

A crowdsourcing study of logical metonymy

Alessandra Zarcone, Sebastian Padó

`< zarconaa@ims.uni-stuttgart.de, pado@cl.uni-heidelberg.de >`

27/07/2011



Outline

- 1 Logical Metonymy**
 - LM Resolution
 - Two takes on LM
 - Elicitation in the study of LM
- 2 A crowdsourcing study of logical metonymy**
 - Three experiments on LM
 - Experiment 1
 - Experiment 2
 - Experiment 3
- 3 Conclusions**



Logical Metonymy (LM)

LM Resolution

Some verb-object pairs require the recovery of covert events (**CE**):

Event-denoting objects (**EV**) vs. entity-denoting objects (**EN**):

EV: begin the **afternoon**

→ ✓ begin(afternoon)

→ ✗ begin(**CE**(afternoon))

EN: begin the **newspaper**

→ ✗ begin(newspaper)

→ ✓ begin(**CE**(newspaper))

→ begin **reading** the newspaper

- **The Trigger Question:** *When* do CEs arise?
- **The Range Question:** *How* CEs are understood?
how is this implicit knowledge retrieved?



Logical Metonymy (LM)

Two takes on LM

The lexical hypothesis (Pustejovsky 1995, McElree et al. 2001):

- ontological trigger hypothesis: CEs are triggered by a type-mismatch (event-subcat. verb + entity-denoting obj.)
- qualia structure hypothesis: CEs from qualia structure in the lexicon
 - Mary began the book: **reading** or **writing**
 - John is a famous wrestler. He really enjoys a good fight (**fighting**)
 - John is a wrestling fan. He really enjoys a good fight (**watching**)

The pragmatic hypothesis (Fodor & Lepore 1998, de Almeida & Dwivedi 2008):

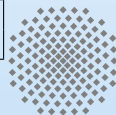
- dynamic inferences based on context and world knowledge
- post-lexical information

Logical Metonymy (LM)

Elicitation in the study of LM

"We cannot assume the parallelism between the frequency of events and the frequency of utterances about events to be perfect [...] Infrequent events may be perceived as more informative or interesting and therefore more worthy of being communicated, [...] frequent events may be perceived as less newsworthy and therefore be mentioned less often than they occur." (U. Pado 2007)

Typical, plausible	plausible, not typical	sel. restr. violations
tapped into by production norms	not elicited	not elicited
expectations in comprehensions	not expected, but not anomalous	not expected anomalous
not always attested (less newsworthy)	attested (more newsworthy)	not attested



Logical Metonymy (LM)

Elicitation in the study of LM

"One can predict that problems with finding evidence for *begin* V NP will occur on the basis of Gricean principles of language production, where the heuristic *be brief* [...] will compel speakers to utter *begin coffee* as opposed to *begin V coffee* if V is one of the plausible interpretations of *begin coffee*" (Lapata & Lascarides 2003)

On-line psycholinguistic studies

- + on-line processing
- + more natural tasks
- CEs in absentia

Crowdsourcing studies

- off-line measures
- metalinguistic analysis
- + CE elicitation



A crowdsourcing study of logical metonymy

Three experiments on LM

Non-expert annotation, intuitions and elicitations from native speakers to study LM

Experiment 1: EN vs. EN nouns

Experiment 2: CE vs. non-CE interpretation

Experiment 3: validation of Experiment 2



A crowdsourcing study of logical metonymy

Experiment 1

10 noun triplets \times 6 conditions = 60 sentences:

EN: Jim **began/spotted the magazine from the camp on the hill.**

EV: Al **began/spotted the ceremony from the camp on the hill.**

EN/EV Nick **began/spotted the conquest from the camp on the hill.**

- aim:**
- check for non-expert annotation of objects (EN, EV, EN/EV)
 - effect of Obj + 1 position on sortal categorization

participants: 14 participants from the US

procedure: crowdsourcing annotation experiment

Jan enjoyed **the automobile**

(possible answers: EN, EV; they could check either or both)

materials: 10 triplets of sentences, in 3 sentential contexts:

no PP: Jan enjoyed the automobile **short PP:** Jan enjoyed the automobile **on**

the premises full PP: Jan enjoyed the automobile **on the premises of the company**



A crowdsourcing study of logical metonymy

Experiment 1

- reasonably good agreement (weighted $\alpha = .52$)
- very good agreement with the Gold Standard ($\alpha = .70$, weighted $\alpha = .79$)
- ruled out effect of sentential context and of the verb on the sortal type assigned to the object for our material sentences
 - Binomial logistic regression model 1: $entity \sim context + verb$
Context: binomial $p = .3621 \rightarrow$ no effect
Verb: $z = 1.491, p = .1359 \rightarrow$ no effect
 - Binomial logistic regression model 2: $event \sim context + verb$
Context: binomial $p = .6138 \rightarrow$ no effect
Verb: $z = -0.504, p = .614 \rightarrow$ no effect



A crowdsourcing study of logical metonymy

Experiment 2

- aim:**
- **the trigger problem** evaluate correlation between EN/EV and CE/noCE
 - **the range problem** elicit CEs and explore their range

participants: 15 participants from the US

materials: same of Experiment 1

procedure: crowdsourcing annotation experiment
Jan enjoyed the automobile

- *does the sentence involve an additional activity that is not mentioned in the sentence?*
(answers: *additional activity* or *no additional activity*)
- when they answered *additional activity*, participants were asked to provide examples



A crowdsourcing study of logical metonymy

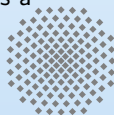
Experiment 2

Analysis 1: CE vs. no-CE

- low agreement ($\alpha = .35$) but good agreement with the Gold Standard ($\alpha = .6$)
- Binomial logistic regression model: $answer \sim obj_type * verb_type$;
 - 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%

The trigger question:
 is the sortal trigger hypothesis a
 tendency or a rule?



A crowdsourcing study of logical metonymy

Experiment 2

condition	V-obj. pair	CE	no-CE
begin,EN	begin the newspaper	89%	11%
begin,EN/EV	begin the breakfast	81%	19%
begin,EN	enjoy the automobile	50%	50%
begin,EN/EV	enjoy the translation	39%	61%
spot,EN	remember the brandy	34%	66%
begin,EV	enjoy the conference	24%	76%
spot,EV	remember the revolt	10%	90%
spot,EN/EV	remember the shower	8%	92%
begin,EV	endure the revolt	3%	97%
spot,EN	approve the automobile	0%	100%
spot,EN/EV	organize the breakfast	0%	100%
spot,EV	organize the afternoon	0%	100%

condition	V-obj. pair	CE	no-CE
begin,EN	begin the newspaper	89%	11%
begin,EN/EV	begin the breakfast	81%	19%
begin,EN	enjoy the automobile	50%	50%
begin,EN/EV	enjoy the translation	39%	61%
spot,EN	remember the brandy	34%	66%
begin,EV	enjoy the conference	24%	76%
spot,EV	remember the revolt	10%	90%
spot,EN/EV	remember the shower	8%	92%
begin,EV	endure the revolt	3%	97%
spot,EN	approve the automobile	0%	100%
spot,EN/EV	organize the breakfast	0%	100%
spot,EV	organize the afternoon	0%	100%



The trigger problem:

A crowdsourcing study of logical metonymy

Experiment 2

The range problem:

- participants tend to give 1 or 2 answers (mean 1.4, range 1-6)
- when they only give one answer: mean 3.2 CEs elicited per VP item
- general mean 5 CEs per VP item (range 1-15)

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%)



A crowdsourcing study of logical metonymy

Experiment 3

aim: validating Experiment 1

participants: 10 participants from the US

materials: 211 elicited sentences from Experiment 2
+ 57 "good" sentences (from materials in McElree et al. 2001 and Traxler et al. 2002)
+ 45 "bad" sentences (e.g. "Alex finished being French")

procedure: Jan enjoyed driving the automobile

- *Rate the acceptability of the sentence on a scale from 1 (not at all acceptable) to 5 (very acceptable)*
- *when they rated a sentence as bad (1, 2 or 3), participants were asked to provide a better formulation for it*



A crowdsourcing study of logical metonymy

Experiment 3

- mean rating 4.05

BAD	Elicited	GOOD
2.19	4.05	4.57

- **28 (13.3%)** "bad" sentences excluded from the elicited sentences
- **183 (86.7%)** sentences selected as "good" (157 "good" sentences + 26 "alternative" sentences)



Conclusions

Crowdsourcing and elicitation:

- fast and affordable collection of linguistic judgements, non-expert annotation, intuitions and elicitations from native speakers
- Importance of elicitation in studying CEs and LM

Logical Metonymy:

- the sortal trigger hypothesis and the qualia structure hypothesis do not seem to be enough
- highly lexically determined CE interpretation
- need for a broader and more dynamic framework to account for LM phenomena
- future work: a plausibility-driven account of LM



Thank you!

This study was supported by a grant from the Deutsche Forschungsgesellschaft (SFB 732/D6).



- de Almeida, R. D. & Dwivedi, V. D. (2008). Coercion without lexical decomposition: Type-shifting effects revisited. *Canadian Journal of Linguistics*, 53(2/3), 301-326.
- Fodor, J. A., & Lepore, E. (1998). The emptiness of the lexicon: Reflections on James Pustejovsky's *The Generative Lexicon*. *Linguistic Inquiry*, 29(2), 269-288.
- McElree, B. et al. (2001). Reading time evidence for enriched composition. *Cognition*, 78, B17-B25.
- McRae, K., et al. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory & Language*, 38(3), 283-312.
- Lapata, M., et al. (2003). Intra-sentential context effects on the interpretation of logical metonymy. *Cognitive Science*, 27(4), 649-668.
- Padó, U. (2007). *The Integration of syntax and semantic plausibility in a wide-coverage model of sentence processing*. Phd thesis. Saarland University.
- Pustejovsky, J. (1995). *The generative lexicon*. MIT Press.
- Traxler, M. J., et al. (2002). Coercion in sentence processing: Evidence from eye-movements and self-paced reading. *Journal of Memory & Language*, 47:530-547.

