturo rischia probabilmente di fare bancarotta!

E' il tipo di produttività più utilizzato negli studi di CL e morfologia quantitativa, ed è chiamato semplicemente P:

- Come si calcola. $P = \text{numero di hapax legomena (costruzione)} / \text{numero di token totali (costruzione)}$

Con un aggiustamento a livello di sotto-corpora, l'indice P è utilizzato nei lavori di Gaeta & Ricca sulla produttività dei suffissi italiani (Gaeta and Ricca 2003, Gaeta and Ricca 2006).

Costruzioni morfologiche: esercizio

Data la (reale) lista di frequenza della costruzione con *-ame* nel corpus La Repubblica, calcolare i tre tipi di produttività.

Costruzioni sintattiche: type/token

Secondo il principio di non sinonimicità, una differenza formale tra due costruzioni implica SEMPRE una differenza di significato: “If two constructions are syntactically distinct, they must be semantically or pragmatically distinct” (Adele Goldberg)

Per quanto riguarda la sintassi, possiamo dunque decidere di studiare in CL una costruzione che può essere espressa in diversi modi formali (type), ciascuno dei quali mostrerà un certo numero di token.

Costruzioni sintattiche: la costruzione di-transitiva inglese

Ad es., l'inglese esprime la costruzione ditransitiva:

Costruzioni sintattiche: la costruzione di-transitiva inglese

Ad es., l'inglese esprime la costruzione ditransitiva:

- Funzione: trasferire qualcosa a qualcuno;

Costruzioni sintattiche: la costruzione ditransitiva inglese

Ad es., l'inglese esprime la costruzione ditransitiva:

- Funzione: trasferire qualcosa a qualcuno;
- ruoli semantici/argomentali: AGENT TRANSFER PATIENT RECIPIENT

Costruzioni sintattiche: la costruzione ditransitiva inglese

Ad es., l'inglese esprime la costruzione ditransitiva:

- Funzione: trasferire qualcosa a qualcuno;
- ruoli semantici/argomentali: AGENT TRANSFER PATIENT RECIPIENT
- dove TRANSFER è un verbo come *to give*, *to bring*, *to tell*, *to play*.

Costruzioni sintattiche: la costruzione ditransitiva inglese

Ad es., l'inglese esprime la costruzione ditransitiva:

- Funzione: trasferire qualcosa a qualcuno;
- ruoli semantici/argomentali: AGENT TRANSFER PATIENT RECIPIENT
- dove TRANSFER è un verbo come *to give*, *to bring*, *to tell*, *to play*.
- in due modi sintatticamente diversi, cioè in due ruoli grammaticali/sintattici diversi:

Costruzioni sintattiche: la costruzione ditransitiva inglese

Ad es., l'inglese esprime la costruzione ditransitiva:

- Funzione: trasferire qualcosa a qualcuno;
- ruoli semantici/argomentali: AGENT TRANSFER PATIENT RECIPIENT
- dove TRANSFER è un verbo come *to give*, *to bring*, *to tell*, *to play*.
- in due modi sintatticamente diversi, cioè in due ruoli grammaticali/sintattici diversi:
- SUBJ TRANSFER DIRECT-OBJECT INDIRECT-OBJECT:
Mary gives a letter to John;

Costruzioni sintattiche: la costruzione ditransitiva inglese

Ad es., l'inglese esprime la costruzione ditransitiva:

- Funzione: trasferire qualcosa a qualcuno;
- ruoli semantici/argomentali: AGENT TRANSFER PATIENT RECIPIENT
- dove TRANSFER è un verbo come *to give*, *to bring*, *to tell*, *to play*.
- in due modi sintatticamente diversi, cioè in due ruoli grammaticali/sintattici diversi:
- SUBJ TRANSFER DIRECT-OBJECT INDIRECT-OBJECT:
Mary gives a letter to John;
- SUBJ TRANSFER DIRECT-OBJECT
DIRECT-OBJECT: Mary gives John a letter.

Costruzioni sintattiche: la costruzione ditransitiva inglese

Ad es., l'inglese esprime la costruzione ditransitiva:

- Funzione: trasferire qualcosa a qualcuno;
- ruoli semantici/argomentali: AGENT TRANSFER PATIENT RECIPIENT
- dove TRANSFER è un verbo come *to give*, *to bring*, *to tell*, *to play*.
- in due modi sintatticamente diversi, cioè in due ruoli grammaticali/sintattici diversi:
- SUBJ TRANSFER DIRECT-OBJECT INDIRECT-OBJECT:
Mary gives a letter to John;
- SUBJ TRANSFER DIRECT-OBJECT
DIRECT-OBJECT: Mary gives John a letter.
- marcando cioè in modo diverso il RECIPIENT

Costruzioni sintattiche: esercizio

Data la (reale) lista di frequenza delle realizzazioni sintattiche della costruzione ditransitiva

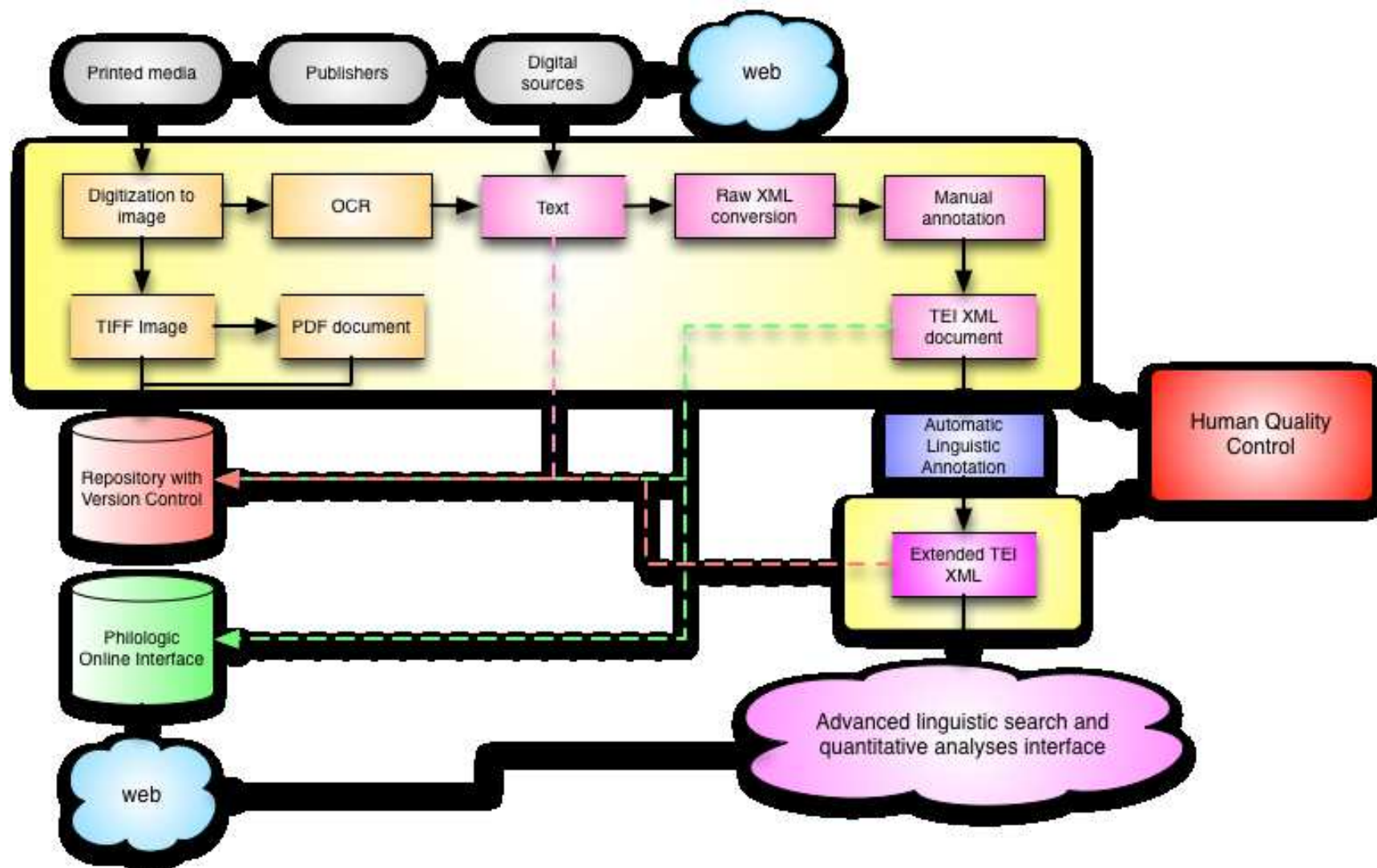
- con quattro verbi inglesi: give, bring, tell, play
- ciascuno nella variante con oggetto diretto o con oggetto indiretto

quali caratteristiche semantiche e funzionali possiamo dedurre dalla distribuzioni di questi otto types sintattici?

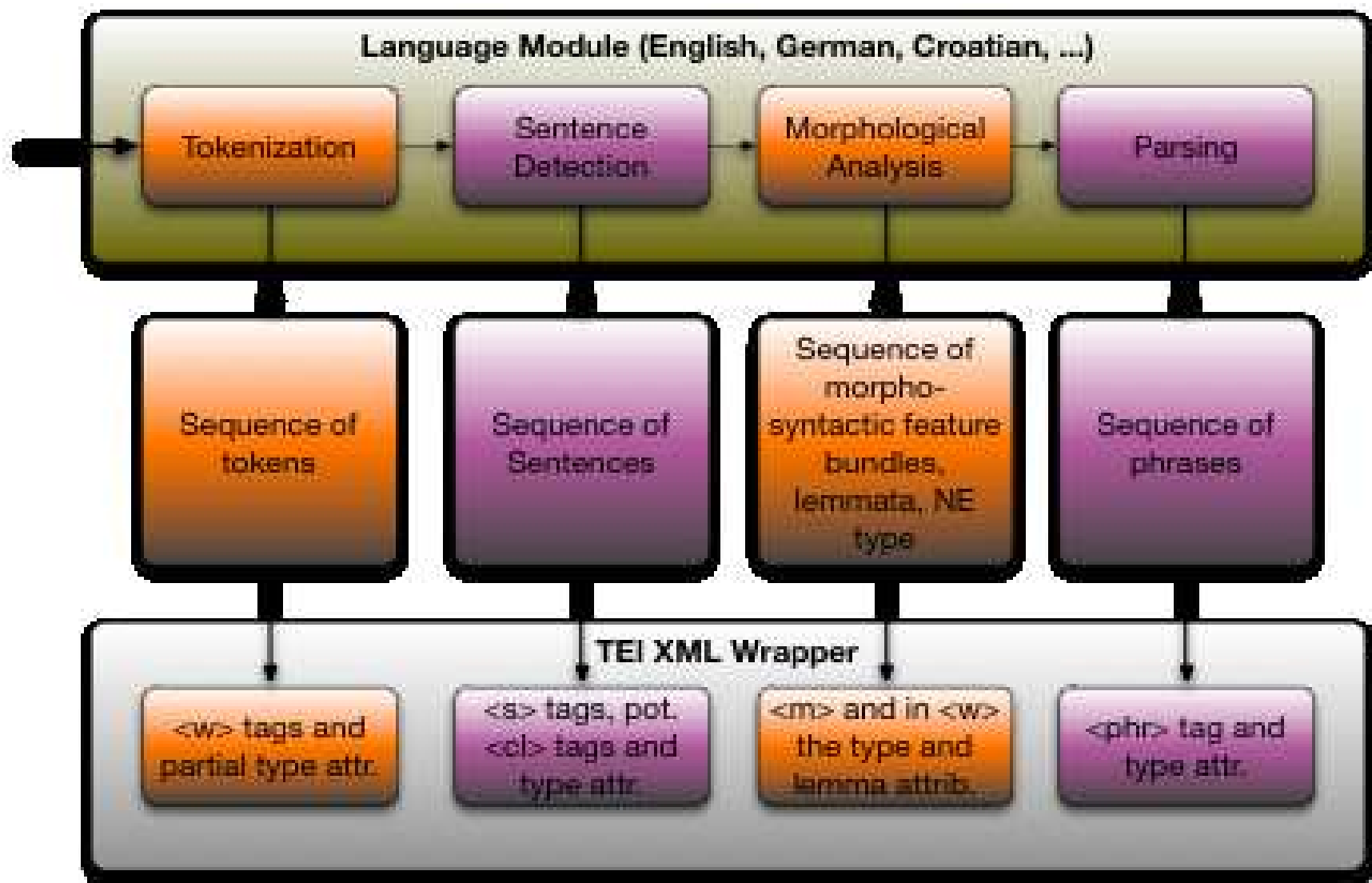


Creazione di un corpus testuale

Tabella di marcia



Annotazione del linguaggio



Annotazione posizionale

Si applica al singolo token e funziona con il modello a colonna:

```
In PRE in
periferia NOUN periferia
fa VER:fin fare
molto ADV molto
caldo ADJ caldo
/ PUN /
mamma NOUN mamma
stai VER:fin stare
tranquilla ADJ tranquillo
,PUN ,
sto VER:fin stare
arrivando VER:geru arrivare
```

Ma come fare ad annotare espressioni con più di un token, ad esempio l'espressione 'auto di piazza' o 'rebbi della forchetta'?

Annotazione strutturale

L'annotazione strutturale 'copre' più di un token e viene implementata attraverso un linguaggio di *markup*. Il linguaggio più utilizzato è XML, disponibile in vari 'dialetti'.

```
<?xml version="1.0" encoding="UTF-8" ?>  
<text num="1" title="Soldi" author="Mahmood">  
<l>
```

```
In PRE in  
periferia NOUN periferia  
fa VER:fin fare  
molto ADV molto  
caldo ADJ caldo  
</l>  
</text>
```

Per codificare un testo linguistico il dialetto XML più utilizzato è lo standard TEI: *Text Encoding Initiative*.

Annotazione TEI-XML

Le annotazioni XML possono in realtà coprire tranquillamente il singolo token, che viene annidato assieme ad altri token in strutture più complesse come i sintagmi, i versi/frasi, fino ad arrivare al testo:

```
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns:off="http://www.tei-c.org/ns/1.0">
  <teiHeader>
  </teiHeader>
  <text>
    <body>
      <l>
        <tok id="w-1" pos="PRE" lemma="in">In</tok>
        <tok id="w-2" pos="NOUN" lemma="periferia">periferia<
        ...
      </l>
    </body>
  </text>
```

Riferimenti bibliografici

- Baayen, H. R. (2009). Corpus linguistics in morphology: morphological productivity. In A. Luedeling and M. Kyto, editors, *Corpus Linguistics. An international handbook.*, pages 900–919. Mouton De Gruyter: Berlin.
- Booij, G. (2010). *Construction Morphology*. Oxford: Oxford University Press.
- Croft, W. (2001). *Radical Construction Grammar: Syntactic Theory in Typological Perspective*. Oxford: Oxford University Press.
- Gaeta, L. and Ricca, D. (2003). Frequency and productivity in Italian derivation: A comparison between corpus-based and lexicographical data. *Rivista di linguistica / Italian Journal of Linguistics*, 15(1), 63–98.
- Gaeta, L. and Ricca, D. (2006). Productivity in Italian word formation: a variable-corpus approach. *Linguistics*, 44(1), 57–91.
- Goldberg, E. A. (2013). Constructionist approaches. In T. Hoffmann and G. Trousdale, editors, *Con-*

struction Grammar Handbook., pages 9–26. Oxford: Oxford University Press.

Gries, S. T. (2009). What is Corpus Linguistics? *Language and Linguistics Compass*, **3**(1), 1–17.