10.12871/CLICIT2014135

Rappresentazione dei concetti azionali attraverso prototipi e accordo nella categorizzazione dei verbi generali. Una validazione statistica.

Gloria Gagliardi
Università degli Studi di Firenze
gloria.gagliardi@unifi.it

Abstract

Italiano. L'articolo presenta i risultati di uno studio volto a valutare la consistenza della categorizzazione dello spazio azionale operata da annotatori madrelingua per un set di verbi semanticamente coesi del database IMAGACT (area semantica di 'girare'). La validazione statistica, articolata in tre test, è basata sul calcolo dell'*inter-tagger agreement* in *task* di disambiguazione di concetti rappresentati mediante prototipi per immagini.

English. *This paper presents the results of a research aimed at evaluating the consistency of the categorization of actions. The study focuses on a set of semantically related verbs of the IMAGACT database ("girare" semantic area), annotated by mother tongue informants. Statistic validation, consisting of three tests, is based on inter-tagger agreement. The task entails the disambiguation of concepts depicted by prototypic scenes.*

1 Introduzione

IMAGACT è un'ontologia interlinguistica che rende esplicito lo spettro di variazione pragmatica associata ai predicati azionali a media ed alta frequenza in italiano ed inglese (Moneglia *et al.*, 2014). Le classi di azioni che individuano le entità di riferimento dei concetti linguistici, rappresentate in tale risorsa lessicale nella forma di scene prototipiche (Rosch, 1978), sono state indotte da *corpora* di parlato da linguisti madrelingua, mediante una procedura *bottom-up*: i materiali linguistici sono stati sottoposti ad una articolata procedura di annotazione descritta estesamente in lavori precedenti (Moneglia *et al.*, 2012; Frontini *et al.*, 2012).

L'articolo illustra i risultati di tre test volti a valutare la consistenza della categorizzazione dello spazio azionale proposta dagli annotatori per un *set* ristretto ma semanticamente coerente di verbi della risorsa: tale scelta è stata dettata dalla volontà di studiare ad un alto livello di dettaglio i problemi connessi alla tipizzazione della variazione dei predicati sugli eventi. La predisposizione di questo *case-study* è inoltre propedeutica alla creazione di una procedura standard, estendibile in un secondo tempo a porzioni statisticamente significative dell'ontologia per la sua completa validazione.

Il paragrafo 2 presenterà i coefficienti statistici adottati, nel paragrafo 3 verranno descritti metodologia e risultati dei test realizzati.

2 Coefficienti statistici

La consistenza della categorizzazione è stata valuta mediante il calcolo dell'*inter-tagger agreement* (I.T.A.). Per l'analisi sono stati utilizzati i seguenti coefficienti[1], presentati in maniera congiunta secondo le indicazioni in Di Eugenio and Glass (2004):

- A_o, "*observed agreement*" o "*Index of crude agreement*" (Goodman and Kruskal, 1954);
- π (Scott, 1955);
- k (Cohen, 1960);
- $2A_o$-1 (Byrt *et al.*, 1993);
- α (Krippendorff, 1980);
- multi-k (Davies and Fleiss, 1982);
- multi-π (Fleiss, 1971).

Tali indici, mutuati dalla psicometria, rappresentano ad oggi uno standard *de facto* in linguistica computazionale (Carletta, 1996).

Per l'analisi dei dati è stato utilizzato il modulo "metrics.agreement" di NLTK - Natural Language Toolkit (Bird *et al.*, 2009).

Il *dataset* è disponibile all'URL http://www.gloriagagliardi.com/miscellanea/.

[1] Nell'articolo viene adottata la terminologia di Artstein & Poesio (2008), lavoro di riferimento sull'argomento. I coefficienti sono illustrati e discussi in Gagliardi (2014); nel medesimo lavoro vengono inoltre esaminati i parametri che influenzano i livelli di accordo raggiungibili (Bayerl and Paul, 2011; Brown *et al.*, 2010) e i valori di significatività dei coefficienti (Landis and Koch, 1977; Krippendorf, 1980; Carletta, 1996; Reidsma and Carletta, 2008; Bayerl and Paul, 2011), in relazione ai principali studi condotti su I.T.A. per l'italiano e l'inglese in domini di tipo semantico.

3 Validazione

3.1 Test 1 (3 categorie)

Con il test 1 si intende valutare il livello di *agreement* raggiungibile nella categorizzazione di occorrenze verbali nelle categorie 'primario' – 'marcato'.

A due annotatori è stato sottoposto un set di 974 concordanze, riconducibili ad un'area semantica coesa ('girare' e lemmi verbali di significato prossimo). Il *task* consiste in un esercizio di disambiguazione "*coarse-grained*": il protocollo di annotazione prevede che ciascun *coder*, dopo aver letto ed interpretato l'occorrenza verbale in contesto, attribuisca il *tag* PRI (primario) o MAR (marcato), ovvero discrimini tra gli usi fisici ed azionali e quelli metaforici o fraseologici. Nel caso in cui non sia possibile per l'annotatore interpretare l'occorrenza o vi sia un errore di *tagging*, l'istanza deve essere annotata con l'etichetta DEL (delete), analogamente a quanto previsto nel *workflow* di IMAGACT. È inoltre richiesto all'annotatore, per le sole occorrenze PRI, di creare una frase standardizzata che espliciti e sintetizzi l'eventualità predicata. Gli annotatori (tabella 1) hanno un alto livello di esperienza nel *task*.

rater	sesso	età	istruzione	professione
A	F	29	dottorando	assegnista
B	M	29	dottorando	assegnista

Tabella 1: Annotatori test 1.

L'intera procedura è svolta dagli annotatori autonomamente ed indipendentemente. In tabella 2 sono sintetizzati i principali parametri descrittivi del test, ed in tabella 3 i risultati.

TEST 1 – 3 categorie	
numero di *rater*	2
tipologia dei dati	occorrenze verbali e relative concordanze
dimensione del *dataset*	974 occorrenze
categorie	3 (PRI - MAR- DEL)
criteri di selezione dei *rater*	Gli annotatori hanno annotato circa il 90% delle occorrenze verbali di IMAGACT-IT
livello di esperienza dei *rater*	esperti
tipo e intensità del training	intenso
coefficienti statistici	A_o, k, π, $2A_o$-1, α

Tabella 2: test 1, parametri descrittivi.

Lemma	A_o	k	π	$2A_o$-1	α
capovolgere	1.0	1.0	1.0	1.0	1.0
curvare	1.0	1.0	1.0	1.0	1.0
girare	0.90	0.84	0.84	0.84	0.84
mescolare	1.0	1.0	1.0	1.0	1.0
rigirare	0.9	0.85	0.85	0.8	0.85
rivolgere	0.79	0.53	0.52	0.57	0.52
ruotare	0.95	0.91	0.91	0.89	0.91
svoltare	1.0	1.0	1.0	1.0	1.0
volgere	1.0	1.0	1.0	1.0	1.0
voltare	0.91	0.66	0.66	0.82	0.66
TOTALE	0.89	0.83	0.83	0.79	0.83

Tabella 3: test 1 (3 categorie), risultati.

I risultati appaiono molto buoni: i coefficienti calcolati sull'insieme delle occorrenze hanno infatti un valore superiore a 0.8. Anche i valori di *agreement* calcolati per i singoli verbi sono alti: l'accordo è addirittura totale per 5 lemmi su 10. Solo i verbi 'rivolgere' e 'voltare' hanno valori di I.T.A. bassi: per il secondo lemma è però osservabile nei dati una forte prevalenza della categoria PRI (corretta dalla misura $2A_o$-1).

3.2 Test 1 (2 categorie)

In seconda battuta si è deciso di rianalizzare i dati scartando gli *item* a cui almeno un annotatore ha assegnato il *tag* DEL, considerando quindi solo le occorrenze che entrambi i *rater* hanno ritenuto interpretabili.[2] I risultati sono sintetizzati in tabella 4.

Lemma	A_o	k	π	$2A_o$-1	α
capovolgere	1.0	1.0	1.0	1.0	1.0
curvare	1.0	/	/	1.0	/
girare	0.98	0.95	0.95	0.96	0.95
mescolare	1.0	1.0	1.0	1.0	1.0
rigirare	1.0	1.0	1.0	1.0	1.0
rivolgere	0.99	0.93	0.93	0.98	0.93
ruotare	1.0	1.0	1.0	1.0	1.0
svoltare	1.0	1.0	1.0	1.0	1.0
volgere	1.0	/	/	1.0	/
voltare	0.93	0.63	0.63	0.87	0.63
TOTALE	0.98	0.96	0.96	0.91	0.96

Tabella 4: test 1 (2 categorie), risultati.

Il livello di I.T.A., già alto, supera grazie alla riformulazione del task la soglia di 0.9. L'unico lemma problematico resta 'voltare', per il problema di prevalenza già evidenziato.

[2] L'annotatore A ha usato il *tag* DEL 232 volte, l'annotatore B 244. 193 *item* hanno ricevuto il *tag* DEL da entrambi gli annotatori (circa l'80% dei casi).

3.3 Test 2

Con il test 2 si intende verificare il livello di *agreement* raggiungibile da annotatori esperti nell'assegnazione delle frasi standardizzate ai tipi azionali IMAGACT, ovvero la solidità e la coerenza della tipizzazione operata sui lemmi verbali oggetto di studio. A tale scopo è stato creato un set di frasi standardizzate a partire dai materiali annotati nel corso del test 1, secondo la seguente procedura:

- selezione dei lemmi per cui è stata identificata in IMAGACT una variazione primaria;[3]
- selezione dei verbi generali per cui, nel corso del test 1, sono state prodotte standardizzazioni primarie;
- raccolta di tutte le standardizzazioni create nel corso del test 1 per i lemmi rimanenti;
- esclusione delle frasi standardizzate uguali.[4]

Mediante questa serie di selezioni successive è stato estratto un set di 169 frasi standardizzate.

A due mesi di distanza dal primo test, agli stessi *coder* (tabella 1) è stato chiesto di assegnare le frasi standardizzate di ciascun lemma ad un inventario dato di tipi, la variazione primaria identificata nel DB IMAGACT.

In tabella 5 sono sintetizzati i principali parametri descrittivi del test, ed in tabella 6 i risultati.

TEST 2	
numero di *rater*	2
tipologia dei dati	frasi standardizzate
dimensione del *dataset*	169 frasi
categorie	da 3 a 11, in base al lemma
criteri di selezione dei *rater*	gli annotatori hanno annotato circa il 90% delle occorrenze verbali di IMAGACT-IT
livello di esperienza dei *rater*	esperti
tipo e intensità del *training*	intenso
coefficienti statistici	A_o, k, π, $2A_o$-1, α

Tabella 5: test 2, parametri descrittivi.

Il livello di I.T.A. è, in generale, buono: il valore dei coefficienti è infatti complessivamente superiore a 0.8. I due annotatori attribuiscono le standardizzazioni alle classi di azioni in modo sostanzialmente condiviso, pertanto la tipizzazione è considerabile fondata e riproducibile.

Lemma	A_o	k	π	$2A_o$-1	α
girare	0.79	0.77	0.76	0.58	0.76
mescolare	0.87	0.8	0.8	0.75	0.80
rigirare	1.0	1.0	1.0	1.0	1.0
ruotare	0.87	0.83	0.83	0.75	0.84
voltare	1.0	1.0	1.0	1.0	1.0
TOTALE	0.83	0.82	0.82	0.66	0.82

Tabella 6: test 2, risultati.

All'interno di un quadro essenzialmente positivo, com'era facilmente immaginabile il verbo più generale 'girare' appare il più difficile da disambiguare. Analizzando qualitativamente il *disagreement*, i dati evidenziano una forte concentrazione del disaccordo (11 casi su 26) in alcune specifiche categorie, la numero 9 e la numero 10, di cui si riportano i video prototipali in figura 1.

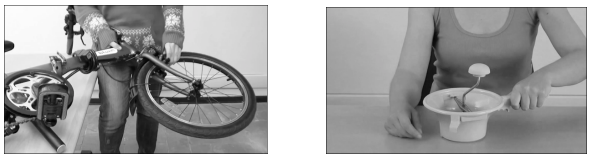

Figura 1: Tipo azionale 9 (a sinistra) e 10 (a destra) del verbo girare.

Vi è un'evidente contiguità tra le due classi di azioni: in entrambi i casi l'agente applica una forza sul tema, imprimendogli movimento rotatorio. Nel tipo 9 il tema è però messo in rotazione mediante un impulso, mentre nel tipo 10 l'agente esercita la forza sull'oggetto in maniera continua. Le tipologie di eventualità, chiaramente distinte sul piano empirico, risultano probabilmente troppo granulari dal punto di vista linguistico, al punto da risultare indistinguibili. Ricalcolando i coefficienti aggregando le due categorie, A_o= 0.879, k = 0.8606, π = 0.8605 ed α= 0.8611.

3.4 Test 3

Si è infine deciso di valutare il livello di *agreement* nell'assegnazione delle frasi standardizzate ai tipi azionali nel caso di annotatori non esperti, per verificare la riconoscibilità e l'effettiva riproducibilità della tassonomia azionale anche per semplici parlanti madrelingua. I *coder* coinvolti non hanno nessuna formazione specifica: gli unici requisiti per la selezione sono stati il livello di istruzione, medio-alto, e la di-

[3] Tali criteri comportano l'esclusione dal *test-set* dei verbi 'capovolgere', 'rivolgere', 'svoltare', 'volgere' e 'curvare'.

[4] La scelta è stata dettata dalla volontà di eliminare, almeno in parte, effetti distorsivi nel campione: alcune frasi, create dagli annotatori da occorrenze del *sub-corpus* di acquisizione LABLITA, si ripetono moltissime volte (es. "Il registratore gira"). Ciò è riconducibile alle modalità espressive del *baby-talk*, non certo ad una maggior frequenza della frase (o del tipo azionale) in italiano standard.

sponibilità a sottoporsi al test senza ricevere alcun compenso. I quattro annotatori reclutati (tabella 7) non sono stati sottoposti ad uno specifico *training*.

rater	sesso	età	istruzione	professione
C	M	32	laurea (LM)	web editor
D	M	28	laurea	web designer
E	M	30	dottorato	insegnante
F	F	26	laurea (LM)	inoccupato

Tabella 7: Annotatori test 3.

Il test segue lo stesso protocollo sperimentale dell'esercizio precedente (tabella 8). In tabella 9 sono sintetizzati i risultati.

TEST 3	
numero di *rater*	4
tipologia dei dati	frasi standardizzate
dimensione del *dataset*	169 frasi
categorie	da 3 a 11, in base al lemma
criteri di selezione dei *rater*	nessuna formazione specifica in linguistica; livello di istruzione medio-alto
livello di esperienza dei *rater*	principianti
tipo e intensità del *training*	Nessun *training*
coefficienti statistici	A_o, multi-k, multi-π, α

Tabella 8: test 3, parametri descrittivi.

Lemma	A_o	Multi-k	Multi-π	α
girare	0.72	0.69	0.69	0.69
mescolare	0.8	0.7	0.7	0.7
rigirare	0.92	0.88	0.88	0.88
ruotare	0.77	0.69	0.69	0.7
voltare	0.82	0.67	0.66	0.67
TOTALE	0.75	0.73	0.73	0.73

Tabella 9: test 3, risultati.

Valori di *agreement* situati intorno alla soglia di 0.7, pur essendo inferiori ai risultati ottenuti dagli annotatori esperti del test 2, sono comunque da ritenersi accettabili, tanto più se si tiene in considerazione la completa assenza di *training*.

Tutti e quattro i *rater* hanno lamentato una maggior difficoltà nell'annotazione del verbo 'girare' rispetto agli altri lemmi, difficoltà che tuttavia non risulta dai dati. A differenza del test 2, l'unificazione dei tipi 9 e 10 in una unica categoria non porta particolari benefici: A_o= 0.7392, multi-k = 0.7064, multi-π = 0.7059, α = 0.7065. Il valore degli indici risulta abbassato, piuttosto, dal comportamento difforme di uno dei *rater*: se, sulla base dei risultati in tabella 9, si selezionassero i migliori tre annotatori (C, E, F) e si ricalcolassero i coefficienti, A_o= 0.8224, multi-k = 0.8078, multi-π = 0.8077, α = 0.8081.

	Pairwise agreement					
Lemma	C-D	C-E	C-F	D-E	D-F	C-D
girare	0.61	0.83	0.76	0.60	0.61	0.75
mescolare	0.65	0.70	0.79	0.65	0.56	0.90
rigirare	1.0	0.77	1.0	0.77	1.0	0.77
ruotare	0.67	0.66	0.65	0.83	0.67	0.66
voltare	0.48	0.85	1.0	0.58	0.48	0.85
TOTALE	0.61	0.83	0.76	0.61	0.61	0.75

Tabella 10: test 3, *pairwise agreement*.

4 Conclusioni

Notoriamente i task di annotazione semantica, ed in particolare quelli dedicati al lessico verbale (Fellbaum, 1998; Fellbaum *et al*., 2001), fanno registrare bassi livelli di I.TA.[5] Nel caso in oggetto la possibilità di ottenere valori alti, anche con annotatori non esperti, è con buona probabilità dovuta alla natura esclusivamente azionale e fisica delle classi usate per la categorizzazione.

In seguito alla validazione è stato possibile utilizzare i dati in applicazioni di tipo psicolinguistico (Gagliardi, 2014): il campionario di verbi dell'ontologia, ampio e al tempo stesso formalmente controllato, se integralmente validato potrebbe rappresentare una fonte inedita di dati semantici per le scienze cognitive. A tale scopo, oltre che per un pieno sfruttamento didattico e computazionale della risorsa,[6] in un prossimo futuro la metodologia illustrata verrà estesa ad una porzione quantitativamente e statisticamente significativa del database.

Acknowledgments

Il progetto IMAGACT è stato finanziato dalla regione Toscana nell'ambito del programma PAR.FAS. (linea di azione 1.1.a.3). Ulteriori ricerche, incluso questo articolo, sono state realizzate grazie al contributo del progetto MODELACT (2013-2016, Futuro in Ricerca).

[5] Per una rassegna dei risultati delle maggiori campagne di valutazione si veda Gagliardi (2014).

[6] Ad esempio per l'arricchimento di risorse semantiche esistenti e Word Sense Disambiguation. Si vedano a questo proposito Bartolini *et al.* (2014) e Russo *et al.* (2013).

Reference

Ron Artstein and Massimo Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4): 555–596.

Roberto Bartolini, Valeria Quochi, Irene De Felice, Irene Russo, Monica Monachini. 2014. From Synsets to Videos: Enriching ItalWordNet Multimodally. In: Nicoletta Calzolari *et al*. (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation – LREC'14*, ELRA – European Language Resources Association, pp.3110-3117.

Petra Saskia Bayerl and Karsten Ingmar Paul. 2011. What determines Inter-Coder Agreement in manual annotations? A meta-analytic investigation. *Computational Linguistics*, 37(4): 699–725.

Steven Bird, Ewan Klein and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Beijing.

Ted Byrt, Janet Bishop, and John B. Carlin. 1993. Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, 46(5): 423–9.

Susan Windisch Brown, Travis Rood and Martha Palmer. 2010. Number or Nuance: Which Factors Restrict Reliable Word Sense Annotation? In: Nicoletta Calzolari *et al*. (eds.), *Proceedings of the Seventh InternationalConference on Language Resources and Evaluation - LREC 2010*. ELRA – European Language Resources Association, pp. 3237-3243.

Jean Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2): 249–254.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1): 37-46.

Mark Davies and Joseph L. Fleiss. 1982. Measuring Agreement for Multinomial Data. *Biometrics*, 38(4): 1047–1051.

Barbara Di Eugenio and Michael Glass. 2004. The Kappa statistic: a second look. *Computational Linguistics*, 30(1): 95–101.

Christiane Fellbaum (ed.). 1998. *WordNet: an electronic lexical database*. MIT Press, Cambridge, MA.

Christiane Fellbaum, Martha Palmer, Hoa Trang Dang, Lauren Delfs and Susan Wolf. 2001. Manual and Automatic Semantic Annotation with WordNet. In: *Proceedings of SIGLEX Workshop on WordNet and other Lexical Resources*.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5): 378–382.

Francesca Frontini, Irene De Felice, Fahad Khan, Irene Russo, Monica Monachini, Gloria Gagliardi and Alessandro Panunzi. 2012. Verb interpretation for basic action types: annotation, ontology induction and creation of prototypical scenes. In: Michael Zock and Reinhard Rapp (eds.), *Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon, CogALex III*, pp. 69–80. The COLING 2012 Organizing Committee.

Gloria Gagliardi. 2014. *Validazione dell'Ontologia dell'Azione IMAGACT per lo studio e la diagnosi del Mild Cognitive Impairment (MCI)*. PhD thesis, Università degli Studi di Firenze, Italia.

Leo A. Goodman and William H. Kruskal. 1954. Measures of association for cross classifications. *Journal of the American Statistical Association*, 49(268): 732– 764.

Klaus Krippendorff. 1980. *Content Analysis: an introduction to its Methodology*. Sage Publications, Newbury Park, CA, prima edizione.

J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1): 159–174.

Massimo Moneglia, Monica Monachini, Omar Calbrese, Alessandro Panunzi, Francesca Frontini, Gloria Gagliardi and Irene Russo. 2012. The IMAGACT cross-linguistic ontology of action. A new infrastructure for natural language disambiguation. In: Nicoletta Calzolari *et al*. (eds.), *Proceedings of the Eigth International Conference on Language Resources and Evaluation – LREC'12*, ELRA – European Language Resources Association, pp. 948-955.

Massimo Moneglia, Susan Windisch Brown, Francesca Frontini, Gloria Gagliardi, Fahad Khan, Monica Monachini and Alessandro Panunzi. 2014. The IMAGACT visual ontology. An extendable multilingual infrastructure for the representation of lexical encoding of action. In: Nicoletta Calzolari *et al*. (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation – LREC'14*, ELRA – European Language Resources Association, pp.3425-3432.

Dennis Reidsma and Jean Carletta, 2008. Reliability measurement without limits. *Computational Linguistics*, 34(3): 319–326.

Eleanor Rosch. 1978. Principles of categorization. In: E. Rosch and B. L. Lloyd (eds.), *Cognition and Categorization*, pp. 27-48. Lawrence Erlbaum Associates, Hillsdale, NW.

Irene Russo, Francesca Frontini, Irene De Felice, Fahad Khan, Monica Monachini. 2013. Disambig-

uation of basic action types through Nouns' Telic Qualia. In: Roser Saurì *et al.* (eds.), *Proceedings of the 6th International Conference on Generative Approaches to the Lexicon (GL2013)*, Association for Computational Linguistics, pp. 70-75.

William A. Scott. 1955. Reliability of Content Analysis: The Case of Nominal Scale Coding. *Public Opinion Quarterly*, 19(3): 321–325.