

---

## Linguistica Computazionale – Lezione 1

# Natural Language Processing (NLP) e Machine Translation (MT)

6 Marzo 2007

Cristiano Chesi, chesi@media.unisi.it

---

## Natural Language Processing e Machine Translation

- Indice
  - Note sul corso
  - Introduzione al Natural Language Processing (NLP)
    - Human Computer Interaction (HCI): dalla fantascienza alla realtà
    - Ubiquitous computing
    - Approccio cognitivo-computazionale
  - Traduzione automatica
    - Cenni storici
    - Modelli teorici
    - Alcuni esempi classici

---

## Note sul corso

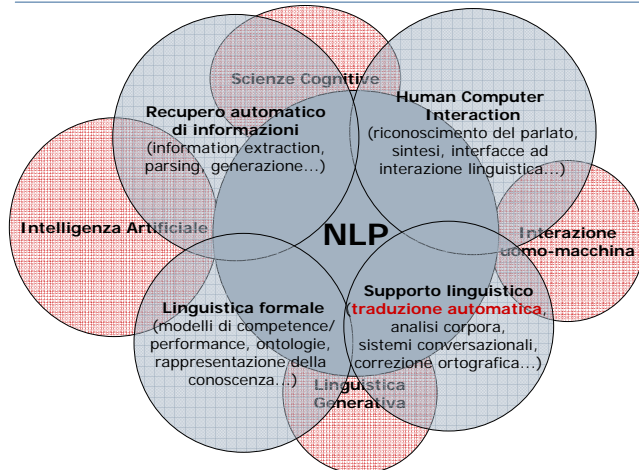
- **Obiettivi formativi**
  - consapevolezza delle necessità/difficoltà del trattamento automatico di lingue diverse
  - concezione e sviluppo di un sistema automatico (completo e complesso) di elaborazione del linguaggio attraverso un lavoro di gruppo
- **Strumenti didattici**
  - lezioni frontali (dispense pdf recuperabili sul sito: [www.ciscl.unisi.it](http://www.ciscl.unisi.it))
  - discussione in classe
  - laboratori
- **Valutazione**
  - **Partecipazione in classe** (20% del voto finale)
  - **Discussioni di gruppo** (40% del voto finale):  
**Mercoledì 4 Aprile** (ciascuno dovrà individuare almeno 1 problema rilevante da discutere in classe)  
**Mercoledì 18 Aprile** (ogni gruppo dovrà illustrare gli intenti/risultati della propria riflessione)
  - **Tesina** (40% del voto finale)  
max 10 pagine in cui si evidenzia il contributo individuale al lavoro di gruppo; discussione orale dell'elaborato per l'appello orale

---

## Lecture, approfondimenti

- **Bibliografia essenziale**
  - Hutchins & Somers (1992) *An introduction to machine translation* London: Academic Press, 1992  
<http://ourworld.compuserve.com/homepages/WJHutchins/IntroMT-TOC.htm>
- **Approfondimenti**
  - Hutchins, J. (2001) *Towards a new vision for MT*. Introductory speech at MT Summit VIII Conference  
<http://ourworld.compuserve.com/homepages/WJHutchins/MTS-2001.pdf>
  - Jurafsky, D. & Martin, J. H. (2000) *Speech and Language Processing*. Prentice-Hall.  
<http://www.cs.colorado.edu/~martin/slp.html>
  - Alcuni esempi di MT:  
<http://www.foreignword.com/Technology/mt/mt.htm>

## Natural Language Processing



5

Linguistica Computazionale A.A. 2006-07 – C. Chesì

## Human-Computer Interaction (HCI)

### Introduzione al 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 2007)

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"*

6

Linguistica Computazionale A.A. 2006-07 – C. Chesì

## Human-Computer Interaction (HCI)

### Introduzione al 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.



7

Linguistica Computazionale A.A. 2006-07 – C. Chesì

## Cosa avrebbe dovuto saper fare HAL 9000:

### Introduzione al NLP

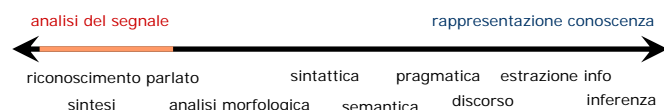
- **speech recognition / synthesis**
  - analisi/produzione del segnale acustico, identificazione delle formanti, sillabazione, suddivisione in parole, identificazione contorni prosodici
- **natural language understanding / generation**
  - **morfologia** – scomposizione delle parole in unità minime 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

8

Linguistica Computazionale A.A. 2006-07 – C. Chesì

## Mappa delle applicazioni di NLP

### Introduzione al NLP



9

Linguistica Computazionale A.A. 2006-07 – C. Chesì

## Cosa si riesce a fare adesso nel 2007:

(inizio ...)

### Introduzione al NLP

#### ■ 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

10

Linguistica Computazionale A.A. 2006-07 – C. Chesì

## Cosa si riesce a fare adesso nel 2007:

(... fine)

### Introduzione al NLP

#### ■ Human Computer Interaction (... continua)

- ricerca "intelligente" su web, corpora, database... (soddisfacente, ma solo se 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 leso (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 servizio all'umanità".  
*La morte nel Villaggio, A. Christie*  
> il protagonista, finito di tagliare il leso, 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... ma poi affronteremo il tema più nel dettaglio)  
[contesto precedente] > I closed the Word window you pointed out.

11

Linguistica Computazionale A.A. 2006-07 – C. Chesì

## Ubiquitous computing

(inizio ...)

### Introduzione al NLP

#### ■ Le idee di base

- Ubiquitous computing Vs. Virtual Reality
- calm technology  
("computer" invisibile)
- interfacce naturali  
(estremizzazione dell'User Friendly)
- integrazione ambientale e contestuale  
(dispositivi sensibili)

12

Linguistica Computazionale A.A. 2006-07 – C. Chesì

## Ubiquitus computing

(... continua ...)

### Introduzione al NLP



#### ■ Dispositivi oggi in commercio

- 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)

13

Linguistica Computazionale A.A. 2006-07 – C. Chesì

## Ubiquitus computing

(... fine)

### Introduzione al NLP

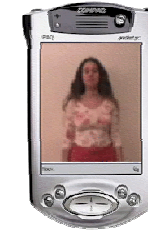
- Traduzione automatica e dispositivi portatili  
**Blue Sign Translator** (<http://bluesign.dii.unisi.it/>)

#### Obiettivo:

- Espressioni in linguaggio verbale  
>  
Lingua Italiana dei Segni (LIS)

#### Cosa occorre minimalmente:

- Lessici delle due lingue
- Regole di "traslazione" da un ordine tra le parole ad un ordine tra segni



ad esempio:

\_\_\_\_\_rel

LIS: CANE GATTO INSEGUIRE, TORNARE A CASA

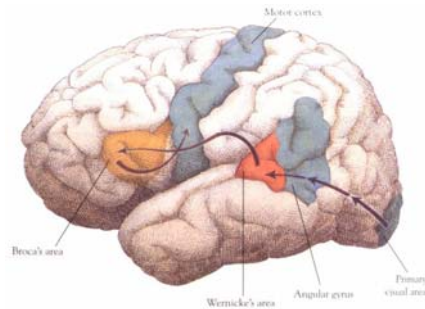
ITA: 'il cane che ha inseguito il gatto è tornato a casa'

14

Linguistica Computazionale A.A. 2006-07 – C. Chesì

## Approccio cognitivo-computazionale

### Introduzione al NLP



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

15

Linguistica Computazionale A.A. 2006-07 – C. Chesì

## Rappresentazione del problema linguistico

(inizio ...)

### Introduzione al NLP

- **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 *g* di *casg* 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...

16

Linguistica Computazionale A.A. 2006-07 – C. Chesì

## Rappresentazione del problema linguistico

(... continua ...)

### Introduzione al NLP

- **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).

17

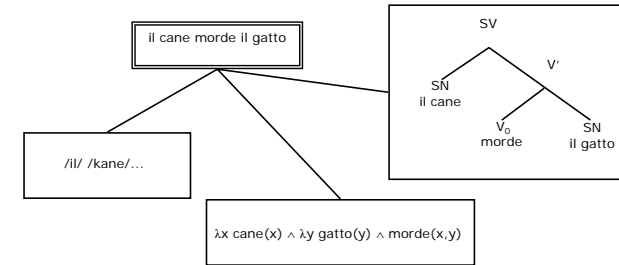
Linguistica Computazionale A.A. 2006-07 – C. Chesì

## Rappresentazione del problema linguistico

(... continua ...)

### Introduzione al NLP

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



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

18

Linguistica Computazionale A.A. 2006-07 – C. Chesì

## Rappresentazione del problema linguistico

(... continua ...)

### Introduzione al NLP

- 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**.

19

Linguistica Computazionale A.A. 2006-07 – C. Chesì

## Rappresentazione del problema linguistico

(... fine)

### Introduzione al NLP

- **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): [<sub>s</sub> [<sub>VP</sub> [<sub>DP</sub> Gianni] [<sub>V</sub> ama [<sub>DP</sub> Maria] <sub>v</sub>] <sub>VP</sub>] <sub>s</sub>]
  - **semantico**: f(agente, paziente) > ama(Gianni, Maria)
  - ...

20

Linguistica Computazionale A.A. 2006-07 – C. Chesì

## Che cos'è la MT

### Traduzione Automatica (MT)

"[Tradurre significa] sostituire il materiale testuale di una lingua (SL) con il materiale testuale di un'altra lingua (TL)"  
(Catford 1965: 20)

"Tradurre consiste nel produrre in una lingua target il più vicino equivalente naturale del materiale testuale della lingua di origine, prima di tutto rispetto al significato, poi allo stile"  
(Nida 1975: 32)

"La traduzione è impossibile in teoria, ma possibile in pratica"  
Mounin (1967)

Catford, J. C. (1965) *A Linguistic Theory of Translation*. Oxford Press, England.

Nida, E. (1975) *A Framework for the Analysis and Evaluation of Theories of Translation*. in Brislin, R. W. (ed) (1975) *Translation Application and Research*. Gardner Press, New York.

Mounin, G. (1967) *Les problèmes théoriques de la traduction*. Paris

21

Linguistica Computazionale A.A. 2006-07 – C. Chesì

## Un po' di storia

(inizio...)

### Traduzione Automatica (MT)

- **Prima del computer**
  - nel 17° secolo si parlava di traduzione automatica (Cartesio e Leibniz);
  - metà anni '30, un franco-armeno, Georges Artsrouni e un russo, Petr Troyanskii, chiesero il brevetto per "macchine traduttrici"
- **I pionieri (1947-1954)**
  - Marzo 1947 lettera di Warren Weaver a Norbert Wiener. T
  - Luglio 1949 Weaver memorandum (successi dei decifраторi di codici, teoria dell'informazione con Shannon, principi linguistici universali)
  - 1954 prima dimostrazione (IBM e Georgetown University). Impressionante nonostante le poche risorse, giustificò fior d'investimenti negli Stati Uniti
- **La grande disillusione (1954-1966)**
  - Dizionari bilingue e regole di riordinamento (troppa complessità e poca generalità)
  - Nuovi paradigmi linguistici (grammatiche generativo-transformativi e a dipendenza)
  - Insormontabili barriere semantiche
  - Alcuni sistemi commerciali: Mark II system (IBM e Washington University), USAF Foreign Technology Division; Euratom in Italia (poco soddisfacenti qualitativamente, ma veloci)
  - 1964 Automatic Language Processing Advisory Committee (ALPAC)
  - 1966 famoso report: la MT è più lenta, meno accurata e 2 volte più cara della traduzione umana. Quindi niente più soldi per la MT, ma solo per lo sviluppo di risorse linguistiche e sistemi di supporto alla traduzione manuale

22

Linguistica Computazionale A.A. 2006-07 – C. Chesì

## Un po' di storia

(... continua ...)

### Traduzione Automatica (MT)

- **Dopo il report ALPAC (1966-1980)**
  - Canada, Francia e Germania continuano le ricerche a dispetto degli USA.
  - Systran (1970), Meteo (Università di Montreal) traduzione automatica bollettini meteo
  - Esponenziale crescita di documenti da tradurre nelle commissioni europee (1976)
  - Dalla metà degli anni '70, nuovi soggetti richiedono rapide traduzioni (semi)automatiche (report tecnici, aziende, articoli scientifici). La miopia del report ALPAC risulta evidente.
- **Gli anni '80**
  - Sistemi mainframe: Systran (varie coppie di lingue); Logos (Tedesco-Inglese, Inglese-Francese); Pan American Health Organization (Spagnolo-Inglese); Métal (Tedesco-Inglese); traduttori Giapponese-Inglese nelle industrie di computer.
  - L'avvento dei microcomputers, primi semplici sistemi di processamento testi.
  - Sistemi interlingua e basi di conoscenza: GETA-Ariane (Grenoble), SUSY (Saarbrücken), Mu (Kyoto), DLT (Utrecht), Rosetta (Eindhoven), Eurotra (Comunità Europea)

23

Linguistica Computazionale A.A. 2006-07 – C. Chesì

## Un po' di storia

(... fine)

### Traduzione Automatica (MT)

- **I primi anni '90**
  - Approcci puramente statistici (Candide, IBM): i risultati sono molto incoraggianti.
  - Traduzione basata su esempi (quindi non più metodi 'rule-based' ma ampi corpora testuali)
  - Sistemi misti (corpus-based, rule-based), speech translation, speech synthesis (ATR, Nara, Giappone e progetto JANUS (ATR, Carnegie-Mellon University, University of Karlsruhe), progetto VerbMobil (Germania)).
  - Nuovi ambiti di impiego della MT: strumenti di supporto per i traduttori professionisti, focalizzazione su linguaggi controllati e domini più ristretti; componenti per il recupero di informazioni multilingua.
- **Fine anni '90, nuovo millennio**
  - software localisation
  - richiesta di MT software per i PC domestici
  - disponibilità di prodotti di traduzione on-line (e.g. Babelfish di AltaVista, Babylon Translator, Google Translator...).
  - Richiesta di traduzione diretta di applicazioni Internet (posta elettronica, pagine web, etc.), risposte in tempo reale, piuttosto che risposte di qualità. Word processing
  - Dal punto di vista della ricerca l'area di maggior espansione è sicuramente quella dell'approccio basato su esempi e l'approccio statistico.

<http://ourworld.compuserve.com/homepages/WJHutchins/Nutshell.htm>

24

Linguistica Computazionale A.A. 2006-07 – C. Chesì

## Aree cruciali per la MT

Traduzione Automatica (MT)

### □ High quality translation Vs. Rough translation

- velocizzare i tempi di traduzione umana (Computer-Aided Human Translation, CAHT o CAT)
- IR cross-linguistico
- filtraggio informazioni (spam)
- marketing
- traduzione di manuali tecnici (sottolinguaggi)

25

Linguistica Computazionale A.A. 2006-07 – C. Chesì

## Problemi per la traduzione

Traduzione Automatica (MT)

- **Eteronimo** = unità lessicale con forma identica ad un'altra unità lessicale ma con significato diverso
- **Divergenze sintattiche**
  - **Divergenze strutturali**  
The man entered the room > L'uomo entrò nella stanza
  - **Divergenze tematiche**  
John likes Mary > A Gianni piace Maria

26

Linguistica Computazionale A.A. 2006-07 – C. Chesì

## Problemi per la traduzione

Traduzione Automatica (MT)

### □ Divergenze sintattiche (... continua)

#### ■ Divergenze categoriali

I'm scared > Ho paura

#### ■ Divergenze di inglobamento

To shelve a book > Riporre su uno scaffale un libro

#### ■ Divergenze lessicali

To take a shower > Fare una doccia

### □ Ambiguità

(vedi slide 19)

27

Linguistica Computazionale A.A. 2006-07 – C. Chesì

## Problemi per la traduzione

Traduzione Automatica (MT)

### □ Divergenze semantiche

#### ■ Divergenze di lessicalizzazione

towel > asciugamano (parola composta)

private > soldato semplice

perifrasi, circonlocuzioni ...

mollica > ?

28

Linguistica Computazionale A.A. 2006-07 – C. Chesì

## La variazione linguistica

Traduzione Automatica (MT)

- **Studi tipologici** (Croft 1990, Comrie 1989)  
similarità e differenze sistematiche tra le lingue
- **Universali di Greenberg** (1931)
- **Parametri Chomskyani** (Chomsky 1981)  
es. parametro testa-complemento

Comrie, B. (1989) *Language universals and linguistic typology: Syntax and morphology*. Oxford: Blackwell, 2nd edn.

Croft, W. (1990) *Typology and universals*. Cambridge: Cambridge University Press.

Greenberg, J.H. (1963) Some universals of grammar with particular reference to the order of meaningful elements. In: *Universals of language*, ed. by J.H. Greenberg, 73-113. Cambridge, Mass.: MIT Press.

Chomsky N. (1981) *Lectures on Government and Binding*. Dordrecht: Foris.

29

Linguistica Computazionale A.A. 2006-07 – C. Chesì

## La variazione linguistica

Traduzione Automatica (MT)

### variazione morfologica

<b>isolanti</b>	<->	<b>polisintetiche</b>
Vietnamita, Cantonese		Esquimese
(1 morfema > 1 parola)		(molti morfemi > 1 parola)

<b>agglutinanti</b>	<->	<b>a fusione</b>
Turco		Russo
(1 morfema > 1 tratto)		(1 morfema > molti tratti)

### variazione sintattica

**SVO** (inglese, italiano, francese ...)

**SOV** (Indi, Giapponese ...)

**VSO** (Irlandese, Arabo classico ...)

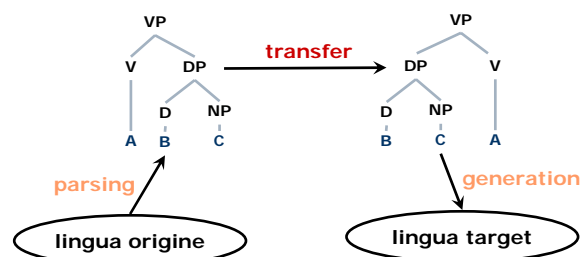
30

Linguistica Computazionale A.A. 2006-07 – C. Chesì

## Modello del Transfer

Traduzione Automatica (MT)

- **Conoscenza contrastiva**  
esplicitare le differenze tra le due lingue è il primo passo verso la traduzione.  
Da questo punto di vista occorre una ristrutturazione sintattica per conformarsi alle regole della lingua target

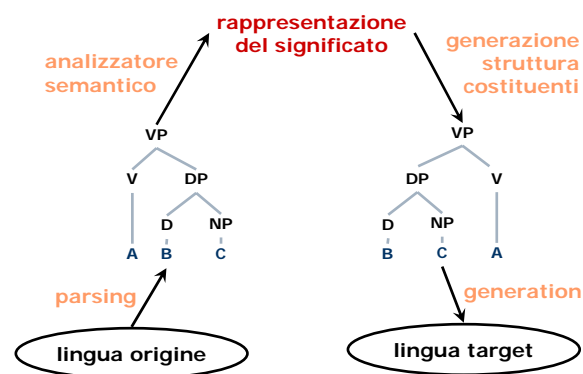


31

Linguistica Computazionale A.A. 2006-07 – C. Chesì

## Modello dell'Interlingua

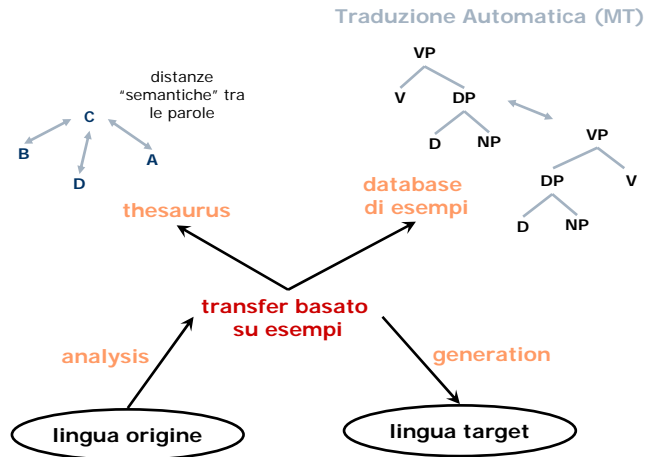
Traduzione Automatica (MT)



32

Linguistica Computazionale A.A. 2006-07 – C. Chesì

## Modello di traduzione basato su esempi



33

Linguistica Computazionale A.A. 2006-07 – C. Chesì

## METEO

(Hutchins & Somers 1992: Cap. 12)

### Traduzione Automatica (MT)

#### ■ Caratteristiche

- produce bollettini meteo in inglese e francese per tutto il Canada (7500 parole nel 1977, implementato su un super computer Control Data's Cyber 7600, oggi traduce circa 80.000 parole al giorno con un'accuratezza del 93% e gira su un network IBM-PC, impiegando circa 4 min. per testo).
- i bollettini standard sono **molto codificati** (stile telegrafico) ed hanno un **lessico limitato**
- la struttura delle frasi in inglese e francese è molto simile (approccio basato sul **transfer**, ma molto limitato)
- sistema di **seconda generazione**: **task-specifico, dominio specifico**, opzioni di **supporto** per interventi umani

34

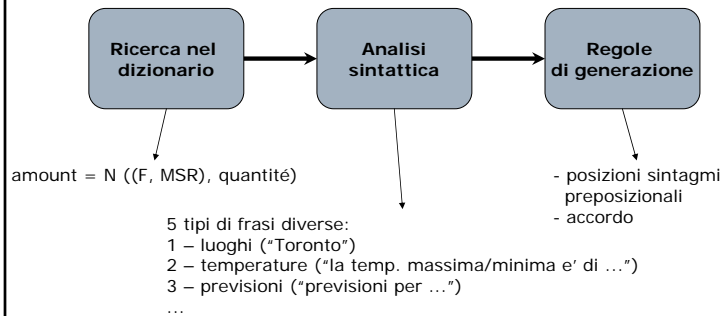
Linguistica Computazionale A.A. 2006-07 – C. Chesì

## METEO

### Traduzione Automatica (MT)

#### ■ Architettura

(notare la debole modularità della struttura computazionale anche se la grammatica può essere modificata significativamente)



35

Linguistica Computazionale A.A. 2006-07 – C. Chesì

## METEO

### Traduzione Automatica (MT)

#### ■ esempio di traduzione

FROM CYZE 311630  
 FORECASTS FOR ONTARIO ISSUED BY ENVIRONMENT CANADA AT 11:30 AM EST WEDNESDAY  
 MARCH 31ST 1976 FOR TODAY AND THURSDAY .  
 METRO TORONTO  
 WINDSOR.  
 CLOUDY WITH A CHANCE OF SHOWERS TODAY AND THURSDAY.  
 LOW TONIGHT 4. HIGH THURSDAY 16.  
 OUTLOOK FOR FRIDAY... SUNNY  
 END

Figure 12.1 Weather report as received

41- STORONTO,2) + FORECASTS + FOR + ONTARIO + ISSUED + BY + ENVIRONMENT + CANADA  
 + AT + 11 + H + 30 + AM + EST + WEDNESDAY + MARCH + 31ST + 1976 + FOR + TODAY + AND  
 + THURSDAY + . - 02./  
 42- STORONTO,3) + METRO + TORONTO + . , + WINDSOR + . - 02./  
 43- STORONTO,4) + CLOUDY + WITH + A + CHANCE + OF + SHOWERS + TODAY + AND +  
 THURSDAY + . - 02./  
 44- STORONTO,5) + LOW + TONIGHT + 4 + . - 02./  
 45- STORONTO,6) + HIGH + THURSDAY + 16 + . - 02./  
 46- STORONTO,7) + OUTLOOK + FOR + FRIDAY + = + SUNNY + . - 02./

Figure 12.2 Formatted weather report

5/  
 FROM CYZE 311630  
 2/  
 PREVISIONS POUR L'ONTARIO EMISES PAR ENVIRONNEMENT CANADA A 11 H 30 JOUR MERcredi  
 LE 31 MARS 1976 POUR AUJOURD'HUI ET JEUDI.  
 1/  
 TORONTO ET BANBEE  
 WINDSOR.  
 8/  
 MIAIGEEZ AVEC POSSIBILITE D'AVERSIS AUJOURD'HUI ET JEUDI  
 6/  
 MINIMUM CE SOIR 4.  
 -0/  
 HIGH THURSDAY 16.  
 2/  
 APERCU POUR VENERDI... ENSOLEILLE

Figure 12.3 Météo output

36

Linguistica Computazionale A.A. 2006-07 – C. Chesì

## Systran

(Hutchins & Somers 1992: Cap. 10)

### Traduzione Automatica (MT)

#### Alcune caratteristiche rilevanti

- Metodologia utilizzata: **Transfer**
- Inglese <-> Russo (1969), molte coppie di lingue dell'UE vengono aggiunte poi (1975)
- Per **analizzare** e **generare** espressioni si usa la stessa base di conoscenza (il cuore del sistema è composto da ampi dizionari bilingui)
- Non esiste un vero e proprio **"modulo di transfer"** (il transfer è realizzato da varie routine in **generazione**)
- Analisi morfo-sintattica parziale (**shallow parsing**) e "scorciatoie" per espressioni idiomatiche o parole composte.
- Semplice **categorizzazione semantica** (umano Vs. inumano), non gerarchica

37

Linguistica Computazionale A.A. 2006-07 – C. Chesì

## Systran

Lezione 4

Lezioni 6-8



Lezione 12

Architettura

38

Linguistica Computazionale A.A. 2006-07 – C. Chesì

## Systran

### Traduzione Automatica (MT)

#### Valutazione

- nonostante gli sforzi di uniformità (esiste un unico nucleo per le lingue romanze), questo sistema risente del maggior difetto dei sistemi di traduzione di prima generazione: mancanza di un adeguato framework linguistico.
- Lexical Routines**
  - He expects to come [ENG]
  - Il s'attende a ce qu'il viene [FR]
  - Eviter que l'argent soit dispensé [FR]
  - Prevent the money being spent [ENG]
  - \*Prevent that the money be spent
- nel complesso:
  - intelligibilità** 47% (1969) -> 73% (1978)
  - tasso di correzione** 40% (1969) -> 36% (1978)

39

Linguistica Computazionale A.A. 2006-07 – C. Chesì

## Rosetta

(Hutchins & Somers 1992: Cap. 16)

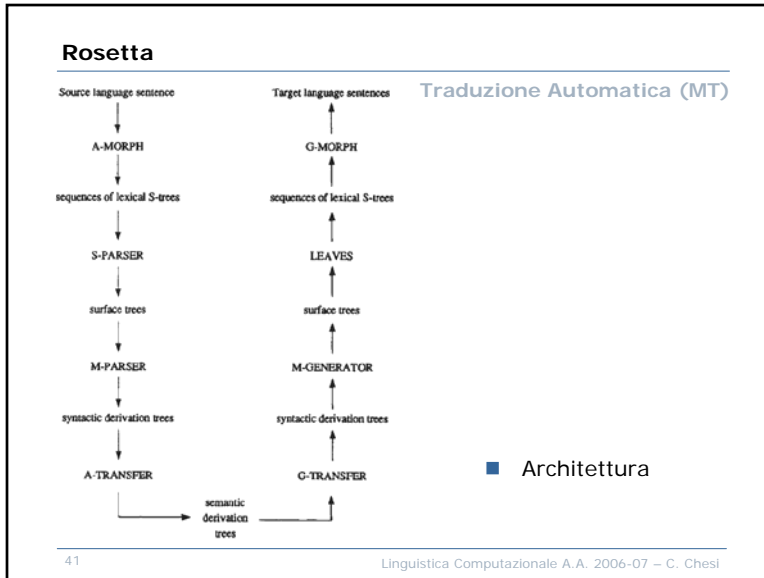
### Traduzione Automatica (MT)

#### Caratteristiche

- Metodologia utilizzata: **Interlingua**
- Approccio semantico: **Grammatica di Montague**  
(il significato di un'espressione è il risultato della composizione del significato delle sue componenti)
- Reversibilità**: la stessa grammatica è usata per analizzare e generare le frasi
- Isomorfismo**: la stessa derivazione semantica deve essere ottenuta per avere una traduzione

40

Linguistica Computazionale A.A. 2006-07 – C. Chesì



### Prossima lezione

(Mercoledì 7 Marzo, ore 16-18, Aula 456, Palazzo S.Niccolò)

- Grammatiche formali
  - Nozioni di base (grammatiche a struttura sintagmatica, PSG)
  - Gerarchia di Chomsky
  - Descrizioni Strutturali e derivazioni
  
- Formalismi applicabili alla MT
  - Regole di Transfer
  - Ontologie
  - Grammatiche ad unificazione
  - Principi e Parametri

42 Linguistica Computazionale A.A. 2006-07 – C. Chesì