


Ciao! **Buon 2021**, un po' in ritardo, a tutte e tutti!

Spero che abbiate passato delle vacanze divertenti, nei limiti di quello che abbiamo potuto fare a causa delle restrizioni di questo periodo. Ha nevicato lì? Qui a Pisa purtroppo non nevicava mai, ma le montagne dell'Appennino sono tutte imbiancate! Ho letto la vostra lettera durante le vacanze, mi avete fatto tantissime domande! Le ho riportate nel testo in corsivo, provo a rispondere, e spero di riuscire a soddisfare la vostra curiosità ;)



La mia scuola, il **liceo classico** M. Pagano. Ho anche pensato a tre aggettivi per descrivermi: socievole, riflessiva, pigra 

Intanto, mi avete chiesto *che scuola ho frequentato e perché*: ho fatto il liceo classico, nella mia città di origine, Campobasso. Non saprei dire di preciso perché l'ho scelta, mi piacevano tante materie ma mi piaceva anche moltissimo **leggere**, e forse per questo ho pensato che al liceo classico avrei trovato persone con la mia stessa passione. Negli anni in realtà mi sono pentita della mia scelta, ho trovato tantissime amici e amiche con cui condividere la lettura e altri interessi fuori dalla mia scuola, e sono proprio alcune di queste amicizie quelle che durano ancora oggi, a più di 10 anni e tanti chilometri di distanza.

Quello che ricordo della fine del liceo è che non avevo assolutamente più voglia di studiare materie letterarie, e anzi avevo voglia di provare qualcosa che fosse il più **nuovo e diverso** possibile: forse proprio per questo ho scelto un corso di laurea composto da così tante anime diverse, e per ironia della sorte ho riscoperto dopo qualche anno una passione per la grammatica, uno dei pilastri del liceo classico.

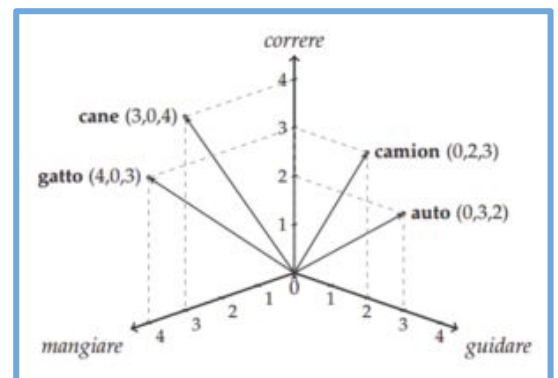
Ma torniamo alla linguistica computazionale: ho visto che avete fatto un bel po' di ricerche!

Provo a rispondere ai vostri dubbi: *intanto, un linguista computazionale è più un linguista o un programmatore?* Una possibile risposta è che dipende dalle domande a cui si sta cercando di rispondere. A volte usiamo il computer solo come **strumento tecnico** per aiutarci ad elaborare i dati in modo più efficace, ma cerchiamo di rispondere a **domande linguistiche** (per esempio, "come nasce una nuova parola?"). Altre volte, invece, dobbiamo inventare nuove soluzioni "computazionali", sviluppando algoritmi per testare ipotesi o per creare applicazioni, e in quel momento siamo forse più programmatrici e programmatori che linguiiste o linguisti. E non sono questi gli unici due campi coinvolti: ci sono anche psicologi, filosofi, traduttori... Le discipline interessate (e soprattutto interessanti) sono tantissime! *Bisogna conoscerle tutte?* No, sarebbe impossibile. Ma bisogna essere pronte e pronti a **confrontarsi** con persone che hanno studiato materie diverse, e che hanno quindi conoscenze ed approcci diversi: è proprio da questi incontri che nascono spesso le domande più interessanti!

Mi avete anche chiesto *che tipo di domande ci poniamo per iniziare una ricerca*.

Ho pensato che il modo migliore per rispondere potesse essere quello di mostrarvi dei veri esempi di domande di ricerca: generalmente vengono individuati **tre filoni principali** nell'ambito della linguistica computazionale, nella prossima pagina proverò a citarvi brevemente un esempio per ognuno di questi. Purtroppo gran parte della produzione scientifica in questo campo viene pubblicata in inglese: questo permette a chi lavora nell'ambito di poter accedere facilmente alle ricerche pubblicate ogni anno, ovunque nel mondo, ma potrebbe rappresentare un ostacolo per chi vuole provare a capirci qualcosa da fuori.

Ho provato a riassumere il nocciolo della questione in italiano, ma vi lascio anche il link agli articoli corrispondenti: non preoccupatevi però se il linguaggio o il contenuto vi risulta difficile da seguire, spesso lo è anche per gli stessi ricercatori e ricercatrici che lavorano in campi diversi!



Questo è un esempio di rappresentazione del **significato delle parole**, simile a quello utilizzato da Katrin Erk e Sebastian Pado nella loro ricerca. In questo caso i sostantivi vengono rappresentati attraverso **vettori in tre dimensioni**, che corrispondono ai tre verbi *mangiare*, *correre*, *guidare*: la lunghezza e l'orientamento dei vettori dipende da quante volte un sostantivo occorre con uno dei verbi scelti. In questo modo, computazionalmente possiamo valutare se due sostantivi sono simili calcolando la **distanza** tra i loro vettori.

## 1. Creazione di risorse:

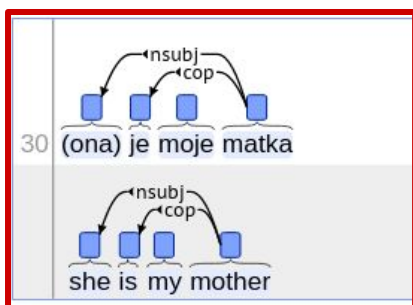
Per fare sì che gli strumenti computazionali funzionino per una certa lingua, è necessario avere a disposizione dei dati relativi a quella lingua, e questi dati devono essere rappresentativi delle diverse varietà e fenomeni che vogliamo studiare (ad esempio: i giornali possono fornire grandi quantità di dati, ma di un genere linguistico molto specifico, la lingua scritta formale, che ci fornisce poche informazioni riguardo allo slang giovanile o al linguaggio medico specialistico).

**Documentare e conservare** è fondamentale anche perché le lingue sono oggetti estremamente mutevoli, la lingua che parliamo quest'anno è un po' diversa da quella che parlavamo qualche anno fa (pensate a tutte le nuove parole introdotte durante la pandemia!) e **tantissime lingue scompaiono ogni anno**, facendo così scomparire anche il mondo culturale ad esse collegato.

È necessario però che le modalità di conservazione dei dati siano condivise da ricercatrici e ricercatori, in modo da confrontare facilmente le informazioni contenute nelle banche dati, potendole anche modificare ed ampliare, generando così nuove conoscenze.

Le raccolte di dati linguistici così formate prendono il nome di **corpus** (*corpora*, al plurale).

Un progetto per la standardizzazione e la condivisione di corpora è quello delle **Universal Dependencies**: un gruppo di ricerca ha creato una lista di categorie linguistiche (es., *nome*, *verbo*, *soggetto*, *complemento diretto*) universali, fornendo anche delle accurate linee guida per rintracciarle nei testi, e questo ha permesso di "tradurre" moltissime risorse già esistenti in questo linguaggio universale, rendendole così confrontabili ed esplorabili in contemporanea.



Ecco un esempio dei dati presenti tra le **Universal Dependencies**: in questo caso vengono mostrate le due categorie **soggetto** (nell'immagine *nsubj*) e **copula** (nell'immagine *cop*) in due lingue, inglese e ceco.

Link:

<https://universaldependencies.org>

## 2. Creazione di modelli:

Il secondo ambito di ricerca è la creazione di modelli computazionali per **simulare alcuni comportamenti linguistici**: poiché la lingua è un oggetto complesso, che usiamo spesso in combinazione con altre nostre capacità (ad esempio i gesti o l'espressione del viso, ma anche la capacità di comprendere le intenzioni altrui a prescindere da ciò che viene pronunciato), attraverso i modelli computazionali è possibile **isolare** meglio alcuni fenomeni.

Un problema molto studiato, per esempio, è quello della **polisemia delle parole**: alcune parole hanno palesemente più significati, come ad esempio *credenza*, che può indicare sia una *convinzione* che un *mobile per riporre le stoviglie*. In casi come questo, i due significati non sembrano avere relazione tra loro e la questione sembra semplice. Ma esistono casi più sottili di polisemia, come ad esempio il caso di *bicchiere*. Se confrontiamo le due frasi "ho riempito un bicchiere di acqua" e "ho bevuto un bicchiere di acqua" ci rendiamo conto che ci riferiamo nel primo caso al contenente (il bicchiere di vetro), nel secondo al contenuto (l'acqua contenuta nel bicchiere). Una cosa simile accade per i verbi: "prendere un bel voto", "prendere la palla", "prendere il raffreddore" indicano azioni piuttosto diverse, anche se il verbo usato è sempre "prendere".

Katrin Erk e Sebastian Pado hanno studiato il problema costruendo un modello del significato di verbi che prevedesse però di essere *adattato* al significato dei loro soggetti ed oggetti: invece che prevedere un solo significato per *prendere*, hanno costruito una rappresentazione in base al complemento oggetto con cui il verbo si compone. Nel caso di "prendere la palla", "prendere" è stato così ritenuto simile a "tirare, scambiarsi", e altre azioni che si possono fare con la palla, ma diverso da *ammalarsi*, che appartiene al campo semantico del *raffreddore*. Hanno dimostrato che costruendo un modello computazionale in questo modo si ottengono sinonimi migliori, e che quindi quando un verbo si compone con un complemento oggetto, c'è un'interazione **tra i due significati**, come se un po' del significato del complemento oggetto venisse trasferito sul verbo stesso.

Link:

<https://www.aclweb.org/anthology/D08-1094.pdf>

## 3. Analisi di dati:

L'ultimo ambito da citare riguarda l'analisi di dati linguistici con strumenti computazionali, per **estrarre conoscenze** di vario genere. Questo può servire a orientare ricercatrici e ricercatori di vari campi verso nuove indagini, e a supportare o smentire le loro ipotesi.

Un esempio è uno studio condotto da Francesca Chiusaroli e vari colleghe e colleghi sul linguaggio scritto sui social network da parte di giovani universitarie e universitari. Studiando il lessico usato nel periodo precedente al primo lockdown di marzo e durante il lockdown stesso, hanno tra le altre cose evidenziato un uso significativamente maggiore di termini legati alla sfera della **nostalgia** (*tornare*, *manca*, *mancano*, *mancate*, *mancare*), un indizio che potrebbe essere utile per svolgere ulteriori analisi, non solo in campo linguistico ma anche sociologico o psicologico.

	gennaio	febbraio	marzo	aprile	maggio
parole della pandemia	0.22%	3.69%	10.33%	9.18%	7.44%
parole della DAD	1.95%	2.99%	18.98%	9.44%	8.68%
verbi della nostalgia	0.8%	0.74%	3.34%	5.61%	6.2%

Le percentuali di vocaboli analizzati nello studio.

Link:

[http://ceur-ws.org/Vol-2769/paper\\_66.pdf](http://ceur-ws.org/Vol-2769/paper_66.pdf)

Da questi studi vi sarà sicuramente chiaro che *non è necessario per un o una linguista conoscere tante lingue*, sarebbe un po' come chiedere a un medico se ha preso tante malattie: sicuramente per tante e tanti di noi le lingue sono una passione, e qualcuno usa anche queste conoscenze anche nella sua ricerca, ma l'obiettivo sia della linguistica che della linguistica computazionale è in generale di studiare la lingua come capacità umana universale, cioè comune a tutte le popolazioni della terra, a prescindere dalle differenze esistenti tra le varie lingue.

Per dimostrarvi come sia possibile, vi propongo un gioco\*: trovate qui di seguito alcune frasi in **Apinayé**, una lingua a rischio d'estinzione parlata in Brasile da circa 2000 persone, in sei soli villaggi.

Immaginate di voler scoprire qualcosa sul funzionamento dell'Apinayé e sulla sua **sintassi** e, per farlo, avete raccolto questi esempi da un parlante nativo.

\*il problema è adattato da uno dei problemi proposti a studenti e studentesse durante le Olimpiadi Internazionali di Linguistica Computazionale, e raccolti in "Radev, Dragomir, and James Pustejovsky. 2013. *Puzzles in Logic, Languages and Computation: The Red Book*. Springer".

**Kukrĩ kokoi.** = La scimmia mangia  
**Ape kra.** = Il bambino lavora  
**Ape kokoi ratš.** = La grande scimmia lavora  
**Ape mĩ mĩtš.** = L'uomo buono lavora  
**Ape mĩtš kra.** = Il bambino lavora bene  
**Ape punui mĩ pišetš.** = L'uomo vecchio lavora male

Per cominciare, riuscireste a tradurre in italiano la frase **Ape ratš mĩ mĩtš**?

Traducendo, potreste aver fatto alcune osservazioni sugli aspetti che tipicamente interessano a linguiste e linguisti quando cercano di studiare la struttura di una lingua. Qui alcuni spunti:

- Siete riusciti a capire quali parole sono **verbi**, **nomi**, **aggettivi**?
- Avete invece trovato degli **articoli** in Apinayé?
- L'**ordine delle parole** nella frase è uguale in Italiano e in Apinayé?
- Avete notato qualcosa di particolare sul comportamento della parola **mĩtš**?

```
# Python program to check
# if a string is palindrome
# or not
st = 'malayalam'
j = -1
flag = 0
for i in st:
    if i != st[j]:
        j = j - 1
        flag = 1
        break
    j = j - 1
if flag == 1:
    print("NO")
else:
    print("Yes")
```

Un esempio di codice in **Python** per controllare se una sequenza di caratteri (in questo caso, *malayalam*) è un palindromo.

È vero però che le linguiste e i linguisti computazionali sanno programmare, e per farlo usano dei **linguaggi di programmazione**, ovvero insiemi di istruzioni che possono essere interpretate da un computer e tramite le quali si possono definire procedure per elaborare i dati.

Mi avete chiesto *qual è la lingua più usata in questi sistemi*, intendevate questi linguaggi?

Voi vi siete mai cimentati con il **coding**? Vi piacerebbe provare?

I linguaggi di programmazione usati per la ricerca sono vari, io di solito uso uno dei più comuni, **Python**. Esistono però anche tantissimi strumenti a disposizione di ricercatrici e ricercatori per esplorare facilmente i corpora e le risorse già esistenti, e che non richiedono la conoscenza di particolari linguaggi di programmazione e che sono spesso usati anche da traduttrici e traduttori per integrare le conoscenze raccolte nel dizionario con usi particolari che spesso non sono riportati nel dizionario, per esempio perché troppo colloquiali.

Ma Python è un linguaggio "moderno" e voi stessi mi avete chiesto a *quando risale la linguistica computazionale*.

La storia della linguistica computazionale ha infatti tante e diverse radici: dal punto di vista internazionale, l'analisi automatica del linguaggio è iniziata proprio quando **Alan Turing**, il padre dell'informatica moderna, ha costruito una macchina per decifrare i messaggi cifrati dell'esercito nazista durante la seconda guerra mondiale. Negli anni immediatamente successivi alla seconda guerra, la fiducia nel fatto che questo tipo di modelli automatici potessero risolvere compiti di **traduzione automatica** (in particolare, importanti per gli statunitensi per "rubare" le ricerche scientifiche russe durante gli anni della guerra fredda) portò tantissimi finanziamenti e sviluppi, che diedero ufficialmente il via a questo ambito di ricerca. Le speranze furono tuttavia ben presto disattese: ci si rese conto che il problema era ben più **complesso** di quello che inizialmente si pensasse, e nacque la "Association for Computational Linguistics" (ACL), l'associazione scientifica internazionale che raccoglie tutte e tutti coloro che si impegnano nello studio computazionale del linguaggio umano.



La macchina **Enigma**, progettata da Alan Turing per decifrare i messaggi nazisti.



Padre Busa con un macchinario della **IBM**, azienda che lo ha supportato nelle sue ricerche e che per anni è stata un punto di riferimento per modelli di linguistica computazionale e di intelligenza artificiale

In Italia, però, la storia internazionale si è incontrata con una ben radicata tradizione locale, iniziata da **Padre Roberto Busa**, un gesuita che per i suoi studi su Tommaso d'Aquino sentì la necessità di effettuare delle analisi più approfondite dei testi che stava esaminando. Nel 1946, scrivendo la sua tesi all'Università Gregoriana, costruì così il primo corpus della storia.

La storia dell'informatica e quella della linguistica computazionale sono, insomma, strettamente collegate 😊

Da Padre Busa ad oggi, tuttavia, ne è passata di acqua sotto i ponti, la vostra domanda su *che risultati sperano di raggiungere i ricercatori* mi ha fatto pensare: abbiamo ancora gli stessi obiettivi? È una domanda davvero difficile, a cui credo sia impossibile dare una risposta soddisfacente. Quando pensiamo a uno scopo, a dei risultati, spesso pensiamo a qualcosa di pratico e concreto, ma non tutta la ricerca prevede dei risultati tangibili. I prodotti che vediamo, perché arrivano alla nostra vita quotidiana come Alexa o un vaccino, sono il risultato di decenni di **lavoro "collettivo"** da parte di studiosi e studiosi, di scambi di idee, di **fallimenti** e di nuove proposte. E, io credo, non tutte le persone che hanno fatto parte del processo avevano un risultato concreto in mente o un vero e proprio obiettivo. Molto più spesso quello che ricercatrici e ricercatori hanno in mente è una domanda, dettata dalla curiosità e dalla voglia di spiegare un pezzettino del mondo che ci circonda. È importante conservare questa curiosità, farsi guidare da essa e fare attenzione a quali siano le domande che spingono la ricerca, piuttosto che ai risultati da raggiungere, perché spesso i risultati ci possono sembrare piccoli e insignificanti rispetto agli sforzi che abbiamo fatto per raggiungerli, ma dobbiamo ricordarci che il vero obiettivo è quello di contribuire a un **processo di conoscenza** collettivo, e che con le nostre domande e i piccoli risultati che abbiamo raggiunto abbiamo gettato le basi perché qualcuno dopo di noi possa porsi nuove e più complesse domande.



Di una cosa comunque possiamo essere sufficientemente certi: **non** vogliamo creare un'intelligenza artificiale *come quelle dei film di fantascienza*.

**HAL 9000**, una delle prime intelligenze artificiali apparse sul grande schermo.

Ciò che vediamo nei film è infatti di solito una rappresentazione di quella che viene chiamata un'**intelligenza artificiale generale**, ovvero un modello che simuli il comportamento umano in tutti i tuoi aspetti. Questo, oltre ad essere al momento impossibile perché sono ancora tantissime le cose che non sappiamo - e che dunque non sapremmo riprodurre - sul funzionamento del nostro cervello, potrebbe non essere neppure utile. Quello che ci aspettiamo dai modelli computazionali non è, infatti, che *si sostituiscano all'uomo* o che lo imitino in tutto e per tutto, come avete notato voi stessi, ma che ci supportino in alcune specifiche situazioni in cui fare lo stesso lavoro, per noi umani, richiederebbe troppo tempo o sarebbe troppo logorante.

È plausibile che nel futuro avremo sempre più "aiutanti robotici", ma questo può avere dei lati positivi: significa che tanti lavori noiosi o pericolosi potranno essere svolti da macchine, mentre noi avremo più spazio per la creatività e il ragionamento. Questo riguarda la robotica in generale, ma include anche la linguistica computazionale: uno degli esempi che mi vengono in mente è quello dei motori di ricerca, che hanno permesso di automatizzare la ricerca di informazioni (pensate di dover andare in biblioteca ogni volta che siete curiose o curiosi su qualcosa, e poi magari la biblioteca non ha il libro che cercate e deve chiederlo in prestito a sua volta a un'altra biblioteca...), aumentando così la quantità di materiale che possiamo raccogliere e processare e lasciandoci più tempo per scegliere, riflettere, integrare, collegare.

C'è un'ultima domanda che mi avete fatto, che richiederebbe una risposta particolarmente lunga, ed è *come funziona esattamente l'NLP*. Per farvi un'idea vi lascio il link a una video-intervista fatta alla dottoressa Giulia Venturi, una ricercatrice dell'Istituto di Linguistica Computazionale del CNR di Pisa, e spero di potervene parlare ancora nella prossima lettera!

**Link:**

<http://www.raiscuola.rai.it/programma-unita/memex-doc-vita-da-ricercatore-pt-16-giulia-venturi/322/43133/default.aspx>

A presto,  
Ludovica

P.S. Datemi del tu! Non sono così vecchia 😊