Event Types in the Mind and in the Corpus

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

Alessandro Lenci
Università di Pisa
Pisa, Italy
alessandro.lenci@ling.unipi.it

Abstract

Event types (ET) have received considerable attention in formal semantics, but their importance in experimental linguistics has developed only recently. The aim of this work is to compare the performance of human annotators and corpus-based models in ET classification of Italian verbs

1 Event Types in experimental linguistics

Event types (ET) play a crucial role in verb semantics, contributing to the temporal constitution of the sentence. We refer here to Vendler's (1967) standard classification of predicates into *states* (STA), *activities* (ACT), *accomplishments* (ACC) and *achievements* (ACH), which can be further cross-classified with respect to the features of dynamicity (DYN), durativity (DUR) and resultativity (RES):

Table 1: Features of Vendler's ETs

ET	[DYN]	[DUR]	[RES]	example
STA	−	+	−	*to know*, *to be tall*
ACT	+	+	−	*to sing*, *to walk*
ACC	+	+	+	*to write a book*, *to walk to the fence*
ACH	+	−	+	*to stumble*, *to die*

The ET of a sentence is the result of a complex interaction between the verb lexical item and the sentence context (arguments, adjunts, verb morphology) (Verkuyl, 1972); contrast for example *to walk* (ACT) and *to walk to the fence* (ACC). Such an interplay emerges very clearly in ET polysemy and ET coercion, which need to be accounted for by any theory of ETs. ET polysemy (Bertinetto, 1986; Lucchesi, 1971) is a fairly regular phenomenon: some verbs show different ETs in different contexts (e.g. ACH/STA in Italian: *impugnare*, "hold"/"get hold of"; *indossare*, "wear"/"put on"). ET coercion (Pustejovsky, 1995; Rothstein, 2004) occurs when contextual features trigger a reinterpretation of a verb into a new ET class: e.g. The student *ate* two sandwiches (ACT ⇒ ACC, countable direct object with numeral modifier); Guests *have been arriving* for hours (ACH ⇒ ACT, bare plural subject, for x time).

ETs have received considerable attention in formal semantics, but their importance in experimental linguistics has developed only recently. We believe that the study of ETs, like a number of other research areas in linguistics, could benefit from a cross-contamination among different fields and methodologies.

Antinucci and Miller (1976) showed that strong correlations between Aspect and ETs emerge in language acquisition, along the axes of telicity/perfectivity/past-reference and atelicity/imperfectivity/present-reference; such correlations also emerged in the distributional model in Li and Shirai (2000). The correlation, though relaxed, can still be detected in adult language, along with other associations between ETs and context features, by computational models such as those in Zarcone and Lenci (2008) and Im and Pustejovsky (2010). Finocchiaro and Miceli (2002) found an effect of ET on the performance of aphasic subjects, showing a double dissociation between STA and ACT and thus supporting the idea that ETs are one fundamental principle of organization of the mental lexicon in the brain. Behavioral studies have been conducted using the paradigms of self-paced reading (Gennari and Poeppel, 2002), ERP (Bott, 2008) and semantic priming (Bonnotte, 2008; Zarcone and Lenci, 2010).

A close interaction between cognitive methods and corpus-based computational methods seems to be promising to explain how this interplay between contextual elements and verb lexical items influences the speakers in determining the ET of a sentence. In this paper we have a twofold goal. First of all, we are going to test the subjects' competence of ETs in a series of cross-modal annotation experiments for English and Italian. Secondly, we are going to compare the performance of human annotators with results from corpus-based models in a task of ET classification of Italian verbs, in order to investigate potentially interesting differences among ET classes and to evaluate the contribution of cognitive and corpus-based methods to the study of ETs.

2 Competence of Event Types

We carried out four web-based annotation experiments: Experiment 1 and 2 for linguistic stimuli (Italian and English), Experiment 3 and 4 for picture stimuli (with Italian speakers and English speakers). Experiments requiring English speakers (2 and 4) were conducted using the crowdsourcing paradigm[1].

2.1 Design and procedure

Experiment 1: Materials for Experiment 1 were 138 Italian predicates (96 transitive VPs (V + Obj) representing all Vendler's classes and 42 intransitive verbs, being 21 ACH and 21 ACT). 20 native Italian-speaking students performed the test in a web-based format, each of them saw all the stimuli. The procedure was inspired by the pilot study in Bonnotte (2008). Per each event, participants were asked to choose one of four pictures, one representative of each ET (figure 1).

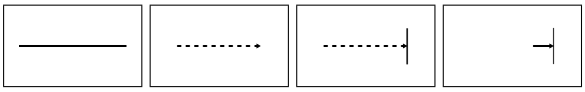

Figure 1: The long continuous line depicts a state that lasts in time, the long dashed arrow depicts a process that develops over a certain period of time, the long dashed arrow ending with a vertical dash depicts a process that develops over a certain period of time and leads to a result, the short arrow ending with a vertical dash depicts a change of state.

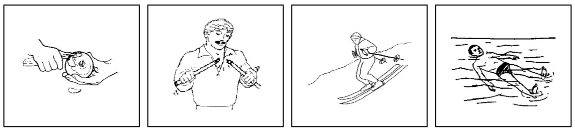

Figure 2: ACC (*to peel*), ACH (*to break*), ACT (*to ski*), STA (*to float*).

Experiment 2: Experiment 2 was conducted with the same modality as Experiment 1, but for English. An effort was made to translate the materials for Experiment 1 into English, taking particular care that each English stimulus showed the same ET and the same low degree of ambiguity of its Italian correlate. Materials for Experiment 2 were 134 predicates (96 transitive VPs (24 ACC, 24 ACH, 24 ACT, 24 STA) and 38 intransitive verbs, being 19 ACH and 19 ACT). 10 of the transitive VPs (2 ACC, 4 ACH, 4 ACT) were also presented together with the particle "up" ("up verbs", e.g. "use the materials"/"use up the materials"); so the total number of stimuli was 144. Our intuition was that the particle added an extra element of telicity to ACT VPs such as "use the materials", or simply made the telicity of ACH and ACC verbs more prominent (e.g. "lock the box"/"lock up the box"). 24 native English speakers took part in Experiment 2; as it is usual for crowdsourcing experiments, not all the participants annotated every stimulus. The minimum number of participants annotating each stimulus was 16, maximum 22, with a mean of 18.

Experiment 3: Experiment 3 was conducted with the same modality as Experiments 1 and 2, but picture stimuli was used instead of word stimuli: 111 pictures (19 ACC, 40 ACH, 40 ACT, 12 STA) were selected from the IPNP database (Bates et al., 2000), see Figure 2. 20 native Italian-speaking students took part in Experiment 3, each of them saw all the stimuli.

Experiment 4: Experiment 4 was conducted with the same modality and stimuli as Experiments 3, but participants were native speakers of English (42). The minimum number of participants annotating each stimulus was 10, maximum 16, with a mean of 13.6.

2.2 Results

Agreement: Agreement was computed with Krippendorff's α and with weighted α_w(Krippendorff, 2004). The latter was com-

[1] Crowdsourcing has been increasingly popular also in linguistics (Snow et al., 2008), allowing for lower economic and logistic costs

		base version				"up" version			
item	ET	ACC	ACH	ACT	STA	ACC	ACH	ACT	STA
draw [up] the map	ACC	6	3	8	0	10	4	4	0
dry [up] the cutlery	ACC	17	0	0	0	17	2	0	0
lock [up] the box	ACH	13	1	3	0	15	3	1	0
swallow [up] the syrup	ACH	12	1	5	0	14	1	3	0
tear [up] the table cloth	ACH	3	13	0	0	4	13	0	0
wake [up] the doorman	ACH	7	10	1	0	12	6	1	0
beat [up] the wife	ACT	5	11	2	0	16	1	2	0
eat [up] the strawberries	ACT	6	0	10	1	15	1	2	0
use [up] the materials	ACT	10	0	5	2	12	1	2	2
wait [up] for the verdict	ACT	4	15	0	0	7	11	0	0
		83	54	34	3	122	43	15	2

Table 2: Answers given for "up" verbs

puted in order to modulate disagreement across categories which are not equally distant[2].

Agreement values were above chance and reasonably good for Experiment 1 ($\alpha = .35; \alpha_w = .43$) and Experiment 2 ($\alpha = .46; \alpha_w = .53$), since the subjects were naive to linguistics and ET classification and no sentence context was given; agreement was lower for Experiment 3 ($\alpha = .22; \alpha_w = .31$) and Experiment 4 ($\alpha = .28; \alpha_w = .39$).

Accuracy: Accuracy values are reported in table 3 (please note that for Experiment 2 "up verbs" were excluded from accuracy computation).

A binomial logistic regression analysis ($correct\_answer \sim ET * valency * sem\_class$) for Experiment 1 yielded a significant effect of ET (binomial $p < 0.05$), a highly significant effect of valency and semantic class[3] (binomial $p < 0.001$), a significant interaction ET*valency and valency*sem_class (binomial $p < 0.05$) and a highly significant interaction ET*sem_class and ET*valency*sem_class (binomial $p < 0.001$). The same analysis for Experiment 2 yielded a highly significant effect of ET, valency and semantic class and semantic class with significant interactions ET*valency (binomial $p < 0.005$) and ET*sem_class (binomial $p < 0.001$). A binomial logistic regression analysis ($accuracy \sim ET$) for both Experiment 3 and 4 yielded a highly significant effect of ET on accuracy (binomial $p < 0.001$).

Certain ETs seem to be easier to identify than others. In particular, within the transitive VPs, ACCs seem easier than ACTs, probably due to their being more prototypically transitive in Italian and English, and ACHs and ACTs seem easier to identify when intransitive (as in Italian and English ACHs and ACTs are more prototypically intransitive).

Also, it seems that the semantic class of the predicate might play an important role in leading the annotators' choices in ET classification. Please note that a straightforward correspondence between ETs and semantic classes (e.g. motion verbs → ACT) was when possible avoided: a special effort was made when building the stimuli, in order to have, within each ET class, representatives of different semantic classes, and, within each semantic class, representatives of different ETs.

As to the 10 transitive VPs (2 ACC, 4 ACH, 4 ACT) which also appeared a second time with the particle "up" (see table 2), the contribution of the particle to the ET of the VPs strengthen their telicity, making it more prominent (for ACC and ACH items) or by changing the value of the RES feature (ACT answers go from 34 for the base version to only 15 for the "up" version).

	all	ACC	ACH	ACT	STA
Exp 1 (it, verbs)	0.63	0.76	0.66	0.61	0.48
Exp 1, transitives	0.59		0.57	0.53	
Exp 1, intransitives			0.76	0.69	
Exp 2 (en, verbs)	0.68	0.81	0.66	0.72	0.51
Exp 2, transitives	0.64		0.60	0.64	
Exp 2, intransitives	0.78		0.73	0.82	
Exp 3 (it, pictures)	0.42	0.34	0.54	0.60	0.34
Exp 4 (en, pictures)	0.54	0.68	0.54	0.50	0.48
MaxEnt	0.85	0.89	0.90	0.74	0.78
SOM	0.50	0.86	0.47	0.50	0.27

Table 3: Accuracy values

[2] Disagreement weights were arranged according to the number of features shared by the ET categories: a disagreement between ACH and ACC, which only differ for the feature of [+/−RES], is not as bad as the one between ACH and STA, which differ for three features ([+/−DUR], [+/−DYN], [+/−RES]).

[3] WordNet top-nodes were used as semantic class labels.

Disagreement with the gold standard: An item-wise analysis showed that, despite our effort to select non-polysemous stimuli (e.g. *passeggiare*, "to stroll", ACT; *montare un gioco*, "to assemble a toy", ACC), the items upon which the participants agreed the least with the gold standard actually allowed for multiple ET interpretations. Consider the following examples from Experiments 1 and 2:

- *formare una fila*, "to form a queue", potentially ACH/STA ambiguous;
- *scegliere il disco*, "to choose the recorder", arguably unspecified for [+/−DUR] (ACC/ACH);
- *conceive the theory*, arguably unspecified for [+/−DUR] (ACC/ACH);
- *tumble*, ACH reading or ACT (iterative) reading;

Some lexical differences emerged between Italian VPs and their English correlate:

- *impiegare i materiali*, "to use the materials" was classified as ACC ($[+RES]$) in Italian, but as ACT ($[-RES]$) in English;
- *precipitare*, "to tumble", was classified as ACH by our Italian participants, but its English correlate seemed to have a more durative (iterative) ACT reading.
- the picture for *to crawl*, was correctly classified as an ACT by English speakers, but 8 out of 20 Italian speakers gave a STA answer; interestingly enough, the speaker of a language lacking of a compact verb for *to crawl* as Italian have also selected a stative reading for the picture;
- *precipitare*, "to tumble", is classified as ACH by our Italian participants, but its English correlate seems to have a more durative (iterative) ACT reading.

Agreement and accuracy were lower for Experiments 3 and 4, which used picture stimuli: a picture offers a sample of reality from which only some parts can be selected. For example, consider the pictures for *to bounce* and *to salute*, both of which showed low accuracy values both for speakers of Italian and English (< 3):

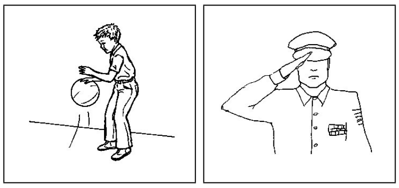

Figure 3: (*to bounce*), (*to salute*).

The picture for *to bounce* was originally labelled as ACH, but the participants interpreted it as an ACT (i.e. repeated acts of bouncing), the picture for *to stand* (STA) was interpreted as ACH (*to stand up*). Also in picture classification tasks ET classes seem far from being comparably evident to metalinguistic judgements.

feature set	distributional feature
adverbial	- temporal adverbs (e.g. in X time, for X time) - intentional adverbs (e.g. deliberately) - frequency adverbs (e.g. rarely, often) - iterative adverbs (e.g. X times)
morphological	- present tense - imperfect tense - future tense - simple past - perfect tenses - progressive periphrasis
syntactic and argument structure	- absence of arguments besides the subj. - presence of direct object, indirect obj. - presence of indirect obj. - presence of a locative argument - presence of a complement sentence - passive diatesis - number, animacy and definiteness of subj. and direct obj.

Table 4: Features

3 Corpus-based models of Event Types

Results from experiment 1 on speakers' metalinguistic judgements of ETs have been compared with the performance of computational models of ET classification trained with linguistically-motivated features extracted from Italian corpora: MaxEnt and SOM from Zarcone and Lenci (2008). MaxEnt is a supervised model which performs ET classification with Maximum Entropy classifiers, SOM is a self-organizing map which identifies ET clusters is an unsupervised way. See accuracy values in table 3[4].

3.1 Linguistically-motivated features

The features used to train the corpus-based models are very well-known in the linguistic literature for being (positively or negatively) correlated with particular event types (Dowty, 1979; Bertinetto, 1986; Pustejovsky, 1995; Rothstein, 2004). Extracted features include the following (see table 4 for a complete list):

adverbial features - they are among the main "event type" diagnostics in ET literature, but they are not very frequent in corpora data;

[4] Accuracy was higher for MaxEnt, but its coverage is limited to only 28 verbs; accuracy for SOM raised to 0.73 when lumping ACH and ACC into a single telic class.

morphological features - although actionality and aspect are independent categories, it is possible to observe typical correlations between some event types and specific aspectual values (Comrie, 1976). This set of features includes verb morphological tense-aspectual values;

syntactic and argument structure features - they include verb morphosyntactic, syntactic and semantic features of verb arguments, which are typically held responsible for ET shifts.

3.2 Corpus-based models vs. behavioral studies

A significant effect of ET on accuracy was yielded by a binomial logistic regression analysis for MaxEnt (binomial $p < 0.001$): significant differences in the pairwise comparisons clearly show a picture where [+RES] ETs (ACC and ACH) are easier to recognize than [−RES] ETs [5] - this seems to be the case also for Experiment 1.

No effect of ET on accuracy was found for SOM (binomial $p > 0.1$), but pairwise comparisons yielded a significant difference between ACC and STA accuracy ($z = -2.17; p < 0.05$). The distance between ACC and STA is comparable to the one found in Experiment 1 and 2: ACC are again the easiest to identify, STA the most difficult. SOM does not perform well on ACH, and this could be due to the sparseness of linguistic indicators for ACH (e.g. “in x time”, punctual temporal indications).

Results from MaxEnt seem to mirror the ones from Experiment 1, showing that [+RES] classes (ACC and ACH) are more prominent and more easily identifiable. Such difference seems to be purely linguistic, since it does not show in Experiment 3. The convergence between the metalinguistic study and the computational models is coherent with the idea that the characterization of ET as “linguistic objects” is strongly related with their corpus distribution. Not only can distributional data capture semantic classes such as ETs, but it seems also to be the case that ET classes which have a clearer distributional characterizations are also easier for the speaker’s to identify.

Similar comparisons between Experiment 2 and computational models trained on English corpora are ongoing.

4 Future experiments

We presented above-chance results from behavioral studies and corpus-based models in event type classification with pictures and lexical items for English and Italian. Materials for the corpus studies and the behavioral studies presented here are not homogeneous: the stimuli for the behavioral experiments were first selected to match criteria for on-line psycholinguistic studies, whereas the corpus-based models were trained with highly frequent verb items, in order to limit the sparseness of the distributional vectors. There is ongoing work to train corpus-based models with a state-of-the-art dependency corpus of Italian (Baroni et al., 2004; Bosco et al., 2009) and to evaluate them using the same dataset of the behavioral experiments presented here. As in Zarcone and Lenci (2008), the contribution of each feature set (adverbial, morphological, syntactic) will be evaluated by running different experiments with different feature sets.

Another battery of experiments is planned to test metalinguistic judgements on small video clips, which promise to be a useful tool in the investigation of event representations, and to better convey features like DUR or RES which are not easily delivered by a picture stimulus.

It has been suggested (Embodied Cognition Framework, Haggard et al. (2007)) that semantic representations are not purely amodal, but rather grounded in our sensorimotor perception. Cross-modal and intra-linguistic differences can provide useful insights to better grasp the very nature of ETs, and to better understand to what extent they are a purely linguistic phenomenon or to what extent they provide us with schemes to interpret reality.

References

F. Antinucci and R. Miller. 1976. How children talk about what happened. *Journal of Child Language*, 3:167–189.

Marco Baroni, Silvia Bernardini, Federica Comastri, Lorenzo Piccioni, Alessandra Volpi, Guy Aston, and Marco Mazzoleni. 2004. Introducing the la Repubblica corpus: a large, annotated, TEI(XML)-compliant corpus of newspaper italian.

[5] Significancies for pairwise comparisons yielded by the binomial logistic regression analysis: ACC > ACT; $z = -6.69, p < 0.001$; ACC > STA; $z = -5.66, p < 0.001$; ACH > ACT, $z = -8, p < 0.001$; ACH > STA, $z = -6.96, p < 0.001$

In Maria Teresa Lino, Maria Francisca Xavier, Ftima Ferreira, Rute Costa, and Raquel Silva, editors, *Proceedings of LREC 2004*, pages 1771–1774, Lisbona. ELDA.

E. Bates, K. Federmeier, D. Herron, and et al. 2000. Introducing the CRL International Picture-Naming Project (CRL-IPNP). *Center for Research in Language Newsletter*, 12.

P.M. Bertinetto. 1986. *Tempo, Aspetto e Azione nel verbo italiano. Il sistema dell'indicativo*. Accademia della Crusca, Firenze.

I. Bonnotte. 2008. The role of semantic features in verb processing. *Journal of Psycholinguistic Research*, 37:199–217.

Cristina Bosco, Simonetta Montemagni, Alessandro Mazzei, Vincenzo Lombardo, Felice DellOrletta, and Alessandro Lenci. 2009. Evalita '09 Parsing Task: comparing dependency parsers and treebanks. In *Proceedings of EVALITA 2009*, Reggio Emilia, Italy.

O. Bott. 2008. *The Processing of Events*. Ph.D. thesis, University of Tübingen. Ph.D. Dissertation.

Bernard Comrie. 1976. *Aspect: An Introduction to Verbal Aspect and Related Problems*. Cambridge University Press.

D. Dowty. 1979. *Word meaning and Montague Grammar. The semantics of verbs and times in generative semantics and in Montague's PTQ*. Reidel, Dordrecht.

C. Finocchiaro and G. Miceli. 2002. Verb actionality in aphasia: data from two aphasic subjects. *Folia linguistica*, 36:335–357.

S. Gennari and D. Poeppel. 2002. Events versus states: Empirical correlates of lexical classes. In *Proceedings of CogSci2002*.

P. Haggard, Y. Rossetti, and M. Kawato, editors. 2007. *Sensorimotor foundations of higher cognition*. Oxford University Press, Oxford, UK.

S. Im and J. Pustejovsky. 2010. Annotating Lexically Entailed Subevents for Textual Inference Tasks. In *Proceedings of FLAIRS 2010*.

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

P. Li and Y. Shirai. 2000. *The Acquisition of Lexical and Grammatical Aspect*. Mouton, New York.

V. Lucchesi. 1971. Fra grammatica e vocabolario. studio sull'aspetto del verbo italiano. *Studi di grammatica italiana*, 1:179–270.

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

S. Rothstein. 2004. *Structuring Events - A Study in the Semantics of Lexical Aspect*. Blackwell Publishing.

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*.

Z. Vendler. 1967. *Linguistics in Philosophy*. Cornell University Press, Ithaca, NY.

H. J. Verkuyl. 1972. *On the Compositional Nature of the Aspects*. D. Reidel Publishing Company, Dordrecht.

A. Zarcone and A. Lenci. 2008. Computational models of event type classification in context. In Andreas Witt, Felix Sasaki, Elke Teich, Nicoletta Calzolari, and Peter Wittenburg, editors, *Proceedings of LREC 2008*, Marrakesh. ELDA.

A. Zarcone and A. Lenci. 2010. Priming effects on event types classication: Effects of word and picture stimuli. Poster presented at CogSci 2010. Portland.