# Fitting, Not Clashing!
# A Distributional Semantic Model of Logical Metonymy

Alessandra Zarcone[1], Alessandro Lenci[2], Sebastian Padó[3], Jason Utt[1]
[1]Universität Stuttgart, [2]Universität Heidelberg, [3]Università di Pisa
[1]zarconaa,uttjn@ims.uni-stuttgart.de,
[2]alessandro.lenci@ling.unipi.it, [3]pado@cl.uni-heidelberg.de

**Abstract**

Logical metonymy interpretation (e.g. *begin the book* → *writing*) has received wide attention in linguistics. Experimental results have shown higher processing costs for metonymic conditions compared with non-metonymic ones (*read the book*). According to a widely held interpretation, it is the type clash between the event-selecting verb and the entity-denoting object (*begin the book*) that triggers coercion mechanisms and leads to additional processing effort. We propose an alternative explanation and argue that the extra processing effort is an effect of thematic fit. This is a more economical hypothesis that does not need to postulate a separate type clash mechanism: entity-denoting objects simply have a low fit as objects of event-selecting verbs. We test linguistic datasets from psycholinguistic experiments and find that a structured distributional model of thematic fit, which does not encode any explicit argument type information, is able to replicate all significant experimental findings. This result provides evidence for a graded account of coercion phenomena in which thematic fit accounts for both the trigger of the coercion and the retrieval of the covert event.

## 1 Introduction

**Type clash in logical metonymy.** Logical metonymy, also known as enriched composition (e.g. *The writer began the novel*), is generally explained in terms of a type clash between an event-selecting metonymic verb (*begin*) and an entity-denoting object (*novel*), triggering the recovery of a covert event (*writing*). Extensive psycholinguistic work (McElree, Traxler, Pickering, Seely, and Jackendoff (2001) and Traxler, Pickering, and McElree (2002), among others) has demonstrated extra processing costs for metonymic constructions. For example, Traxler et al. (2002) combine metonymic and non-metonymic verbs with entity-denoting and event-denoting nouns (*The boy [started/saw]$_V$ [the puzzle/fight]$_{NP}$*) and report significantly higher processing costs for the coercion combination (metonymic verb combined with entity-denoting object, e.g. *The boy started the puzzle*).

Building on this and similar experiments, Frisson and McElree (2008) ascribe the extra processing cost to "the deployment of operations to construct a semantic representation" of the event (*writing the novel*) that is supposed to be triggered by the type clash. However, this explanation remains problematic. Notably, metonymic interpretations are also possible for event-denoting objects given suitable contexts (e.g. *John is a wrestling fan, he really enjoyed the fight last night* → *watching the fight*).

**Thematic fit in logical metonymy.** Another pervasive aspect of language processing is thematic fit in the shape of selectional preferences, that is, expectations of predicative lemmas about plausible fillers for their argument slots (e.g., the fact that *eat* requires a *[+edible]* object or that *crook* is a more fitting object for *arrest* than *cop*). While early formalizations of selectional preferences aimed at modeling a binary distinction between "sensical" and "nonsensical" predicate-argument combinations, later work such as Wilks (1975) adopted a graded notion of selectional preference. In psycholinguistics, thematic fit has emerged as a pivotal concept to explain effects on expectations about upcoming input in language

comprehension (McRae, Spivey-Knowlton, and Tanenhaus 1998; Ferretti, McRae, and Hatherell 2001; Matsuki, Chow, Hare, Elman, Scheepers, and McRae 2011).

Concerning logical metonymy, there is considerable behavioral as well as modeling evidence that thematic fit plays an important role in metonymy *interpretation*, that is, the retrieval of covert events for metonymical constructions. Behavioral studies (Zarcone and Padó 2011; Zarcone, Padó, and Lenci 2012) as well as computational models (Lapata and Lascarides 2003; Zarcone, Utt, and Padó 2012) found that the retrieved event will be the event most compatible with our knowledge about typical events and their participants (as captured, e.g. by generalized event knowledge, (McRae and Matsuki 2009)), that is the interpretation with the highest thematic fit with the context. This is in contrast to traditional accounts of logical metonymy (Pustejovsky 1995) which ascribe covert event retrieval to complex lexical entries associating entities with events corresponding to their typical function or creation mode (qualia: *book* → *read / write*). The advantage of thematic fit-based accounts is that they can account for the influence of context on interpretation. For example, given *baker* and *child* as subjects of *finish the icing*, *baker* will cue *spread* as a covert event, while *child* will cue *eat*, even though it is possible that bakers eat icing or that children spread it.

**Thematic fit as a trigger of logical metonymy.** In this paper, we propose that thematic fit can explain not only the interpretation phase of metonymy (that is implicit event recovery), but also to the *triggering phase*. We claim that thematic fit can provide a convincing explanation for the triggering of the coercion operation: metonymic verbs prefer event-denoting objects, and sentences that have traditionally been analysed involving a coercion operation have a low thematic fit between the verb and the object. This account preserves the observation that metonymic verbs disprefer entity-denoting objects, but can explain it purely in terms of standard graded accounts of selectional preferences[1]. A thematic-fit based account of logical metonymy would bring a clear advantage of theoretical economy: it accounts for two phenomena (triggering and interpretation) with a single mechanism, that is, generalized event knowledge (quantified in terms of thematic fit). Furthermore, generalized event knowledge / thematic fit operate in any type of predicate-argument composition (McRae and Matsuki 2009). Thus, an explanation in terms of thematic fit would bring metonymy closer to "normal" online language comprehension process.

We test this hypothesis by modeling three datasets from well-known experimental studies on metonymy. We compute thematic fit predictions of all items relying solely on distributional information, without any information about predicate semantic types, and compare differences in thematic fit across conditions with corresponding differences in processing cost from the experiments. As we show below, our distributional model of thematic fit predicts all significant effects that have been found in the experiments previous experiments and that were interpreted as type-clash effects.

# 2 Experimental Setup

## 2.1 A Distributional Model of Thematic Fit

Distributional semantic models (Turney and Pantel 2010) build on the Distributional Hypothesis (Harris 1954; Miller and Charles 1991) which states that the meaning of a word can be modelled by observing the contexts in which it is used. Current distributional models are typically built by collecting contexts of word occurrences in large corpora, where "context" can be defined in many possible ways. Pairwise word similarity is then computed by comparing the similarity between the vectors which record the word co-occurrences in the data. Distributional models have been successful in modelling a range of cognitive tasks, including lexical development (Li, Farkas, and MacWhinney 2004), category-related deficits (Vigliocco, Vinson, Lewis, and Garrett 2004), and thematic fit (Erk, Padó, and Padó 2010).

Distributional Memory (DM, Baroni and Lenci (2010)) is a general framework for building distributional semantic models from syntactically analysed corpora. It constructs a three-dimensional tensor of

[1] Similarly, thematic-fit based accounts of selectional preferences encompass binary distinctions (e.g., *eat* requires a *[+edible]* object), while still including more fine-grained differences (e.g., *crook* is a more fitting object for *arrest* than *cop*).

weighted word-relation-word tuples each tuple is mapped onto a score by a function $\sigma\colon \langle w_1\ r\ w_2\rangle \rightarrow \mathbb{N}$, where $w_2$ is a target word, $r$ as a relation and $w_1$ an argument or adjuct of the target word. For example, $\langle$*marine subj shoot*$\rangle$ has a higher weight than $\langle$*teacher subj shoot*$\rangle$. The set of relations can be defined in different ways, which gives rise to different flavors of DM. We use TypeDM, which uses generic syntactic relations as well as lexicalized relations (see Baroni and Lenci (2010) for details).

For our experiments, we project the DM tensor onto a $W_1 \times RW_2$ matrix, and we represent each target word $W_1$ in terms of a vector with dimensions corresponding to pairs of context words and their relations ($R \times W_2$). On this matrix, we compute a verb's expectations for its most typical object with a method similar to Erk et al. (2010) and Lenci (2011): For each verb $v$, we determine the 20 highest-scoring nouns in object relation and compute the centroid $c_o(v)$ of their context vectors. The thematic fit of a new noun $n$ for $v$'s object position is then defined as the cosine of the angle between $n$'s own context vector and $c_o(v)$. Since all the vectors' components are positive, the thematic fit values range between 0 and 1.

## 2.2 Datasets

As stated above, we model three datasets from psycholinguistic experiments. The datasets fall into two categories: sentence triplets, and sentence quadruplets. Please refer to the corresponding psycholinguistic studies (McElree et al. 2001; Traxler et al. 2002) for further details about how the datasets were built.

**Sentence Triplets** The two datasets in this group consist of sentence triplets: one sentence with a metonymic verb ("type-shift" in the original papers), paired with two non-metonymic sentences, where one condition shows high thematic fit and one low thematic fit ("preferred" and "non-preferred" in the original papers). An example is *the writer [finished / wrote / read]$_V$ the novel*.

**McElree dataset.** This dataset is composed of 31 triplets of sentences from the self-paced reading experiment in McElree et al. (2001), for a total of 93 sentences. We excluded two triplets from the original dataset for problems of coverage, as they included low-frequency words.

**Traxler dataset 1.** This dataset is composed of 35 triplets of sentences from the eye-tracking experiment (experiment 1) in Traxler et al. (2002), for a total of 105 sentences. We excluded one triplet from the original dataset for problems of coverage, as it included low-frequency words.

The finding for these materials was a main effect of verb type on reading times (McElree et al. 2001) and eye tracking times (Traxler et al. 2002). Pairwise comparisons in both studies yielded (a) higher processing costs for the metonymic condition and (b) no significant differences between the high- and low-typicality condition.

**Sentence Quadruplets** The dataset in this group consists of sentence quadruplets which cross two factors: (i) metonymic verb vs. non-metonymic verb and (ii) event-denoting object vs. entity-denoting object. As an example, consider *The boy [started / saw]$_V$ [the fight / puzzle]$_{NP}$*.

**Traxler dataset 2.** This dataset is composed of 32 sentence quadruplets from experiments 2 (eye-tracking) and 3 (self-paced reading) in Traxler et al. (2002), for a total of 120 sentences. We exclude one triplet from the original dataset for problems of coverage.

The findings for this dataset were a main effect of object type on eye tracking times (experiment 2) and on reading times (and experiment 3), and a significant verb$*$object interaction, with higher processing costs for the metonymic condition (metonymic verb combined with entity-denoting object).

## 2.3 Evaluation Method

For each test sentence, we compute the verb-object thematic fit in DM. We assume that processing load increases with lower thematic fit and that processing cost (reading time and eye tracking time) corresponds to $1 -$ thematic fit. We manipulate thematic fit in DM as a dependent variable, and we employ linear

| | | high-typicality | low-typicality | metonymic |
|---|---|---|---|---|
| triplets from McElree et al. (2001) | reading times at the obj. + 1 position | 360 | 361 | 385 |
| | $1 -$ thematic fit | 0.484 | 0.571 | 0.763 |
| triplets from Traxler et al. (2002) | eye tracking (total time) | 397 | 405 | 444 |
| | $1 -$ thematic fit | 0.482 | 0.576 | 0.744 |

Table 1: Comparing behavioral data from McElree et al. (2001) and Traxler et al. (2002) and thematic fit data from the computational model

regression analyses to test for main effects of factors (object type, verb type) on the dependent variable, as well as Wilcoxon rank sum task to test the significance of pairwise differences between conditions in terms of thematic fit. We then verify if the same main effects and significant pairwise differences were yielded by the psycholinguistic models and by the computational model.

# 3 Results

**Sentence Triplets** On the McElree dataset, the model mirrors all effects reported by the experimental studies, namely by yielding a main effect of the object type ($F = 20.247, p < 0.001$), and significant differences between the metonymic condition and both high-typicality ($W = 877, p < 0.001$) and low-typicality ($W = 740, p < 0.001$) conditions, but no significant difference between the high- and low-typicality conditions ($W = 595, p > 0.5$).

The model also mirrors the experimental effects found for the Traxler dataset 1. It yields a main effect of the object type ($F = 18.084, p < 0.001$), and significant differences between the metonymic condition and both high-typicality ($W = 1050, p < 0.001$) and low-typicality ($W = 889, p < 0.01$) conditions, but only a marginal difference between the high- and low-typicality conditions ($W = 780.5, p = 0.5$).

**Sentence Quadruplets** On the Traxler dataset 2, the model mirrors the main effect of the object type ($F = 8.0039, p < 0.01$) and the verb$*$object type interaction ($F = 8.3455, p < 0.01$) reported both by the self-paced reading and by the eye tracking studies. This result is visualized in Figure 1, which shows the close correspondence between experimental results from self-paced reading and modeling results.

The model also yields the same pair-wise differences reported by the eye-tracking study: within the sentences with entity-denoting objects, metonymic verbs yield lower thematic fit compared to non-metonymic sentences ($W = 300, p < 0.05$). Within the sentences with metonymic verbs, entity-denoting objects also yield lower thematic fit compared to entity-denoting objects ($W = 208, p < 0.01$).

# 4 Discussion and Conclusions

The distributional models successfully replicate the pattern of results of the psycholinguistic experiments. For the triplet dataset, the model yields a main effect of verb type, and produces the lowest thematic fit for the metonymic condition (corresponding to the longest reading times and eye fixations in the experiments); also, the significant differences detected by the model are the same as those in the experimental studies: high-typicality vs. metonymic and low-typicality vs. metonymic. Interestingly enough, while the computational model does not reveal a significant difference between what we called high- and low-typicality conditions (preferred and non-preferred conditions according to the paper's terminology), the psycholinguistic experiments were not able to show one either, suggesting that in the experimental materials that were employed, the preferred and non-preferred conditions do not differ significantly with regard to typicality.

For the quadruplet dataset, the model yields a main effect of object type and a significant verb$*$object interaction, producing the lowest thematic fit for metonymic verbs combined with entity-denoting objects.

| | | metonymic verb | | non-metonymic verb | |
|---|---|---|---|---|---|
| | | EN obj. | EV obj. | EN obj. | EV obj. |
| quadruplets from Traxler et al. (2002) | self-paced reading times at obj. + 1 | 512 | 427 | 467 | 455 |
| | 1 − thematic fit | 0.230 | 0.336 | 0.283 | 0.287 |

Table 2: Comparing behavioral data (self-paced reading) from Traxler et al. (2002) and thematic fit data from the computational model. Eye-tracking data mirror the findings of the self-paced reading study.

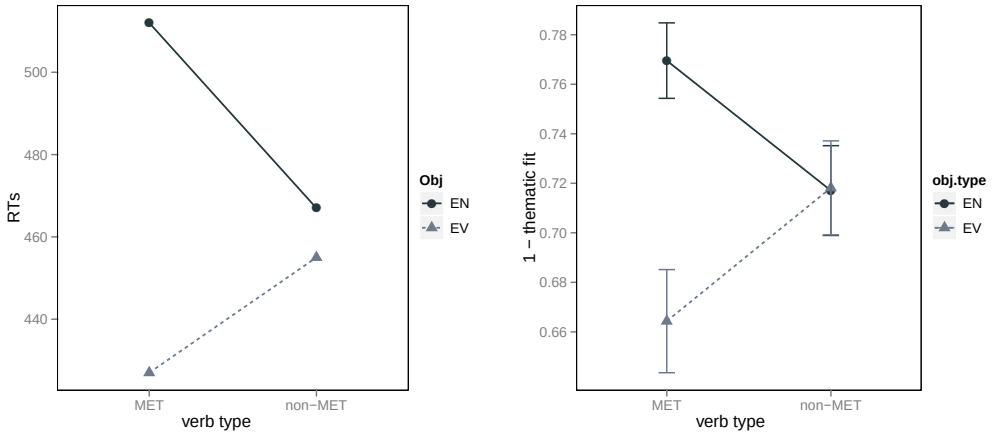


Figure 1: Comparing reading times in Traxler et al. (2002) (left, experiment 3) and results from thematic fit from our model (right).

The significant differences detected by the model were the same as those in the experimental studies: for the entity-denoting objects, metonymic verbs yielded lower thematic fit than non-metonymic verbs, whereas this difference was not significant for event-denoting objects; for the metonymic verbs, entity-denoting objects yielded lower thematic fit than event-denoting objects, whereas this difference was not significant for non-metonymic verbs.

The model's success in replicating the results from the psycholinguistic experiments shows that distributional similarity-based estimates of thematic fit are sufficient to account for the behavioral correlates of logical metonymy processing (such as longer reading and eye-tracking times) found so far in cognitive experiments. More precisely, the model, which did not encode any explicit type information, was able to replicate all significant experimental findings which were formulated based on type distinctions. This result supports the following general claims that we are currently testing on further experiments: i.) the two phenomena of triggering and interpretation in logical metonymy can be explained with a single mechanism relying on general event knowledge activation and integration; ii.) metonymy is actually much more closer to "normal" online predicate-argument composition processes, both being based on thematic fit computation. The structured distributional model of the latter proves to be an interesting tool to critically reanalyze psycholinguistic datasets, by (a) highlighting possible implicit lexical-semantic biases which may influence the results and (b) providing alternative explanations for them.

**Acknowledgements** The research for this paper was funded by the German Research Foundation (Deutsche Forschungsgemeinschaft) as part of the SFB 732 "Incremental specification in context" / project D6 "Lexical-semantic factors in event interpretation" at the University of Stuttgart.

## References


Baroni, M. and A. Lenci (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics 36*(4), 1–49.

Erk, K., S. Padó, and U. Padó (2010). A flexible, corpus-driven model of regular and inverse selectional preferences. *Computational Linguistics 36*(4), 723–763.

Ferretti, T. R., K. McRae, and A. Hatherell (2001). Integrating verbs, situation schemas and thematic role concept. *Journal of Memory and Language 44*, 516–547.

Frisson, S. and B. McElree (2008). Complement coercion is not modulated by competition: Evidence from eye movements. *Journal of Exp. Psychology: Learning, Memory, and Cognition 34*(1), 1.

Harris, Z. S. (1954). Distributional structure. *Word 10*(23), 146–162.

Lapata, M. and A. Lascarides (2003). A probabilistic account of logical metonymy. *Computational Linguistics 29*(2), 263–317.

Lenci, A. (2011). Composing and updating verb argument expectations: A distributional semantic model. In *Proceedings of CMCL*, Portland, Oregon, pp. 58–66.

Li, P., I. Farkas, and B. MacWhinney (2004). Early lexical development in a self-organizing neural network. *Neural Networks 17*, 1345–1362.

Matsuki, K., T. Chow, M. Hare, J. L. Elman, C. Scheepers, and K. McRae (2011). Event-based plausibility immediately influences on-line language comprehension. *Journal of Exp. Psychology: Language, Memory, and Cognition 37*(4), 913–934.

McElree, B., M. Traxler, M. Pickering, R. Seely, and R. Jackendoff (2001). Reading time evidence for enriched composition. *Cognition 78*(1), B17–B25.

McRae, K. and K. Matsuki (2009). People use their knowledge of common events to understand language, and do so as quickly as possible. *Language and Linguistics Compass 3/6*, 1417–1429.

McRae, K., M. Spivey-Knowlton, and M. Tanenhaus (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language 38*(3), 283–312.

Miller, G. A. and W. G. Charles (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes 6*(1), 1–28.

Pustejovsky, J. (1995). *The Generative Lexicon*. MIT Press.

Traxler, M. J., 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.

Turney, P. D. and P. Pantel (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research 37*, 141–188.

Vigliocco, G., D. P. Vinson, W. Lewis, and M. F. Garrett (2004). Representing the meanings of object and action words: The featural and unitary semantic space hypothesis. *Cognitive Psychology 48*(4), 422–488.

Wilks, Y. (1975). A preferential, pattern-seeking, semantics for natural language inference. *Artificial Intelligence 6*(1), 53–74.

Zarcone, A. and S. Padó (2011). Generalized event knowledge in logical metonymy resolution. In *Proceedings of CogSci*, Austin, TX, pp. 944–949.

Zarcone, A., S. Padó, and A. Lenci (2012). Inferring covert events in logical metonymies: a probe recognition experiment. In *Proceedings of CogSci*, Austin, TX.

Zarcone, A., J. Utt, and S. Padó (2012). Modeling covert event retrieval in logical metonymy: Probabilistic and distributional accounts. In *Proceedings of CMCL*, Montreal, Canada.