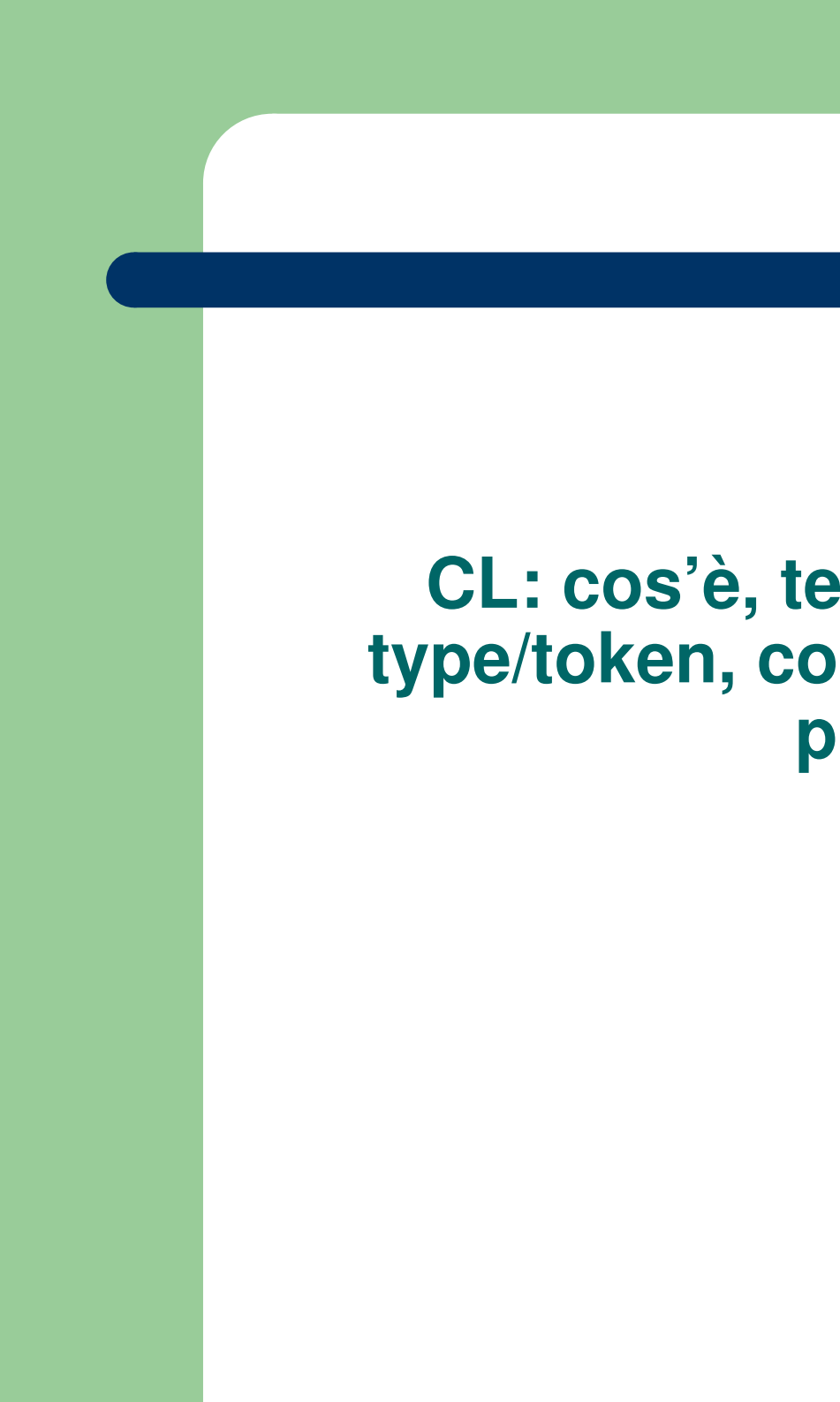


Linguistica dei corpora: una breve introduzione

Luigi Talamo (luigi.talamo@unibg.it)

Dottorato in Scienze Linguistiche: Università degli Studi di Bergamo e Università degli Studi di Pavia

A green vertical bar on the left side of the slide, with a light green rounded rectangle at the top left and a dark blue horizontal bar extending from the left edge.

**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 questo seminario è largamente basato.

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) basandoci solo sui dati 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)

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: 263 token del type 'paludato' sul corpus La Repubblica

2071736: E dunque, in una Londra devastata dai bombardamenti, nasce una storia — collage che ? a mezzo tra Bulli e pupe e l' Opera da tre soldi i nella prima versione di Strehler : due ragazze pepate della Salvation Army (bravissima Cristina Noci , in coppia con Rosalba Caramoni) si muovono sgombrando e battendo i tacchi , tra austeri poliziotti , tra ladruncoli affiliati in piccole bande (ecco il titolo Pick-Pocket che non deriva da Bresson , e molto brevi sono il poliziotto Roberto Stocchi e soprattutto i quattro balordi , Stefano Onofri , Roberto Tedesco , Bruno Burbi , e Leonardo Amato) , con comparsa brechtiana di un re <paludato> , in un susseguirsi di motivi che par di avere già sentito e che Marcucci ha allegramente ripescato dal suo bagaglio musicale .

6200164: Ogni riga del libro ha fatto un lungo cammino : ha filtrato le impurità , le scorie della storia <paludata> , della relazione ufficiale del secondo " ottomila " della terra .

14721831: Con i soldi strappati al fisco indiano , nel 1981 Rajneesh accompagnato dai suoi discepoli <paludati> in sete arancioni , era arrivato negli Stati Uniti per comprarsi sessantaquattromila acri di terra per un milione di dollari , tutti fango e colline nel cuore dell' Oregon , in un ranch che era servito a John Wayne per girare uno dei suoi western .

16698815: MENTRE a teatro fervono , o almeno si fanno dibattere , certe riletture che hanno un senso specchiante , vale a dire , talvolta , antinomico come ad esempio L' onesto Iago , la riproposta attuale de Il governo di Verre già elaborato per la scena nel '65 da Mario Prosperi e Renzo Giovampietrino traendo spunto dalle sette orazioni " verrine " di Cicerone , dopo anche gli allora critici rilievi per una ridondante , manichea contrapposizione di virtù <paludata> e corruzione nequissima , non ? che dovesse qui oggi compensarci con l' esatto risvolto della medaglia , e cioè con un Cicerone (come in parte fu , lo sappiamo) paladino opportunista della morale pubblica , già aspirante " monstre " dell' arte forense .

17040686: Questa e altre , molte altre storie della psicoanalisi — alcune più piccanti , altre più <paludate> , come del resto ? tipico di questa disciplina ancor oggi sospesa tra la più germanica professorialità e un certo pizzicore di trasgressione , o addirittura di ciarlataneria — verranno raccontate per quattro giorni , a partire da oggi , a Trieste .

21508194: Cuore e cervello , ragione e passione hanno coabitato spesso , e non sempre in armonia , tra queste mura <paludate> del Quartiere Latino , confinanti con la più agitata Sorbona .

21860974: Il palazzo di boulevard Raspail ? quanto di meno <paludato> si possa immaginare .

23356030: Superati gli ostacoli linguistici con la convenzione che , nella finzione come in realtà , ognuno reciti nella sua lingua , in mezzo a rovine romane va avanti la prova della commedia , che gli italiani sono portati a rendere il più possibile allegra e volgare , e invece il nobile spagnolo vorrebbe <paludata> e severa .

Collocazioni di *paludato*

Cosa scopriamo dalle (poche) collocazioni di *paludato* 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 *paludato*

Cosa scopriamo dalle (poche) collocazioni di *paludato* viste sopra? Abbastanza!

- è un aggettivo: si accorda per genere e numero col nome a cui si riferisce ed è gradabile;

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 *paludato*

Cosa scopriamo dalle (poche) collocazioni di *paludato* viste sopra? Abbastanza!

- è un aggettivo: si accorda per genere e numero col nome a cui si riferisce ed è gradabile;
- si può riferire sia ad un essere umano che ad un oggetto;

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 *paludato*

Cosa scopriamo dalle (poche) collocazioni di *paludato* viste sopra? Abbastanza!

- è un aggettivo: si accorda per genere e numero col nome a cui si riferisce ed è gradabile;
- si può riferire sia ad un essere umano che ad un oggetto;
- può reggere un argomento: *in sete arancioni...*;

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 *paludato*

Cosa scopriamo dalle (poche) collocazioni di *paludato* viste sopra? Abbastanza!

- è un aggettivo: si accorda per genere e numero col nome a cui si riferisce ed è gradabile;
- si può riferire sia ad un essere umano che ad un oggetto;
- può reggere un argomento: *in sete arancioni...*;
- ha un (generico) significato di ‘solenne’...

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 *paludato*

Cosa scopriamo dalle (poche) collocazioni di *paludato* viste sopra? Abbastanza!

- è un aggettivo: si accorda per genere e numero col nome a cui si riferisce ed è gradabile;
- si può riferire sia ad un essere umano che ad un oggetto;
- può reggere un argomento: *in sete arancioni...*;
- ha un (generico) significato di ‘solenne’...
- ...ma anche di ‘gozzo, inadatto’.

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')

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

Introduzione

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:

Introduzione

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;

Introduzione

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;

Introduzione

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?

Introduzione

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'? e *paludato* al posto di 'solenne' o di 'goffo'?

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 *paludato* 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).

rottame 1572 bestiame 1036 salame 716 legname
349 liquame 315 ciarpame 130 vasellame 122 pella-
me 109 fogliame 102 scatolame 73 culturame 49 fat-
tame 32 cordame 28 fasciame 27 velame 20 pelame
18 pietrame 14 pentolame 12 collettame 11 novel-
lame 10 valvolame 9 ossame 7 brulicame 5 spiccio-
lame 5 mosciame 4 girolame 4 cespugliame 3 no-
tabilame 3 catename 3 contadiname 3 servitorame
3 tavolame 2 parentame 2 Scatolame 2 criticame 2
comportame 1 cingolame 1 uccellame 1 notiziame 1
Cialtroname 1 frascame 1 erbame 1 cantautorame 1
pellicolame 1 ciottolame 1 vetrame 1 cassetname 1
gabinettame 1 frondame 1 rudimentame 1 lastrame
1 carname 1

Costruzioni sintattiche: type/token

Per quanto riguarda la sintassi, abbiamo una certa costruzione che può essere espressa in diversi modi formali (type), ciascuno dei quali ha un certo numero di token.

Ad es., l'inglese possiede due modi di esprimere la costruzione sintattica ditransitiva, cioè AZIONE QUALCOSA A QUALCUNO, dove AZIONE è un verbo come dare, portare, ecc. :

- Mary gives a letter to John;
- Mary gives John a letter.

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.