

# Metodi di crowdsourcing nello studio del linguaggio

Alessandra Zarcone  
Institut für Maschinelle Sprachverarbeitung, Stuttgart  
( `zarconaa@ims.uni-stuttgart.de` )

Metodi sperimentali per la raccolta e l'analisi dei dati linguistici  
Pisa, 6-7 dicembre 2011

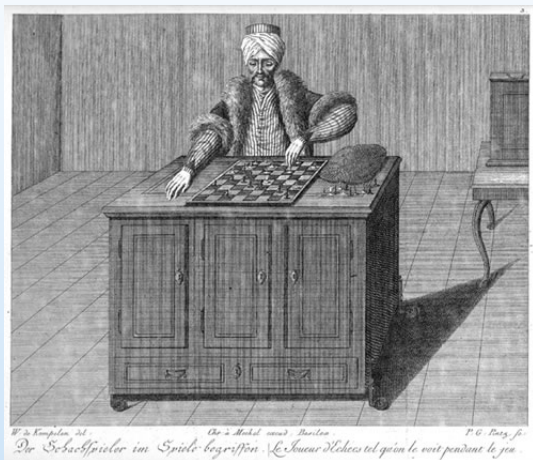


# Sommario

- 1 Cos'è il crowdsourcing**
  - Il turco meccanico
  - Intelligenza artificiale artificiale
  - Come funziona il crowdsourcing
  - Amazon Mechanical Turk e Crowdfunder
- 2 Crowdsourcing e studio del linguaggio**
  - Dall'azienda alla ricerca
  - Chi sono i "lavoratori"?
- 3 Esperimenti di crowdsourcing**
  - Crowdsourcing e classificazione azionaria
  - Crowdsourcing e metonimia
  - Crowdsourcing e linguaggi artificiali
  - Crowdsourcing e raccolta di corpora
- 4 Conclusioni**



## Il turco meccanico intelligenza artificiale?

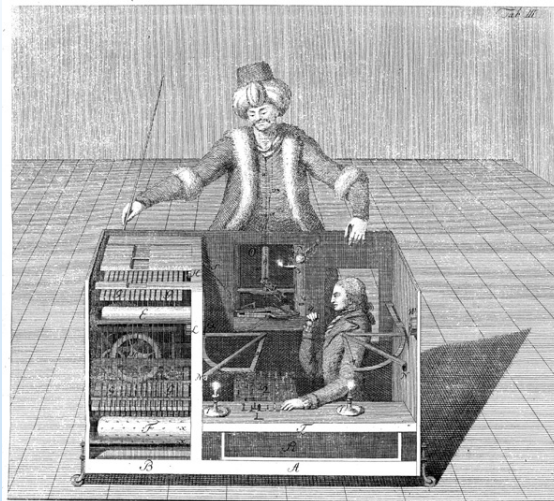


"Il Turco", Wolfgang von Kempelen, 1769  
(incisione di Karl Gottlieb von Windisch)



# Dentro il turco

intelligenza artificiale artificiale



# Intelligenza artificiale artificiale

Esempi: parsing, categorizzazione di immagini, traduzione, elicitazione..

- Intelligenza artificiale

**Come:** sistemi di machine learning (test, training, valutazione)

- Intelligenza artificiale artificiale

**Come:** invece di chiedere "alla macchina", chiedo "all'uomo dentro la macchina"



# Cos'è il crowdsourcing

**(Uno o pochi) annotatori umani "esperti"**

tempo, denaro, esperienza



**CROWDSOURCING** (crowd + outsourcing)

Crowd = tanti lavoratori, "gente comune"

Outsourcing = "esportare" alcuni compiti affidandoli a risorse esterne



# Come funziona il crowdsourcing

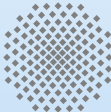
## Un "modello di business":

- "datore di lavoro" → "open call"
- un portale che raccoglie le calls
- lavoratori con accesso internet

## Microtasking:

- i compiti sono relativamente facili (per un essere umano) e rapidi  
→ pochi centesimi a "task"
- i dati provenienti da una "crowd" di lavoratori
- il lavoratore "lavora" in momenti di pausa o nel tempo libero, svolgendo diversi "micro-compiti"

**Micro task** = qualunque cosa si possa "incorporare"  
in un browser o in una app



# Amazon Mechanical Turk (AMT)

**Il "turco meccanico" di Amazon** - <https://www.mturk.com>  
uno dei portali più grandi e più usati per il crowdsourcing

Lato lavoratore:

- più di 183 mila "compiti"
- non necessariamente tutti "aperti"
- può scegliere in base alla retribuzione e/o in base a cosa preferisce

Lato datore di lavoro:

- creare nuovi compiti
- controllo dei compiti assegnati
- "accettare" o "rifiutare" i compiti svolti dai lavoratori
- assegnare dei bonus





# Amazon Mechanical Turk

Cosa vede il lavoratore:

"HITs" (Human Intelligence Tasks) - "a question that needs an answer"

## All HITs

1-10 of 1630 Results

Sort by:  60

Show all details | Hide all details

1 2 3 4 5 > Next >> Last


INV_B_TYPE	Requester	HIT Expiration Date:	Time Allotted:	Reward:	HITs Available:	Request Qualification (Why?)	View a HIT in this group
	<a href="#">rehzit0d</a>	Dec 16, 2011 (1 week 5 days)	12 hours	\$0.00	22797		
Are these two pictures of the same kind of place?	<a href="#">Michelle Greene</a>	Dec 1, 2012 (51 weeks 6 days)	10 minutes	\$0.02	16445	Request Qualification Take Qualification test (Why?)	View a HIT in this group
What Services Does this Wellness Salon/Spa Provide?	<a href="#">Dolores Labs</a>	Dec 8, 2011 (4 days 11 hours)	60 minutes	\$0.05	9339		View a HIT in this group
Write a Brief Description About a Local Business - \$0.40 for 40-60 words	<a href="#">WorldsOyster</a>	Dec 7, 2011 (3 days 5 hours)	60 minutes	\$0.40	8572	Not Qualified to work on this HIT (Why?)	View a HIT in this group
Categorize these products from Amazon.com Electronics (New Version! Simple yes/no questions, BinaryQualified )	<a href="#">Amazon Requester Inc - browse classification</a>	Dec 12, 2011 (1 week 1 day)	5 minutes	\$0.06	8404	Request Qualification Request Qualification (Why?)	View a HIT in this group

# Amazon Mechanical Turk

Cosa vede il datore di lavoro:

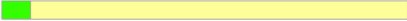
Thematic-based event generation - Patient-D' @ 07 Dec 00:04 Results Delete

Created:	December 07, 2010	Assignments Completed:	108 / 108
Time Elapsed:	14 days	Estimated Completion Time:	COMPLETE
Average Time per Assignment:	9 minutes 13 seconds	Effective Hourly Rate:	\$2.604

Batch Progress:   
100% submitted 100% published

Thematic-based event generation - Patient-D' @ 03 Dec 07:12 Results Delete

Created:	December 03, 2010	Assignments Completed:	7 / 101
Time Elapsed:	14 days	Estimated Completion Time:	Not yet available
Average Time per Assignment:	7 minutes 39 seconds	Effective Hourly Rate:	\$1.569

Batch Progress:   
7% submitted 100% published

# Crowdfower

<http://crowdfower.com/>

"the worlds largest enterprise crowdsourcing platform"

- 1.5 milioni di lavoratori
- "gold answers" per controllo qualità
- i task sono poi distribuiti ai lavoratori attraverso varie piattaforme (ad esempio, AMT)
- "interfaccia" per AMT ma non solo



# Crowdsourcing e studio del linguaggio

Dall'azienda alla ricerca

- 2007: AMT apre a task esterni ad AMT stesso e Powerset inizia a usarlo per raccogliere dati per sviluppare un motore di ricerca in linguaggio naturale
- per più di un anno la maggior parte delle HIT che compaiono su AMT sono create da Powerset
- dall'NLP alla traduzione automatica fino ad esperimenti veri e propri

(Munro et al. 2010)



# Crowdsourcing e studio del linguaggio

## Dall'azienda alla ricerca

Dal "microtasking" in azienda alla ricerca:  
raccolgere dati sperimentali e condurre esperimenti

Vantaggi:

- grande disponibilità di partecipanti
- campione molto variegato
- rapidità e convenienza

Svantaggi:

- campione molto variegato  
(e poco controllo su chi sono realmente i partecipanti)
- serve "creatività" nel formulare il compito e nel controllo qualità  
(non ci sono "risposte giuste")
- scarso controllo sulle liste  
(ma è possibile, cfr. Watts e Jaeger 2011)



# Crowdsourcing e studio del linguaggio

## Chi sono i "lavoratori"?

### Sondaggio del 2010

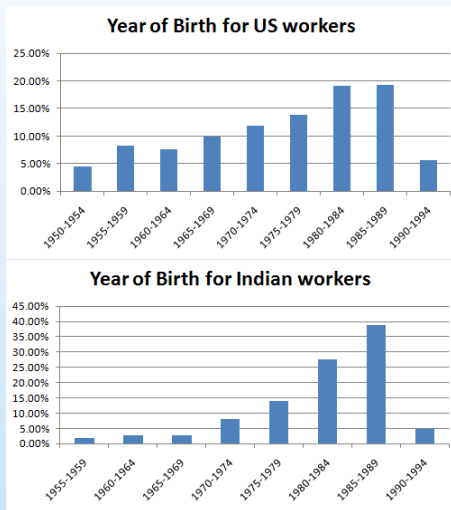
- United States: 47%
- India: 34%
- Miscellaneous: 19%

fonte: <http://www.behind-the-enemy-lines.com/2010/03/new-demographics-of-mechanical-turk.html>



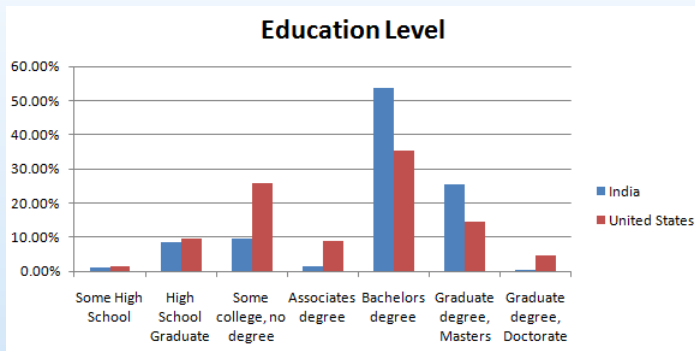
# Crowdsourcing e studio del linguaggio

## Chi sono i "lavoratori"?



# Crowdsourcing e studio del linguaggio

## Chi sono i "lavoratori"?



fonte: <http://www.behind-the-enemy-lines.com/2010/03/new-demographics-of-mechanical-turk.html>





# Crowdsourcing e studio del linguaggio

"Is it good?"

Snow et al. 2008 "Cheap and Fast - But is it Good? Evaluating Non-Expert Annotations for Natural Language Task"

- Diversi domini di ricerca (affect recognition, word similarity, recognizing textual entailment, event temporal ordering, word sense disambiguation)
- Alti ITA (inter-tagger agreement, "indici di concordanza" tra annotatori) tra non-esperti su AMT e "professionisti"

Ma anche:

Fort et al. 2011 "Amazon Mechanical Turk: Gold Mine or Coal Mine?"

- i risultati di Snow et al. non persistono su studi di larga scala
- stima: 80% delle HITs svolte dal 20% dei "turkers" più attivi (> 15 ore a settimana a persona)
- media di meno di 2 dollari l'ora



# Crowdsourcing e studio del linguaggio

## Strategie

- non pagare troppo né troppo poco
- creare un rapporto "fiduciario" con i lavoratori
- creare compiti "divertenti", o quantomeno "interessanti"
- creare dei sistemi di "controllo" dei dati



# Crowdsourcing e studio del linguaggio

Workshops nel 2011

- July 2011, Boston MA:  
Tutorial at CogSci 2011 on "How to use Mechanical Turk for Cognitive Science Research" (W. Mason e S. Suri)
- July 2011 Boulder CO:  
Workshop on Crowdsourcing Technologies for Language and Cognition Studies at the LSA Institute (R. Munro e H. Tily)  
*"the first time that researchers had come together for a workshop dedicated wholly to crowdsourcing technologies as a tool for empirical studies"*



# Crowdsourcing e studio del linguaggio

"Armchair linguistics is changing"



# Crowdsourcing e studio del linguaggio

## "Armchair linguistics is changing"

### Workshop on Crowdsourcing Technologies for Language and Cognition Studies

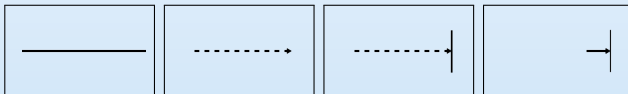
- Artificial language learning (Jaeger et al.)
- Interazione e dialogo (Clausen e Potts)
- Logical metonymy (Zarcone e Pado)



# Esperimenti di crowdsourcing

## Crowdsourcing e classificazione azionale

	[dyn]	[dur]	[res]	
<b>STA</b>	—	+	—	<i>essere olandese</i>
<b>ACT</b>	+	+	—	<i>cantare</i>
<b>ACC</b>	+	+	+	<i>disegnare un ritratto</i>
<b>ACH</b>	+	—	+	<i>scompare</i>

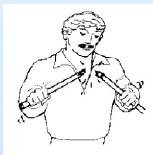


# Esperimenti di crowdsourcing

## Crowdsourcing e classificazione azionale

Quattro esperimenti (Zarcone e Lenci 2010)

- classificazione di immagini e di VP per l'italiano (esperimento via web più "tradizionale")
- classificazione di immagini e di VP per l'inglese (crowdsourcing)



# Esperimenti di crowdsourcing

## Crowdsourcing e classificazione azionale

		$\alpha$	$\alpha_w$	accuracy	
Experiment 1	IT, verbs	0.35	0.43	0.63	sito web
Experiment 2	EN, verbs	0.46	0.53	0.68	crowdsourcing
Experiment 3	IT, img	0.22	0.31	0.42	sito web
Experiment 4	EN, img	0.28	0.39	0.54	crowdsourcing





# Esperimenti di crowdsourcing

## Crowdsourcing e metonimia

"Logical metonymy" e "covert events" (CE)

### oggetti che denotano eventi (EV) vs. entità (EN):

EV: begin the **afternoon**

→ ✓ begin(afternoon)

→ ✗ begin(**CE**(afternoon))

EN: begin the **newspaper**

→ ✗ begin(newspaper)

→ ✓ begin(**CE**(newspaper))

→ begin **reading** the newspaper

Generative Lexicon (Pustejovsky 1995)

- i CE vengono interpretati solo con oggetti del tipo EN
- i CE sono recuperati dai qualia dell'oggetto (quale agentivo, quale telico)



# Esperimenti di crowdsourcing

## Crowdsourcing e metonimia

**partecipanti:** 15 partecipanti dagli USA

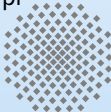
**materiali:** 60 frasi:

**EN:** Jim began/spotted the magazine

**EV:** Al began/spotted the ceremony

**EN/EV** Nick began/spotted the conquest

- procedura:**
- *does the sentence involve an additional activity that is not mentioned in the sentence?*  
(*additional activity* o *no additional activity*)
  - quando rispondevano *additional activity*,  
ai partecipanti veniva chiesto di fornire degli esempi



# Esperimenti di crowdsourcing

## Crowdsourcing e metonimia

- agreement basso ( $\alpha = .35$ ) ma buon agreement con gli annotatori ( $\alpha = .6$ )
  - Obj\_type: binomial  $p < .001 \rightarrow$  significant effect
  - Verb\_type:  $z = -8.322, p < .001 \rightarrow$  significant effect
  - Interaction: binomial  $p < .001 \rightarrow$  significant effect

condition	CE	no-CE
begin,EN	63%	37%
spot,EN	11%	89%
begin,EN/EV	39%	61%
spot,EN/EV	6%	94%
begin,EV	18%	82%
spot,EV	6%	94%



# Esperimenti di crowdsourcing

## Crowdsourcing e metonimia

- 1 or 2 risposte per partecipante (media 1.4, min. 1, max. 6)
- chi ha dato una risposta sola: media 3.2 CE per ogni VP
- media generale di 5 CE per VP (min 1 max 15)

**EN:** start the portrait → 9 CEs: paint (x20), draw (x4), critique (x3), hang (x2), model (x2), sketch (x2), admire, pose for, review

**EN/EV:** finish the harvest → 15 CEs: gather (x5), collect (x4), plan (x3), reap (x3), sell (x3), load (x2), store (x2), cook, eat, enjoy, jar, package, pick, pull, ship

**EV:** enjoy the conference → 4 CEs: attend (x3), hold (x2), participate in, watch

	tot	Qualia-structure CEs		other CEs
		agentive	telic	
elicited CEs (tokens)	542	132 (24.3%)	162 (29.9%)	248 (45.8%)
elicited CEs (types)	205	31 (15.1%)	25 (12.2%)	149 (72.7%)



# Esperimenti di crowdsourcing

## Crowdsourcing e linguaggi artificiali

Jaeger et al., "A web-based (iterated) language learning paradigm with human participants"

sa ent aw shnoo lodi

Progress

Learning Names

Which Person?

Actions 1

1 out of 12

Which Video?

More Actions

Which Video? 2

Speaking

Replay

Continue

"gioco" di apprendimento linguaggi artificiali → usato per replicare studi di laboratorio in Hudson Kam, C., and Newport, E. (2005)





# Conclusioni

"A tool is only as good as the people who use it."

- necessario conoscerne caratteristiche e limiti
- potenzialità diverse da quelle dei metodi più tradizionali
- richiede di "pensare" l'esperimento in modo diverso
- necessità di controlli di qualità
- questioni etiche



Grazie





- Fort, K., Adda, G., e Cohen, K. B., 2011. *Amazon Mechanical Turk: Gold Mine or Coal Mine?*. Computational Linguistics 37, 413420.
- Munro, R., Bethard, S., Kuperman, V., Tzuyin Lai, V., Melnick, R., Potts, C., Schnoebelen T., e Tily, H., 2010. *Crowdsourcing and language studies: the new generation of linguistic data. Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, Los Angeles, CA.
- Munro, R. e Tily, H., 2011. *The Start of the Art: Introduction to the Workshop on Crowdsourcing Technologies for Language and Cognition Studies*, Boulder.
- Snow, R., O'Connor, B., Jurafsky, D., Ng, A. Y., 2008. *Cheap and Fast - But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks*. Proceeding of EMNLP '08.
- Watts, A. e Jaeger, T. F., 2011. *Balancing experimental lists without sacrificing voluntary participation*. Workshop on Crowdsourcing Technologies for Language and Cognition Studies. Boulder, Colorado.
- Zarcone, A. e Padó, S.. 2011. *A crowdsourcing study of logical metonymy*. Workshop on Crowdsourcing Technologies for Language and Cognition Studies. Boulder, Colorado.

