Introduzione alla linguistica computazionale e alle sue possibili applicazioni

Introduzione

L'ambizione a creare esseri artificiali, meccanici, che svolgessero azioni solitamente svolte da esseri viventi, animali e uomini è sempre stata nella fantasia dell'uomo.
L'informatica è definita come la disciplina che si occupa della raccolta e del trattamento delle informazioni […] (De Mauro, 2000) e dato che la linguistica si occupa dell'analisi e dello studio dello strumento più potente che l'uomo possieda per esprimere informazioni (cioè il linguaggio), il loro incontro era inevitabile: un calcolatore elettronico che deve trattare e usare informazioni dovrà dunque servirsi di una qualche forma di linguaggio; ed è la possibilità di rappresentare la lingua come un sistema regolato, ordinato e coerente, quindi simile a calcoli, ad aver spinto ad ipotizzare la possibilità di una riproducibilità delle competenze linguistiche e delle loro manifestazioni da parte di una macchina, contribuendo allo sviluppo della linguistica computazionale.

I modelli

La possibilità di costruire macchine che producano comportamenti linguistici è fondata sulla presenza di un modello, che contempli tutte le possibilità dell'interazione, e che può essere diverso dal modello usato dagli uomini ma al tempo stesso può far luce su di esso.

Vi sono due principali approcci. Il primo concepisce una regola linguistica come la descrizione di una pratica, ossia una tendenza probabilistica e statistica di comportamento. Questo approccio è definito di tipo statistico-probabilistico, largamente basato sui dati estratti da testi reali e volto all'imitazione.
La posizione di Noam Chomsky differisce da questa e propone un approccio basato su un insieme di regole sintattiche innate, facenti parte della competenza linguistica, che da un insieme finito e astratto di regole (tra l'altro due caratteristiche chiave dei linguaggi di programmazione[1]) possa permettere la formazione di un'infinità di frasi ben formate (Chomsky, 1957). Questo modello è invece detto *rule-based.*
Ad una prima analisi l'approccio *rule-based* sembra il più facilmente sviluppabile dato il numero finito di regole da computare, eppure rimangono dei fattori di incalcolabilità nella lingua, che costituiscono un ostacolo non indifferente. Questi fattori sono l'infinitezza non solo di nuove produzioni ma anche di nuovo lessico, la presenza di sinonimie (*alunno, studente, allievo)*, polisemie (*tagliare la torta / tagliare il prato*), omonimie (*faccia: v. fare / parte del corpo)*, le poliematiche (*presidente del consiglio*), le collocazioni (*compilare un modulo*), i significati organizzati in accezioni (*ho dato un calcio / il calcio è un elemento chimico*), la differenza di accezioni tra lessemi simili in lingue diverse (*kick, calcium*), l'estensibilità di significato di una parola (*navigare nel web*), le ambiguità sintattiche disambiguabili solo grazie al contesto e alla prosodia (*una vecchia porta la sbarra*), i riferimenti anaforici e cataforici (*il giovane ha ringraziato l'amico e sua moglie)*, in alcuni contesti anche i quantificatori (*tutti gli studenti hanno letto due libri*). Importante anche la metalinguistica, che rappresenta una forte differenza tra lingue naturali e calcoli e che pone in grossa difficoltà in quanto rende grammaticali determinati tipi di frasi ma non altri con struttura simile (**con mangia le mani / con è una preposizione*). Da non dimenticare anche la pragmatica (*Sai che ore sono? #Sì / Sono le tre)*.
Ultimo fattore da tenere a mente, ma non meno importante, è l'individuazione di tutte le regole del linguaggio da implementare nel sistema. Lavoro ben più difficile di quanto si pensi.

Chomsky mette in dubbio l'utilità teorica dei procedimenti empirici basati sulla raccolta di corpora linguistici data la loro finitezza e incompletezza (che ci priva di tutte le frasi prodotte ma non catalogate e di quelle non

[1] È definito linguaggio di programmazione un particolare insieme di istruzioni interpretabili dal computer.

ancora prodotte ma possibili). Per lui infatti un corpus non permetterebbe di fare delle predizioni sulla grammaticalità delle frasi ma darebbe solo indicazioni sulla frequenza d'uso.
Dalla sua posizione, per opera di altri, è nato il Natural Language Processing[2], con gli obiettivi di far produrre alla macchina frasi ben formate della lingua (previo implementazione di regole), e un'analisi delle frasi stesse. L'NLP è a sua volta diviso in due settori: uno di analisi (Understanding o Analysis) e l'altro di generazione (Generation). Il Natural Language Understanding mira a rappresentare i testi in modo formale e univoco, mentre il Natural Language Generation si occupa della costruzione di automi capaci di produrre frasi grammaticali.

Il corpus

Un corpus (corpora al plurale) è una raccolta completa e ordinata di scritti di uno o più autori e si può considerare come una collezione di esecuzioni linguistiche.
Definendo come *popolazione* l'insieme di tutte le possibili osservazioni e come *campione* una selezione, una parte della popolazione, che include solo alcune delle possibili osservazioni; un campione dovrebbe essere scelto in modo che ogni possibile osservazione abbia la stessa probabilità di occorrere nel campione e nella popolazione. Il campione quindi deve rappresentare la popolazione.
Altra caratteristica fondamentale è la sua estensione, ossia la sua ampiezza, correlata in base alle necessità.
L'utilità di un corpus si misura in base alle possibilità che ci fornisce di accedere a informazioni linguistiche in modo sintetico e rapido. Fondamentale diventa quindi il formato elettronico dei corpora.
Molti di essi sono usufruibili gratuitamente online, ma può essere il web esso stesso un corpus? È meglio non considerarlo tale date le sue caratteristiche: le lingue vi sono rappresentate in modo ineguale, non presenta le diverse varietà della lingua, l'accesso ai materiali avviene attraverso motori di ricerca che prediligono i siti maggiormente cliccati (e non quelli più aderenti alla richiesta), ci sono frequenti duplicazioni, vi è un elevato tasso di errori e non può considerarsi né statico né dinamico poiché nuove pagine vengono aggiunte ed eliminate ogni giorno, senza un criterio fisso temporale rendendo le analisi aleatorie. Al di là di ciò, con le dovute cautele e approfittando di ulteriori applicativi web, la rete può senz'altro essere usata come corpus.

CHILDES[3]

Il sistema CHILDES (acronimo per CHild Language Data Exchange System, letteralmente Sistema per lo Scambio di Dati sul Linguaggio Infantile) è uno tra gli strumenti più completi e diffusi nel mondo per la trascrizione, la codifica, l'archiviazione e l'analisi computerizzata di dati del linguaggio infantile. È stato ideato nel 1984 da Brian MacWhinney e Catherine Snow per lo studio interlinguistico del linguaggio infantile tramite procedure informatiche. CHILDES non è certamente l'unico strumento disponibile per l'analisi informatizzata di dati linguistici ma, a differenza di altri sistemi, è indubbiamente quello che presenta le caratteristiche di maggiore generalità e flessibilità ai diversi livelli di analisi linguistica (pragmatico-discorsiva, lessicale, morfosintattica, fonologica). Inoltre, CHILDES rappresenta il primo e il più esplicito tentativo di rendere pubblici e disponibili[4] per la comunità scientifica i dati empirici su cui si basano le indagini di singoli ricercatori.
Usando la banca dati CHILDES, un ricercatore può avere accesso a dati tratti da un gran numero di progetti, potendo dedicarsi fin da subito allo studio, senza dover spendere parte del suo tempo raccogliendo dati.
Nell'ambito degli studi sull'acquisizione del linguaggio è ormai palese che per quanto riguarda la condivisione di dati, i vantaggi superano di gran lunga i potenziali svantaggi. Un archivio computerizzato può diventare un insieme di dati in continuo sviluppo, arricchito da ogni utente! La questione, adesso, non è più se condividere o meno i dati, bensì come condividerli nel modo più affidabile e responsabile. La computerizzazione delle trascrizioni allarga le possibilità a condivisioni e analisi di vario genere che altrimenti sarebbero impossibili. Ma

[2] Ridotto in NLP, trattamento automatico del linguaggio naturale
[3] Il capitolo è tratto da (Bortolini & Pizzuto, 1997)
[4] http://childes.talkbank.org/

per sfruttare appieno questa opportunità è necessario sviluppare un sistema standardizzato di trascrizione e analisi.
Con un sistema computerizzato per lo scambio di dati linguistici si possono perseguire tre obiettivi fondamentali:

1. fornire una quantità maggiore di dati su più bambini, su più fasce di età, su più lingue;
2. migliorare la qualità dei dati grazie a un sistema di trascrizione omogeneo e con una documentazione esauriente;
3. automatizzare il processo di analisi dei dati;

Il sistema CHILDES ha mirato a questi tre obiettivi sviluppando tre strumenti distinti ma integrati. Il primo è il CHAT, ossia le norme standardizzate per la trascrizione e codifica dei dati. Il secondo è il CLAN, il pacchetto dei software per l'analisi automatizzata di trascrizioni, e il terzo è la banca dati che include un'ampia varietà di campioni di linguaggio su una gamma altrettanto ampia di età e situazioni. I tre strumenti sono complementari. Il sistema utilizza procedure standardizzate in cui vengono esplicitate le metodologie di trascrizione, codifica e analisi di dati. Ciò consente di confrontare anche sul piano interlinguistico i dati e i risultati ottenuti da ricercatori diversi, e quindi di vagliare in modo rigoroso la validità e/o la generalizzabilità dei risultati conseguiti. Un sistema di trascrizione deve garantire la chiarezza delle categorie che utilizza, insieme alla leggibilità e alla facilità d'uso. Deve rispondere alle esigenze di due tipi di interlocutori: gli interlocutori umani che trascrivono, analizzano e leggono una trascrizione, e il computer e i suoi programmi. Per soddisfare le esigenze di questi due diversi tipi di interlocutori, un sistema di trascrizione computerizzata deve raggiungere i seguenti obiettivi:

1. chiarezza: ciascun simbolo utilizzato nel sistema di codifica deve avere un riferimento chiaro e ben definibile. La relazione fra il simbolo e l'oggetto o il concetto a cui si riferisce deve essere consistente e affidabile;
2. leggibilità: come una lingua umana deve essere facile da elaborare, allo stesso modo una trascrizione deve essere facilmente leggibile;
3. facilità nell'immissione dei dati: l'aumento delle distinzione e relative codifiche all'interno di un sistema di trascrizione rende sempre più difficile l'immissione dei dati, mentre aumenta la possibilità di commettere errori. Il programma CLAN ci viene in aiuto in questo processo.

Esempio di trascrizione[5]

```
@Loc: Romance/Italian/Roma/010817.cha
@PID: 11312/c-00030284-1
@Begin
@Languages: ita
@Participants: CHI Francesco Target_Child, MOT Mother, VIR Virginia
Observer
@ID: ita|Roma|CHI|1;8.17||||Target_Child|||
@ID: ita|Roma|MOT|||||Mother|||
@ID: ita|Roma|VIR|||||Observer|||
@Date: 14-OCT-1970
[…]
*CHI:   otta [*] .
%err:   otta = rotta
*VIR:   cosa è successo ?
*CHI:   otta [*] .
%err:   otta = rotta
*VIR:   cosa hai fatto ?
*CHI:   bum@o !
%com:   rumore dei fuochi d' artificio
```

[5] L'esempio è stato preso da http://childes.talkbank.org/browser/index.php?url=Romance/Italian/Roma/010817.cha

```
*VIR:   che cosa bum@o ?
*CHI:   uti [*] .
%err:   uti = luci
*VIR:   le luci non fanno bum@o .
*VIR:   questa fa bum@o ?
*MOT:   non fanno bum@o le luci .
*CHI:   non fanno 0 [*] .
%err:   0 = bum@o [?] .
*VIR:   e questa cos' è ?
*CHI:   bù@o .
*CHI:   baba@o [=? cane] .
*VIR:   ma dorme , ha l' occhio chiuso !
*VIR:   chi ti porti nel letto tu a fare la nanna ?
*CHI:   pitto@c [=? micio] .
*VIR:   ah !
*VIR:   il gatto .
[…]
@End
```

Nella trascrizione possiamo identificare chiaramente chi sono gli interlocutori, descritti ad inizio testo insieme ad altre importanti informazioni preliminari. Seppur breve, in questo tratto di trascrizione notiamo l'uso di onomatopee, aferesi fonetiche e omissioni, elementi distintivi del linguaggio infantile.

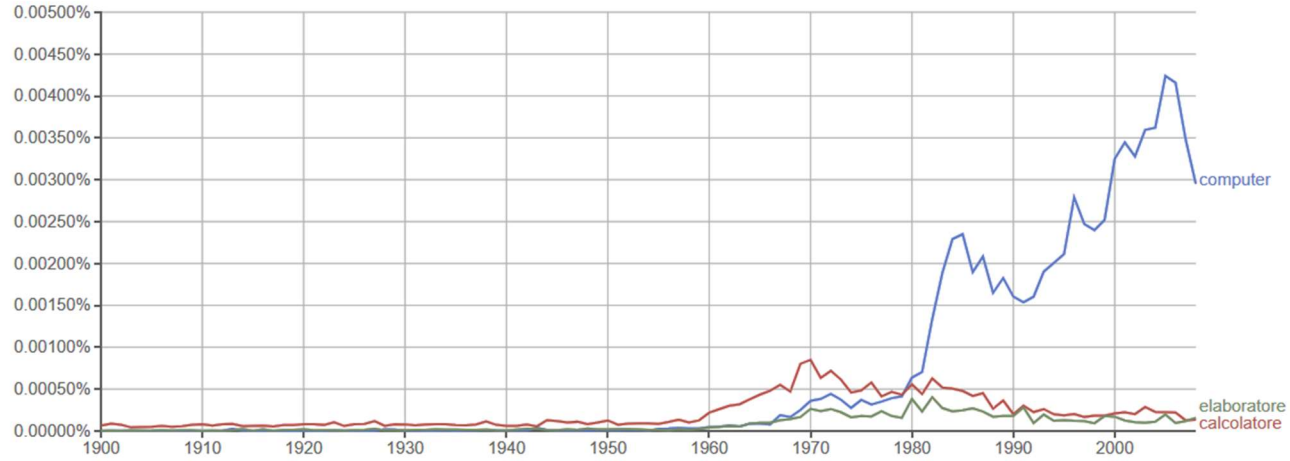
Google Books Ngram Viewer

Un interessante strumento alla portata di connessione è "Google Books Ngram Viewer"[6], grazie al quale, inserendo nel box di ricerca una parola o un insieme di parole, si può vedere la presenza della stessa nel corpus linguistico memorizzato in Google Books, con la frequenza negli anni. I libri analizzati datano dal 1500 al 2008 e raccolgono inglese, spagnolo, francese, tedesco, russo, italiano, cinese ed ebraico.
Il dato è normalizzato data la disparità tra libri stampati nell'ultimo secolo ed i secoli precedenti.
C'è da dire che i dati potrebbero non essere accurati (OCR[7] errata) e che la tipologia prevalente catalogata negli ultimi anni è di testi tecnico scientifici (con il conseguente innalzamento del valore di certi vocaboli specialistici). Inoltre trattandosi di una catalogazione di un libro per titolo, indipendentemente dal numero di copie lette, un vocabolo poco usato potrebbe avere lo stesso valore o addirittura superare un altro vocabolo letto da molte più persone. Insomma, non è paragonabile a corpus come CORIS/CODIS[8] e CHILDES e per questo è rischioso usarlo per veri e propri studi linguistici di tipo diacronico o culturale, ma può essere analogamente interessante come una prima analisi superficiale o come punto di partenza per sviluppi successivi.

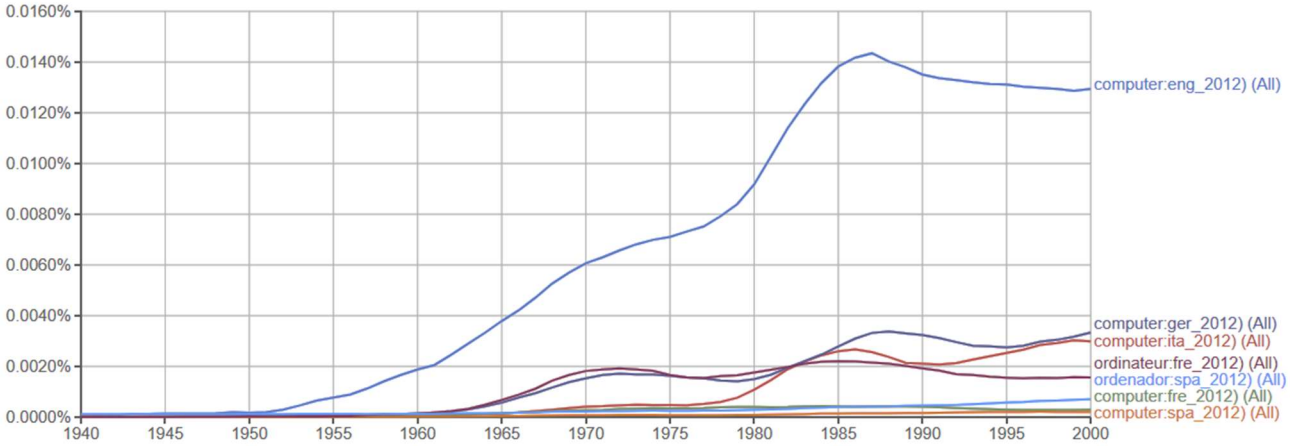
Esempi

Nei grafici l'asse *x* rappresenta l'anno mentre l'asse *y* la frequenza di occorrenza nel corpus. Inoltre ho messo come valore di smoothing[9] 0.

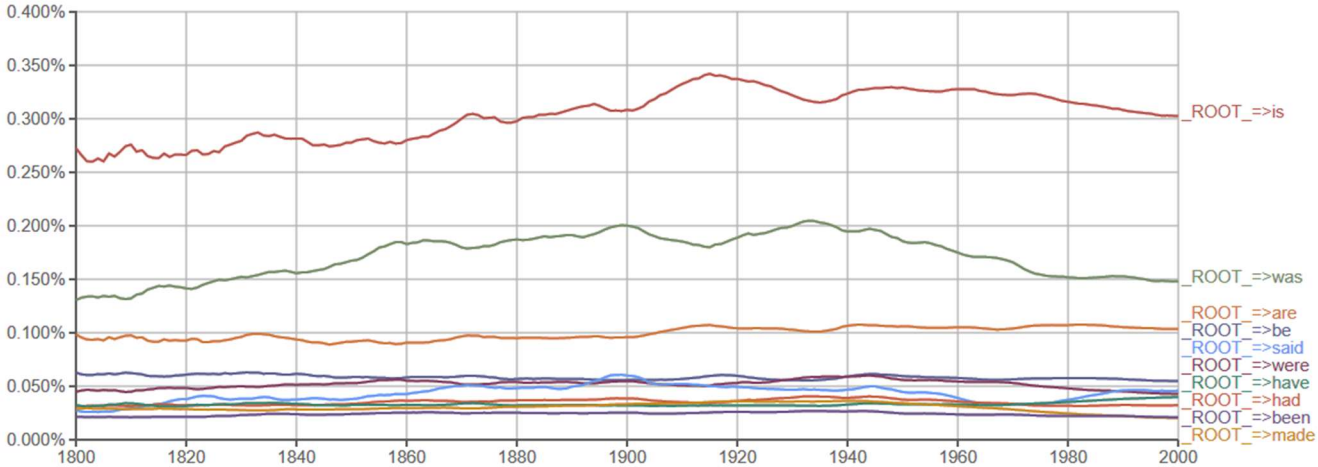
[6] https://books.google.com/ngrams
[7] OCR: Optical Character Recognition, ossia la conversione di testo cartaceo in digitale tramite strumenti elettronici o meccanici.
[8] Corpus di italiano scritto contemporaneo.
[9] Inserendo un valore diverso da 0, come ad esempio 1, il valore mostrato per il 1950 sarebbe la media tra l'anno in questione, un anno precedente e un anno successivo, quindi dei valori annui del 1949, 1950 e 1951.



La parola "calcolatore" nel corpus italiano nel 1970 superava "computer", ma quest'ultima dal 1980 in poi, ha subito un forte aumento d'uso. Può essere che ci sia correlazione con la presentazione del primo PC IMB, datata 1981.



Confrontando varie lingue si può vedere l'uso della parola "computer" e delle corrispettive traduzioni. Si ricorda però che l'inglese è la lingua con il maggior numero di libri catalogati.



Inserendo la query "_ROOT_=>" cercando nel corpus inglese si possono vedere le dieci più frequenti occorrenze di verbo principale di frase.*

Ovviamente queste ricerche sono lungi dall'essere scientifiche, e Google Ngram Viewer è ben lontano dall'essere un metodo affidabile di ricerca, ma se non altro è una interessante strada da percorrere, che si propone alla portata di tutti e perché no, anche ludica e stimolante.

(Google, 2013)

L'unione

Nonostante la posizione di Chomsky, la linguistica dei corpora ha avuto un notevole sviluppo, soprattutto con il progresso dei sistemi informatizzati, che ne ha permesso uno sfruttamento ottimale.
Grazie all'informatica infatti è possibile memorizzare enormi quantità di testo, analizzarle velocemente, modificarle facilmente e renderle disponibili a moltissime persone.
Gli approcci statistico-probabilistici hanno quindi raggiunto una certa robustezza (resistenza al rumore[10] e ad input parziali o incompleti) e proprio questa caratteristica gli ha permesso di passare in vantaggio.
Altro punto a loro favore è la portabilità, ossia la possibilità di applicare il modello a diversi domini (lingue, linguaggi specifici) modificandone al minimo la struttura.
Ma in realtà per capire il valore di questo metodo basta pensare a tutte quelle caratteristiche incalcolabili elencate precedentemente, che guardando statisticamente all'occorrenza di determinate parole nei corpora, non come singole, ma prese nel contesto, permette di analizzarle correttamente e facilita compiti di traduzione automatica.

Oggi si è arrivati all'integrazione dei due metodi nello Statistical Natural Language Processing, in cui insieme alle regole grammaticali vengono anche considerate le statistiche di frequenza per sopperire a tutte le situazioni di difficoltà in cui la semplice applicazione di regole non è sufficiente.

Un esempio applicativo è Google Translate che nel 2006 è stato lanciato come un programma di traduzione automatica statistica, basato su modelli di traduzione creati dai corpora (Och, 2006). Nel settembre 2016 Google ha poi annunciato il cambiamento di tecnologia passando ad un sistema chiamato Google Neural Machine Translation (GNMT) che traduce intere frasi alla volta e si basa su un sistema di auto-apprendimento (Turovsky, 2016).
Questo esempio rafforza ancora di più il fatto che la parola venga interpretata nel contesto, più che come singola unità.

Applicazioni

Sembra quasi superfluo fare una lista delle possibili applicazioni e dei benefici derivanti dalla linguistica computazionale. Oltre alla precedentemente citata traduzione automatica e alle tecnologie già largamente in uso come dizionari elettronici e sistemi di correzione ortografica, si stanno diffondendo sempre più tecnologie come il riconoscimento vocale (ad esempio l'assistente Google[11]) che può avere una funzione marginale ("perché *parlare* allo smartphone quando posso scriverci?") ma diventa fondamentale per tutte le persone con disabilità, ad esempio i non vedenti, che possono quindi usare un computer con audio-comandi o ascoltare un libro, un documento o una pagina web letta dal terminale; per le persone con problemi uditivi, per le quali poter vedere per iscritto un testo orale (in una sorta di dettato) può davvero fare la differenza, o più semplicemente può essere interessante l'implementazione di queste tecnologie in determinati ambienti e situazioni, come ad esempio nelle automobili in cui può avere l'effetto di contribuire a distrarre meno il conducente. Senza dimenticare ambienti come i motori di ricerca, che basandosi sul linguaggio, ad una maggiore comprensione di quest'ultimo farebbero corrispondere una individuazione più efficace delle informazioni richieste.

Conclusioni

È forse proprio il lavoro in simbiosi di metodologia *rule-based* e statistica, la strada migliore da percorrere, e potrebbe anche essere quella che adoperiamo noi nei nostri linguaggi naturali ed essere quindi la chiave per una completa comprensione delle nostre tecniche.

[10] Con rumore si intendono errori e aberrazioni nel linguaggio, ossia un disturbo.
[11] Basato sull'NLP e disponibile in italiano dal 1 novembre 2017 (Chow, 2017).

In molte situazioni infatti probabilmente anche noi ci basiamo su tecniche statistiche o di frequenza. Basti pensare alla distinzione tra lingue diverse che viene fatta fin dai primi giorni basandosi sul ritmo (Guasti, 2007), alla distinzione di parole diverse attraverso un'analisi del flusso del parlato (frequenza di determinati suoni vicini e struttura della parola correlata con la complessità sillabica), al fatto che pochissime parole compongono la stragrande maggioranza dei testi[12] e proprio in tutte quelle situazioni "incalcolabili" in cui il contesto la fa da padrone.

Inoltre il fatto che la linguistica computazionale sia dominata dalla contrapposizione, in parte ora mediata, tra modelli formali di tipo *rule-based* e modelli probabilistici, orientati sui dati empirici, riproduce per molti aspetti la distinzione chomskyana tra competenza ed esecuzione. E forse proprio con l'ausilio di queste tecnologie, o con lo studio e la ricerca che ne conseguiranno, si scoprirà la spaccatura e/o l'unione tra di esse.

[12] In numero variabile, le prime 100 parole coprono il 60% e le prime 1000, l'85% (Guiraud, 1954).

Bibliografia

Bortolini, U., & Pizzuto, E. (1997). *Il progetto CHILDES-Italia : contributi di ricerca sulla lingua italiana.* Tirrenia, Pisa: Edizioni del Cerro.

Chiari, I. (2005). *Informatica e lingue naturali.* Roma: Aracne.

Chiari, I. (2007). *Introduzione alla linguisitica computazionale.* Bari: Laterza.

Chomsky, N. (1957). *Le strutture della sintassi.*

Chow, E. (2017, novembre 1). *The Google Assistant on phones, now in Spanish and Italian*. Tratto da https://blog.google/products/assistant/google-assistant-on-phones-now-in-new-languages/

De Mauro, T. (2000). *Dizionario della lingua italiana.* Milano: Paravia.

Google. (2013). *What does the Ngram Viewer do?* Tratto da https://books.google.com/ngrams/info

Guasti, M. T. (2007). *L'acquisizione del linguaggio, un'introduzione.* Raffaello Cortina.

Guiraud, P. (1954). *Les caractéres statistiques du vocabulaire.* Presses Universitaires de France.

Och, F. (2006, aprile 28). *Statistical machine translation live*. Tratto da https://research.googleblog.com/2006/04/statistical-machine-translation-live.html

Turovsky, B. (2016, novembre 15). *Found in translation: More accurate, fluent sentences in Google Translate*. Tratto da https://blog.google/products/translate/found-translation-more-accurate-fluent-sentences-google-translate/