Introduzione alla Linguistica Computazionale

Salvatore Sorce
Dipartimento di Ingegneria
Chimica, Gestionale, Informatica e Meccanica
Ludici Adattati da Alessandro Lenci
Dipartimento di Linguistica “T. Bolelli”

InformaticaUmanistica

Elaborazione del Parlato e del Linguaggio (Speech and Language Processing)

- L'idea di dare ai computer la capacità di elaborare il *linguaggio naturale* è molto antica
- In "A Space Odyssey" (Kubrick, 1969) HAL 9000 è un agente artificiale che può dialogare con gli uomini utilizzando il linguaggio umano
- in "Star Wars" (Lucas, 1977) C-3PO è un robot protocollare cioè si occupa proprio delle comunicazioni tra umani e robot

Elaborazione del Parlato e del Linguaggio

- Quali conoscenze linguistiche dovrebbero possedere tali agenti?
  - articolare e decodificare i suoni di una lingua
    - fonetica articolatoria e acustica, fonologia, prosodia, ecc.
  - conoscere le parole di una lingua, la loro struttura e la loro organizzazione
    - lessico e morfologia
  - comporre le parole in espressioni linguistiche complesse (sintagmi, frasi, ecc.)
    - sintassi
  - assegnare significati alle espressioni linguistiche semplici e complesse
    - semantica (lessicale e composizionale)
  - usare le frasi nei contesti, situazioni e modi appropriati agli scopi comunicativi
    - pragmatica

Quali e Quante Discipline Scientifiche?

- La progettazione di HAL 9000 coinvolge molte discipline ad esempio:
  - robotica
  - intelligenza artificiale
  - linguistica
  - scienze cognitive
  - …
- Noi ci occuperemo di *rappresentazione* e *analisi* del testo
- Alla fine del corso avrete conoscenze di base, ma fondamentali, di **linguistica computazionale**
  - Sarà fornito lo **strumentario di base** per il linguista computazionale

Linguistica Computazionale

- La Linguistica Computazionale è lo studio di sistemi informatici per la comprensione e la produzione di linguaggio naturale.

  R. Grishman, ``Computational Linguistics - An Introduction'', 1986

- La Linguistica Computazionale si occupa dello sviluppo di una teoria computazionale del linguaggio, sfruttando le nozioni di algoritmi e strutture dati provenienti dall'Informatica.

  J. Allen, ``Natural Language Understanding'', 1994

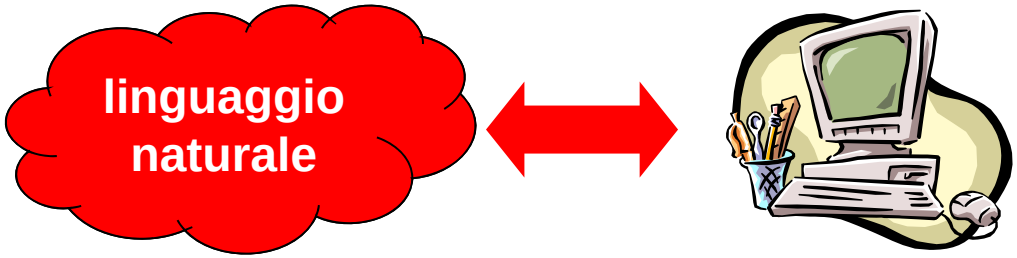
Cosa è la linguistica computazionale?

La linguistica computazionale si occupa dell'analisi ed elaborazione del linguaggio naturale attraverso l'uso di metodologie informatiche

La linguistica computazionale si concentra sullo sviluppo di formalismi descrittivi del funzionamento del linguaggio naturale, tali che si possano trasformare in programmi eseguibili dai computer.

Cosa è la linguistica computazionale?

- Come si applica il trattamento automatico dell'informazione (Informatica) al linguaggio naturale?
  1. esplorazione ed analisi dei "dati linguistici" con strumenti informatici
  2. sviluppo di sistemi informatici dotati di conoscenze linguistiche e in grado di esibire capacità linguistiche "comparabili" a quelle umane
  3. elaborazioni di modelli computazionali e simulazioni della competenza linguistica umana, della sua acquisizione e del suo uso



Cosa può fare il computer per il linguaggio?

- Usare il linguaggio implica produrre dati linguistici
  - libri, giornali, pagine web, conversazioni, e-mail, chat, ecc.
- Il computer come gestore ed elaboratore di dati:
  - immagazzinare
  - calcolare
  - ordinare
  - comparare
  - ricercare
- Analisi computazionali dei dati linguistici
  - le potenzialità “standard” del computer possono essere usate per la gestione e l’analisi avanzata dei dati linguistici

Analisi computazionale dei dati linguistici

- La linguistica computazionale permette di affrontare queste ricerche attraverso
  - metodi e strumenti informatici per la rappresentazione e gestione di grandi quantità di dati linguistici
    - digitalizzazione di testi
    - trascrizioni del parlato
    - dati linguistici raccolti sul campo
    - ecc.
  - ricerche ed esplorazioni avanzate del testo
  - metodi matematici e statistici per analizzare i dati linguistici ed elaborare modelli del linguaggio

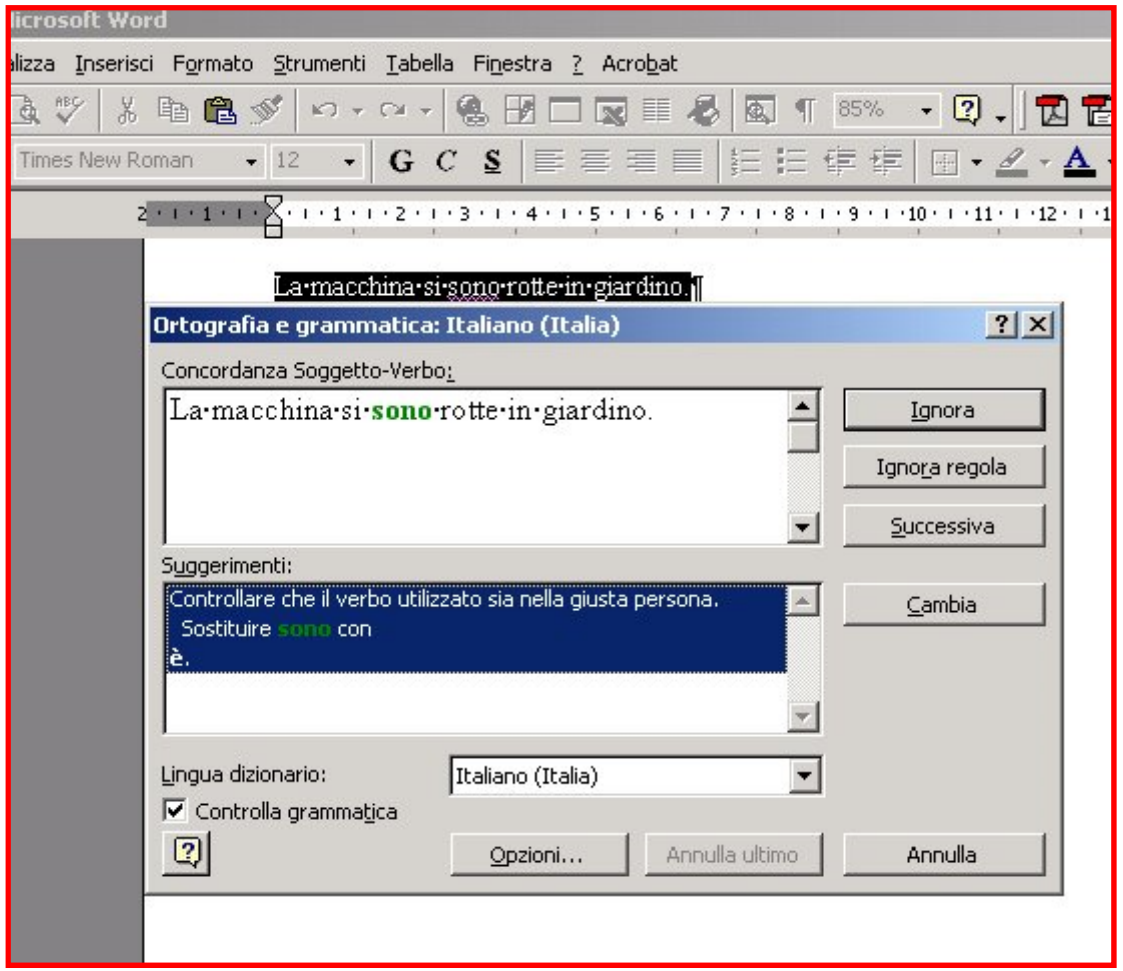
Cosa può fare il computer per il linguaggio?

- Il Natural Language Processing (NLP) o Trattamento Automatico del Linguaggio (TAL) cerca di dotare il computer di conoscenze linguistiche allo scopo di:
  - progettare programmi e sistemi informatici che assistano l'uomo in "compiti linguistici"
    - traduzione
    - gestione dei documenti e della conoscenza, ecc.
  - sviluppare sistemi informatici che usano il linguaggio naturale per:
    - interagire con essere umani in maniera "naturale"
    - estrarre automaticamente informazioni da testi o da altri media
    - estendere dinamicamente la propria competenza linguistica

Natural Language Processing (NLP)
Alcune applicazioni

- Correttori ortografici, grammaticali, ecc.
- Recupero “intelligente” di documenti
  - Information Retrieval
- Riconoscimento automatico del parlato
  - Automatic Speech Recognition (ASR)
- Sintesi automatica della voce
  - Text-To-Speech (TTS)
- Estrazione automatica di informazione da testi
  - Information Extraction (IE)
- Interrogare documenti attraverso domande in linguaggio naturale
  - Question Answering (QA)
- Traduzione (semi)-automatica di testi
  - Machine translation
- Interazione (conversazione) uomo-macchina multimodale
  - Agenti conversazionali complessi

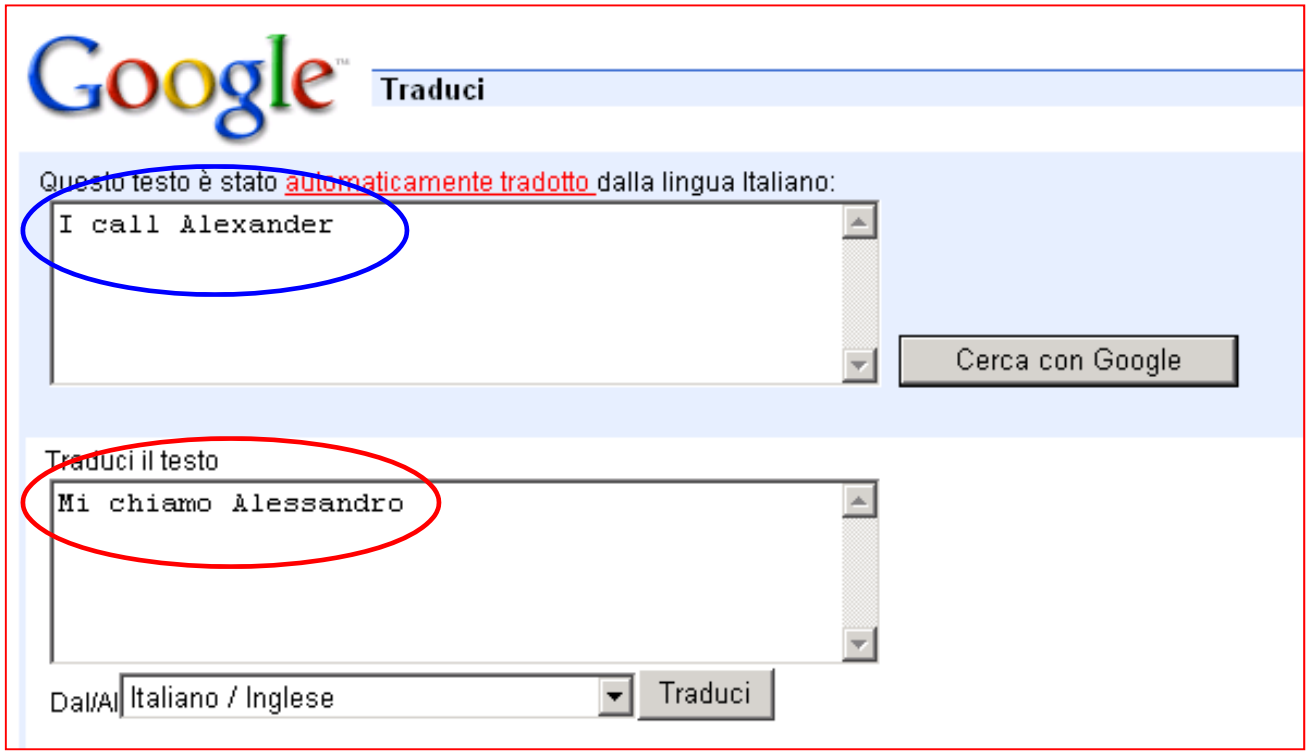
NLP - correttori grammaticali



Word

Correttore grammaticale

NLP - traduzione automatica

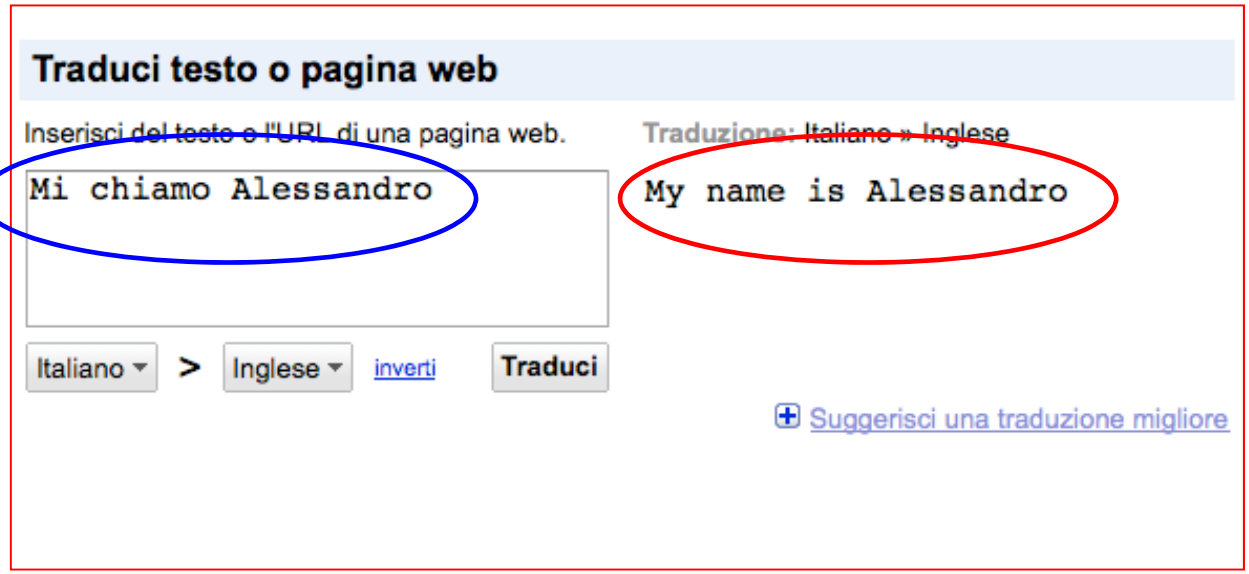


Google

traduzione automatica

2008

NLP - traduzione automatica

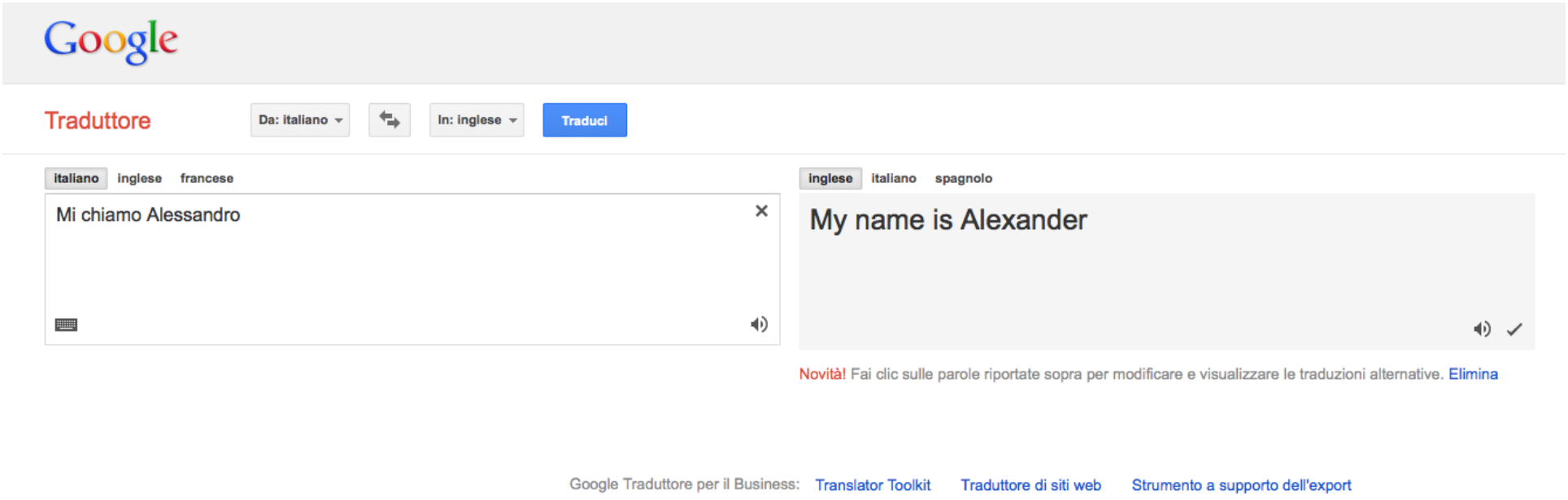


Google

traduzione automatica

... 2009

NLP - traduzione automatica

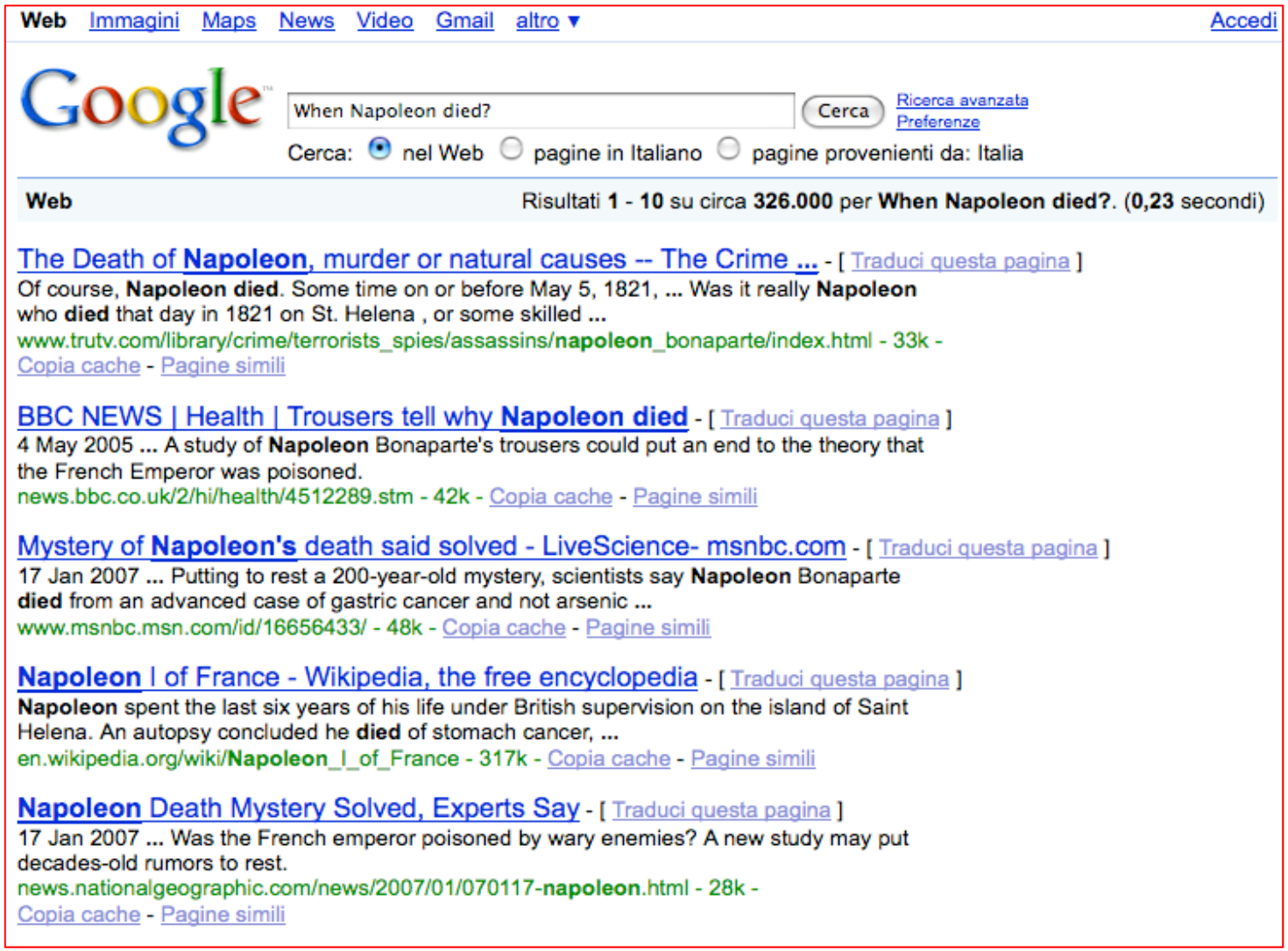


Google

traduzione automatica

... 2012

NLP - question-answering



Natural Language Processing (NLP)

Un'applicazione "vicina": l'Atlante Linguistico Siciliano

- L'ALS si occupa di sociolinguistica cioè di come i fattori sociali ed economici influenzano gli elementi della lingua.
- L'aspetto principale che viene investigato è come la popolazione si rapporta all'utilizzo dell'italiano e del dialetto siciliano sia nell'uso che nella percezione.
- Aspetti linguistici
  - pronuncia
  - Cadenza
- Aspetti traduttivi
  - passaggio da un codice all'altro
- Aspetti percettivi
  - quando si usa un codice piuttosto di un altro
  - come viene percepito l'uso di un codice in un determinato contesto

Natural Language Processing (NLP)

Un'applicazione "vicina": l'Atlante Linguistico Siciliano

- Lo strumento per la raccolta dei dati è l'intervista ad una popolazione
  - modello di raccolta dati: 17 interviste per centro suddiviso in 5 famiglie composte da 3 generazioni differenti più 2 adolescenti extra (nonno, padre, figlio)
  - dati memorizzati su file audio
  - necessità di strumenti specifici
    - strumenti di trascrizione (dall'audio al testo su supporto informatico)
    - strumenti di annotazione (dal testo al testo arricchito di informazioni sui fenomeni sociolinguistici)
    - strumenti di visualizzazione dei risultati (mappe georeferenziate)

Natural Language Processing (NLP)

Un'applicazione "vicina": l'Atlante Linguistico Siciliano

- Annotazione: processo di arricchimento delle informazioni su un testo
- viene realizzato con strumenti XML secondo la seguente modalità:
  - input: file trascritto e salvato come XML (passaggio da testo puro a XML → definizione di un modello di strutturazione delle informazioni)
  - arricchimento secondo un insieme specifico di possibili annotazioni (TAG di un determinato schema)
  - recupero delle informazioni memorizzate nel DB XML

La linguistica computazionale ieri e oggi