

La classificazione azionale del verbo italiano

primi esperimenti computazionali

Alessandra Zarcone

Università di Pisa - Corso di Laurea in Informatica Umanistica

12 Ottobre 2006



Indice

- 1 Introduzione
- 2 Il modello linguistico
- 3 Il modello computazionale
- 4 Un modello stocastico della classificazione azionale
- 5 Primi esperimenti computazionali
- 6 Conclusioni e sviluppi futuri



Un'analisi computazionale del verbo italiano

task: l'azionalità: un compito semantico **complesso**

ipotesi: indicatori contestuali \Rightarrow categoria azionale

precedente: Siegel e McKeown 2000

- scopi:**
- 1 costruzione di un modello stocastico della classificazione azionale
 - 2 valutazione del ruolo degli indicatori contestuali nella scelta della categoria azionale
 - 3 analisi computazionale del problema

modello teorico linguistico: categorie vendleriane

modello computazionale: Maximum Entropy



L'Azionalità

L'Azionalità

L'Azionalità è un tratto della semantica del verbo che individua il tipo di evento descritto da ogni verbo in un contesto

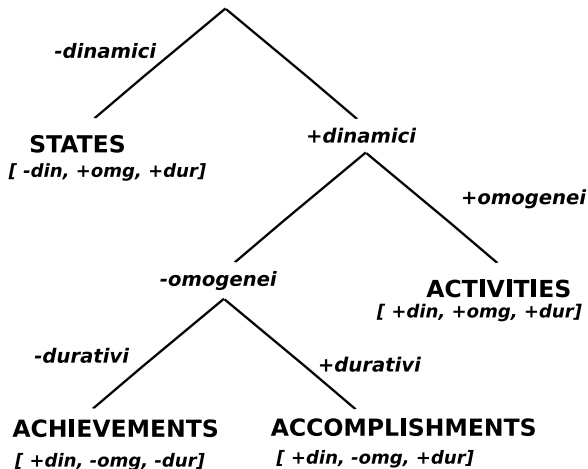
(modelli di Azionalità: Vendler 1967, Bertinetto 1986, Rothstein 2004)

Le classi azionali sono descrivibili tramite tratti (omogeneità, duratività, dinamicità)

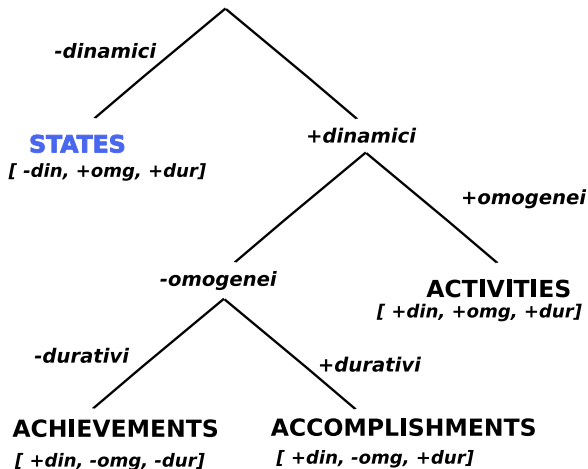
Categorie vendleriane:	[omg]	[dur]	[din]
States (STA)	+	+	-
Activities (ACT)	+	+	+
Accomplishments (ACC)	-	+	+
Achievements (ACH)	-	-	+



Categorie vendleriane



States



States

States (STA)

Indicano qualità, abitudini, abilità e altre caratteristiche, permanenti o transitorie, possedute dal soggetto:

- Si trovava in Calabria in compagnia della moglie.

Gli states sono tipicamente **durativi**, **non dinamici** e **omogenei**

omogeneità

un evento omogeneo è descrivibile come una serie omogenea di elementi (istanti o "piccole azioni") uguali



States

States (STA)

Indicano qualità, abitudini, abilità e altre caratteristiche, permanenti o transitorie, possedute dal soggetto:

- Si trovava in Calabria in compagnia della moglie.

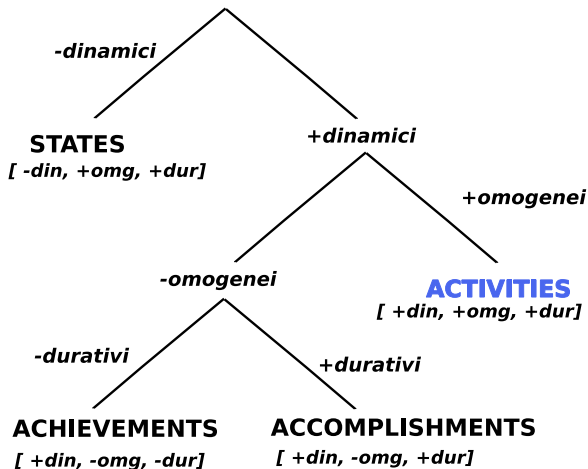
Gli states sono tipicamente **durativi**, **non dinamici** e **omogenei**

omogeneità

un evento omogeneo è descrivibile come una serie omogenea di elementi (istanti o “piccole azioni”) uguali



Activities



Activities

Activities (ACT)

Sono verbi di attività:

- Di notte **scrive** alla sua fidanzata

Si tratta di verbi **durativi**, **dinamici** e **omogenei**

omogeneità

un'activity è omogenea perchè descrivibile come una "catena di eventi" (camminare, ad esempio, è una serie di eventi minimi descrivibili come "fare un passo")



Activities

Activities (ACT)

Sono verbi di attività:

- Di notte **scrive** alla sua fidanzata

Si tratta di verbi **durativi**, **dinamici** e **omogenei**

omogeneità

un'activity è omogenea perchè descrivibile come una "catena di eventi" (camminare, ad esempio, è una serie di eventi minimi descrivibili come "fare un passo")



omogenei vs. telici

- States e Activities sono classi di verbi omogenei e atelici
- Accomplishments e Achievements sono classi di verbi non omogenei e telici

omogeneità

un evento omogeneo è descrivibile come una serie omogenea di elementi (istanti o “piccole azioni”) uguali

telicità

un evento telico esprime intrinsecamente il **raggiungimento** di un certo risultato

l'azione descritta da un evento telico tende quindi verso un “compimento”: non è una catena omogenea di eventi uguali (“camminare”) ma una catena progressiva di eventi incrementali che tendono verso un risultato (“dipingere un quadro”)



omogenei vs. telici

- States e Activities sono classi di verbi omogenei e atelici
- Accomplishments e Achievements sono classi di verbi non omogenei e telici

omogeneità

un evento omogeneo è descrivibile come una serie omogenea di elementi (istanti o “piccole azioni”) uguali

telicità

un evento telico esprime intrinsecamente il raggiungimento di un certo risultato

l'azione descritta da un evento telico tende quindi verso un “compimento”: non è una catena omogenea di eventi uguali (“camminare”) ma una catena progressiva di eventi incrementali che tendono verso un risultato (“dipingere un quadro”)

omogenei vs. telici

- States e Activities sono classi di verbi omogenei e atelici
- Accomplishments e Achievements sono classi di verbi non omogenei e telici

omogeneità

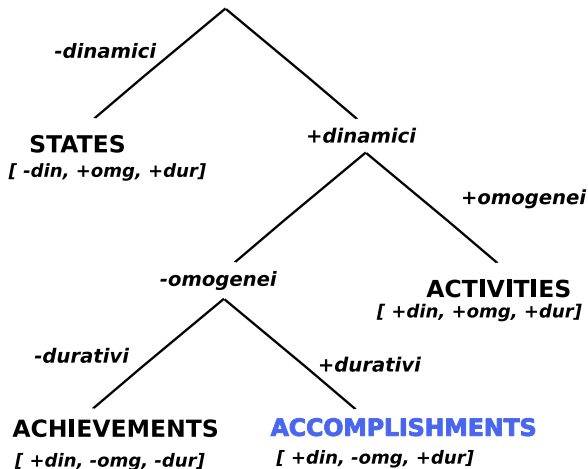
un evento omogeneo è descrivibile come una serie omogenea di elementi (istanti o “piccole azioni”) uguali

telicità

un evento telico esprime intrinsecamente il **raggiungimento** di un certo risultato

l'azione descritta da un evento telico tende quindi verso un “compimento”: non è una catena omogenea di eventi uguali (“camminare”) ma una catena progressiva di eventi incrementali che tendono verso un risultato (“dipingere un quadro”)

Accomplishments



Accomplishments

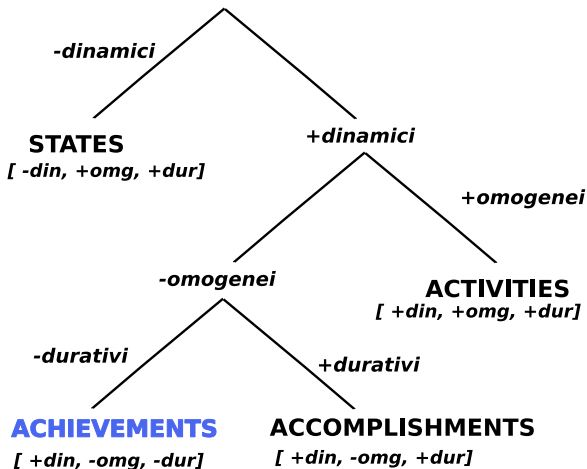
Accomplishments (ACC)

Esprimono il raggiungimento di un risultato, sono verbi **dinamici**, **durativi** e **telici** (quindi **non omogenei**). Il tratto di telicità li distingue dalle *ACT* (spesso gli *ACC* sono *ACT* “telicizzate” da un complemento):

- mangiare (*ACT*) \iff mangiare una mela (*ACC*)
- trascinare (*ACT*) \iff portare qualcosa fino a qui (*ACC*)
- leggere (*ACT*) \iff leggere tutto “Guerra e Pace” (*ACC*)
- Di notte **scrivo** alla mia fidanzata (*ACT*)
- Giorgio Celli **ha scritto** un romanzo giallo (*ACC*)



Achievements



Achievements

Achievements (ACH)

Sono verbi **dinamici** e **telici** (quindi **non omogenei**), come gli ACC, ma **non durativi**. Esprimono generalmente un cambio di stato, una trasformazione:

- Il treno è **partito** alle quattro
- Ho **scoperto** il loro nascondiglio



Ibridismo azionale, commutazione azionale

Ibridismo azionale e opposizioni infralessicali

Naturale polisemia statisticamente significativa tra due interpretazioni azionali. Un verbo come *conoscere*, ad esempio, può equivalere a 'fare la conoscenza' o a 'essere a conoscenza', dove l'uno indica 'l'attuarsi dell'azione indicata dall'altro' (Lucchesi 1971).

- Ieri ho conosciuto Chiara (ACH)
- Conosco Chiara da diversi anni (STA)

Commutazione azionale

Può al contrario coinvolgere un verbo fortemente prototipico. Particolari elementi del contesto possono "forzare" la classe azionale prototipica di un verbo e modificarla

- A un certo punto Mangiafuoco starnutì (ACH)
- Mangiafuoco starnutì per mezz'ora (ACT)

Ibridismo azionale, commutazione azionale

Ibridismo azionale e opposizioni infralessicali

Naturale polisemia statisticamente significativa tra due interpretazioni azionali. Un verbo come *conoscere*, ad esempio, può equivalere a 'fare la conoscenza' o a 'essere a conoscenza', dove l'uno indica 'l'attuarsi dell'azione indicata dall'altro' (Lucchesi 1971).

- Ieri ho conosciuto Chiara (ACH)
- Conosco Chiara da diversi anni (STA)

Commutazione azionale

Può al contrario coinvolgere un verbo fortemente prototipico. Particolari elementi del contesto possono "forzare" la classe azionale prototipica di un verbo e modificarla

- A un certo punto Mangiafuoco starnutì (ACH)
- Mangiafuoco starnutì per mezz'ora (ACT)

Apprendimento automatico

tipo: supervisionato

task: assegnare la categoria azionale corretta
(\Rightarrow la più probabile dato un contesto)

modello: Maximum Entropy (Berger et al. 1996)

algoritmo: AMIS (Miyao and Tsujii 2005)

test: i test sono stati condotti secondo il metodo della **k-fold crossvalidation** (nel nostro caso, 30-fold crossvalidation)



Maximum Entropy

- frase = verbo + contesto linguistico (insieme di marche contestuali)
- il modello costruisce la **distribuzione di probabilità** $p(r | c)$
(dove r = particolare classe azionale e c = caratterizzazione del contesto)
- risposta corretta = $\operatorname{argmax}(p(r | c))$
- secondo la Maximum Entropy $p(r | c)$ viene stimata come:

$$p(r | c) = \frac{1}{Z_c} \prod_{i=1}^k r_i^{m_i(r,c)}$$

dove Z_c sarà un fattore di normalizzazione, e $m_i(r, c)$ i parametri di ogni marca m per (r, c) (i pesi assegnati alle marche: ovvero quanto la marca m del contesto c è rilevante per la risposta r)

- 1 **fase di training**: il modello calcola le distribuzioni di probabilità e restituisce i pesi delle marche
- 2 **fase di test**: il modello utilizza i pesi calcolati per trovare la categoria azionale di contesti nuovi



Precision, Recall, F-measure

Sugli stativi:

precision: $P_{STA} = \frac{(\text{STA corretti})}{(\text{tot. risposte STA})}$

(percentuale di STA corretti trovati dal sistema su tutti gli STA individuati dal sistema)

recall: $R_{STA} = \frac{(\text{STA corretti})}{(\text{tot. STA nel corpus})}$

(percentuale di STA corretti trovati dal sistema rispetto a tutti gli STA presenti nel test corpus)

f-measure: $F_{STA} = \frac{2(P_{STA} \cdot R_{STA})}{P_{STA} + R_{STA}}$

(media armonica tra precision e recall)

Precision, recall e f-measure sono calcolati su ognuna delle quattro categorie. Un sistema perfetto avrà precision e recall uguali a 1, e quindi anche f-measure uguale a 1.



L'annotazione

- 3429 occorrenze (di 33 verbi particolarmente frequenti) con relativi contesti (italiano scritto contemporaneo), estratte dal corpus TreSSI (Montemagni et al. 2003)
- le 3429 frasi sono state annotate secondo le quattro categorie azionali descritte come vettori di tratti
- *cambiare, crescere, ridurre, salire, scendere* (per un totale di 300 occorrenze): sono stati considerati appartenenti a una quinta categoria (**incrementativi**) e “messi da parte” per una fase successiva del lavoro
- totale di 3129 frasi annotate (**gold standard**):

verbi	tot.occorrenze	ACT	STA	ACH	ACC
28	3129	430	583	1294	822



Ibridi e prototipici

lemma	tot. occorrenze del lemma	tot. occorrenze per classe azionale					percentuale	classe prototipica
		STA	ACT	ACH	ACC	INC		
chiamare	93	36	13	0	44	0	47%	ACC
cambiare	72	4	5	28	0	35	49%	INC
scendere	60			23	4	33	55%	INC
mettere	166			66	94	0	57%	ACC
trovare	173			99	0	0	57%	ACH
capire	81			47	0	0	58%	ACH
indicare	97	37	1	58	1	0	60%	ACH
conoscere	68			26	0	0	62%	STA
presentare	143			15	97	0	68%	ACC
prendere	131			90	24	0	69%	ACH
passare	129			0	92	0	71%	ACC
portare	174	29	18	0	127	0	73%	ACC
ridurre	51			0	12	39	76%	INC
scrivere	106			0	82	0	77%	ACC
costituire	62			14	0	0	77%	STA
salire	67			11	2	52	78%	INC
comprendere	64	53	0	11	0	0	83%	STA
trattare	113			0	2	0	83%	STA
lasciare	121			04	1	0	86%	ACH
vendere	73			24	0	0	88%	ACH
controllare	65			0	1	0	89%	STA
entrare	90	8	0	82	0	0	91%	ACH
arrivare	185	6	8	171	0	0	92%	ACH
chiedere	152	0	11	0	141	0	93%	ACC
spiegare	115	7	0	0	108	0	94%	ACC
partire	78	4	0	74	0	0	95%	ACH
parlare	181	9	172	0	0	0	95%	ACT
morire	87	0	0	83	4	0	95%	ACH
chiudere	100	4	0	96	0	0	96%	ACH
tornare	117	1	1	113	2	0	97%	ACH
lavorare	84	0	82	0	2	0	98%	ACT
vincere	81	0	0	81	0	0	100%	ACH
crescere	50	0	0	0	0	50	100%	INC



Baseline

- Siegel e McKeown (2000): “fundamental aspectual class” (non considerano l'ibridismo azionale, solo la commutazione)
- proviamo a legittimare il concetto di “fundamental aspectual class”
- data un'occorrenza, il sistema assegna la classe azionale più probabile per il lemma corrispondente:

Gruppi	Baseline
Primo gruppo	56,1%
Secondo gruppo	60%
Terzo gruppo	64,6%
Quarto gruppo	69,6%
Tutto il corpus	79,8%



Selezione di marche contestuali rilevanti

- Il modello ha bisogno di una selezione di marche contestuali rilevanti per la caratterizzazione azionale dei contesti verbali
- Marche estratte semi-automaticamente
- Tipologie di marche utilizzate:
 - ① **tratti morfologici tempo-aspettuali e diatesi** della testa verbale;
 - ② **modificatori della testa verbale** (di varie tipologie);
 - ③ **argomenti e complementi**, retti dalla testa verbale, e loro tratti morfologici, sintattici e semantici (lessicalizzati e non lessicalizzati).



Revisione ciclica del modello



Configurazione migliore

	ACT	STA	ACC	ACH	
ACT	319	21	50	40	⇒ tot. act: 430
STA	20	454	30	79	⇒ tot. sta: 583
ACC	33	15	733	41	⇒ tot. acc: 822
ACH	10	55	63	1166	⇒ tot. ach: 1294
precision:	0,84	0,83	0,84	0,88	occorrenze: 3129
recall:	0,74	0,78	0,89	0,9	errori: 14,6%
f-measure:	0,79	0,8	0,86	0,89	corrette: 85,4%



Confronto con la baseline

Gruppi	Baseline	Conf. migliore
Primo gruppo	56,1%	69,34%
Secondo gruppo	60%	72,37%
Terzo gruppo	64,6%	75,47%
Quarto gruppo	69,6%	78,43%
Tutto il corpus	79,8%	85,39%



Errori comuni del sistema

- 1 usi idiomatici (soprattutto verbi *mettere, prendere, portare*)
 - Umberto Eco ha potuto divertirsi a prendere un po' tutti per il bavero, sia gli autori gialli sia i lettori di quelli, ma non solo i gialloman.
- 2 infiniti, participi, gerundi
 - Basta chiedere a Poulidor, infaustamente chiamato Poupou, che segue il Tour con un'aria da clown.
 - Tutto ciò lascia capire come il gruppo Lazard abbia già mangiato la foglia.



Inter-Tagger Agreement

Problema dell'Inter-Tagger Agreement (ITA)

- problema tipico di ogni approccio automatico alla semantica
- disaccordo in letteratura tra diverse teorie sull'azionalità

Un esperimento di stima dell'ITA per valutare la complessità del task dal punto di vista di un parlante nativo

- 1 selezione di 100 frasi dal secondo gruppo di ibridi
- 2 le frasi vengono sottoposte ad altri tre annotatori umani, che devono annotarle secondo le quattro categorie vendleriane

	annot.A	annot.B	annot.C	sistema
accordo osservato	73%	44%	67%	76%

Media di k-statistic: 0,51

Media di accordo osservato in percentuale: 65%



Decisione su tratti

- non più decisione su classi,
ma su singoli tratti
- baseline diversa
(..e più alta)
- maggiori difficoltà
con il tratto di duratività

Gruppi	Baseline	Conf. migliore
DUR		
Primo gruppo	60,9%	72,8%
Secondo gruppo	62,2%	74,3%
Terzo gruppo	70%	79,1%
Tutto il corpus	87,7%	90,6%
DIN		
Primo gruppo	63,9%	79,9%
Secondo gruppo	68,3%	84,9%
Terzo gruppo	75,5%	85,4%
Tutto il corpus	88,3%	92%
OMG		
Primo gruppo	56,9%	79,5%
Secondo gruppo	66,8%	81,7%
Terzo gruppo	71,9%	83,2%
Tutto il corpus	84,4%	89,9%

Conclusioni

- Dalla teoria al modello stocastico
- Dal modello stocastico alla teoria
- Una questione interessante: gli incrementativi
- Modelli non supervisionati
- Psicolinguistica

