



# Linguistica computazionale e strumenti digitali applicati allo studio del linguaggio

*Sara Tonelli, Rachele Sprugnoli*

# Cos'è la Linguistica Computazionale?

---

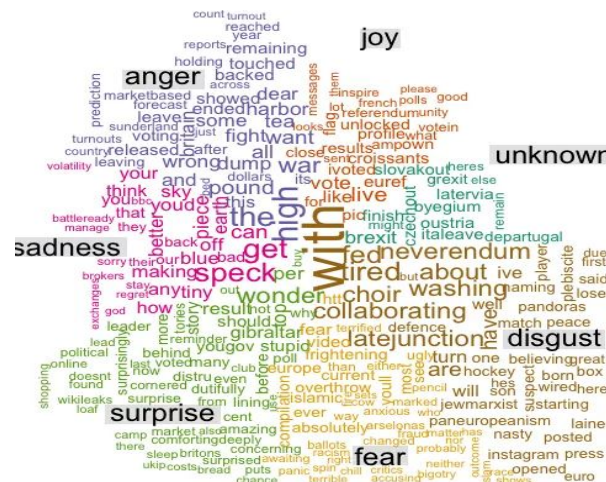
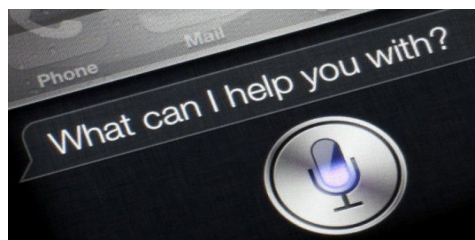
**Testo e Computer**, 2005

“L’obiettivo centrale della Linguistica Computazionale (LC) è quello di sviluppare modelli computazionali della lingua, cioè modelli del funzionamento del linguaggio naturale che possano essere tradotti in programmi eseguibili dal calcolatore e che consentano a quest’ultimo di acquisire le competenze necessarie per comunicare direttamente nella nostra lingua”



# Cos'è la Linguistica Computazionale?

- Comprende molte discipline:
  - linguistica
  - informatica
  - intelligenza artificiale
  - statistica
  - apprendimento automatico (machine learning)
- Parte della nostra vita quotidiana



# Dove si fa ricerca nella LC?



# Moduli di Analisi del Testo

---

Il lupo perde il pelo ma non il vizio

# Moduli di Analisi del Testo

---

Il lupo perde il pelo ma non il vizio

Token



# Moduli di Analisi del Testo

---

Il lupo perde il pelo ma non il vizio

Token



POS

A N V A N C R A N

# Moduli di Analisi del Testo

	Il lupo perde il pelo ma non il vizio								
Token	1	2	3	4	5	6	7	8	9
POS	A	N	V	A	N	C	R	A	N
Lemma	Il	lupo	perdere	il	pelo	ma	non	il	vizio



# Moduli di Analisi del Testo

Il lupo perde il pelo ma non il vizio

Token

1 2 3 1 4 5 6 1 7

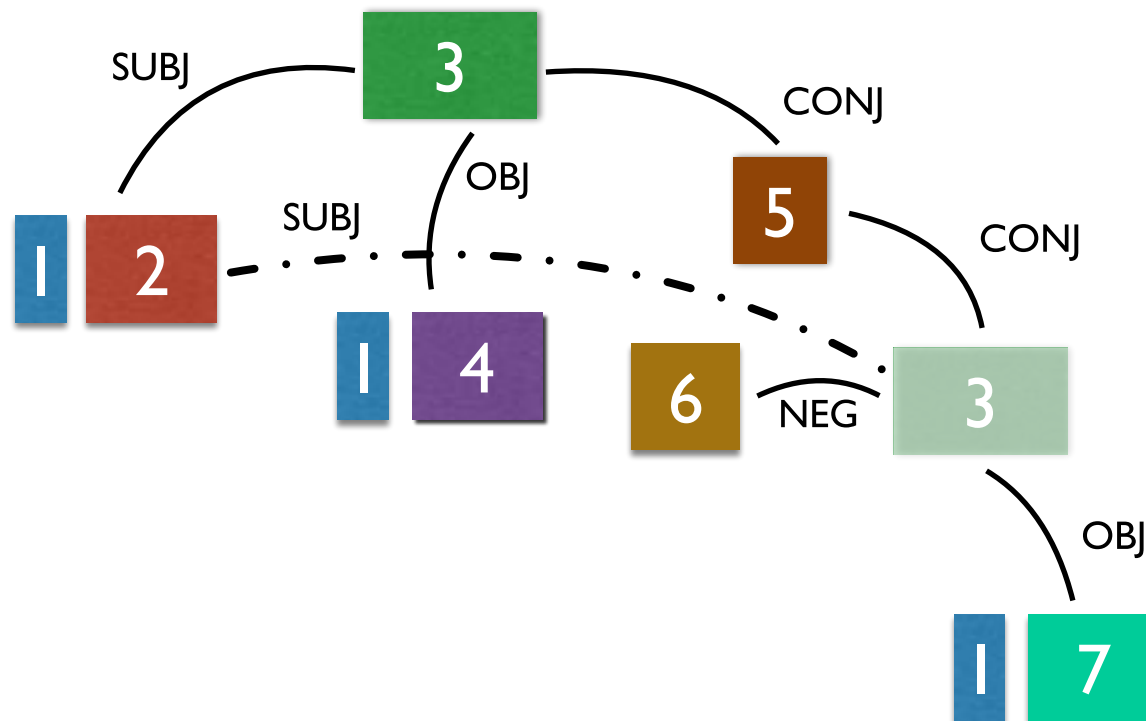
POS

A N V A N C R A N

Lemma

Il lupo perdere il pelo ma non il vizio

Parsing



# Moduli di Analisi del Testo

Il lupo perde il pelo ma non il vizio

1 2 3 4 5 6 7

Token

POS

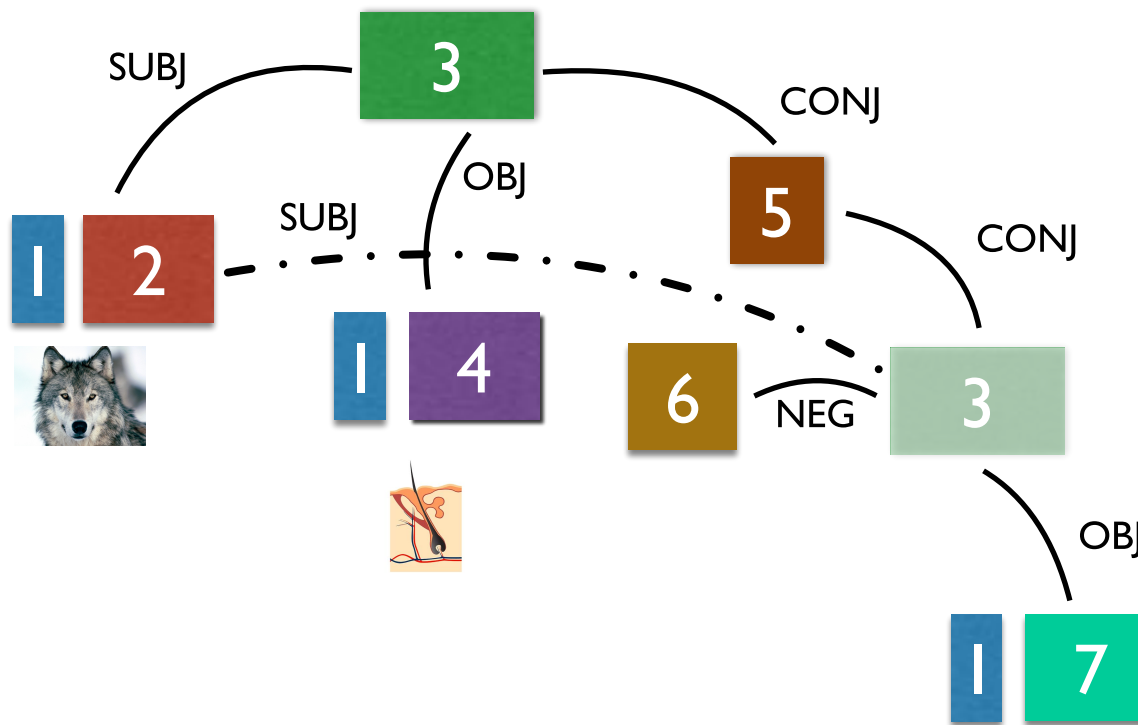
A N V A N C R A N

Lemma

Il lupo perdere il pelo ma non il vizio

Parsing

Linking



# Moduli di Analisi del Testo

Token

1 2 3 4 5 6 7

POS

A N V A N C R A N

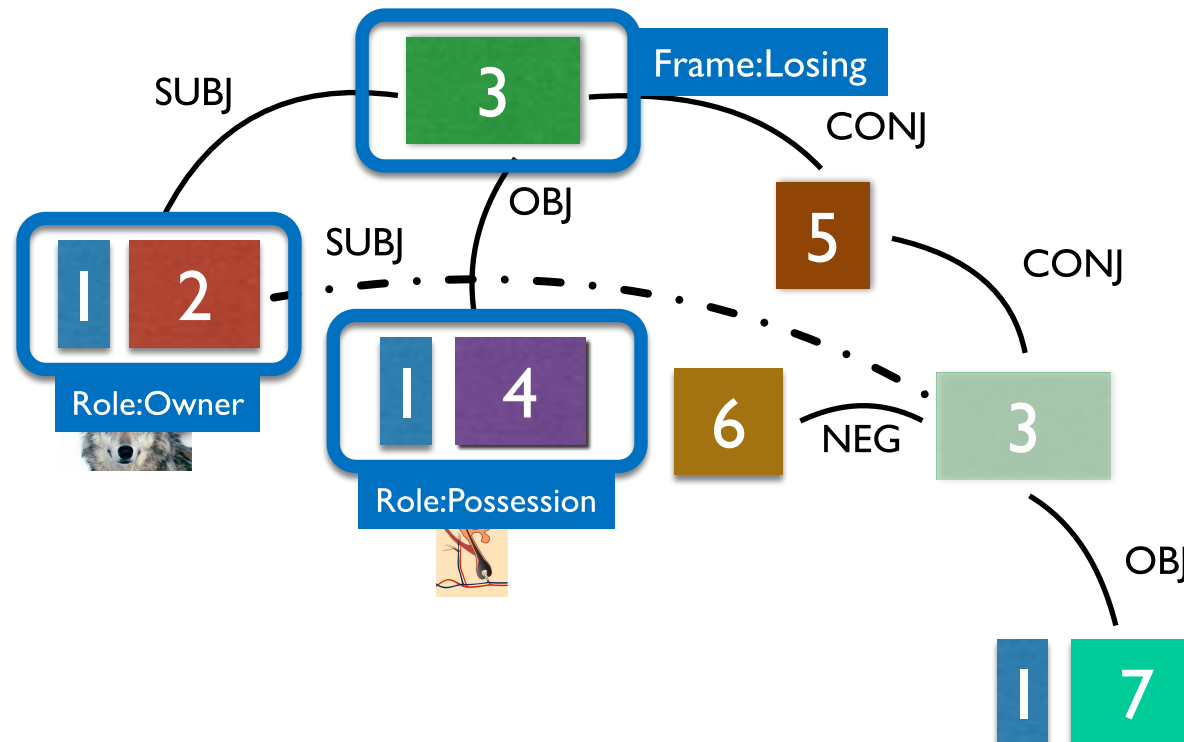
Lemma

Il lupo perdere il pelo ma non il vizio

Parsing

Linking

Frames



# Moduli di Analisi del Testo

---

- Demo online per l'analisi linguistica automatica dell'italiano
  - UDPipe: <https://lindat.mff.cuni.cz/services/udpipe/>  
moduli disponibili: sentence splitting, tokenizzazione, PoS tagging, lemmatizzazione, analisi morfologica, dependency parsing
  - Tint: <https://dh.fbk.eu/tint-demo/#/home/>  
moduli disponibili: sentence splitting, tokenizzazione, PoS tagging, lemmatizzazione, analisi morfologica, dependency parsing, NER , analisi dei verbi composti, keyphrase extraction, analisi dei derivati, leggibilità,

# Topic Modeling

---

- Quali sono gli argomenti contenuti in un corpus?

But to fix our immigration system, we must change our leadership in Washington and we must change it quickly. Sadly, sadly there is no other way. The truth is our immigration system is worse than anybody ever realized. But the facts aren't known because the media won't report on them. The politicians won't talk about them and the special interests spend a lot of money trying to cover them up because they are making an absolute fortune. That's the way it is. Today, on a very complicated and very difficult subject, you will get the truth. The fundamental problem with the immigration system in our country is that it serves the needs of wealthy donors, political activists and powerful, powerful politicians.

Trump, 31 August 2016

# Topic Modeling

---

- Quali sono gli argomenti contenuti in un corpus?

But to fix our immigration system, we must change our leadership in Washington and we must change it quickly. Sadly, sadly there is no other way. The truth is our immigration system is worse than anybody ever realized. But the facts aren't known because the media won't report on them. The politicians won't talk about them and the special interests spend a lot of money trying to cover them up because they are making an absolute fortune. That's the way it is. Today, on a very complicated and very difficult subject, you will get the truth. The fundamental problem with the immigration system in our country is that it serves the needs of wealthy donors, political activists and powerful, powerful politicians.

Trump, 31 August 2016

- IMMIGRAZIONE

- POLITICA

- ECONOMIA

# Topic Modeling

But to fix our must change Washington and quickly. Sadly other way. immigration s anybody ever aren't known report on their talk about interests spent to cover the making an abso way it is. complicated subject, you fundamental immigration sy that it serve donors, poli powerful, powe	As secretary Clinton allow criminal alie because their to take them. They were too them back. Who would do this? this? A weak a would do thi described Hill most radical in United States summary of wh support sanctu Security, Med welfare for al by making them which will die immigrants.	Social Secu lifetime we immigrants b citizens. And being treated veterans. Reme going to all illegal immigr visa overstay release on the hey, go ahead It's called Expanding unconstitution including ins millions of i even more crim Obama's non- And she wants in Syrian refu country .	All Americans country, in wonderful, p immigrants are jobs and wag totally protect our nation are people living everybody. An erased -- it lawful immigr if you look a the borders, a are erased, borders, we r And that's r And I have t endorsed by th 16,500. By IC First time anybody for pr	As I mentioned, Pueblo is filled with wonderful, hard-working immigrants. It's these hard-working immigrants who stand to lose the most from our open border immigration policy. Illegal immigration and broken Visa programs take jobs directly from Latino and Hispanic workers living here lawfully today -- you know that. They're taking your jobs. Illegal immigration also brings with it massive crime and massive drugs, including a terrible heroin problem right here in Colorado -- you have a big problem. So we're going to build the border wall and we are not -- what? We're going to build the wall and we're going to stop the drugs, the gangs, the violence from pouring into Colorado.
---	--	--	--	---

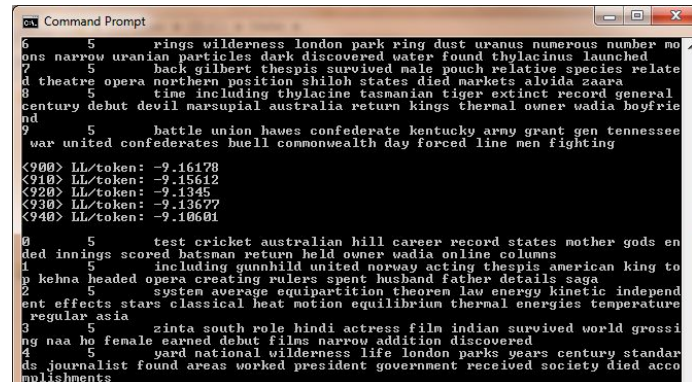
*“That’s how topic modeling works in practice. You assign words to topics randomly and then just keep improving the model, to make your guess more internally consistent, until the model reaches an equilibrium that is as consistent as the collection allows.”*

Ted Underwood, 2012

# Topic Modeling : Strumenti

- MALLET:

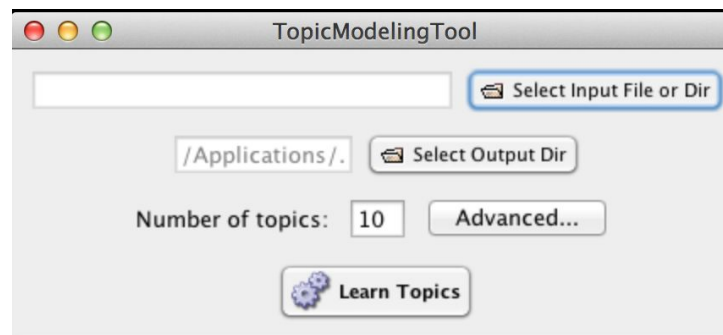
<http://mallet.cs.umass.edu/>



```
Command Prompt
6      5      rings wilderness london park ring dust uranus numerous number mo
ons narrow uranian particles dark discovered water found thylacinus launched
7      5      back gilbert thespis survived male pouch relative species relate
d theatre opera northern position shiloh states died markets alvida zaara
8      5      time including thylacine tasmanian tiger extinct record general
century debut devil marsupial australia return kings thermal owner wadia boyfrie
nd
9      5      battle union hawes confederate kentucky army grant gen tennessee
war united confederates buell commonwealth day forced line men fighting
<900> LL/token: -9.16178
<910> LL/token: -9.15612
<920> LL/token: -9.1345
<930> LL/token: -9.13677
<940> LL/token: -9.10601
0      5      test cricket australian hill career record states mother gods en
ded innings scored batsman return held owner wadia online columns
1      5      including gunnhild united norway acting thespis american king to
p kehna headed opera creating rulers spent husband father details saga
2      5      system average equipartition theorem law energy kinetic independ
ent effects stars classical heat motion equilibrium thermal energies temperature
regular asia
3      5      zinta south role hindi actress film indian survived world grossi
ng naa ho female earned debut films narrow addition discovered
4      5      yard national wilderness life london parks years century standar
de journalist found areas worked president government received society died acco
mplishments
```

- Topic-modeling-tool:

<https://code.google.com/archive/p/topic-modeling-tool/>





# Topic Modeling : Pro e Contro

---



- Difficile valutare l'output
- Non c'è un modo definito per decidere il numero di topic
- Troppo ambiguo
- Solo su grandi quantità di dati



- Buon punto di partenza per esplorare un corpus
- Crea nuovi modi per vedere gradi quantità di dati

# Estrazione di concetti-chiave

---

- I concetti-chiave catturano i concetti principali di un documento

But to fix our **immigration system**, we must change our **leadership in Washington** and we must change it quickly. Sadly, sadly there is no other way. The truth is our **immigration system** is worse than anybody ever realized. But the facts aren't known because the **media** won't report on them. The **politicians** won't talk about them and the special interests spend a **lot of money** trying to cover them up because they are making an absolute **fortune**. That's the way it is. Today, on a very complicated and very difficult subject, you will get the truth. The fundamental problem with the **immigration system** in our **country** is that it serves the needs of **wealthy donors**, **political activists** and powerful, **powerful politicians**.

Sia parole singole che  
espressioni

Funziona anche se documenti  
brevi

# Sistema di estrazione di concetti-chiave

- KD = Keyphrase Digger

[http://dhlab.fbk.eu:8080/KD\\_KeyDigger/](http://dhlab.fbk.eu:8080/KD_KeyDigger/)



## Altre possibili analisi: La complessità del linguaggio

---

- **Tecnologie di analisi automatica del testo** per capire quanto un testo è difficile e quali sono gli elementi che influiscono sulla sua leggibilità
- Nasce in ambito educativo: Thorndike (1921) *Teacher's Wordbook*
- Utilizza **metriche di leggibilità** che misurano la complessità di un testo su diversi livelli: lessicale, sintattico, del discorso, ecc.

## Il Gulpease

---

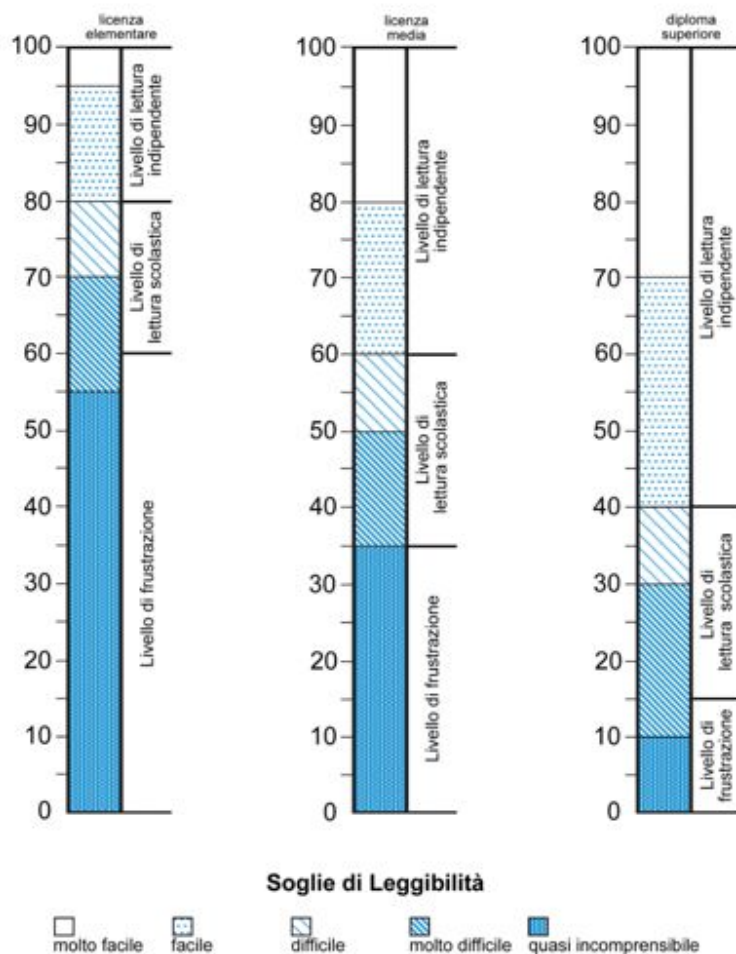
- **GULP** (Gruppo Universitario Linguistico Pedagogico, Università La Sapienza), 1988.

Considera lunghezza delle frasi rispetto al numero di parole: più le parole e le frasi sono lunghe, più un testo sarà difficile

$$89 + \frac{300 * (\text{numero delle frasi}) - 10 * (\text{numero delle lettere})}{\text{numero delle parole}}$$

# Il Gulpease

Indice Gulpease: scala dei valori



< 80: Testo difficile per chi possiede la licenza elementare

< 60: Testo difficile per chi possiede la licenza media

< 40: Testo difficile per chi possiede un diploma di scuola superiore

# Oltre il Gulpease: Quali metriche?

---

Dipende da cosa si vuole misurare:

- a. leggibilità a livello di **documento** o di **frase**?
- b. si vuole un **unico indice** (es. Gulpease) o diversi indicatori?
- c. ci sono **aspetti particolari** di un testo che si intendono misurare?

# Livelli di analisi automatica

---

Livello puramente **statistico**: conteggi sulla lunghezza delle parole e delle frasi, Gulpease

*“Vorrei andare in bicicletta, ballare, fischiettare, guardare il mondo, sentirmi giovane, sapere che sono libera, eppure non devo farlo notare perché, pensa un po’, se tutti e otto ci mettessimo a lagnarci e a far la faccia scontenta, dove andremmo a finire?” (G 47)*

*“Ho voglia di andare in bicicletta, di ballare, di fischiettare, di guardare il mondo, di sentirmi giovane e libera. Ma non dico queste cose ai miei familiari o ai miei amici. Cosa succederebbe se tutti iniziassimo a piangere, protestare, a fare una faccia infelice?” (G 61)*



## Livelli di analisi automatica

---

- Livello **lessicale**: ricchezza del vocabolario, presenza di termini tecnici

*“I bei tempi finirono nel maggio 1940; prima la guerra, la **capitolazione**, l'**invasione** tedesca, poi cominciarono le **sventure** per noi ebrei.”*

*“I tempi felici finirono nel maggio 1940; dopo la guerra, la **sconfitta**, e l'**arrivo** dei soldati tedeschi, cominciarono i **problemi** per noi ebrei.”*

# Livelli di analisi automatica

---

- Livello **sintattico**: numero di coordinate e subordinate, profondità dell'albero sintattico

[http://www.ilc.cnr.it/dylanlab/apps/texttools/?tt\\_user=guest](http://www.ilc.cnr.it/dylanlab/apps/texttools/?tt_user=guest)

# Livelli di analisi automatica

---

Livello del **discorso**: struttura del documento corrisponde alla struttura temporale del racconto, presenza di anafora, informazioni implicite ed esplicite

*“La nostra cameretta, con i suoi muri nudi, era assai disadorna; grazie al babbo **che fin da prima** aveva portato qui la mia **collezione di stelle del cinema** e di cartoline illustrate, ho trasformato la stanza, **dopo** aver spennellato di colla le pareti, in una fitta mostra di figurine. Così ha un’aria più allegra.”*

*“La nostra cameretta era molto vuota e aveva i muri bianchi. Per fortuna il papà aveva portato qui la mia collezione di immagini di attori e attrici famosi e le mie cartoline illustrate, così ho trasformato la mia stanza: ho messo la colla sulle pareti e ho appeso tutte le mie figurine. Ora la **stanza** è più allegra.”*

## Esempi di analisi

---

- A. Barbagli, P. Lucisano, F. Dell'Orletta, S. Montemagni, G. Venturi *“Il ruolo delle tecnologie del linguaggio nel monitoraggio dell'evoluzione delle abilità di scrittura: Primi risultati”* Italian Journal of Computational Linguistics, Vol. 1, N. 1, 2015.
- Usare tecnologie per studiare come evolvono le abilità di scrittura di studenti madrelingua
- 109 testi, prove di scrittura somministrate nel primo e secondo anno in diverse scuole secondarie di primo grado di Roma

## Esempi di analisi

---

- Nel confronto tra i due anni, **diminuisce la lunghezza media** dei periodi e il **numero di pronomi** ma aumenta l'uso della punteggiatura.
- **Diminuisce** l'uso dell'**indicativo**
- **Diminuiscono** le dipendenze sintattiche “lunghe” all'interno delle frasi
- **Aumenta** la ricchezza **lessicale**

# Esempi di analisi

- R. Sprugnoli, S. Tonelli, A. Palmero Aproso, G. Moretti “*Analysing the Evolution of Students’ Writing Skills and the Impact of Neo-Standard Italian with the help of Computational Linguistics*”, submitted.
- Usare tecnologie per studiare come evolvono le abilità di scrittura di studenti madrelingua analizzando 2.500 temi di maturità su un arco temporale di 15 anni, 28 diversi tratti linguistici

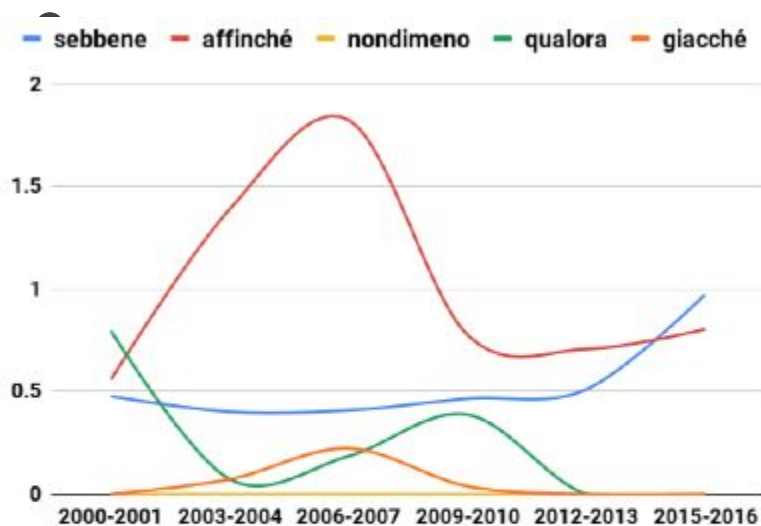


Figure 1: Observed relative frequency of complex connectives per 10,000 words.

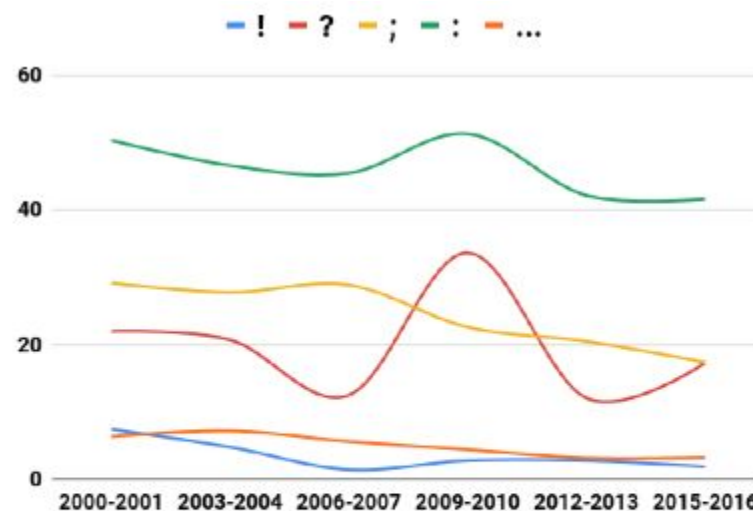
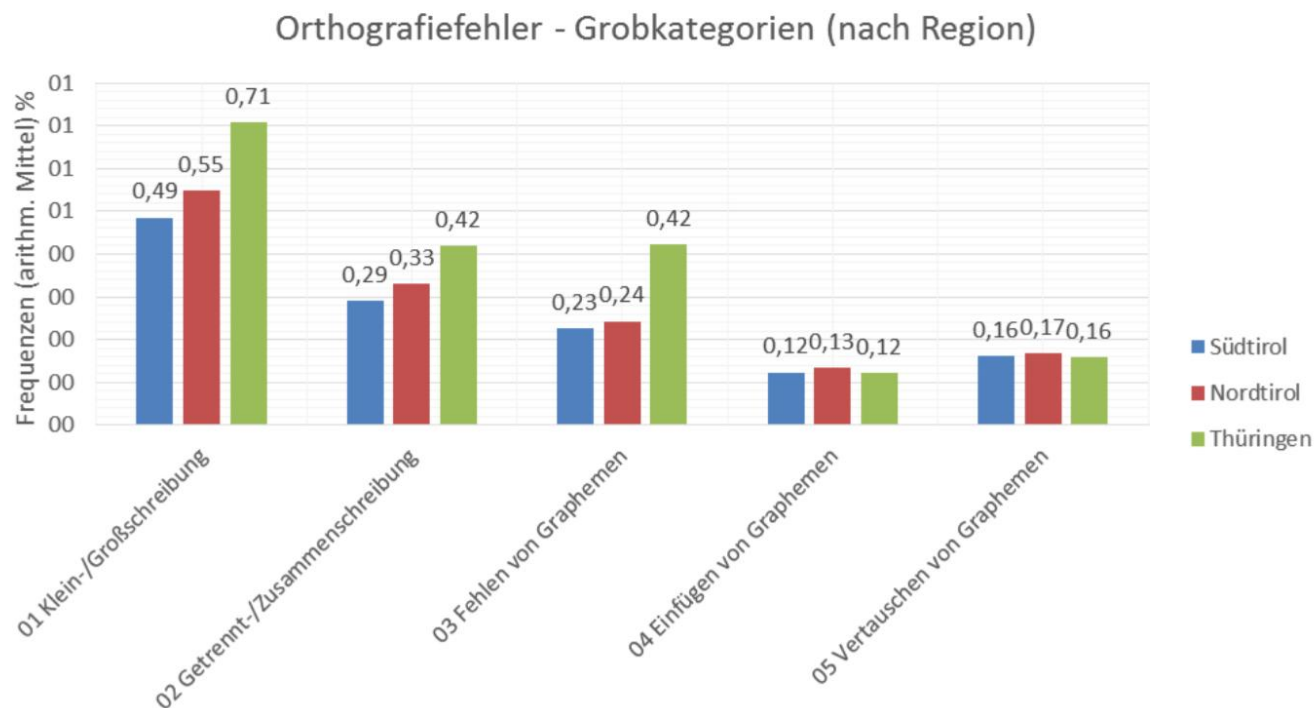


Figure 2: Observed relative frequency of punctuation per 10,000 words.

# Esempi di analisi

- A. Abel, A. Glaznieks, L. Nicolas & E. Stemle, Egon. (2014). *KoKo: an L1 Learner Corpus for German*. Proceedings of LREC 2014, International Conference on Language Resources and Evaluation
- Studio del tedesco scritto, 1511 studenti, 66 scuole da Alto Adige, Austria e Germania. Temi argomentativi e relativo questionario.

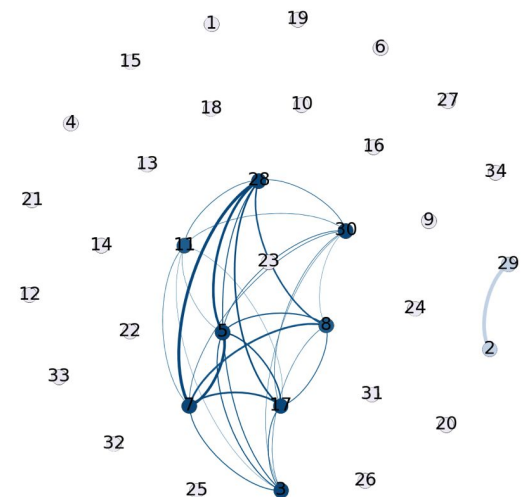
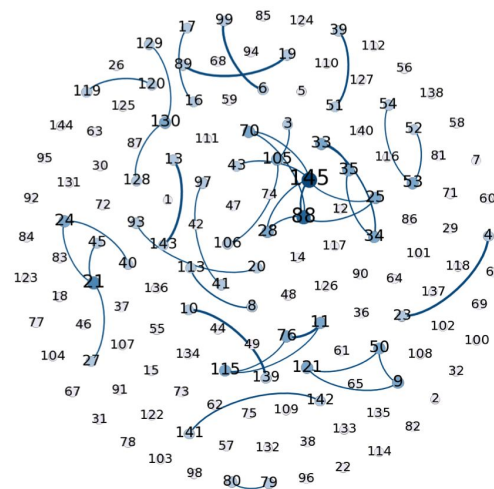
<http://www.eurac.edu/it/research/projects/Pages/projectdetails.aspx?pid=7639>



# Altri casi d'uso: Similarità semantica

Analisi delle **sovrapposizioni** tra le proposte di progetto: progetti diversi presentati dalla stessa scuola, progetti presentati da scuole in rete e progetti presentati da scuole diverse e distanti geograficamente

Sovrapposizioni possono essere indicatori di collaborazione, oppure attenzione per le stesse fonti informative





# Visualizzazione Dati

---

- Alcuni strumenti:
  - Carto: <https://carto.com/>  
Esempio: mappatura dei nomi di luoghi
  - Easy Linavis: <https://ezlinavis.dracor.org/>  
Esempio: rete dei personaggi dell'Ottavia di Alfieri
  - Gephi: <https://gephi.org/>  
Esempio: rete dei personaggi delle tragedie
  - RAW Graphs: <http://rawgraphs.io/>  
Esempi: topic modeling, keyphrase extraction, domain identification

# Materiale/Strumenti

---

- Cartella da scaricare per le esercitazioni: <http://bit.do/ev6Km>
- <http://app.rawgraphs.io/>
- <https://mimno.infosci.cornell.edu/jsLDA/jslda.html>
- <https://voyant-tools.org/>
- [http://dhlab.fbk.eu:8080/KD\\_KeyDigger/](http://dhlab.fbk.eu:8080/KD_KeyDigger/)