Linguistica Computazionale – Lezione 3

Introduzione al NLP: Applicazioni e Problemi

10 Marzo 2004

Cristiano Chesi, chesi@media.unisi.it

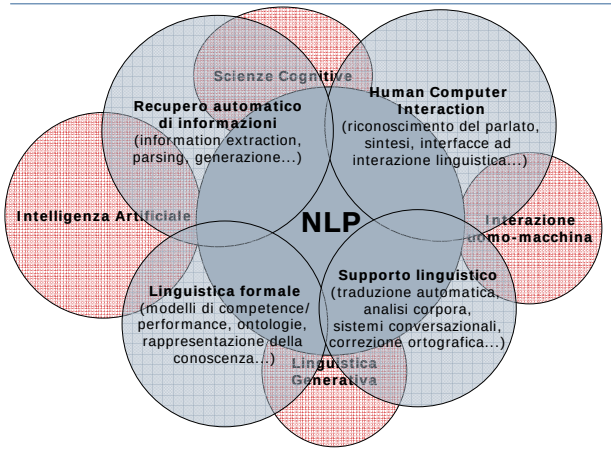
Introduzione al NLP: Applicazioni e Problemi

- Indice
  - Alcuni esempi di Natural Language Processing (NLP)
    - Human Computer Interaction (HCI): dalla fantascienza alla realtà
    - Correzione ortografica
    - Ubiquitous computing
      - input
      - rappresentazione dati lato server
    - Approccio cognitivo-computazionale
  - Aspetti linguistici che vorremmo descrivere computazionalmente
    - lessico
    - parsing

Letture, approfondimenti

- **Bibliografia essenziale**
  - Jurafsky & Martin (2000) *Speech & Language Processing*. Prentice Hall, NJ (Cap. 1)
  - Ferrari G. (1991) *Introduzione al Natural Language Processing*. Calderini Bologna (Cap. 1, 2)
- **Approfondimenti**
  - Silfverberg, MacKenzie Korhonen (1999) *Predicting Text Entry Speed on Mobile Phones*. Ms.
  - Shank (2001) *I'm sorry Dave, I'm afraid I can't do that*. in G. Stork "HAL's Legacy". MIT Press
  - Speech Recognition Grammar Specification Version 1.0 http://www.w3.org/TR/speech-grammar/

Natural Language Processing



Human-Computer Interaction (HCI)

Alcuni esempi di NLP

Eliza (Weizenbaum, **anno 1966**):
Utente: *il mio ragazzo dice che sono sempre depressa*
Eliza: *sono spiacente di sapere che sei depressa*

HAL 9000 (Kubrick & Clarke, 2001 Odissea nello spazio; **anno 1968**):
David: *Apri la saracinesca esterna, Hal.*
Hal: *Mi dispiace David, purtroppo non posso farlo.*

Correttore Grammaticale di Microsoft Word eXPerience
(Expert System, **anno 2005**)
Utente: *"voglio veduto Mario al posto mio"*
(intendendo "voglio vedere Mario al posto mio")
Arturo: *questa forma dialettale deve essere sostituita con l'equivalente in italiano. Sostituire "voglio veduto" con "devo essere veduto" oppure "vado veduto"*

Human-Computer Interaction (HCI)

Alcuni esempi di NLP

Espressioni da evitare

L'uso di termini dialettali è generalmente sconsigliato perché rende il testo incomprensibile alla maggior parte delle persone; purtroppo la radio, la televisione e la stampa quotidiana fanno spesso un uso eccessivo del dialetto, al fine di dare maggiore vivacità al linguaggio parlato. Anche scrittori di fama e di successo utilizzano certe voci dialettali per rendere più colorita la loro prosa.

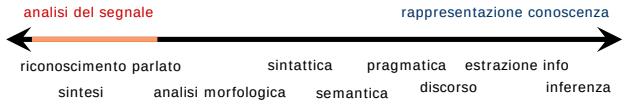
Cosa avrebbe dovuto saper fare HAL 9000:

Alcuni esempi di NLP

- speech recognition / synthesis
  - analisi/produzione del segnale acustico, identificazione delle formanti, sillabazione, suddivisione in parole, identificazione contorni prosodici
- natural language understanding / generation
  - **morfologia** – scomposizione delle parole in unità minime di significato (dogs = dog + s)
  - **sintassi** – definizione delle relazioni strutturali tra parole
  - **semantica** – attribuzione del significato delle espressioni
  - **pragmatica** – attribuzione di intenti in base agli usi/convenzioni linguistiche
  - **discorso** – recupero di relazioni tra unità linguistiche più ampie della singola frase
- information extraction / retrieval
  - identificazione delle porzioni di testo/conoscenza in cui risiede l'informazione rilevante e rielaborazione di tale informazione
- inferenza
  - trarre le adeguate conseguenze dalle informazioni recuperate

Mappa delle applicazioni di NLP

Alcuni esempi di NLP



Cosa si riesce a fare adesso nel 2005:

(inizio …)

- **word processing**
  - **sillabazione** (soddisfacente)
    es. casa > ca-sa
  - **correzione ortografica** (soddisfacente)
    es. caza > casa
  - **correzione grammaticale** (povera)
    es. lo casa > la casa
  - **correzione stilistica** (pessima)
    es. mi trovai per una selva oscura > ero in un bosco buio
- **Human Computer Interaction**
  - **riconoscimento del parlato** (soddisfacente)
    es. /kasa/ > casa
  - **filtraggio/recupero di informazioni** (dipende dal contesto. In contesti ristretti in genere è soddisfacente)
    es. "nel 2005 Volare Web è fallita" > società: Volare Web; stato: fallimento; periodo: 2005
  - **rispondere a domande** (dipende dal contesto. In contesti ristretti in genere è soddisfacente)
    es. dove si trova la penna? > sul tavolo

Cosa si riesce a fare adesso nel 2005:

(… fine)

- **Human Computer Interaction (… continua)**
  - **ricerca "intelligente" su web, corpora, database…** (soddisfacente, ma solo usa euristiche extralinguistiche)
    es. capire se il tipo di ricerca è espansiva (raccolta di molte informazioni) o puntuale (risposta ad una domanda precisa "quanto è alto il monte Everest?")
  - **riassunto automatico e classificazione di un testo** (spesso insoddisfacente)
    es. "Avevo appena finito di tagliare il lesso (parecchio filaccioso, a dire la verità), quando nel rimettermi a sedere osservai, con una disposizione di spirito poco in carattere col mio abito, che chiunque si fosse preso la briga di eliminare il colonnello Protheroe avrebbe reso un gran servigio all'umanità".
    La morte nel Villaggio, A. Christie
    > il protagonista, finito di tagliare il lesso, pensò che pochi si sarebbero dispiaciuti della morte del colonnello P.
  - **pseudo-comprensione** (dipende dal contesto. In contesti ristretti in genere è soddisfacente)
    potresti chiudere questa finestra?
    > [chiusura della finestra di Word in questione]
  - **generazione del linguaggio naturale** (dipende dal contesto. In contesti ristretti in genere è soddisfacente)
    [contesto precedente] > ho chiuso la finestra di Word a cui ti riferivi
  - **traduzione automatica** (soddisfacenti traduzioni parola per parola; grosse difficoltà di disambiguazione)
    [contesto precedente] > I closed the Word window you pointed out.

La correzione ortografica

- **correzione ortografica** è diversa dal **controllo ortografico**: mentre il controllo può limitarsi semplicemente ad accettare/rifiutare una stringa di testo, la correzione deve proporre una forma corretta in alternativa.

Esempio di approccio **ingegneristico**:

1. **definizione** precisa del **problema**
2. **raccolta dati** rilevanti
3. **classificazione** degli errori
4. **ricerca di soluzioni adeguate** ed **efficienti**
   (relativamente alle classi di errori)

La correzione ortografica – classificazione errori

(inizio …)

- all'identificazione degli errori tipici segue solitamente una **categorizzazione** su quattro livelli:
  - **lessicale**
  - **sintattico**
  - **semantico**
  - **pragmatico**
- Va ricordato che ogni errore può essere riconosciuto come tale sia perché è un vero errore (**malformatezza assoluta**), sia perché il sistema non è in grado di trattare, per la limitatezza delle risorse linguistiche utilizzate, la forma che in realtà sarebbe corretta (**malformatezza relativa**)

La correzione ortografica – classificazione errori

(… continua …)

Alcuni esempi di NLP

- **malformatezze lessicali**
  - **Relative**
    - parole non presenti nel lessico del sistema
  - **Assolute**
    - tipografiche (omissioni, sostituzioni, inserzioni involontarie di lettere)
    - cognitive (errata credenza sull'ortografia della parola)
    - fonetiche (errata credenza sull'ortografia della parola in base alla sua pronuncia)

La correzione ortografica – classificazione errori

(… continua …)

Alcuni esempi di NLP

- **malformatezze sintattiche**
  - **Relative**
    - inadeguatezza della teoria sintattica implementata (poche regole > ipergeneralizzazione; troppe regole > inconsistenza, esclusione di strutture in realtà corrette)
    - forme colloquiali o dialettali (espressioni idiomatiche, indicativo al posto del congiuntivo…)
    - pronomi di ripresa (pro-sintagmi ripetuti impropriamente)
  - **Assolute**
    - pronome sbagliato (es. me sono andato)
    - mancanza di accordo tra:
      soggetto - verbo (es. Loro è andati…)
      modi - tempi (es. Voglio vado; vorrei andato)
      determinanti - nomi (es. Lo casa)
      aggettivi - nomi (es. Il mare verdi)
    - omissioni di argomenti obbligatori (es. ho messo sul tavolo _)

La correzione ortografica – classificazione errori

(inizio …)

Alcuni esempi di NLP

- **malformatezze semantiche**
  - **Relative**
    - relazione non presente (relazioni tra gli oggetti non disponibili nella base di conoscenze del sistema)
    - violazione delle restrizioni di selezione (uso di espressioni che violano le restrizioni della base di conoscenze del sistema)
    - sinonimia (mancanza del collegamento semantico tra due sinonimi)
    - polisemia (significati alternativi non presi in considerazione dal lessico di macchina)
  - **Assolute**
    - violazione delle restrizioni di selezione (es. il telescopio nuotò)
    - logica spaziale (es. vieni là)
    - logica temporale (es. domani sono andato a ballare)

La correzione ortografica – classificazione errori

(… fine)

Alcuni esempi di NLP

- **usi figurativi**
  - metafora (es. “con un filo di voce” per “con voce flebile”)
  - metonimia (es. “quel ferro vecchio va rottamato” per “quella macchina”)
  - sineddoche (es. “il mondo ci è nemico” per “si percepisce una certa ostilità”)
  - antonomasia (es. “il divino poeta” per “Dante”)
  - perifrasi (es. “quel coso per asciugare i capelli” per “asciugacapelli”)
  - eufemismo (es. “passare a miglior vita” per “morire”)
  - litote (es. “non è certo un’aquila” per “non è molto intelligente”)
  - iperbole (es. “l’ho detto mille volte” per “l’ho già detto molte volte”)
  - idioma (es. “il dado è tratto” per “ormai la decisione è stata presa”)

La correzione ortografica – trattamento malformatezze

(inizio …)

Alcuni esempi di NLP

- Le varie tecniche che permettono di gestire le malformatezze si basano principalmente su un sistema di **pattern matching** con le forme archiviate nel **lessico** di cui dispone il sistema e su una serie di **euristiche** per decidere le correzioni possibili alle forme errate
  - metodi **simbolici**
    (buona rappresentazione del problema)
  - metodi **subsimbolici**
    (rappresentazione del problema insufficiente)

La correzione ortografica – trattamento malformatezze

(… continua …)

Alcuni esempi di NLP

- **Distanza minima**

  Il sistema inventato da Damerau (Damerau 64) e perfezionato da Wagner (Wagner 74) tratta l'errore come una forma che si differenzia da quella corretta per un numero minimo di operazioni di **inserimento**, **cancellazione**, **sostituzione** e **scambio** di caratteri.

  Il metodo consiste nel calcolare attraverso una funzione diversa da sistema a sistema, la minima distanza di correzione tra le stringhe ortograficamente scorrette e le parole presenti nel vocabolario. Se questa distanza è considerata accettabile il vocabolo è considerato come possibile correzione della forma non standard.

  Il grave difetto di questo approccio è l'**inefficienza**: l'elaborazione richiede un numero *n* di confronti, con *n* uguale al numero delle parole del vocabolario.

La correzione ortografica – trattamento malformatezze

(… continua …)

Alcuni esempi di NLP

- **Chiave di somiglianza** (algoritmo SOUNDEX, Odell e Russel 1918, correzione di errori fonetici, migliorato ed esteso da Davidson 1962)

  Questa tecnica associa ad ogni stringa una chiave costruita in modo che tutte le parole scritte o pronunciate in un modo simile abbiano una chiave uguale o molto somigliante.

  Confrontando, non le parole, ma solo le chiavi si ottengono le candidate alla correzione della parola scorretta.

  chiave = prima lettera della parola + sequenza di numeri associati secondo certe regole e statistiche di frequenza
  Gli zero e i numeri ripetuti vengono eliminati

  Esempio:

vocali	b, f, p, v	altre consonanti
0	1	2

  casa = c020 > c2; *csa* = c20 > c2

La correzione ortografica – trattamento malformatezze

(… continua …)

Alcuni esempi di NLP

- **Chiave di somiglianza** – migliorata
  (Pollock e Zamorra, SPEEDCOP, 84)

  migliorano il metodo della chiave di somiglianza attribuendo due tipi di chiavi ad ogni parola del vocabolario, basandosi sulle seguenti osservazioni riguardo alla distribuzione degli errori:

  1. l'ordine delle vocali è spesso mantenuto invariato
  2. raramente viene sbagliata la prima lettera, statisticamente gli errori si situano verso la fine della parola

  - **skeleton key** = prima lettera della parola + consonanti nell'ordine in cui si presentano nella parola senza ripetizioni + vocali (sempre nell'ordine e sempre senza ripetizioni) (es. gambero = gmbraeo);
  - **omission key** = consonanti, senza ripetizione in un ordine di frequenza (determinato staticamente) e poi dalle vocali, senza ripetizioni, nell'ordine in cui si presentano nella parola.

  Gestiti il 94% degli errori singoli e tra il 74% e l'88% degli errori complessivi presenti nel testo

La correzione ortografica – trattamento malformatezze

(… continua …)

Alcuni esempi di NLP

- **Regole**

  La tecnica basata su regole utilizza algoritmi ed euristiche per rappresentare la conoscenza necessaria per determinare quali sono le regole che il termine sbagliato ha violato e le correzioni necessarie per correggerlo (es. restrizioni fonotattiche + informazioni sull'ordine delle lettere sulla tastiera).

  Una volta applicate tutte le regole a disposizione, i risultati vengono presentati all'utente secondo una stima di probabilità.

  Il sistema realizzato da Yannakoudakis e Fawthrop (83) permette una precisione intorno al 76% di errori rilevati. Means (Means 88) affina la tecnica inserendo nel suo correttore oltre alle regole della morfologia inglese altre regole di abbreviazione e flessione non standard migliorando in parte i risultati del primo prototipo.

La correzione ortografica – trattamento malformatezze

(… continua …)

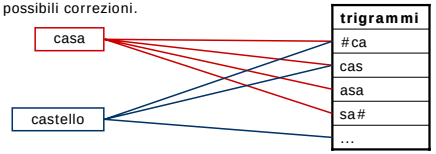
Alcuni esempi di NLP

- **N-grammi** (Kohonen 80; DeHer 82; Angell et al. 83; DeSmedt e VanBerkel 88)

  parola = insieme di sottostringhe (n-grammi) che si sovrappongono

  esempio:
  casa = #c + ca + as + sa + a# (bi-grammi)
  strumento = #st str tru rum ume men ent nto to# (tri-grammi)

  vocabolario = tabella di n-grammi indicizzati; ogni indice rinvia ad un determinato termine nel vocabolario di macchina.
  L'insieme dei rinvii determina il campo d'attivazione delle parole e seleziona le possibili correzioni.



La correzione ortografica – trattamento malformatezze

(… continua …)

Alcuni esempi di NLP

- **N-grammi** - la procedura di correzione degli errori

  1. ogni parola non corretta viene scomposta nei suoi n-grammi
  2. tali n-grammi vengono utilizzati come indici nella tabella per individuare le possibili parole candidate alla correzione
  3. i vocaboli candidati alla correzione saranno tutti quelli che presentano un livello soglia di n-grammi in comune con il termine sbagliato.

  - Un esempio d'implementazione di questo metodo è il programma ACUTE realizzato da Angell e al. (83). Il sistema utilizza una tabella a tri-grammi
  - DeSmedt e VanBerkel (88) propongono una diversa analisi chiamata triphone analysis che permette di correggere errori nel riconoscimento del parlato.
  - Le prestazioni di questo sistema variano a seconda dei vocabolari utilizzati e nessun test standardizzato ha paragonato questo approccio agli altri presentati.

La correzione ortografica – trattamento malformatezze

(… continua …)

Alcuni esempi di NLP

- **Analisi probabilistica**

  - utilizzato per migliorare le prestazioni del precedente metodo con n-grammi.
  - due indici che vengono solitamente assegnati alle possibili parole di correzione:
    - **probabilità di transizione** (la probabilità che ha una determinata lettera di seguire una sequenza di caratteri)
    - **probabilità di confusione** (stima della probabilità di sostituzione tra una lettera e l'altra)
  - I primi studi fatti hanno evidenziato come questa sola tecnica non sia sufficiente per ottenere risultati soddisfacenti. Kashyap e Oommen (84) hanno utilizzato questo metodo probabilistico per correggere parole con meno di sei caratteri (svantaggiate dal precedente approccio per n-grammi). Church e Gale (91) propongono con il loro sistema, CORRECT, un approccio ancora più complesso utilizzando quattro matrici di confusione contenenti 44 milioni di parole errate tratte da vari testi.

La correzione ortografica – trattamento malformatezze

(… continua …)

Alcuni esempi di NLP

- **Reti neurali**
  - L'applicazione delle reti neurali a questo campo cerca di sfruttare la versatilità che caratterizza questi sistemi per approssimare funzioni euristiche implicite: vista l'intrinseca difficoltà nel definire "regole di violazione", si cerca di far apprendere alla rete ad associare forme errate con forme presenti nel lessico attraverso cicli di addestramento in cui si mostrano "associazioni cognitivamente plausibili".
  - Rumelhart, Burr, Matan (Rumelhart 86; Burr 87; Matan 92) hanno adottato questo approccio in sistemi di correzione che, secondo una stima di Kukich (Kukich 92), possono raggiungere una capacità di correzione che si aggira intorno al 75% dei termini errati.
  - l'efficacia dell'approccio è strettamente dipendente dal tipo di input che si sceglie di dare in pasto alla rete (stringhe di caratteri semplici, n-grammi, sequenze fonetiche…); il problema di una correzione efficiente viene perciò semplicemente spostato, ma non risolto e una riflessione "simbolica" sulla natura del problema sembra sempre comunque fondamentale per il trattamento del problema.

La correzione ortografica – trattamento malformatezze

(… continua …)

Alcuni esempi di NLP

- **Espressioni dipendenti dal contesto**
  - Vari autori (Thompson 80, Eastman e McLean 81; Young 91) hanno messo in evidenza che gli errori prodotti, dipendenti dal contesto, sono tra il 25% e il 50% degli errori totali, e di questi circa il 75% è di ordine sintattico.
  - Esistono due principali tipi di approccio:
    - **simbolico** – necessita di un robusto parser e degli analizzatori morfologici e sintattici (richiede una solida teoria linguistica e una efficiente implementazione software)
    - **probabilistico** – utilizza delle tabelle di probabilità per determinare le sequenze di termini consentite (richiede una mole consistente di dati)

La correzione ortografica – trattamento malformatezze

(… continua …)

Alcuni esempi di NLP

- **Espressioni dipendenti dal contesto** (Microsoft Word XP)
  - Regole grammaticali:
    - **Punteggiatura** (dopo aver mangiato, decise di lasciare la tavola)
    - **Maiuscole** (le scarpe di paola sono molto costose)
    - **Genere-Numero** (Franco ha comprato dei pantaloni nuove)
    - **Concordanza Soggetto-Verbo** (Il cane e il gatto ha mangiato i resti del pranzo; Io speriamo di vincere un premio. Gli scolari sono uscito alcuni minuti prima del solito)
    - **Frasi** (segnala i più comuni errori relativi alla frase e alla sua costruzione. Esempi di errori rilevati: La donna disse sarebbe andata in città)
    - **Verbi** (segnala gli errori relativi all'uso di un verbo con l'ausiliare sbagliato; L'aereo ha arrivato con parecchi minuti di ritardo sull'orario previsto. Io ho potuto partire per la Francia grazie all'aiuto di mio padre)
    - **Aggettivi** (segnala gli usi impropri degli aggettivi. Esempi di errori rilevati: lavoro molto poco in primavera; corregge in "pochissimo")
    - …

La correzione ortografica – trattamento malformatezze

(… continua …)

Alcuni esempi di NLP

- **Espressioni dipendenti dal contesto** (Microsoft Word XP)
  - Regole grammaticali:
    - **…**
    - **Articoli** (Il yogurt è un alimento molto indicato per i bambini)
    - **Elementi della frase** (segnala un insieme di errori commessi con una certa frequenza e che coinvolgono diversi elementi della frase. Esempi di errori rilevati:
      La torre di Pisa è tanto alta come bella.
      ma anche: ho mangiato tanto cioccolato come quando ero bambino > sostituire come con quanto)
    - **Preposizioni** (segnala l'esattezza nell'uso delle preposizioni insieme con sostantivi, aggettivi, pronomi, verbi ed avverbi, e segnala alcune tra le più comuni forme del parlato che sono errate nei testi scritti. Esempi di errori rilevati: Il nonno si è addormentato come al solito. La nuova macchina stampa 100 copie all'ora. Con domani inizieremo la costruzione della seconda ala dell'edificio

La correzione ortografica – trattamento malformatezze

(… fine)

- **Espressioni dipendenti dal contesto** (Microsoft Word XP)
  - Regole di stile:
    - **Espressioni da evitare / parole ridondanti** (Ed è per questo che abbiamo deciso di modificare i piani di produzione, Per potere avere una promozione, bisogna meritarsela. Quella maionese è lievemente acidula. Le domande devono essere presentate entro e non oltre le ore 17 del 12 ottobre)
    - **Leggibilità** (l'arciere non sapeva scegliere fra frecce rosse e frecce verdi. Il treno arrivò a Ascoli con due ore di ritardo. Il di lui cane è molto affettuoso)
    - **Termini ripetuti** (La casa vicina al ponte è più bella della casa di tuo padre. Per eliminare un problema, abbiamo eliminato anche molte cose utili)
    - **Uso errato** (Questi ragazzi hanno un gran spirito d'iniziativa. Abbiamo deciso di comprarlo sia lui che io. Malgrado tutto, siete riusciti ad arrivare in tempo a scuola)

Ubiquitus computing

(inizio …)

- **Le idee di base**
  - **Ubiquitus computing** Vs. **Virtual Reality**
  - **calm technology**
    ("computer" invisibile)
  - **interfacce naturali**
    (estremizzazione dell'User Friendly)
  - **integrazione ambientale** e **contestuale**
    (dispositivi sensibili)

Ubiquitus computing

(… continua …)

- **Alcuni dispositivi in commercio adesso**

- **alta connettività**
- **input/output audio**
- **display ridotto**
- **tastiere limitate**
- **limitate risorse computazionali**
- **necessità d'uso immediato (anche in contesti in cui la modalità visiva è occupata, tipo durante la guida)**

Ubiquitus computing – un problema di input

(… continua …)

- **Approcci realistici ai dispositivi attualmente in commercio:** SMS e cellulari
  1. **definizione precisa del problema**
     composizione il più veloce possibile dei messaggi di testo tenendo conto dei vincoli della tastiera
  2. **raccolta dati**
     esempi di messaggi, parole utilizzate, struttura delle parole
  3. **classificazione**
     problemi probabilistici, semplicemente combinatori, morfologici
  4. **ricerca di soluzioni adeguate ed efficienti**
     modelli di selezione per numero minimo di pressioni, modelli probabilistici

Ubiquitus computing – un problema di input

(… continua …)

- **Vincoli della tastiera e metodi di composizione di SMS** (Silfverberg e al. 1999)

Ubiquitus computing – un problema di input

(… continua …)

- **Alcune soluzioni possibili:**

	C	A	S	A	tot
Multi-press	2-2-2	2	7-7-7-7	2	**8**
two-key	2-3	2-1	7-4	2-1	**8**
T9	2	2	7	2	**4**

Ubiquitus computing – T9

(… fine)

	abc	abc	pqrs	abc
T9	2	2	7	2

- **Risorse linguistiche necessarie per il T9**
  - Vocabolario
  - Indici di frequenza (es. premendo 6-6 in inglese “ON” viene selezionata prima di “NO” sulla base di osservazioni statistiche basate su corpora, in questo caso il British National Corpus, si calcola che il lavoro di disambiguazione non superi il 5% delle produzioni)
- **Risorse non linguistiche per valutare i modelli**
  - Legge di Fitts (modello quantitativo di valutazione dei movimenti rapidi diretti ad un fine)
- **Risultati** (in Words Per Minutes, wpm)
  - Multi-press: 25-27 wpm
  - Two-key: 22-25 wpm
  - T9: 41-46 wpm

Ubiquitus computing – Accesso a risorse dati

(inizio …)

- **Linguaggi naturali e linguaggi “taggati”**
  - parentesi [[A] [B C]]
  - HTML 123 Mario Rossi
  - XML
```
<studente> <id> 123 </id>
<nome> Mario Rossi </nome>
</studente>
```
  - voiceXML
```
<vxml version="2.0">
<form>
<block>123, Mario Rossi</block>
</form>
</vxml>
```

- **Il VoiceXML** permette di standardizzare la creazione di dialoghi (uomo-servizio informatizzato) basati sulla modalità vocale in particolare (Interactive Voice Response, IVR):
  - sintesi del parlato (o utilizzo di audio digitalizzato)
  - toni della tastiera (Dual Tone Multi Frequency, DTMF)
  - riconoscimento di comandi vocali
  - specificazione di grammatiche

- **Un documento VoiceXML** (applicazione):

```
<vxml version="2.0">
    <form>
        <field name="prezzo" type="boolean">
        <prompt> Vuoi sapere il prezzo del biglietto? </prompt>
        <filled> Ok!
        <if cond="prezzo"> Ecco qua:
                <goto next="prezzo.vxml" />
        <else /> Allora torniamo alla lista…
                <goto next="lista.vxml" />
        </if>
        </filled>
        </field>
    </form>
</vxml>
```


- **Specificazione di una grammatica in VoiceXML** (Speech Recognition Grammar Specification, SRGS):

```
<grammar xml:lang="IT" type="application/srgs+xml" version="1.0"
    mode="voice">
    <rule id="yes_no_cancel" scope="public">
        <one-of>
            <item tag="no">no</item>
            <item tag="yes">sì</item>
            <item tag="yes">va bene</item>
            <item tag="cancel">annulla</item>
        </one-of>
    </rule>
</grammar>
```



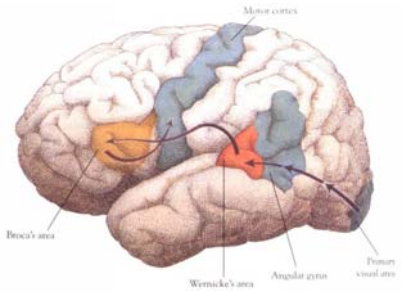
```
<grammar version="1.0" mode="voice" root="basicCmd">

<rule id="basicCmd" scope="public">
    <example> please move the window </example><example> open a file </example>
    <ruleref uri="#command"/>
</rule>
<rule id="command">
    <ruleref uri="#action"/> <ruleref uri="#object"/>
</rule>

<rule id="action"> <one-of>
    <item weight="10"> open <tag>TAG-CONTENT-1</tag> </item>
    <item weight="2"> close <tag>TAG-CONTENT-2</tag> </item>
    <item weight="1"> delete <tag>TAG-CONTENT-3</tag> </item>
    <item weight="1"> move <tag>TAG-CONTENT-4</tag> </item>
    </one-of> </rule>
<rule id="object">
    <item repeat="0-1">
        <one-of> <item> the </item> <item> a </item> </one-of>
    </item>
        <one-of> <item> window </item> <item> file </item> <item> menu </item>
        </one-of>
</rule>
</grammar>
```

Approccio cognitivo-computazionale

Alcuni esempi di NLP



- Come ogni modulo cognitivo (tatto, equilibrio, movimento, visione…) il linguaggio esprime una qualche forma di **competenza** (data-structure)
- **processing**
- **performance** (risorse limitate)

Rappresentazione del problema linguistico

(inizio …)

Aspetti linguistici da descrivere

- **Competence** (data-structure, natura del problema)
  - di che tipo di struttura dati ha bisogno la conoscenza linguistica?
    - una parola può iniziare per *ma…* (*mare*) ma non per *mr…*
    - la e di case ha un valore diverso da quella di mare
    - "le case sono sulla collina" Vs. *"case le collina sono sulla"
    - il gatto morde il cane > sogg: gatto(agente); verbo: morde(azione); ogg: cane(oggetto)
    - ?il tostapane morde il gatto
    - l'espressione "le case" si riferisce ad un gruppo di case evidente dal contesto (Vs. "delle case")
  - ad ogni livello si devono specificare delle primitive elementari:
    - **fonemi** - tratti segmentali e soprasegmentali
    - **morfemi** - identificazione delle regole combinatorie
    - **parole** - gruppi di morfemi significativi
    - **sintagmi** - gruppi tipizzati di parole che esprimono relazioni
    - **elementi tematici** - paziente, agente…
    - **elementi discorsivi** - convenzioni, relazioni pragmatiche pertinenti…

Rappresentazione del problema linguistico

(… continua …)

Aspetti linguistici da descrivere

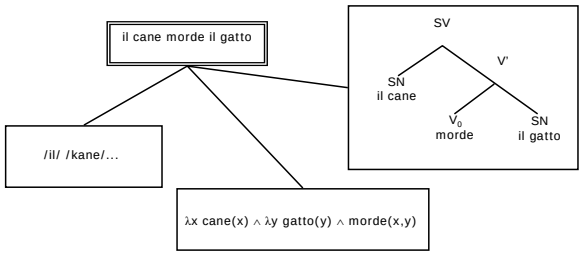
- **processing** (competence in uso)
  - precise specifiche di combinazione; come si usa la conoscenza codificata dalla struttura dei dati:
    - **livello fonologico -** restrizioni fonotattiche che impediranno la combinazione di certe concatenazioni di tratti fonemici o la riduzione di determinate sequenze in altre,
    - **livello morfologico** - regole di combinazione morfofonemiche che permetteranno, ad esempio in italiano, di flettere "mangiare" in "mangiato" e "sapere" in "saputo"…
  - Un esempio storico (probabilmente il primo): Panini (400-600AC) descrive il sanscrito usando una serie di **regole di produzione** sotto forma di aforismi (**sutra**): partendo da circa 1700 elementi base suddivisi in classi (nomi, verbi ecc.) e indicando le regole di combinazione (circa 4000), si riusciva (almeno teoricamente) a derivare ogni forma accettabile in sanscrito.
  - **processing** può essere diverso da **performance**, cioè dallo studio dell'uso delle risorse linguistiche dato un accesso limitato (realistico) a certe risorse (es. memoria a breve termine).

Rappresentazione del problema linguistico

(… continua …)

Aspetti linguistici da descrivere

- **Lessico**
  - il modello dello **spiral notebook**



  - ogni livello deve poter essere mappabile con gli altri livelli.

Rappresentazione del problema linguistico

(… continua …)

Aspetti linguistici da descrivere

- La **complessità del problema** deriva dal fatto che la mappatura non è sempre univoca:
  - **ambiguità lessicale** (la vecchia legge la regola)
  - **ambiguità sintattica** (ho visto il ragazzo nel parco con il cannocchiale)
  - **ambiguità semantica** (la pesca non è stata fruttuosa)
- morale: un problema è più difficile se contemporaneamente devo valutare più possibilità, tutte ugualmente plausibili. Scelte multiple tra cui non ho euristiche di scelta portano al **non-determinismo**.

Rappresentazione del problema linguistico

(… fine)

Aspetti linguistici da descrivere

- **Parsing**: accettare/rifiutare un input e, in caso di accettazione, assegnare a tale input un'appropriata descrizione strutturale
  - **lessicale** (tagger): casa = Part-of-speech (Nome comune)
  - **morfologico**: casa = {N, sing, fem}
  - **sintattico** (parser): $[_S\ [_{VP}\ [_{DP}$ Gianni$]\ [_{V'}$ ama $[_{DP}$ Maria$]\ _{V'}]\ _{VP}]\ _S]$
  - **semantico**: f(agente, paziente) > ama(Gianni, Maria)
  - …

Concetti chiave della lezione di oggi

- Applicazioni di NLP
  - aree di ricerca e relazioni con altre discipline
  - sogni (HAL9000) e realtà (Correttore ortografico di Word)
  - il caso della correzione ortografica (come si riflette su un problema, varie possibili euristiche per risolverlo, reverse engineering applicato al correttore di Word)
  - il caso dell'ubiquitous computing (specificità e utilità delle interfacce in NL, miglioramento delle possibilità di input, struttura dell'informazione lato server)
  - modellizzazione di un aspetto dell'intelligenza umana: il linguaggio (Competence Vs. Processing (Vs. Performance))
  - accenni al lessico, al parsing e al non-determinismo

Prossima lezione

(Domani, Venerdì 11 Marzo, ore 10-13, Aula Workshop 2, S. Francesco)

- Corpora linguistici
  - motivazioni e struttura
  - database e corpora
- Lessici computazionali
  - struttura generale
  - un esempio: Wordnet
- Analisi automatiche
  - analisi morfologica
  - introduzione all'acquisizione del linguaggio