"I like work: I can sit and look at it for hours" Type clash vs. plausibility in covert event recovery

Alessandra Zarcone, Sebastian Padó
University of Stuttgart
Stuttgart, Germany
zarconaa,pado@ims.uni-stuttgart.de

Abstract

A range of event-subcategorizing verbs can combine with entity-denoting nouns, like *begin the newspaper*. The interpretation of such sentences typically involves the recovery of covert events (CE) which are not realized on the surface, as in *begin reading the newspaper*. We report on an ongoing study that scrutinizes two assumptions made by traditional accounts: (a) that the triggering of CEs can be ascribed to the object's ontological type; and (b), that one or two CEs can be retrieved for each noun. Preliminary evidence against both assumptions is presented.

1 Covert Events

There is a substantial class of more than a dozen verbs whose members have been argued to subcategorize for an event (Pustejovsky, 1995; Jackendoff, 1997), but which can also combine with an entity. This class comprises a number of high-frequency verbs, such *enjoy* or *begin*. These verbs do not pose problems when combined with event-denoting objects (EV, e.g. *the afternoon*), but when combined with entity-denoting objects (EN, e.g. *the newspaper*) they constitute a challenge for traditional compositional accounts of sentence meaning, because their interpretation seems to require the recovery of covert events (CE) which are not realized on the surface (*begin the newspaper* → *begin reading the newspaper*). The interpretation of such pairs seems to involve at least two specification steps: (1) the triggering of (the need for) a CE; (2) the recovery of a specific CE.

The main determinant of step (1) has been argued to be the ontological type of the object (EN vs. EV objects) and its type-clash with the event-subcategorizing verb (Pustejovsky, 1995; Jackendoff, 1997; Traxler et al., 2002). Step (2) is traditionally assumed to result in one or at most two CEs retrieved from the *qualia structure* (QS) of the lexical entry for the object (Pustejovsky, 1995).

Behavioral studies have grounded this binary distinction in higher processing costs for conditions that involve CE recovery (see Pylkkänen and McElree (2006) for a review). Traxler et al. (2002) compared EN conditions ("began the book") with EV conditions ("began the fight"), using both eye-tracking and self-paced reading, and detected higher processing costs for EN objects with event-subcategorizing verbs both at the target position (the object itself) and at the post-target position.

2 Open Issues

The goal of our work is to scrutinize two assumptions of the traditional account: the nature of the "trigger" and the range of possible CEs.

The trigger problem. The following examples illustrate our intuition that a type clash between verb and object cannot be the only factor responsible for evoking CEs:

(1) I like work: it fascinates me. I can sit and look at it for hours.[1]

(2) Mary began the translation → began the translation process (EV) OR began reading/revising/typing the translation (EN).

(3) a. John is a famous wrestler. He really enjoyed the fight last night.
b. John is a wrestling fan. He really enjoyed the fight last night. → enjoyed watching the fight.

The twist that turns (1) into a joke is exactly the interpretation of work as an event, which is nevertheless later modifier by the recovery of a CE inserted between the verb and the object. The second example introduces a whole category of cases which are problematic for an ontological trigger, namely sortally ambiguous nouns that can assume

[1] J. K. Jerome, *Three men in a boat*, 1889

Noun Type	Example		Interpretation	Paraphrase
EV	begin the **afternoon**	→	✓ begin(afternoon)	
		→	× begin(**CE**(afternoon))	
EN	begin the **newspaper**	→	× begin(newspaper)	
		→	✓ begin(**CE**(newspaper))	begin **reading** the newspaper
EN/EV	begin the **breakfast**	→	? begin(breakfast)	
		→	? begin(**CE**(breakfast))	begin **eating** the breakfast

Table 1: Interpretation of different noun types after event-subcategorizing verbs

both an EN and an EV reading (cf. Table 1). One possible prediction would be that if a reading without type clash (i.e., an EV reading) is available, it will be chosen. This prediction is contradicted by Example (3), which shows that preceding discourse context can determine the choice between EN and EV.

Evidence against the type clash hypothesis also comes from work on metonymy resolution (Markert and Hahn, 1997; Markert and Hahn, 2002), which rejects this hypothesis on the basis of computational and cognitive arguments, and from MEG studies (Pylkkänen and McElree, 2007; Pylkkänen et al., 2009), which showed different brain activity correlates for semantic anomaly and for CE constructions.

Corpus sentence	Interpretations
If you are going hungry, seek **help with food** right away	obtain, buy, get
One friend works in the kitchen, **helping with food**	prepare, cook
I need **help with** dog **food**	select, choose

Table 2: Examples of verb+EN noun pairs

The range problem. Another issue concerns the retrieved CEs. If we equate CEs with qualia roles, there should be one or two CEs associated with each noun. However, the examples in Table 2 indicate that a wider range of CEs might be available, as Vendler (1968) had also observed.

Also, as observed by Lapata and Lascarides (2003) and Shutova and Teufel (2009), CEs are to be considered not as single verb lexical items but rather as classes of events sharing semantic similarities: each entry in Table 2 can be interpreted with a set of synonymous verbs rather than with a single lexical item.

2.1 An alternative mechanism: Plausibility

The alternative hypothesis that we want to explore is that interpretation is basically *plausibility-driven*. This hypothesis is coherent with the results obtained by probabilistic models of logical metonymy (Lapata and Lascarides, 2003; Shutova and Teufel, 2009).

The trigger problem. Probabilistic models yielded interesting results in predicting CE interpretations, but they did not distinguish between contexts in which CE are retrieved and contexts in which they are not. In order to account for the trigger problem, we suggest that CEs are retrieved when the plausibility of the standard verb/noun combination is small compared to the plausibility of the verb/CE/noun combination[2].

The range problem. Assuming an important role of plausibility, there is also no reason why the range of CEs should be limited a priori; rather, the CE could be sampled from distributional knowledge about plausible predicate-argument structures (Padó et al., 2007); more than one or two clusters of meaning can be retrieved and ranked for their plausibility (Lapata and Lascarides, 2003).

Steps of interpretation. These are the operations that we assume to take place when a potentially metonymic construction v, o is processed, given a previous context c:

1. candidate retrieval: a number of CE interpretations ce are activated, showing high plausibilities $Plaus(v, ce, o|c)$;

2. CE triggering: $Plaus(v, e, o|c)$ for the selected interpretations are compared to $Plaus(v, o|c)$; if $Plaus(v, o|c)$ is high enough to warrant non-CE interpretation, then no CE is retrieved; if instead the most plausible interpretation involves a CE, then the CE interpretation is selected;

3. CE range: the most plausible CE interpretation for v, o given c is selected and the meaning of e is integrated into the sentence meaning.

[2] The plausibility of the verb/CE/noun combination (v, e, o) can be estimated as the joint probability of $P(v)$, $P(v|e)$ and $P(o|e)$ (Lapata and Lascarides, 2003).

We make four observations. (a) Traxler et al. (2005) and Frisson and McElree (2008) showed that higher processing costs in CE conditions are not due to the retrieval of the CE, but to the "building of an extended event sense of the complement", so the plausibility comparisons in step 2 alone do not lead to higher processing costs. (b) The model does not imply a strong rejection of the type-clash model, but rather its predictions capture "tendencies" of the model: EV nouns tend to show higher plausibilities for the verb/noun combination, EN nouns show an opposite tendency and therefore more often require the recovery of CEs. (c) The range in 3 can be wider or narrower depending on the skewedness of the distribution over covert events given the previous context c. (d) Less plausible interpretations can remain available, in case following context falsifies the selected interpretation.

Predictions from the model. EN/EV ambiguous nouns as objects provide a suitable test object for our hypothesis: with both readings available, we can test to what extent plausibility considerations can account for differences in reading times. We expect EN nouns to show longer reading times than EV nouns in metonymic contexts; as to EN/EV nouns, we expect their behavior to be highly lexically-determined and to correlate with plausibility estimations. We therefore plan a self-paced reading study involving EN/EV ambiguous nouns, which is described in Section 3.

As to the range problem, reading time studies cannot help us in regard to it, as the CEs do not form part of the information acquired from the subjects. Section 4 therefore presents web-based elicitation methods that serve both to select materials for the reading time study and to explore the correlation between speaker's categorization of objects into EN / EV and their CE interpretation.

3 A self-paced reading study

Our design mirrors the study in Traxler et al. (2002), with an additional level: together with EN and EV objects, we are going to analyze the interaction between entity-subcategorizing verbs and EN/EV ambiguous nouns. 10 triplets of EN - EV - EN/EV ambiguous nouns were selected. For each triplet, two verbs were chosen: an event-subcategorizing verb (*begin-verb*), and a verb which could categorize both for an event and an entity (*spot-verb*). See an example triplet:

EN: Keith enjoyed/approved the automobile on the premises of the company.

EV: Daniel enjoyed/approved the conference on the premises of the company.

EN/EV: Walter enjoyed/approved the translation on the premises of the company.

Objects were matched within each triplets for length, frequency (Francis and Kucera, 1967), and co-occurrence frequency with the begin-verb and the spot-verb (ukWaC corpus, Ferraresi et al. (2008)), as a rough indicator of plausibility (Lapata and Lascarides, 2003). The 10 triplets were selected after threefold annotation, to evaluate our annotation of the nouns as EN, EV or EN/EV. Non-weighted Krippendorff's α (Krippendorff, 2004) for the selected triplets was .71, or good agreement. We also computed the weighted version of α, which incorporates the idea that EN vs. EV is a stronger disagreement compared to either of the types vs. the ambigous EN/EV type.[3].

Weighted α is =.79 – that is, determining the appropriate reading is not trivial, but doable.

4 Web experiments

The experiments were delivered using the crowdsourcing paradigm (Snow et al., 2008), for fast and affordable collection of judgments.

4.1 Experiment 1

In Experiment 1, 14 annotators from the US re-annotated the 30 nouns from the 10 triplets selected for the self-paced reading study for their readings (EN, EV, EN/EV). The aim of Experiment 1 was to check for non-expert annotation of the materials for the self-paced reading study, and to verify that this annotation did not change with different PP contexts.

Each noun appeared with a begin-verb and with a spot-verb and in three contexts: without the PP ("Keith enjoyed the automobile"), with the first part of the PP ("Keith enjoyed the automobile on the premises"), and with the complete sentence ("Keith enjoyed the automobile on the premises of the company"). We found a reasonably good agreement among annotators for a crowdsourcing experiment (weighted $\alpha = .52$)[4] and were able to rule out potential meaning changes caused by the

[3] We assigned a weight of 1 to EN-EV and a weight of 0.5 to EN-EN/EV and EV-EN/EV.

[4] Note that 14 annotations allow us to compute a reliable "majority vote" so that the practical reliability is higher.

post-nominal PPs: higher processing costs in the self-paced reading study will only be ascribed to CE recovery.

4.2 Experiment 2

It is not unusual for works on logical metonymy to include off-line norming studies, which can involve estimation of plausibilities for given CEs in a metonymical construction (Lapata and Lascarides, 2003) or the elicitation of a CE in a cloze completion task (McElree et al., 2001; Lapata et al., 2003). Nevertheless, the very same design of these experiments neglected the two aspects we are focusing on: cloze completion and plausibility estimation do not explore differences betweeb CE and no-CE interpretation (trigger problem) and limit the range of elicitations to only one CE (range problem). The aim of Experiment 2 is to evaluate the role of EN, EV, EN/EV nouns in triggering CE interpretations, to elicit more than one CE interpretation and to explore their range.

4.2.1 Experiment 2: materials and design

Experiment 2 was conducted with the same materials and procedure of Experiment 1, but this time participants were asked to choose between a CE interpretation and a simple compositional interpretation (*does the sentence involve an additional activity that is not mentioned in the sentence?*). Two options were given (*additional activity* vs. *no additional activity*), some examples are provided, and when a participant answered *additional activity*, she or he was asked to provide instances of possible activities. EN and EV interpretations were not mentioned in the experiment's instructions.

4.2.2 Experiment 2: results

The results from Experiment 2 involve two aspects 1) the CE/no-CE answer; 2) the elicited CEs.

CE/no-CE. Agreement for Experiment 2 was rather low ($\alpha = .35$)[5], but the majority vote showed a good agreement with the Gold Standard ($\alpha = .60$).

A binomial logistic regression on the CE/no-CE answers ($answer \sim obj\_type * verb\_type$) yielded a significant effect of the type of the object (binomial $p < 0.001$), and of the verb type ($z = -8.322; p < 0.001$), with interaction (binomial $p < 0.001$). These effects seem to confirm the type-clash hypothesis, but consider Table 3: 38% of begin-verb/EN-noun combinations did not elicit CEs, while 18% of begin-verb/EV-noun combinations did.

condition	% CE	% no-CE
begin,EN	0.63	0.38
spot,EN	0.11	0.89
begin,EN/EV	0.39	0.61
spot,EN/EV	0.06	0.94
begin,EV	0.18	0.82
spot,EV	0.06	0.94

Table 3: CE and no-CE answers in Experiment 2

condition	V-N pair	% CE	% no-CE
begin,EN	begin the newspaper	0.89	0.11
begin,EN/EV	begin the breakfast	0.81	0.19
begin,EN	enjoy the automobile	0.50	0.50
begin,EN	endure the brandy	0.42	0.58
begin,EN/EV	enjoy the translation	0.39	0.61
spot,EN	remember the brandy	0.34	0.66
begin,EV	enjoy the conference	0.24	0.76
begin,EV	begin the afternoon	0.20	0.80
spot,EV	remember the revolt	0.10	0.90
spot,EN/EV	remember the shower	0.08	0.92
begin,EN/EV	endure the shower	0.07	0.93
spot,EV	approve the conference	0.07	0.93
begin,EV	endure the revolt	0.03	0.97
spot,EN	approve the automobile	0.00	1.00
spot,EN/EV	approve the translation	0.00	1.00
spot,EN	organize the newspaper	0.00	1.00
spot,EN/EV	organize the breakfast	0.00	1.00
spot,EV	organize the afternoon	0.00	1.00

Table 4: CE and no-CE answers for single items in Experiment 2

The type-clash hypothesis seems to capture a tendency in the data rather than to predict the participants' answers in every single case. As shown by examples in Table 4, an item-wise analysis shows a continuum of behaviors rather than clear-cut separate categories: 1) EN nouns tend to have a strong majority of CE answers with begin-type verbs; 2) EV nouns tend to have a strong majority of no-CE answers with begin-type and spot-type verbs, but exceptions are possible (e.g. *enjoy the conference*) 3) not all the spot-type verbs block CE interpretations (e.g. *remember the brandy*); 4) the behavior of EN/EV ambiguous nouns is highly lexically determined (contrast ad example *begin the breakfast*, *enjoy the translation* and *endure the shower*).

Range of CEs. Per each V-Obj combination each participant elicited on average 1.4 CEs (range 1-6). Although we did not limit the number of CEs to be elicited, eliciting only one CE appears to be a common behavior. Nevertheless, if we only look

[5] $\alpha = .36$ when excluding EN/EV ambiguous nouns, showing that the low agreement was not due to their presence

at the cases when participants elicited not more than one CE, a variety of different CEs per VP was given (average 3.2, range 1-7). In several cases the elicited CEs cover a broader set than the one given by the telic and agentive qualia:

EN: consider the butter → 8 CEs: eat (x4), add, buy, churn, cook with, eat, make, melt

EN/EV: prefer the collection → 6 CEs: view (x3), buy, discuss, polish, study, watch

EV: start the semester → 3 CEs: spend, teach, join

Even within a theory of extended qualia (Busa et al., 2001), CEs like *buy* or *melt* are difficult to account for with the QS of *butter*.

The average of elicited CEs per each verb-object combination across all participants was 5 (range 1-15). Consider the following examples from the elicited CEs:

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

Again, ascribing the sets of verbs for an EN-noun like *portrait* to the QS of the noun seems to be an unsatisfying solution, at least if the qualia are understood as specific verbs, rather than concepts (like, e.g., the agentive quale of *portrait* is *to paint*): the sets of elicited CEs form *semantically motivated verb classes* structured by semantic relations (synonymy, hyponymy), which can be understood as classes of plausible events. Among the elicited CEs there are also events which do not fall under the categories of agentive quale or telic quale: *hang*, *model*, *review*. As to EV objects (e.g. *conference*), they can also elicit CE readings (*enjoy attending/holding a conference*), and for EN/EV ambiguous objects like *harvest* both readings often give rise to elicited events. Note also that the elicited CEs include not only light verbs (*performing a translation*), which would be semantically largely transparent, but also full verbs (*reading / completing a translation*).

Table 5 reports on the amount of CEs which can be accounted for by a QS-based theory. The annotation was performed by the authors by assigning an agentive quale and a telic quale to each noun and comparing them with the elicited CEs. We considered qualia as classes of meaning, in order to cover also synonyms of the annotated qualia. Almost half of the elicited CEs did not fall in either the agentive quale category or in the telic quale category.

	tot	QS 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%

Table 5: CEs accounted for by a QS-based theory vs. other CEs

5 Conclusions

We are proposing an alternative mechanism for the recovery of covert events, according to which CEs are activated when the overt form cannot be given a plausible interpretation. We use a combination of self-paced reading and web-based elicitation to explore our hypothesis: the former detects processing costs differences, while the latter provides access to the range of CEs understood by speakers.

Results from a web elicitation study showed that the type-clash and the QS hypothesis are not enough to predict elicited CEs in a given context: CEs are elicited also for EV and EN/EV nouns, and in general the triggering of a CE seems to be highly lexically determined. Recovered CEs seems to fall in a wider range than those captured by the QS, and this range is also fairly wide when participants only give one answer.

While challenging the type-clash model, a plausibility-driven model can still retain the descriptive power of the sortal trigger hypothesis by subsuming it as a general tendency: EV nouns “tend to” show higher plausibilities for the verb/CE/noun combination, EN nouns show an opposite tendency and therefore more often require the recovery of CEs. Also, in a plausibility-driven model there is no reason why the range of CEs should be limited a priori: more than one of two clusters of meaning can be retrieved and ranked for plausibility.

Acknowledgments

The authors thank Berry Claus and Ulrike Padó for helpful discussions concerning this work.

References

F. Busa, N. Calzolari, and A. Lenci. 2001. Generative Lexicon and the SIMPLE model; developing semantic resources for NLP. In F. Busa and P. Bouillon, editors, *The Language of Word Meaning*, pages 333–349. Cambridge University Press.

A. Ferraresi, E. Zanchetta, M. Baroni, and S. Bernardini. 2008. Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *Proceedings of the WAC4 workshop at LREC 2008*, Marrakech. ELRA.

W. N. Francis and H. Kucera. 1967. *Computational Analysis of Present-Day American English*. Brown University Press, Providence.

Steven Frisson and Brian McElree. 2008. Complement coercion is not modulated by competition: evidence from eye movements. *Journal of Experimental Psychology*, 34:1–11.

R. Jackendoff. 1997. *The Architecture of the Language Faculty*. MIT Press.

K. Krippendorff. 2004. *Content Analysys: An introduction to its methodology (2nd ed.)*. Sage, Thousand Oaks, CA.

Mirella Lapata and Alex Lascarides. 2003. A probabilisitic account of logical metonymy. *Computational Linguistics*, 29(2):263–317.

Mirella Lapata, Frank Keller, and Christoph Scheepers. 2003. Intra-sentential context effects on the interpretation of logical metonymy. *Cognitive Science*, 27(4):649–668.

Katja Markert and Udo Hahn. 1997. In support of the equal rights movement for literal and figurative language: A parallel search and preferential choice model. In *Proceedings of CogSci 1997*, pages 289–294.

Katja Markert and Udo Hahn. 2002. Understanding metonymies in discourse. *Artificial Intelligence*, 135:145–198.

B. McElree, M. J. Traxler, M. J. Pickering, R. E. Seely, and R. Jackendoff. 2001. Reading time evidence for enriched composition. *Cognition*, 78:B17–B25.

S. Padó, U. Padó, and K. Erk. 2007. Flexible, corpus-based modelling of human plausibility judgements. In *Proceedings of EMNLP*, pages 400–409, Prague, Czech Republic.

J. Pustejovsky. 1995. *The Generative Lexicon*. MIT Press, Cambridge MA.

L. Pylkkänen and B. McElree. 2006. The syntax-semantic interface: On-line composition of sentence meaning. In Mattew Traxler and Morton Ann Gernsbacher, editors, *Handbook of Psycholinguistics*, pages 537–577. Elsevier, 2nd edition.

Liina Pylkkänen and Brian McElree. 2007. An MEG study of silent meaning. *Journal of Cognitive Neuroscience*, 19(11):1905–1921.

Liina Pylkkänen, Andrea E. Martin, Brian McElree, and Andrew Smart. 2009. The anterior midline field: Coercion or decision making? *Brain and Language*, 108(3):184–190.

E. Shutova and S. Teufel. 2009. Logical metonymy: Discovering classes of meanings. In *Proceedings of DiSCo 2009 Workshop*, Prague, Czech Republic.

R. Snow, B. O'Connor, D. Jurafsky, and A.Y. Ng. 2008. Cheap and fast - but is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of EMNLP 2008*.

M. J. Traxler, M. J. Pickering, and B. McElree. 2002. Coercion in sentence processing: Evidence from eye-movements and self-paced reading. *Journal of Memory and Language*, 47:530–547.

Matthew J. Traxler, Brian McElree, Rihana S. Williams, and Martin J. Pickering. 2005. Context effects in coercion: Evidence from eye movements. *Journal of Memory and Language*, 53:1–25.

Z. Vendler. 1968. *Adjectives and Nominalizations*. Mouton, The Hague, The Netherlands.