



Centro per la Riforma dello Stato

# LE NUOVE FABBRICHE DEL LINGUAGGIO

COME FUNZIONANO, CHI LE POSSIEDE, COME OCCUPARLE

Guido Vetere

15/11/21



## FACEBOOK IN MYANMAR, 2018

*«La velocità con cui vengono segnalati contenuti inappropriati [bad] in birmano, che si tratti di incitamento all'odio o disinformazione, è bassa. [...] **Quindi stiamo investendo molto nell'intelligenza artificiale in grado di segnalare in modo proattivo i post che infrangono le nostre regole**»*

<https://about.fb.com/news/2018/08/update-on-myanmar/>

# LINGUAGGIO E INTELLIGENZA ARTIFICIALE

Il mezzo **è** il messaggio:

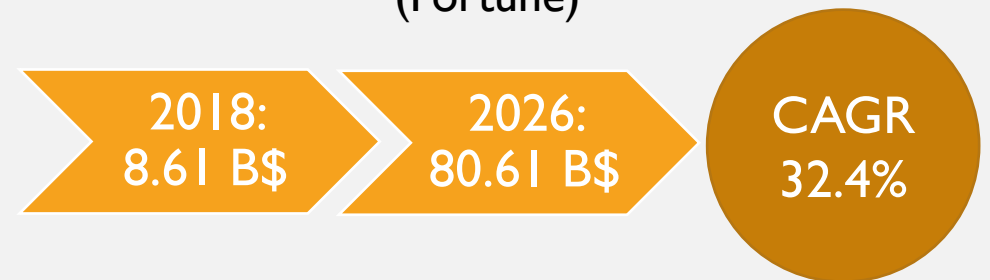
- XX sec: **è** = definisce le condizioni
- XXI sec: **è** = definisce i contenuti



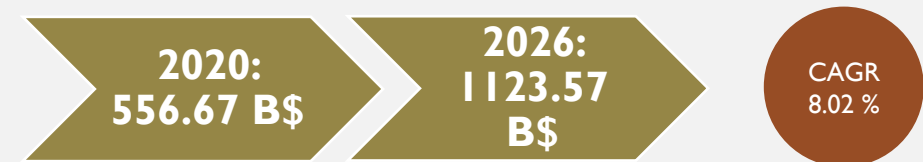
# L'INDUSTRIA DEL LINGUAGGIO

Ricerca di informazione	<ul style="list-style-type: none"><li>• Motori di ricerca</li><li>• Gestione dei contenuti</li><li>• Business intelligence</li></ul>
Traduzione automatica	<ul style="list-style-type: none"><li>• Commercio elettronico</li><li>• Editoria</li><li>• Relazioni istituzionali</li></ul>
Analisi e classificazione	<ul style="list-style-type: none"><li>• Sicurezza</li><li>• Reti sociali</li><li>• Pubblicità</li></ul>
Interazione	<ul style="list-style-type: none"><li>• Supporto alla clientela</li><li>• Relazioni col pubblico</li><li>• Assistenza personale</li></ul>
Generazione di testi	<ul style="list-style-type: none"><li>• Gioco e intrattenimento</li><li>• Editoria</li><li>• Lavoro creativo</li></ul>

## Global Natural Language Processing Market (Fortune)



## Global IT Services Market (Research and Markets)



# USI SOCIALI DELLE TECNOLOGIE LINGUISTICHE (ESEMPIO)

## Discover New Insights About the Novel Coronavirus

Quickly explore the latest literature using these open tools built by the team at Allen Institute for AI.



### Download CORD-19

The COVID-19 Open Research Dataset (CORD-19) is a growing resource of scientific papers on COVID-19 and related historical coronavirus research.

[Download →](#)



### Adaptive Research Feed

Personalize your free AI-powered Research Feed to get coronavirus research recommendations.

[Stay Up To Date →](#)



### Recent Research

Query the Semantic Scholar corpus for the latest CORD-19 research sorted by recency.

[View Research →](#)



### SPIKE-CORD

A powerful sentence-level, context-aware, linguistically informed system for extracting important information from a large corpus of COVID-19-related text.

[View SPIKE-CORD →](#)



### SciSight

Visually investigate associations between concepts appearing in the scientific literature contained in CORD-19.

[View SciSight →](#)



### SciFact

Find out whether published scientific research supports or contradicts claims about COVID-19.

[View SciFact →](#)

# USI PERVERSI DELLE TECNOLOGIE LINGUISTICHE

## Microsoft sacks journalists to replace them with robots

Users of the homepages of the MSN website and Edge browser will now see news stories generated by AI



From July, the MSN homepage will no longer feature news stories produced by journalists at PA Media, formerly the Press Association. Photograph: Alamy

Dozens of journalists have been sacked after [Microsoft](#) decided to replace them with artificial intelligence software.



Jim Waterson, 30 Maggio 2020

[www.theguardian.com/technology/2020/may/30/microsoft-sacks-journalists-to-replace-them-with-robots](https://www.theguardian.com/technology/2020/may/30/microsoft-sacks-journalists-to-replace-them-with-robots)

# POSSIBILI USI DELLE TECNOLOGIE LINGUISTICHE ANCORA PIÙ PERVERSI



CNET Highlights

212.000 iscritti

28/10/21: A Facebook Connect, il CEO Mark Zuckerberg svela la visione della sua azienda su come gli utenti interagiranno e socializzeranno all'interno del Metaverso.

<https://www.youtube.com/watch?v=b9vVShsmE20>

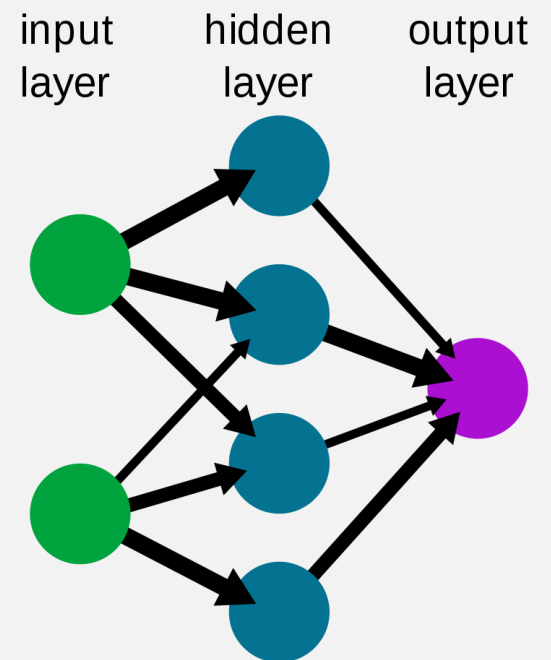


# TECNOLOGIE LINGUISTICHE E INTELLIGENZA ARTIFICIALE

Le tecnologie della lingua (NLP) si basano oggi principalmente sulle capacità di apprendimento delle «reti neurali»

- Le reti neurali sono **classificatori** addestrati su campioni di dati (*training set*)
  - «Imparano» a produrre un *output* a fronte di un *input* mediante il rinforzo (o l'indebolimento) delle connessioni tra elementi atomici detti «neuroni»
  - Richiedono una specifica codifica dei dati, una complessa architettura interna, e una funzione di calcolo dell'errore nella predizione del risultato (*loss function*)

A simple neural network



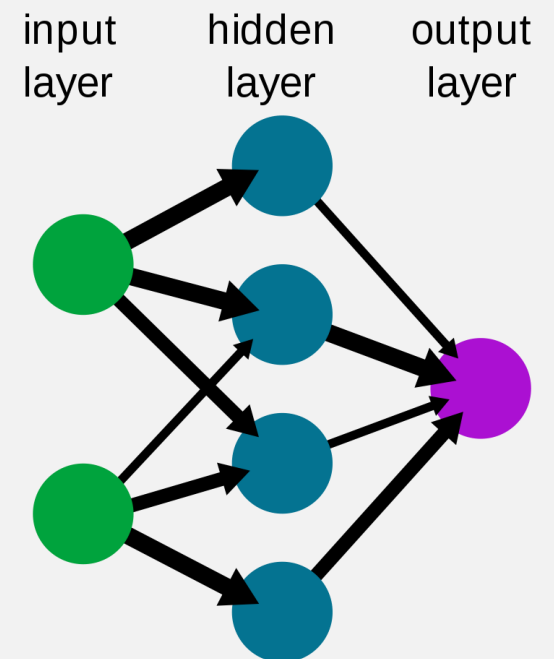


# TECNOLOGIE LINGUISTICHE E INTELLIGENZA ARTIFICIALE

Vi sono diversi modi per addestrare le «reti neurali», i quali rientrano in due classi:

- Apprendimento supervisionato
  - Il *training set* è annotato da esseri umani, i quali forniscono esempi (positivi e/o negativi) di quale *output* si attende a fronte di quale *input*
- Apprendimento non supervisionato
  - Il *training set* è costituito da dati «grezzi», l'associazione tra *input* e *output* è intrinseca

A simple neural network

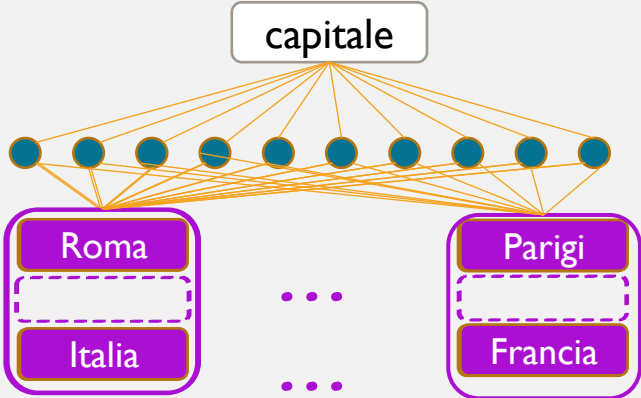


**Il salto di qualità delle tecnologie linguistiche è dovuto principalmente all'apprendimento non supervisionato**

# MODELLAZIONE LINGUISTICA NON SUPERVISIONATA

Le unità del linguaggio sono rappresentate numericamente

- Per ogni token (parola o frammento), una rete apprende a predire il contesto in cui appare (es. *skip-gram model*)
- Dalle connessioni (parametri) della rete addestrata vengono estratti vettori numerici chiamati *embedding*
- L'insieme degli *embedding* stimati sui testi di una lingua è chiamato «modello linguistico» (*language model*)



Fill-Mask Mask token: [MASK]

Roma è la [MASK] d'Italia. Compute

Computation time on cpu: cached.

capitale	0.932
Capitale	0.042
città	0.004
regina	0.001
porta	0.001

capitale	0.5	0.1	0.9	0.7	0.0	0.0	0.6	0.3	1.0	.
città	0.4	0.2	0.9	0.5	0.1	0.0	0.5	0.5	0.9	.

Le nuove fabbriche del linguaggio, Guido Vetere, CRS - CC-BY

<https://huggingface.co/>

# USO DEI MODELLI LINGUISTICI

*downstream*

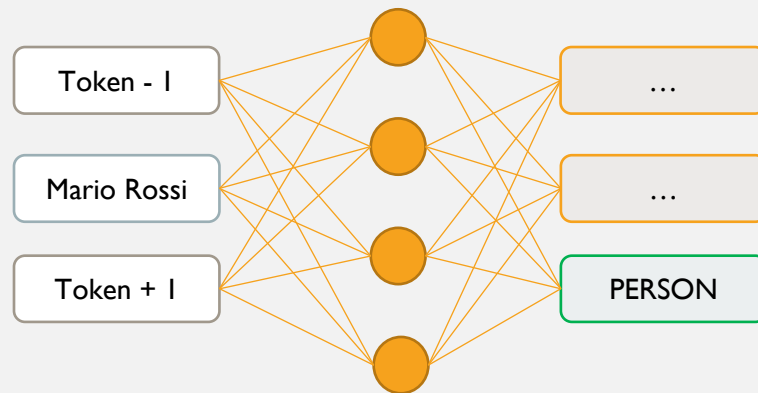
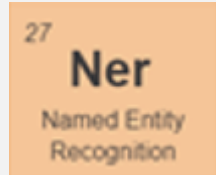
*task*



			29 <b>Pri</b> Price Parser											63 <b>Nex</b> Next Token Prediction	69 <b>Rel</b> Relation Extraction
14 <b>Tok</b> Tokenization	19 <b>Ste</b> Stemming	24 <b>Ngr</b> N-grams	30 <b>Geo</b> Geocoding							43 <b>Trn</b> Training Models	48 <b>Spa</b> Spam Detection	53 <b>Key</b> Keyword Extraction	58 <b>Syn</b> Wordnet Synsets	64 <b>Rep</b> Report Writing	70 <b>Qan</b> Question Answering
15 <b>Voc</b> Vocabulary Building	20 <b>Lem</b> Lemmatization	25 <b>Phr</b> Rulebased Phrasematcher	31 <b>Tmp</b> Temporal Parser	35 <b>Sen</b> Sentencizer	39 <b>Ded</b> Deduplication					44 <b>Tst</b> Evaluating Models	49 <b>Sed</b> Sentiment and Emotion Detection	54 <b>Esu</b> Extractive Summarization	59 <b>Dst</b> Distance Measures	65 <b>Tra</b> Machine Translation	71 <b>Cha</b> Chatbot Dialogue
16 <b>Mor</b> Morphological Tagger	21 <b>Nrm</b> Normalization	26 <b>Chu</b> Dependency Nounchunks	32 <b>Nel</b> Named Entity Linking	36 <b>Par</b> Paragraph Segmentation	40 <b>Raw</b> Raw Tekst Cleaning					45 <b>Exp</b> Explaining Models	50 <b>Int</b> Intent Classification	55 <b>Top</b> Topic Modeling	60 <b>Sim</b> Document Similarity	66 <b>Asu</b> Abstractive Summarization	72 <b>Sem</b> Semantic Search Indexing
17 <b>Pos</b> Part-of-Speech Tagger	22 <b>Spl</b> Spell Checker	27 <b>Ner</b> Named Entity Recognition	33 <b>Crf</b> Coreference Resolution	37 <b>Grm</b> Grammar Checker	41 <b>Met</b> Meta-Info Extractor					46 <b>Dpl</b> Deploying Models	51 <b>Cls</b> Text Classification	56 <b>Tre</b> Trend Detection	61 <b>Dis</b> Distributed Word Representations	67 <b>Prp</b> Paraphrasing	73 <b>Kno</b> Knowledge Base Population
18 <b>Dep</b> Dependency Parser	23 <b>Neg</b> Negation Recognizer	28 <b>Abr</b> Abbreviation Finder	34 <b>Anm</b> Text Anonymizer	38 <b>Rea</b> Readability Scoring	42 <b>Lng</b> Language Identification					47 <b>Mon</b> Monitoring Models	52 <b>Mlc</b> Multi-Label Multi-Class Classification	57 <b>Out</b> Outlier Detection	62 <b>Con</b> Contextualized Word Representations	68 <b>Lon</b> Long Text Generation	74 <b>Edi</b> E-Discovery

R. van Zoest: 6 NLP Tasks for Training Data Generation (medium.com)

# ANALISI DI RIFERIMENTI E CONCETTI (NAMED ENTITY RECOGNITION)



- Le reti sono addestrate per classificare ciascun «token» sulla base del contesto in cui appare
- L'addestramento è supervisionato mediante corpora annotati e si avvale normalmente di modelli linguistici
- I corpora annotati si possono ottenere tramite *crowdsourcing*, o sfruttando risorse esistenti (es. Wikipedia)



Il numero delle etichette è in genere molto limitato (<20)

# USO DEI MODELLI LINGUISTICI


*zero-shot*

Modelli  
generali

- Infosfera (multilingue)
- Apprendimento non supervisionato

Generatori

- *Pre-trained transformer*

 OpenAI GPT-3



**Completion**

Generate or manipulate text and code



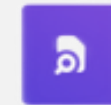
**Fine-tuning** Beta

Train a model for your use case



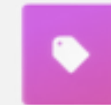
**Question answering** Beta

Generate high-accuracy answers



**Semantic search**

Score text based on relevance



**Classification** Beta

Classify text into different categories

**RETE NEURALE: 175 MLD PARAMETRI**  
**COSTI SVILUPPO: 11 < M\$ > 27 (STIME)**  
**ENERGIA: 190K kWh (85T CO<sub>2</sub>)**

B. Dickson, The GPT-3 economy

<https://bdtechtalks.com/2020/09/21/gpt-3-economy-business-model/>

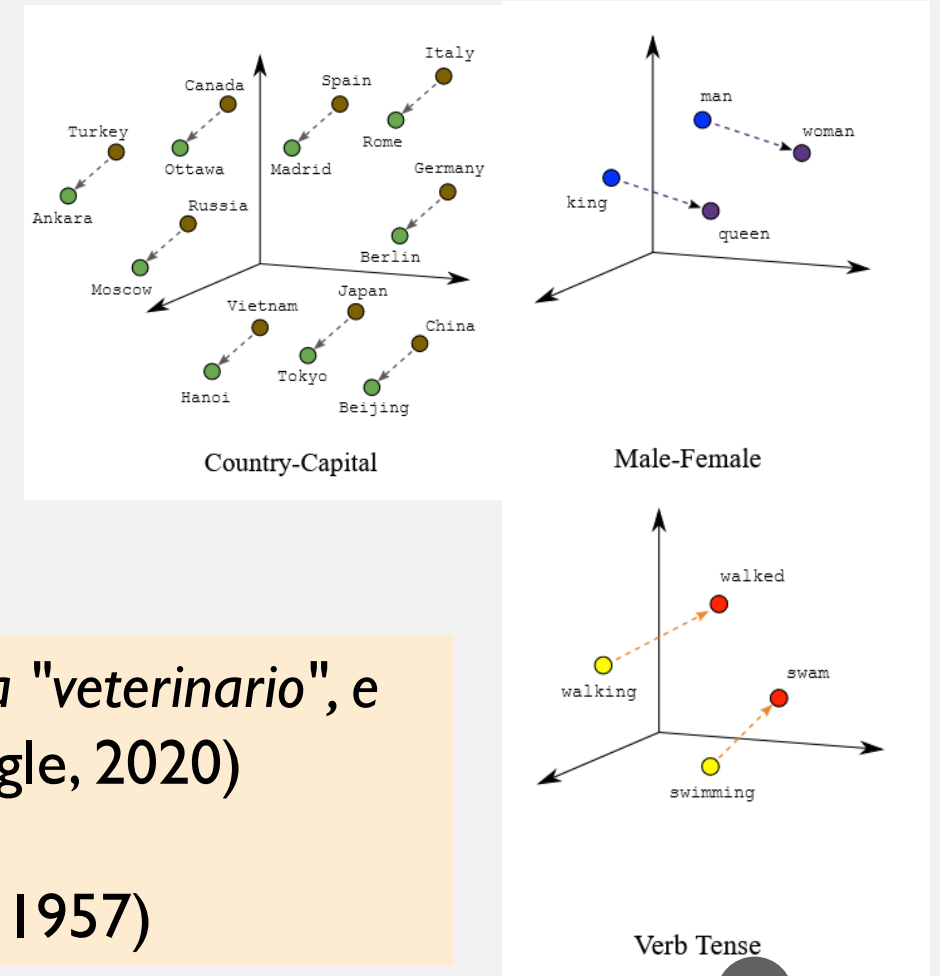
# LA SEMANTICA DEI MODELLI NEURALI

I modelli linguistici neurali considerano il significato in modo **distribuzionale**

- Gli *embedding* di parole che compaiono negli stessi contesti risultano simili
- I loro vettori vengono mappati in uno «spazio» n-dimensionale
- Le relazioni semantiche sono viste come misure di distanza in questo spazio

“Sia "cane" che "gatto" appaiono spesso vicini alla parola "veterinario", e questo fatto riflette la loro somiglianza semantica” (Google, 2020)

“You shall know a word by the company it keeps” (Firth, 1957)



# LA SEMANTICA DEI MODELLI NEURALI



I modelli distribuzionali hanno accesso solo al «significante» del segno linguistico

- La «semantica» delle reti neurali fa emergere solo le relazioni concettuali che hanno effetto sulla distribuzione delle parole nei testi forniti
- Ciò che è fuori dal testo (*hors-texte*) per l'automa non esiste
- Più in generale, non esiste, per l'automa, alcun **processo di significazione**
- Denotazione e connotazione dunque collassano, da cui (tra l'altro) il noto problema dei pregiudizi (*bias*)



# ARCHITETTURE E MODELLI DI BUSINESS

## *cloud*

- Software e modelli linguistici gestiti nei server dei fornitori
- Licenze per l'uso remoto (API)
- Funzionalità *out-of-the-box*
- *Vendor-lock, protezione dati*



Centralizzazione • Monopolio • *Knowledge divide*

## *on-premise*

- Software e modelli linguistici gestiti dagli utilizzatori

### *Sviluppo in-house*

- Open source
- Modelli aperti
- *Competenze*



### *Terze parti*

- Piattaforme Enterprise
- Modelli custom
- *Vendor-lock*



Decentralizzazione • Mercato • *Knowledge commons*

# LINGUE DIVERSE DALL'INGLESE



Lingua	Modelli	Dataset
en	2486	309
es	269	72
fr	261	68
de	199	68
it	76	48

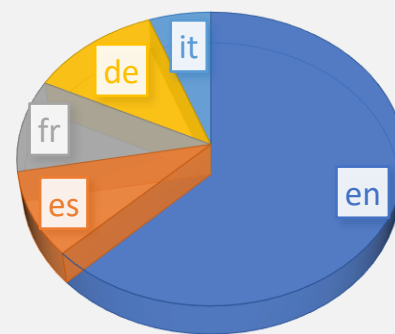
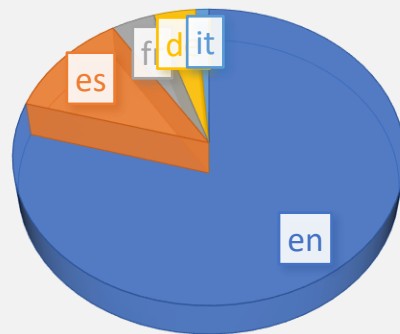
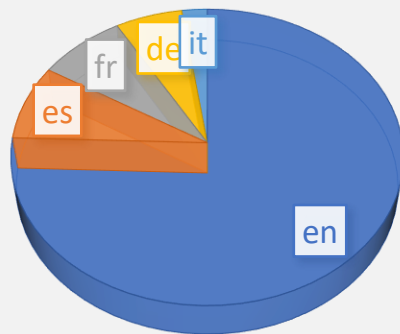


Lingua	Risorse
en	524
es	77
fr	26
de	25
it	8



Lingua	Corpora	Onto\Lex
en	1535	464
es	229	298
fr	250	144
de	295	121
it	139	68

- Le tecnologie linguistiche dipendono oggi dalla disponibilità di dati grezzi o pre-elaborati (modelli)
- Tra i temi del *knowledge divide* c'è lo squilibrio nella produzione di risorse linguistico-computazionali



Studio completo (2012)  
<http://www.meta-net.eu>

# LINGUE DIVERSE DALL'INGLESE

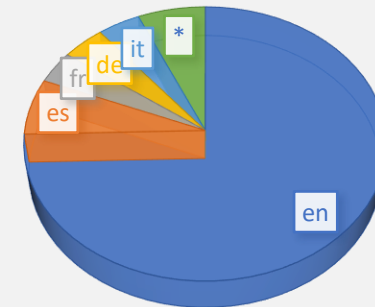
Il «paradigma open» non risolve di per sé il problema del predominio dell'inglese nelle tecnologie linguistiche

Spark NLP models & pipelines	en	es	fr	de	it	*
Named Entities	38	8	5	5	5	
Text Classification	15	0	0	0	0	
Sentiment Analysis	5	0	0	0	0	
Translation	3	0	(1)	(1)	0	
Question Answering	5	0	0	0	0	
Summarization	4	0	0	0	0	
Sentence Detection	1	0	0	0	0	9
Embeddings	101	0	0	0	0	10
Part-of-speech	7	5	2	3	3	
Lemmatization	19	6	3	3	3	
Relation Extraction	12	0	0	0	0	
Spell check	4	0	0	0	1	
Totale	214	19	11	12	12	19



Spark NLP è la piattaforma aperta più usata per scopi industriali (Gradient Flow: NLP Industry Survey, 2020)

- Basato su Apache Spark
- Programmabile in Python e Scala
- 1426 modelli e *pipeline* «out of the box»



# L'INGLESE DA *LINGUA* A *LINGUAGGIO*

- L'AI anglofona rischia oggi di ridefinire la **facoltà del linguaggio** su scala globale
- Non più una lingua tra le altre, ma il presupposto stesso della vita linguistica

## Colonizzazione

- Trasferimento di risorse verso i monopoli digitali
- Influenza nelle economie locali

## Omologazione

- Pressione dell'inglese sulle lingue «di minoranza»
- Perdita di diversità linguistica \ culturale



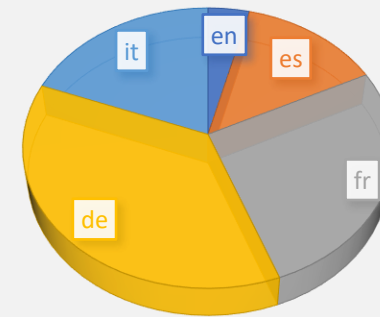
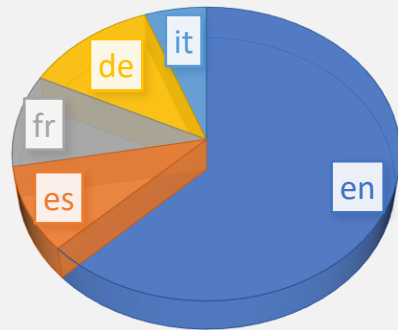
# COSA STA FACENDO L'EUROPA



**MULTILINGUAL  
EUROPE  
TECHNOLOGY  
ALLIANCE**



*Developing an agenda and a roadmap  
for achieving full digital language  
equality in Europe by 2030*



*Towards the Primary Platform for  
Language Technologies in Europe*

Lingua	PIL 2020
en	306
es	1058
fr	2061
de	2832
it	1479

GDP 2020 per area linguistica  
delle maggiori economie EU +  
l'Irlanda

# COSA STA FACENDO L'EUROPA



21\4\21: Proposta  
per una  
regolamentazione  
dell'approccio  
europeo  
all'Intelligenza  
Artificiale

*Di fronte al rapido sviluppo tecnologico dell'IA e a un contesto politico globale in cui sempre più paesi investono fortemente nell'IA, l'UE deve agire unitariamente per sfruttare le numerose opportunità e affrontare le sfide dell'IA in modo adeguato agli sviluppi futuri.*

<https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-european-approach-artificial-intelligence>

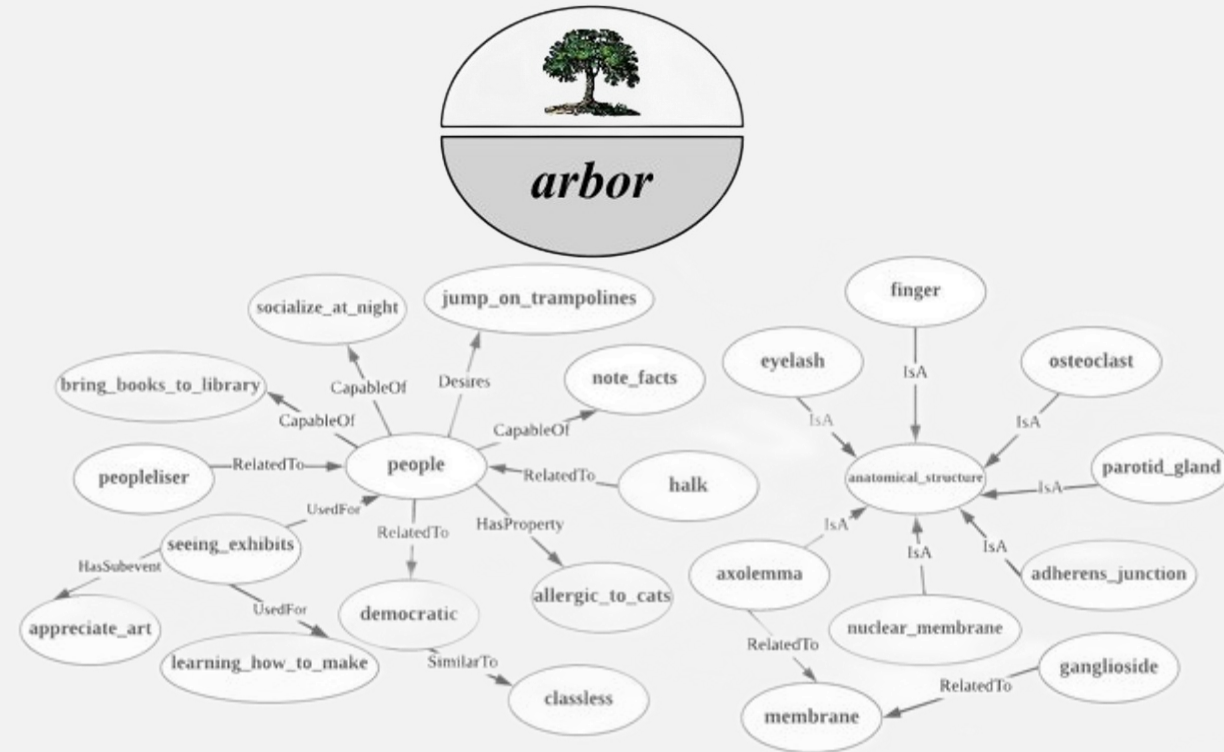
- Trasparenza e verificabilità dei dati di addestramento
- Considerazione della specificità delle diverse culture
- Eliminazione dei pregiudizi (bias)
- Trasparenza dei processi di generazione dei risultati (auditability)
- Contrasto agli usi manipolativi

# PROSPETTIVE DI RICERCA: BENTORNATO SEGNO!

Gli *embedding* ottenibili da risorse lessico \ ontologiche (aka Knowledge Graph, es. WordNet, ConceptNet, BabelNet) sono competitivi per molti *task*

- Controllo sociale, plasmabilità
- Sostenibilità eco-tecnologica
- Supporto a lingue in scarsità di risorse
- Supporto a lingue morfologicamente complesse

Anche sui Knowledge Graph , tuttavia, c'è da colmare un grande divario tra l'inglese e le altre lingue



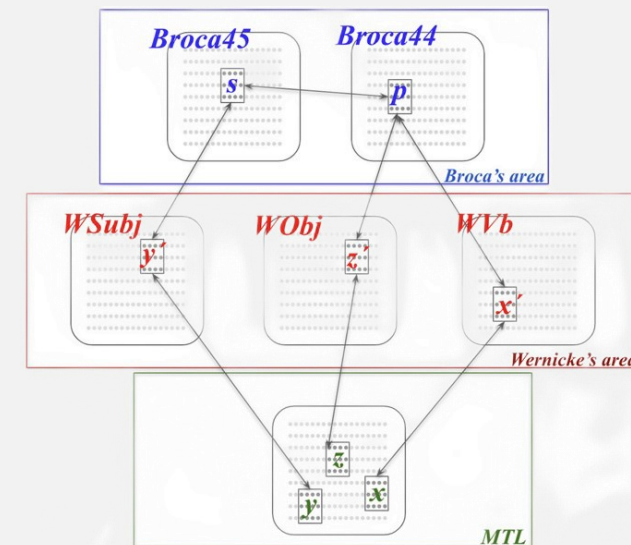
ConceptNet Knowledge Graph

A Data-Driven Study of Commonsense Knowledge using the ConceptNet Knowledge Base, Ke Shen, Mayank Kejriwal, 2020



# PROSPETTIVE DI RICERCA: BENTORNATO PENSIERO!

- Modellazione **plausibile** delle strutture neurali che sono alla base delle funzioni cognitive complesse
- Assembly Calculus: proiezione, associazione, fusione
- *Pattern completion*
- Principale campo applicativo: il linguaggio (primi esperimenti con la sintassi già in corso)



Linguaggio come fondamento della cognitività umana nelle pratiche sociali (Lev S. Vygotskij, *Pensiero e linguaggio*, Laterza 1990)



## CONCLUSIONE

- L'AI moderna (reti neurali) ha rivoluzionato le tecnologie linguistiche e ne ha aumentato significativamente la portata
- Tuttavia, molte attuali tecnologie sradicano il linguaggio dai processi di significazione che costituiscono il loro fondamento sociale
- Inoltre, per come sono prodotte e usate, tali tecnologie aumentano le concentrazioni e il divario in favore delle economie anglofone
- Questo divario esercita una pressione sulla vita linguistica dei Paesi europei
- Per favorire uno sviluppo equilibrato delle tecnologie linguistiche, l'Europa dovrà diffondere competenze e investire risorse (regolamenti e «marketplace» non bastano)
- La ricerca di modelli alternativi a quelli basati sulla «forza bruta» delle reti neurali è promettente e va sviluppata