# Modulo A

## Valutare i sistemi automatici di interazione linguistica

11. Valutare in pratica il corpus CLIPS

9 novembre 2016

Linguistica italiana II
Mirko Tavosanis
A. a. 2016-2017

# Una proposta di lavoro

Ci sono le risorse per fare 24 lavori diversi:

- Scegliere una delle 12 città di CLIPS
- Per quella città, scegliere se esaminare il riconoscimento vocale con il Letto o con il Dialogico (tra i 20 e i 25 minuti di registrazione)
- Procedere nel modo descritto nella prossima diapositiva

# Procedimento

1. Individuare e scaricare i materiali
2. Controllare le trascrizioni originali
3. Sottoporre i file audio a Google e Dragon
4. Registrare l'output di Google e Dragon
5. Convertire i file (originali e output) per il controllo con SCLITE
6. Calcolare il WER
7. Commentare dal punto di vista linguistico i risultati e gli errori di trascrizione

# 1. Individuare e scaricare i materiali

- Il corpus CLIPS presenta materiali complessi
- Ho fatto un primo inventario, ma non escludo che mi sia sfuggito qualcosa (ce ne accorgeremo lavorando, penso)
- Soprattutto nel caso dei file audio: molte cose si possono controllare solo ascoltando tutta la registrazione

# Struttura del corpus

- Il corpus è diviso nei sottocorpora:

- DIALOGICO
- LETTO
- ORTOFONICO
- RTV
- TELEFONICO

- Prendiamo LETTO 

Diviso per città:

- BARI
- BERGAMO
- CAGLIARI
- CATANZARO
- FIRENZE
- GENOVA
- LECCE
- MILANO
- NAPOLI
- PALERMO
- PARMA
- PERUGIA
- ROMA
- TORINO
- VENEZIA

Prendiamo BARI

# BARI

Contiene:

- Corpus (contiene le sottocartelle LF, LM, LT)
- etichettato
- LF.zip («Lettura frasi»)
- LM.zip («Letto Map»)
- LT.zip («Letto Task»)

C’è anche un avviso non molto chiaro: “I file .zip presenti in alcune cartelle contengono l'insieme dei files contenuti in quell'estratto di corpus”

In realtà, per esempio, LF.zip di Letto – Bari non contiene tutti i file Wav contenuti nella sottocartella LF di Corpus

# BARI

Il corpus contiene i file audio, i testi delle trascrizioni e un file XML. Nel caso del LETTO-BARI per le Frasi troviamo:

- Audio
  - Due «parlanti» (p1 e p2) leggono 2 «mappe» (A e B) a testa; per ogni mappa ci sono 4 dialoghi (01-04)
  - Ogni «parlante» legge 8 dialoghi e ci sono in tutto 16 file
  - In realtà p1 e p2 corrispondono a parlanti diversi per ogni dialogo; può darsi che per le altre città sia diverso
- Trascrizioni: sono solo 6, tre per i file p1 (A02B, A03B) e tre per i file p2 (A02B, A03B, B02B)
- File XML: vuoto

Attenzione! ha senso fare la verifica file per file, mantenendo i dati separati: la diversità dei parlanti può spiegare la diversità dei risultati

# File su cui ho fatto la valutazione

Bari – Letto, Frasi:

- LFp1A02B.txt
- LFp1A03B.txt
- LFp1B02B.txt
- LFp2A02B.txt
- LFp2A03B.txt
- LFp2B02B.txt

Per ogni file di testo è disponibile anche la registrazione .wav; finché non saranno esauriti i file, è consigliabile lavorare sulle registrazioni di cui esiste la trascrizione

# Oltre Bari

In tutte le 12 città la composizione del corpus è:

- LF (Lettura frasi): 16 file Wav lunghi circa 2 minuti, 6 .txt, 1 .xml (vuoto)
- LM (Lettura oggetti map task): 16 file Wav lunghi circa 30 secondi, 6 .txt, 1 .xml (vuoto)
- LT (Lettura oggetti test delle differenze): 16 file Wav lunghi circa 1 minuto, 6 .txt, 1 .xml (vuoto)

Tuttavia, la certezza si può avere solo aprendo i file, per controllare che non ci siano per esempio errori nel modo in cui sono stati nominati.

Quindi, per il Letto, per ogni città ci sono 18 file accompagnati da una trascrizione:

- 6 LF = 12 minuti circa
- 6 LM = 3 minuti circa
- 6 LT = 6 minuti circa

La lunghezza complessiva del Letto per una città dovrebbe aggirarsi intorno a **21 minuti**

# Dialogico

Scegliendo Dialogico si va alle città. Ogni città contiene le sottocartelle:

- corpus
- etichettato

La cartella corpus contiene 2 sottocartelle:

- mt (Map Task: 8 file Wav di lunghezza attorno ai 12 minuti, 3 .txt, 1 .xml)
- td (test delle differenze: 8 file Wav di lunghezza attorno ai 12 minuti, 3 .txt, 1 .xml)

Ogni file di **mt** dura circa **12 minuti**, quindi la durata complessiva dei dialoghi di cui esiste una trascrizione è di **36 minuti** per ogni città

Ogni file di **td** dura circa **10 minuti**, quindi la durata complessiva dei dialoghi di cui esiste una trascrizione è di **30 minuti** per ogni città

Proposta: rivedere un blocco di 2 file = un po' più di **20 minuti**, con riduzioni nel caso di difficoltà (per esempio, sovrapposizioni di turno)

# Altri corpora?

- **Ortofonico** e **Radiotelevisivo** ci interessano meno, ma sono comunque utili per confronto
- **Telefonico** sarebbe di gran lunga il più interessante, ma il problema è la qualità delle registrazioni
- Quindi: se vi interessano, potete scegliere anche questi

# 2. Controllare le trascrizioni originali

- Le trascrizioni di CLIPS sono piuttosto accurate, ma non perfette
- Un controllo permette di ridurre il tasso di errore
- Per un lavoro su CLIPS sarà obbligatorio almeno un ascolto di controllo (= ascoltare il file audio tenendo sott'occhio il testo della trascrizione): la qualità del lavoro di controllo contribuisce alla qualità complessiva
- Gli errori trovati devono essere:
  - documentati nel modo che vedremo subito
  - corretti nel testo da confrontare con i risultati della trascrizione
- La trascrizione, **corretta**, deve naturalmente diventare il **file di controllo**, da confrontare con i risultati della trascrizione automatica per ottenere il WER

# p1

**p1A02B** #10: nella realtà il parlante non pronuncia la *-d* eufonica indicata nella trascrizione (30 secondi)

è possibile che non trovi mai un momento per scrivere a**d** un amico lontano ?

**p1A03B** #14: nella realtà la parlante forse  dice «chiamai **un** medico» (50 secondi), ma sono incerto e lascio il testo originale

<tongue-click> chiamai **il** medico perché avevo male agli occhi <inspiration> ma quando uscii mi accorsi di non sentire neanche i rumori

**p1A03B** #17: nella realtà la parlante sicuramente non legge il secondo «mai» (1 minuto e 20 secondi)

quel ragazzo non dice mai la verità <inspiration> ma tu non fargli **mai** vedere che pensi sia un bugiardo

**p1A03B** #19: nella realtà la parlante sicuramente non legge il «ne» (1 minuto e 45 secondi)

Lucio era certo che sarebbe diventato una persona importante <inspiration> un uomo politico o magari un mini+ un ministro <inspiration> aveva a cuore il bene della società <inspiration> rispettava la legge <sp> se **ne** teneva un discorso trovava le parole adatte ad ogni situazione , <inspiration> si sentiva proprio un buon italiano

**p1B02B** #19: nella realtà la parlante forse non legge la «o» (1 minuto e 30 secondi); intervengo

#19:  Lucio {<creacky-voice> era} certo che sarebbe diventato una persona importante <sp> un uomo politico **o** magari un ministro

# p2

**p2A02B** #19: nella realtà la parlante sicuramente non dice «rispettva» e non legge il «ne» (1 minuto e 30 secondi)

#19:  Lucio era certo che sarebbe diventato una persona importante , un uomo politico o magari un ministro , aveva a cuore il bene della società , **rispettva** la legge se **ne** teneva un discorso trovava le parole adatte ad ogni situazione

**p2A03B** #17: nella realtà il parlante sicuramente non legge il secondo «mai» (1 minuto e 15 secondi)

quel ragazzo non dice mai la verità <inspiration> ma tu non fargli **mai** vedere che pensi sia un {<NOISE> bugiardo} <NOISE>

**p2A03B** #19: nella realtà il parlante sicuramente non legge il «ne» (1 minuto e 35 secondi)

Lucio era certo che sarebbe diventato una persona importante un uomo politico o magari un un ministro <inspiration> aveva a cuore il bene della società <sp> rispettava la legge se **ne** teneva un discorso trovava le parole adatte ad ogni situazione

**P2B02B**: mi sembra tutto corretto

In generale, in tutte le registrazioni ho dubbi su «uscì» / «uscii»

# 3. Sottoporre i file audio a Google e Dragon

- Per Dragon posso fornirvi le trascrizioni io
- Può essere utile fare diverse prove di trascrizione, variando i parametri audio… sono curioso di vedere per esempio se c'è davvero differenza tra missaggio interno e uso degli altoparlanti
- Il lavoro deve essere fatto bene: non ha gradazioni

# 4. Preparare l'output di Google e Dragon

- Per il controllo con SCLITE i file devono essere segmentati in modo corrispondente alle trascrizioni disponibili
- Sarebbe possibile anche fare accorpamenti (per esempio, unire tutte le trascrizioni usate in un unico file)
- Tuttavia, consiglio di mantenere la granularità delle trascrizioni fornite da CLIPS
- Il lavoro deve essere fatto bene: non ha gradazioni

# 5. Convertire i file per il controllo con SCLITE

Per ogni testo: tutti e due i file (trascrizione originale e output di Google e Dragon) devono essere registrati in formato trn

Occorre quindi:

- Eliminare tutto ciò che non rappresenta le parole del parlato (note, segni di interpunzione)
- separare con a capo (newline) ogni enunciazione – tipicamente ogni frase, ma qui si può anche scegliere di usare altre unità
- Inserire a fondo enunciazione un codice tra parentesi tonde con identificativo del parlante, trattino basso e numero progressivo: «(a_1)»

Il lavoro deve essere fatto bene: non ha gradazioni

# Sostituzione automatica

- <sp> <lp> <inspiration> <tongue-click> <creacky-voice> <NOISE>
- [screaming]
- ,
- !
- ?
- { }
- *
- /
- Soprattutto, gli spazi doppi!
- Non il + (vedremo come mai)

# 6. Calcolare il WER

Nel campione che ho verificato io i risultati sono stati:

**Dragon**
LFp1A02B: 13,7 (sostituzioni 6,3, cancellazioni 7,3)
LFp1A03B: 8,1 (sostituzioni 3,7, cancellazioni 4,1, inserimenti 0,3)
LFp1B02B: 8,4 (sostituzioni 4,4, cancellazioni 4,1)
LFp2A02B: 7,1 (sostituzioni 3,7, cancellazioni 3,4)
LFp2A03B: 8,7 (sostituzioni 4,3, cancellazioni 4,0, inserimenti 0,3)
LFp2B02B: 9,0 (sostituzioni 4,3, cancellazioni 4,7)
Media: **9,2**

**Google**
LFp1A02B: 4,7 (sostituzioni 2,7, cancellazioni 1,3, inserimenti 0,3)
LFp1A03B: 4,1 (sostituzioni 2,0, cancellazioni 1,4, inserimenti 0,7)
LFp1B02B: 7,1 (sostituzioni 3,0, cancellazioni 3,0, inserimenti 1,0)
LFp2A02B: 4,7 (sostituzioni 3,7, cancellazioni 1,0)
LFp2A03B: 6,7 (sostituzioni 3,7, cancellazioni 2,0, inserimenti 1,0)
LFp2B02B: 5,7 (sostituzioni 3,3, cancellazioni 2,0, inserimenti 0,3)
Media: **5,5**

Il lavoro deve essere fatto bene: non ha gradazioni

# 7. Commentare dal punto di vista linguistico i risultati e gli errori di trascrizione

Per esempio (LFp1A02B):

**Originale**
Lucio era certo che sarebbe diventato una persona importante un uomo politico o magari un ministro aveva a cuore il bene della società rispettava la legge se ne teneva un discorso se teneva un discorso trovava le parole adatte ad ogni situazione si sentiva proprio un buon italiano (p1_19)

**Google**
**lui c'ho** era certo che sarebbe diventato una persona importante un uomo politico o magari un ministro aveva a cuore il bene della società rispettava la legge se ne teneva un discorso se teniamo un discorso trovava le parole adatte ad ogni situazione **[]** sentiva proprio un buon italiano (p1_19)

Questa è l'attività meno prevedibile! La variabilità nei risultati può essere molto forte