Event Knowledge and Models of Logical Metonymy Interpretation

Von der Fakultät Informatik, Elektrotechnik und Informationstechnik der Universität Stuttgart zur Erlangung der Würde einer Doktorin der Philosophie (Dr. phil.) genehmigte Abhandlung

Vorgelegt von

Alessandra Zarcone

aus Palermo, Italien

Hauptberichter:	Prof. Dr. Sebastian Padó
1. Mitberichter:	Prof. Dr. Ken McRae
2. Mitberichter:	Prof. Dr. Jonas Kuhn

Tag der mündlichen Prüfung: 9. Mai 2014

Institut für Maschinelle Sprachverarbeitung der Universität Stuttgart

2014

Hiermit erkläre ich, dass ich unter Verwendung der im Literaturverzeichnis aufgeführten Quellen und unter fachlicher Betreuung diese Dissertation selbständig verfasst habe.

(Alessandra Zarcone)

The TV scientist who mutters sadly, “The experiment is a failure; we have failed to achieve what we had hoped for,” is suffering mainly from a bad script-writer. An experiment is never a failure solely because it fails to achieve predicted results. An experiment is a failure only when it also fails adequately to test the hypothesis in question, when the data it produces don’t prove anything one way or another.

Skill at this point consists of using experiments that test only the hypothesis in question, nothing less, nothing more. If the horn honks, and the mechanic concludes that the whole electrical system is working, he is in deep trouble. He has reached an illogical conclusion. The honking horn only tells him that the battery and horn are working. To design an experiment properly he has to think very rigidly in terms of what directly causes what.

R. Pirsig, *Zen and the Art of Motorcycle Maintenance.*

I like work: it fascinates me. I can sit and look at it for hours.

J. K. Jerome, *Three men in a boat.*

Abstract

During language understanding, people do not only rely on what they read or hear, but they also exploit implicit information. For example, when they process the expression *begin the book*, they understand it involves an event which is not explicitly mentioned (e.g. *begin reading the book*). This thesis looks at these constructions, known as logical metonymies, which combine an event-selecting verb and entity-denoting object and involve covert events. Logical metonymies are an interesting challenge for theories of lexical semantics: they need to be reconciled with compositionality, they require the integration of context (writers typically *write books*, students typically *read* them), and they lie at the interface between lexicon and world knowledge (is the information that *books are read* stored in our mental lexicon or in our world knowledge?).

I critically analyze previous hypotheses on logical metonymy with regard to the answer they provide to two core problems: the *source problem* (what events are retrieved? what type of event knowledge is assumed?) and the *trigger problem* (why do some constructions trigger a metonymic interpretation and others do not?). Lexicalist approaches claim that the metonymy arises from a type clash between the event-selecting verb and an entity-denoting object, and posit complex lexical items, encoding event information about artifacts (e.g. *book* → *read*), to explain the recovery of covert events. Pragmatic-based approaches argue against the idea that lexical items have an internal structure, suggesting that covert events arise from the underspecification of a logical metonymy and are inferred via non-lexical knowledge. I look with particular attention at the role of event knowledge, which lexicalist approaches place in our mental lexicon, while pragmatic-based approaches place it in our world knowledge.

I propose a third hypothesis, based on thematic fit and generalized event knowledge of typical events and their participants, which have been shown to guide efficient incremental processing: I argue that contextual elements cue generalized event knowledge, which plays a key role in determining the covert event for a logical metonymy. I explore

this hypothesis from an interdisciplinary perspective, employing both psycholinguistic experiments and computational models, in order to seek converging evidence and confront it with the theoretical investigation. The results from the psycholinguistic experiments and from the computational (distributional) models support the hypothesis that covert event retrieval is guided by generalized event knowledge. I also employ the computational models to analyze previous experimental results and to explore the hypothesis that thematic fit, informed by generalized event knowledge, is ultimately responsible for the trigger of the logical metonymy. I then report on more psycholinguistic evidence showing that a notion of type is indeed necessary to account for differences between metonymic and non-metonymic constructions, and that both type and thematic fit play a role in logical metonymy interpretation. Lastly, I argue for a context-sensitive model of logical metonymy interpretation that exploits an information-rich lexicon, but needs to rethink the notion of type and reconcile it with the notion of thematic fit.

Zusammenfassung

Zum Sprachverständnis nützen Menschen nicht nur die Bestandteile der Eingabe, die sie explizit lesen oder hören, sondern auch implizite Informationen. Wenn man beispielsweise eine Äußerung wie *das Buch beginnen* verarbeitet, versteht man, dass diese Äußerung ein Ereignis evoziert, das nicht explizit verbalisiert wird (z.B. *das Buch zu lesen beginnen*). Diese Dissertation beschäftigt sich mit solchen Konstruktionen, die logische Metonymien genannt werden. Diese Konstruktionen kombinieren ein Ereignis-selegierendes Verb mit einem Objekt, das eine Entität beschreibt, und evozieren implizite Ereignisse. Logische Metonymien sind für Theorien der lexikalischen Semantik besonders interessante Herausforderungen: sie müssen mit der Kompositionalität von Sprache vereinbart werden, sie verlangen Integration von vorangegangenem Kontext (z.B. Schriftsteller *schreiben* typischerweise *Bücher*, während Studenten sie typischerweise *lesen*), und sie sind ein Phänomen an der Schnittstelle von Lexikon und Weltwissen (ist die Information, dass *man Bücher liest*, in unserem mentalen Lexikon oder in unserem Weltwissen gespeichert?).

Ich führe eine kritische Analyse der existierenden Ansätze zur logischen Metonymie durch, insbesondere im Hinblick auf die Antworten zu zwei primären Problemen: das *Quelle-Problem* (was für Ereignisse werden abgerufen? was für Ereigniswissen wird verwendet?) und das *Auslöser-Problem* (warum evozieren manche Konstruktionen implizite Ereignisse, andere aber nicht?). Lexikalistische Ansätze besagen, dass die Metonymie aus einem Typ-Konflikt zwischen dem Ereignis-selegierenden Verb und dem Entitäts-beschreibenden Objekt entsteht. Um den Abruf impliziter Ereignisse erklären zu können, postulieren diese Ansätze komplexe lexikalische Einheiten, die Ereignis-Informationen über Artefakte (z.B. *Buch* → *lesen*) kodieren. Pragmatik-basierte Ansätze argumentieren gegen diese Vorstellung, dass lexikalische Einheiten eine innere Struktur haben, und postulieren, dass implizite Ereignisse aus der Unterspezifikation einer logischen Metonymie entstehen und durch nicht-lexikalisches Wissen determiniert werden. Zusammengefasst besteht der Unterschied zwischen

lexikalistischen und pragmatischen Ansätzen also in den Annahmen über die Rolle von Ereigniswissen: Lexikalistische Ansätze verorten es in unserem mentalen Lexikon, Pragmatik-basierte Ansätze dagegen in unserem Weltwissen.

Ich vertrete eine dritte Hypothese, die auf generalisiertem Wissen über Ereignisse und ihre Mitspieler basiert („generalized event knowledge"). Solches Wissen, das als die Typikalität eines Arguments für eine thematische Rolle („thematic fit") berechenbar ist, ist bereits als entscheidend für eine effiziente inkrementelle Sprachverarbeitung bekannt. Die zugrundeliegende Hypothese besagt, dass kontextuell gegebene Elemente generalisiertes Ereigniswissen aktivieren, welches eine zentrale Rolle in der Bestimmung eines impliziten Ereignisses für eine logische Metonymie spielt. Die Hypothese wird aus einer interdisziplinären Perspektive untersucht - sowohl durch psycholinguistische Studien als auch durch Computermodelle - um konvergente Evidenz zu erhalten und diese mit theoretischen Untersuchungen zu vergleichen. Die Ergebnisse der psycholinguistischen Studien und der distributionellen Computermodelle unterstützen die Hypothese, dass der Abruf der impliziten Ereignisse durch generalisiertes Ereigniswissen determiniert ist. Die Computermodelle kommen auch zum Einsatz, um vorangegangene experimentelle Ereignisse zu analysieren und um die Hypothese zu untersuchen, dass „thematic fit", geprägt durch generalisiertes Ereigniswissen, letztlich auch für den Auslöser der Metonymie verantwortlich ist. Anschließend berichte ich über weitere psycholinguistische Evidenz, dass das Konzept von Typ dennoch benötigt wird, um zwischen metonymischen und nicht-metonymischen Konstruktionen zu unterscheiden, und dass sowohl Typ als auch „thematic fit" eine zentrale Rolle bei der Interpretation der logischen Metonymie spielen. Zuletzt argumentiere ich für ein kontextabhängiges Modell der Interpretation logischer Metonymien, das sich auf ein informationsreiches Lexikon stützt, aber das auch erfordert, das Konzept für Typ neu zu durchdenken und mit dem Konzept von „thematic fit" vereinbar zu machen.

Acknowledgements

All professors were once students, but only few of them remember it: I would like to thank my supervisor Sebastian Padó, who certainly has not forgotten how it feels to be a student. His understanding of the issues that a grad student encounters has been precious and has helped me feel more confident on my journey to become a mature researcher. I am particularly grateful for having been hired to work on the D6 project and for having experienced a perfect balance of guidance and freedom during these years.

I would also like to thank Ken McRae and Jonas Kuhn for agreeing to review my dissertation and for their help and advice. Ken's methodological soundness, thought-provoking ideas and clear and engaging presentation style have been a great source of inspiration. I am thankful to Jonas and to Sabine Schulte im Walde for always making sure I had a room to run my experiments. I am also grateful to Sabine for taking over project D6 and thus allowing the project to go on for the whole funding period and for always providing interesting feedback and precious encouragement.

My work as a PhD student has been funded by the project D6 within the SFB 732 "Incremental Specification in Context". I am thankful to the Deutsche Forschungsgemeinschaft for funding the project, and to the Scuola Normale Superiore funding my undergrad years. Both institutions have established in me the deep conviction that free and state-founded education is not only useful but necessary. I would like to thank the SFB for funding my conference trips and for allowing me to invite several guest speakers. Talking to them (in particular to Berry Claus, Francesca Delogu, Kazunaga Matsuki and Petra Schumacher) has been extremely fruitful for the work presented in this dissertation. I am grateful to Hans Kamp, Alessandro Lenci and James Pustejovsky for inviting me to the Dagstuhl Seminar 13462 ("Computational Models of Language Meaning in Context"), allowing me to take part in an exciting and thought-provoking week. I would like to thank Alessandro as well as Pier Marco Bertinetto, for sparking my

interest in linguistics during my freshman year and for throwing me (metaphorically speaking) in deep waters several times to see if I could swim.

During my years at the IMS I have profited from a fantastic work environment, which is an enormous privilege. I would like to thank Jason Utt for the productive cooperation and for being a great office mate, Nadja Schauffler and Michael Walsh for helping me finding participants for my experiments and everybody else at the IMS (in particular Florian Laws, Lukas Michelbacher, Özlem Çetinoğlu, Gabriella Lapesa, Kyle Richardson, Christian Scheible, Anders Björkelund, Stefan Roller, Charles Jochim, Wolfgang Seeker and Antje and Katrin Schweitzer) and my conference buddies Diego Frassinelli, Moreno Coco and Nick Gaylord for being such awesome colleagues and friends (you would not imagine how profitable an exchange can stem from just posting a research question on Facebook — I surely had not). I would also like to thank all the students who took part in the study, for their availability and enthusiasm.

I would also like to thank my family, for supporting me and for teaching me the passion and the respect for scientific research, and Mike, for proofreading this dissertation, for patiently adapting to my variable work schedule, for supporting me with love (and coffee) in the worst moments of the PhD (late night deadlines, journal rejections, the final stages of the dissertation), for helping me disconnect from work when I needed a break, and for being the spouse every grad student would wish to have by her side.

to Irene, a little woman who at the age of 3 and a half
already doesn't let anyone or anything get in her way

Contents

Acronyms and Abbreviations

AMT	Amazon Mechanical Turk
ANN	Annotator
ANOVA	Analysis of Variance
AQ	Agentive Quale
B	Baseline
CE	Covert Event
DM	Distributional Memory
DSM	Distributional Semantic Model
ECU	Expectation Composition and Update
EN	entity
EV	event
GL	Generative Lexicon
INSCTXT	Insufficient context
ISI	Inter-Stimulus Interval
L	Link
LM	Logical Metonymy
LF	Long Form
LMI	Local Mutual Information
LSA	Latent Semantic Analysis
M	Mean
meton.	metonymic
N	Noun
NLP	Natural Language Processing
NP	Noun Phrase
Obj	Object
RT	Reaction Time
SD	Standard Deviation
SO	Subject-Object
SOV	Subject-Object-Verb
Subj	Subject
TQ	Telic Quale
UNDET	Undeterminate qualia structure
V	Verb
VP	Verb Phrase
W	Word

List of Tables

List of Figures

Part I.

Introduction: Problems, Methods and Models

1. Introduction

Language understanding requires not only processing sequences of words that we hear or read, but also interpreting implicit information. An interesting example is logical metonymy, which is the subject of this dissertation: when we understand the meaning of *begin the book*, we integrate an event (e.g. *reading* or *writing*) which is not explicitly mentioned. In this chapter I will delimit the field of my research by defining what logical metonymy is, why it is an interesting phenomenon to study for models of language understanding, what its most problematic aspects are, and what other (different but related) phenomena are often associated with it.

1.1. Covert Events in Logical Metonymy

As long as people have reflected upon language, they have realized that, when we listen or read language, we combine sequences of words in a meaningful way. The Principle of Compositionality, formulated by Frege (1892)[1], states that the meaning of a complex utterance can be decomposed into the meaning of its parts. However, the Principle of Compositionality only takes explicit parts of the utterance into account: for example, the meaning of *The philosopher sits* is a function of the meaning of *philosopher*, combined with the meaning of the verb phrase *sits*. Our understanding, though, goes beyond what is merely said or written and often relies on what is unsaid: we integrate implicit elements into our interpretation. For example, if we hear *The woman stirred her coffee*, we understand that she probably used a *spoon* for stirring (Ferretti et al., 2001; Matsuki et al., 2011).

Likewise, if we hear that a writer *began a book*, we understand that he probably began *writing* it:

[1] The foundations of the principle can be found in the work of several ancient philosophers (e.g. Plato and Al-Farabi, see Hodges, 2012 and Werning et al., 2012).

(1.1) a. Jack Kerouac **began the book** around 1949 in New York.
→ writing the book

b. Jack Kerouac **began writing the book** around 1949 in New York.

1.1.b is a very plausible interpretation of 1.1.a: in 1.1.a, the verb (*begin*) would require an event-denoting complement (something you can begin, for example *writing the book*) but is instead combined with an entity-denoting object[2] (*the book*) and involves the interpretation of an implicit event (**covert event**).

An extra event is not required when the object of *begin* is itself an event-denoting object:

(1.2) Jack Kerouac **began the journey** that would take him back and forth across America.

Constructions like 1.1.a are known as instances of **logical metonymy** (Pustejovsky, 1991, 1995). The term **metonymy** (Stern, 1931; Nunberg, 1979) is typically used to indicate a (part-for-the-whole) transfer of meaning (*the book* → *writing the book*), and in this case it indicates that the semantic type selected by the verb (the event: *writing the book*) is denoted by a subpart of it (the object: *the book*). **Logical** refers to the fact that the verb's syntactic argument is not the same logical argument in the semantic relation (Pustejovsky, 1995, p. 54); being a regular alternation phenomenon, logical metonymy is seen as a particular case of regular polysemy (Weinreich, 1966; Apresjan, 1974; Nunberg, 1978, 1979; Pustejovsky, 1991, 1995), because it arises systematically from the combination of an event-selecting verb and its arguments (specifically, when the direct object denotes a non-event). I will refer to constructions like 1.1.b, which can be considered paraphrases of logical metonymies where the event is explicit, as **long forms**.

Different covert events can be retrieved, depending on intra-sentential context and discourse context (Fodor and Lepore, 1998; Lascarides and Copestake, 1998; Markert and Hahn, 2002; de Almeida and Dwivedi, 2008; Asher, 2011). For example,

[2] I follow here the broad distinction between "events" and "objects" (Casati and Varzi, 2010) exemplified by the WordNet ontology (Fellbaum, 1998), and refer with "entity" to the ontological class of "object" as opposed to "event". I will question the clear-cut distinction between (arguably metonymic) constructions combining event-selecting verbs with entity-denoting objects and (arguably non-metonymic) constructions combining them with event-denoting objects in the course of this dissertation.

the combination *begin* + *book* gives rise to two different interpretations in 1.1 and in 1.3, simply because writers typically *write* books and book lovers typically *buy* and *read* them:

(1.3) I found *On the Road* in a second-hand bookstore. I **began the book** as soon as I got home.
→ reading the book

The information required to understand the covert events (for example, basic information about books being *written* and *read*) may be included in our model of the lexicon (Pustejovsky, 1991, 1995), but at least in some cases (1.4) world knowledge is required to interpret the metonymy (Lascarides and Copestake, 1998; Asher, 2011):

(1.4) The goat **began Jack Kerouac's book**.
→ eating the book
(example adapted from Lascarides and Copestake, 1998)

As we know that goats do not read, we understand that the goat probably began *eating* the book.

1.2. Logical Metonymy: the Problematic Aspects

Logical metonymy constitutes an interesting challenge for theories of meaning, for several reasons:

1. logical metonymy challenges compositionality (Partee et al., 1990; Baggio et al., 2012), in regard to both its structural aspect and its content aspect: its structure shows a mismatch between the overt syntactic structure of the metonymy (VP → V NP) and its interpretation (where a complement verb — the covert event — is needed); its interpretation requires the processing of some implicit content (the covert event) which is not overtly realized and which can not be retrieved just by lexical lookup, but requires at least some degree of world knowledge integration;

2. logical metonymy touches on the problem of polysemy and regular sense alternation:

a) event-selecting verbs may be considered polysemous because they allow for a "multiple number of syntactic and semantic contexts, such as Verb Phrase, Gerundive Phrase, or Noun Phrase" (Pustejovsky, 1995, pp. 32-33), with regular meaning alternations depending on the complement they select;

b) logical metonymy systematically arises in cases of combination between event-selecting verbs and entity-denoting objects, thus showing at least some degree of regularity;

c) logical metonymies are underspecified with regard to their covert event interpretation, because multiple covert event interpretations are often possible (one covert event interpretation is then selected depending on context);

3. logical metonymy lies at the interface between semantics and syntax (and world knowledge), as the covert event interpretation involves lexical semantic information (e.g., for *the author began the book*, the information that *books are written*) as well as world knowledge information (e.g., for *the author began the book*, the information that *goats don't read books*).

A number of approaches, in linguistics, psycho/neurolinguistics and computational linguistics, have addressed logical metonymy, identifying at least two main problems and providing different explanations for them. These challenges can be summarized into two questions:

- **the trigger problem:** *what triggers the logical metonymy?*
  Logical metonymies have been defined as a combination of an event-selecting verb and an entity-denoting object, which gives rise to a type clash and therefore requires the interpretation of a covert event. But is this a satisfactory definition of the trigger of the metonymic interpretation? I will show that this definition may be problematic.

- **the source problem:** *what is the source of the covert event?*
  The label "metonymy" is given to the meaning transfer from an entity to an event, but given that covert events are implicit, what is the range of the retrieved events? Where is the event information stored? What cognitive resources are

involved in the event retrieval? Theories of logical metonymy differ greatly with regard to the status of event knowledge (in the lexicon or outside the lexicon) and with regard to how this information is retrieved.

1.3. Related Phenomena

For the purposes of this thesis, in order to better restrict the scope of my analysis, I have only focused on logical metonymy stricto sensu, that is verb [+ preposition] + object constructions where a covert event is implied and which can be described as type shifts (entity → event), thus distinguishing logical metonymy from other (although related) phenomena. I prefer using the term **logical metonymy**, as this (rather theory-neutral) term underlines two rather uncontroversial key aspects of the phenomenon: its regularity (stemming from a type incongruence: *logical*) and the part-for-the-whole relationship between the direct object and the (transitive) covert event (*metonymy*).

Other terms have been used to refer to this phenomenon, which denote a broader set of phenomena or sub-processes of logical metonymy interpretation, or are more theory-loaded:

Metonymy: this broader term also includes entity-to-entity metonymies which do not coerce an entity into an event, for example:

(1.5) As it became fashionable to be beat, it became less fashionable to **read Jack Kerouac**.
→ read the books of Jack Kerouac [author-to-book metonymy]

Enriched composition: Jackendoff (1997) used this term to refer to the integration of extra implicit material (the covert event) that is not overtly realized, but "must be present in [the] conceptual structure" (Jackendoff, 1997, p. 49), in contrast with simple compositional expressions whose meaning results directly from the meaning of their parts. Jackendoff's account preserves the compositionality of logical metonymies and groups them together with similar (regular) phenomena such as telic adjectives (Pustejovsky, 1991; Jackendoff, 1997) which also involve covert events:

(1.6) Jack Kerouac, a **fast typist**, typed On the Road on a roll of paper so he would not be interrupted by having to change the paper.
→ someone who typed fast [enriched composition: telic adjective]

Type coercion / Type shift: these terms (Pustejovsky, 1993) refer to the shift of type (entity → event) arguably occurring in logical metonymies, but is also used to refer to other type shifts such as aspectual coercion, for example the coercion of the semelfactive *click* into the structure of a durative event, determined by the temporal modifier "all night" (Pustejovsky and Bouillon, 1995; Rothstein, 2008):

(1.7) We spent some time trying to sleep on the bench at the railroad ticket office, but the telegraph **clicked all night** and we couldn't sleep.
(Kerouac, 1957)
→ +durative [aspectual coercion]

1.4. Thesis Contributions

This thesis addresses the problem of the role of event knowledge in models of logical metonymy, a phenomenon at the interface between lexicon and world knowledge. Despite having narrowed the scope of my investigation, I embrace what one could call an "equal rights movement"[3] for metonymic and non-metonymic constructions, aiming at identifying what central properties metonymic processes have in common with "normal" cases of online language processing, notably being incremental, efficient, and fast. The account of logical metonymy I propose is based on thematic fit and generalized event knowledge, which have been proved to play a crucial role in efficient incremental processing but whose contribution to the interpretation of logical metonymy and in general to the integration of implicit semantic content had not previously been explored.

My research exploits psycholinguistic and computational methods, crowdsourcing and corpus analyses to investigate the role of thematic fit and event knowledge in logical metonymy, seeking converging evidence from the corpus and crowdsourcing data, the psycholinguistic results and the computational model and confronting it with the theoretical investigation.

[3] See also the "equal rights movement for literal and figurative language" (Hahn and Markert, 1997).

I report on the first corpus study of logical metonymy for German, which was conducted on a much larger corpus than previous studies on English and resulted in the publication of a German Logical Metonymy Database (1886 metonymies and 2661 long forms annotated by two native speakers), publicly available for scientific research purposes.

I present the first similarity-based account of logical metonymy, contributing to the recently expanding field of compositional distributional semantics with (a) a treatment of implicit semantic content (covert events) and (b) a discussion on the representation of semantic types, which traditionally fall within the domain of formal semantics.

I present four psycholinguistic experiments, which provide evidence for rich lexical knowledge intervening early during expectation-driven language processing. Touching on a broader debate on the representation of lexical knowledge, I question the existence of a sharp distinction between lexicon and world knowledge and argue in favor of an enriched lexicon where generalized event knowledge is widely used during processing, both to predict upcoming input and understand implicit events. The psycholinguistic experiments were carried out for German, which unlike English is not commonly used in studies of logical metonymy.

1.4.1. Thesis Plan

This Part (I) introduces the research methods employed in my research, including a discussion on methodological aspects and on the intrinsic interdisciplinary character of my work. The different methodologies and their interest for studies on logical metonymy are presented in Chapter 2. Chapter 3 provides an overview on the predominant takes on logical metonymy interpretation (the Lexical Hypothesis and the Pragmatic Hypothesis) and proposes a new hypothesis (the Words-as-cues Hypothesis).

Part II and part III expand on the two main issues that a model of logical metonymy should address, respectively the source problem and the trigger problem, and on our hypotheses and experiments regarding those two problems.

Part II presents results from a study of the usage of logical metonymies in corpora and in a crowdsourcing study (Chapter 4), a psycholinguistic study of the source of the covert event (Chapter 5), and a computational model of covert event interpretation

(Chapter 6). The experimental studies presented speak in favor of a Words-as-cues Hypothesis of covert event retrieval.

Part III tackles the trigger problem, this time moving from computational evidence (Chapter 7) and then testing our hypothesis on a psycholinguistic study (Chapter 8), which evaluates a Words-as-cues Hypothesis of type clash.

Part IV discusses a revised model of logical metonymy and its implications for theories of lexical meaning (Chapter 9), and concludes this dissertation with some remarks and discussion on open issues and further directions of research (Chapter 10).

Scientific research is rarely the work of a single individual working alone. The computational modeling was carried out together with Jason Utt, who was responsible for the implementation of all models. Also, I have included references to the relevant publications every time the work presented in this thesis was published elsewhere and benefited from cooperation with others. When presenting those results I have used "we" rather than "I", not as a pluralis maiestatis, but rather to point out when the results arose from a collaboration.

1.4.2. Publications

Parts of this thesis (results and discussions) have been previously reported in the following publications:

> Rüd, S. and Zarcone, A. (2011). Covert events and qualia structures for German verbs. In *Proceedings of the Metonymy 2011 Workshop*, pages 17–22, Stuttgart, Germany
>
> Utt, J., Lenci, A., Padó, S., and Zarcone, A. (2013). The curious case of metonymic verbs: A distributional characterization. In *Proceedings of the IWCS Workshop "Towards A Formal Distributional Semantics"*, Potsdam, Germany
>
> Zarcone, A., Lenci, A., Padó, S., and Utt, J. (2013). Fitting, not clashing! a distributional semantic model of logical metonymy. In *Proceedings of the 10th International Conference on Computational Semantics*, Potsdam, Germany
>
> Zarcone, A. and Padó, S. (2010). "I like work: I can sit and look at it for hours" - Type clash vs. plausibility in covert event recovery. In *Proceedings of Verb 2010* -

Interdisciplinary Workshop on Verbs, pages 209–214, Pisa, Italy

Zarcone, A. and Padó, S. (2011). Generalized event knowledge in logical metonymy resolution. In *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*, pages 944–949, Boston, MA

Zarcone, A., Padó, S., and Lenci, A. (2012b). Inferring covert events in logical metonymies: a probe recognition experiment. In *Proceedings of the 34th Annual Meeting of the Cognitive Science Society*, pages 1215–1220, Sapporo, Japan

Zarcone, A., Padó, S., and Lenci, A. (2014). Logical metonymy resolution in a words-as-cues framework: evidence from self-paced reading and probe recognition. *Cognitive Science*, 38(5):973–996

Zarcone, A. and Rüd, S. (2012). Logical metonymies and qualia structures: an annotated database of logical metonymies for German. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 1799–1804, Istanbul, Turkey

Zarcone, A., Utt, J., and Padó, S. (2012d). Modeling covert event retrieval in logical metonymy: probabilistic and distributional accounts. In *Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics*, pages 70–79, Montréal, Canada

Some parts have been presented as talks and posters at conferences:

Zarcone, A., Lipenkova, J., and Michelbacher, L. (2012a). Easy / difficult constructions as triggers of implicit content: comparing covert event elicitations and events extracted from a very large corpus. Poster presented at Linguistic Evidence 2012

Zarcone, A. and Padó, S. (2013). Logical metonymy: Disentangling object type and thematic fit. Poster presented at the 19th Conference on Architectures and Mechanisms for Language Processing, Marseille, France

Zarcone, A., Utt, J., and Lenci, A. (2012c). Logical metonymy from type clash to thematic fit. Poster presented at the 18th Conference on Architectures and Mechanisms for Language Processing, Riva del Garda, Italy

2. Methods for Studying Logical Metonymy

The work reported in this thesis has employed a range of techniques, from psycholinguistic to computational methods. Psycholinguistic experiments were carried out to gain insights into the way we process logical metonymies, and more broadly into language comprehension and the mental lexicon. Computational models were applied both in a predictive way (simulating behavioral results) and as an explanation method in their own right, in order to explore the influence of contextual factors in logical metonymy interpretation, looking for converging results from the theoretical investigation, the psycholinguistic experiments and the computational modeling. Each method comes with a linking hypothesis (Crocker, 2010), which is necessary to link empirical data (coming from the experimental measurements) to a theory or model and specifies the relationship between them.

I will now introduce the methods and techniques employed, along with their relevance for the study of logical metonymy interpretation, in order to justify my methodological choices and to explain the different linking hypotheses (from the theory to the data and from the data to the theory) which these methods are based on. I will then discuss the relation between computational models and psycholinguistic data.

2.1. Studying the Usage of Logical Metonymies

2.1.1. Corpus Analyses

Corpora are a valuable resource to investigate language usage in naturally-occurring text, both for qualitative and quantitative analyses (McEnery and Hardie, 2012): specific constructions can be searched, in order to study examples of their usage, and their relative frequencies can be estimated; frequencies of contexts of use can then

be contrasted with those of other (less marked) constructions. Features of different constructions can be contrasted with descriptive and inferential statistics.

Regarding logical metonymies, one can for example investigate their usage in written language, by looking for instances of metonymic verbs in a corpus and analyzing the frequencies of their different subcategorization frames and the object fillers in the NP subcategorization frame (as in the seminal work of Briscoe et al., 1990 and Verspoor, 1997a,b).

Covert events are by definition not attested in the corpus, because they are implicit: it is then necessary to first annotate the metonymies with their covert event in order to study the extent of their interpretations. Also, corpus analyses can contrast logical metonymies with constructions where the event is explicit (long forms), in order to evaluate differences between the two. Intuitively, we may expect implicit events to be more obvious, whereas explicit events will probably correspond to non-default interpretations. A corpus-based study of such constructions can corroborate these claims with usage data.

We have performed an extensive corpus-based analysis of German metonymic verbs, to be discussed in Chapter 4, Section 4.1.

2.1.2. Offline Web Experiments: Crowdsourcing

Offline studies in the form of questionnaires are widely used to collect native speakers' judgments and ratings and to elicit linguistic production to investigate usage. They are called **offline** experiments, to distinguish them from online methodologies (such as self-paced reading or eye tracking) which study sentence processing while it unfolds over time. Another common term is **norming studies**, when the aim is to collect norms, that is average norm values (from ratings, frequencies or other data) to be made available for other studies (Wurm and Cano, 2010). Offline studies have also been successfully conducted on web platforms, and more recently on crowdsourcing platforms like Amazon Mechanical Turk (AMT — `http://www.mturk.com`) and Crowdflower[1] (`http://crowdflower.com/`).

Crowdsourcing has sped up researchers' access to speakers' judgments, allowing for fast and affordable collection of reliable and vast linguistic data, such as, for example,

[1] Crowdflower provides an interface to Amazon Mechanical Turk and other crowdsourcing platforms with easier access for non-US citizens.

semantic associations (Schulte im Walde et al., 2008), paraphrases of words (McCarthy and Navigli, 2009), task-oriented dialogues (Potts, 2012), compositionality ratings (Roller et al., 2013), and many more[2]. However, the fidelity of web experiments and AMT experiments for use in behavioral studies and NLP annotation tasks (Wurm and Cano, 2010; Fort et al., 2011) has been questioned. Studies comparing crowdsourced data and lab data have shown promising correlations between the two (see for example Snow et al., 2008; Callison-Burch, 2009; Munro et al., 2010), and have supported the feasibility of reading time and reaction time studies on web platforms (Keller et al., 2009), but have also highlighted potential issues and necessary caveats, encouraging experimenters to carefully consider the characteristics of the interface they are using, such as difficulties in spreading complex tasks over multiple pages or in assigning participants to different lists, the need to keep scammers from interfering with the study[3], the lack of control over the experimental setting (which is more controlled in a lab where distractions are kept to a minimum, Wurm and Cano, 2010), and last but not least ethical issues regarding wages (Fort et al., 2011).

As we have seen, covert events are by definition not attested in the corpus, so corpus studies rely either on long forms as paraphrases of logical metonymies (where the event is explicitly realized) or on semantic annotations of logical metonymies (where the event has to be annotated). Web elicitation studies on crowdsourcing platforms like Amazon Mechanical Turk and Crowdflower can be very useful to study semantic interpretation and to go beyond what is found in corpora (for example, McCarthy and Navigli, 2009 used paraphrases as an empirical correlate of word senses); specifically for logical metonymies, elicitations can be exploited as a window into the covert event interpretation. We used offline web experiments to (a) elicit paraphrases of covert events (Chapter 4, Section 4.2), and also (b) as norming studies to collect ratings on a Likert-scale (Fabrigar et al., 2005) and (c) build experimental materials for the psycholinguistic studies (Chapters 5 and 8).

[2]See also the 2011 Workshop on Crowdsourcing Technologies for Language and Cognition Studies held in Boulder, Colorado, `http://www.crowdscientist.com/workshop/`.

[3]Geographic origin is usually controlled by IP address checking.

2.2. In the Lab: Psycholinguistic Experiments

2.2.1. The Time Course of Logical Metonymy Interpretation: Self-paced Reading and Eye Tracking

While offline methods do not offer insights about the time course of interpretation, **online** studies are aimed at studying language processing while it unfolds over time. Self-paced reading and eye-tracking methods use **reading time** as a correlate of online processing cost: in regions (words or phrases) which are considered harder to process, longer reading times are expected.

In studies using the moving-window **self-paced reading** paradigm (McConkie and Rayner, 1975; Just et al., 1982), each sentence is represented as a sequence of dashes, and participants read at their own pace by pressing a button, revealing one word at a time, while the rest of the sentence remains represented by dashes; reading times per word are measured as the time it takes the participant to press a button to move to the next word and are used as a correlate of processing cost: the more difficult (costly) the word or phrase, the longer the reading time. To make sure that participants do indeed read the sentences, these are often followed by a comprehension question.

Eye-tracking studies measure reading times in terms of eye movements on a region. A number of measures of eye movements are typically reported as relevant for language processing: first fixation, first-pass time, regression path time, total time, regression probability. *First fixation* is simply the duration of the first fixation in a region; *first-pass time* is the sum of all fixations in a region beginning with the first fixation until the gaze leaves the region (either going back or going forward) for the first time; *regression path time* is the total time from the first fixation to the first eye movement past the region (forward), including re-reading other regions; the *total time* is the sum of all fixations in the region, including secondary fixations; *regression probability* is computed per trial across trials (across items and participants) and per region, as the percentage of regressions out of a region (usually, first-pass regressions only). Sometimes second-pass time is also reported, being the time spent re-reading the region following first-pass fixations, including the time spent in the region after regressions and the time spent in the region after reading past it (Carreiras and Clifton, 2004; Boland, 2004). Eye movements are also considered a correlate of processing cost (the more costly the word, the longer the eye fixations), but different measures are associated with

different processes: first-fixation and first-pass times are usually considered early measures, capturing early language processes, for example lexical access and retrieval difficulties, whereas regression path times and total reading times are considered late measures, capturing late language processes, for example integration into a discourse representation (Carreiras and Clifton, 2004; Boland, 2004; Pickering et al., 2004; Staub and Rayner, 2007).

Previous studies on logical metonymy have employed both self-paced reading and eye-tracking techniques (see Pylkkänen and McElree, 2006, for a review), mostly addressing the trigger problem and thus contrasting metonymic sentences (*The journalist began the article*) with non-metonymic (*The journalist wrote the article*) and anomalous constructions (*The journalist astonished the article*), and have reported extra processing costs for the type shifts (but see Fodor and Lepore, 1998; de Almeida and Dwivedi, 2008 for contrasting results).

We have carried out two self-paced reading experiments: Experiment 1 (Chapter 5), addressing the source problem, and Experiment 3 (Chapter 8), addressing the target problem. In particular, Experiment 1 faces the same drawback of using logical metonymies as test sentences to investigate covert events, and our solution was to exploit paraphrases of German logical metonymies (long forms), which differ from the metonymies only in that they contain a clause-final target verb, that is the — now explicit — covert event (*Der Konditor begann die Glasur → Der Konditor begann die Glasur aufzutragen* / *The baker began the icing → The baker began spreading the icing*).

The original work reported in this dissertation does not include eye-tracking experiments, but data from eye-tracking experiments carried out by others have been used for comparison and their material sentences have been used in the modeling study reported in Chapter 7.

2.2.2. Probe Recognition

Comprehension experiments have also used the **probe recognition** paradigm (Sternberg, 1966, 1975; McKoon and Ratcliff, 1986) to measure the accessibility of information in working memory: after a comprehension phase (reading or listening to a sentence), participants are presented with a probe word, sometimes after a fixed time interval called Inter-Stimulus Interval or ISI (e.g. *With his exam coming up, the student*

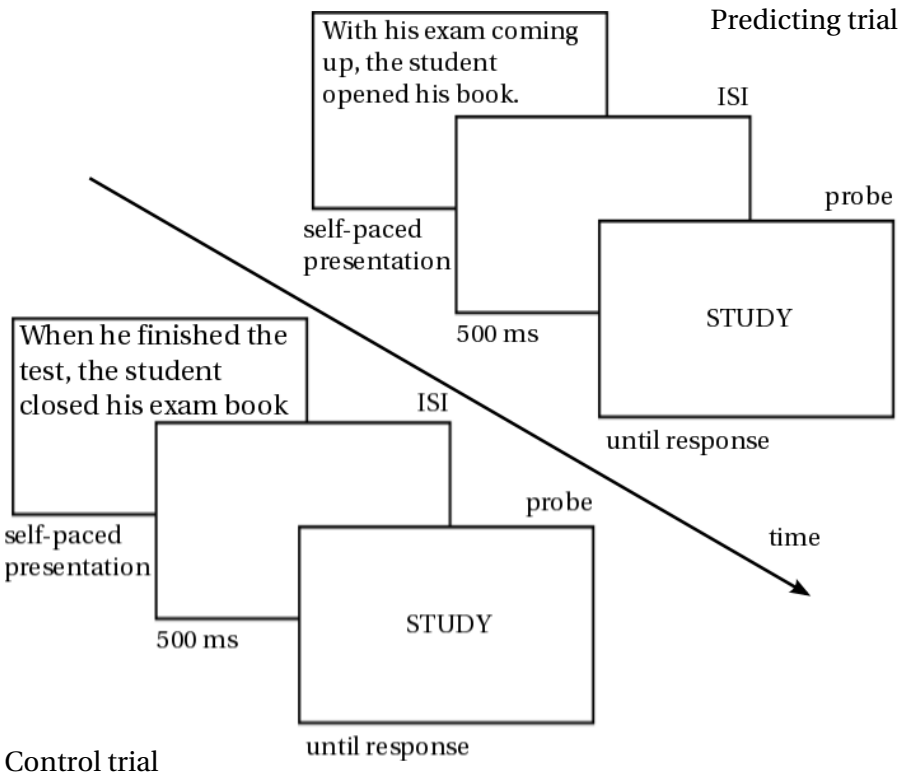

Figure 2.1.: Example of a trial structure in a probe recognition experiment (materials from McKoon and Ratcliff, 1986: predicting trial vs. control trial).

opened his book. → *[ISI]* → *STUDY*, McKoon and Ratcliff, 1986); they are then required to perform a lexical decision on a probe (*Is STUDY a word or a non-word?*) or to say if the probe was part of the sentence (*Was STUDY mentioned in the sentence?*) by pressing one of two buttons (see an example of a trial structure in Figure 2.1). Decision latencies are measured as the time between the presentation of the probe and the moment the participant presses a button to respond, and are used as a correlate of memory retrieval and of the spontaneous activation of certain concepts during text comprehension.

A control condition is typically used, to estimate facilitation or inhibition effects on decision latencies determined by the experimental manipulation with respect to the baseline. In lexical decision tasks the assumption is that the less accessible a target word is in working memory, the longer the participant will take to recognize it; when the task is to answer if the probe appeared in earlier discourse, the assumption is that

cued / predicted probes will interfere with the decision, leading to longer reaction times. Different ISI conditions are often contrasted (e.g. short ISI at 100 ms and long ISI at 900 ms), and are expected to influence memory retrieval, as concept activation degrades over time.

Previous studies on logical metonymy have not used probe recognition methods, but we have introduced this paradigm (Experiments 2 and 2b, Chapter 5) as a strategy to investigate implicit linguistic content by presenting the covert events as probes following a metonymic sentence (*Der Konditor hörte mit der Glasur auf* → [ISI] → *AUFTRAGEN* / *The baker finished the icing* → [ISI] → *SPREAD*). If the probe is mentioned in the sentence and is still active at the time of the response, the probe recognition or lexical decision should be facilitated (shorter decision latencies). Conversely, if the probe is implicit (as in the case of covert events: *The author began the book.* → *[ISI]* → *WRITE*) the probe recognition should be inhibited (longer decision latencies): if the concept is active, then it is challenging to quickly respond that it was NOT present in the sentence (giving a *no* response to the probe recognition).

2.3. Computational Models of Logical Metonymy

Another important source of evidence for the study of cognitive processes (and of language) is computational modeling. Computational simulation can help uncover the interplay of different variables in accounting for language data (for example, corpus data or human annotations / judgments). Also, computational models constitute an important complement to behavioral studies and to theoretical investigation, as they force us to think of the architectural constraints of the models we adopt (e.g. *what representations are assumed? how are they acquired? how is new information processed?*), and then allow us to evaluate such constraints (through the predictions resulting from them) on behavioral data (e.g. *what is the impact of each parameter? what combination of parameters can best explain a linguistic phenomenon? what are the implications for the theory?*, see Dijkstra and de Smedt, 1996).

Computational models of language processes can also be aimed at improving NLP applications (for instance, providing one unambiguous interpretation for *the author began the book* → *reading, writing*), and previous computational treatment of logical metonymy (see Lapata and Lascarides, 2003; Lapata et al., 2003; Roberts and

Harabagiu, 2011; Shutova, 2009; Shutova and Teufel, 2009; Shutova et al., 2013) has mainly been NLP-oriented, but this aspect is beyond the scope of my work. The interest for computational models in this dissertation on the other hand arises from the crucial theoretical contributions that they can make in the study of logical metonymy (more in the spirit of the work presented in Lapata et al., 2003). For example, in Chapter 7 I will address the question of what role object type and object thematic fit play in logical metonymy interpretation. If a model that is only informed about thematic fit can mirror experimental results, we may then consider questioning the role played by type.

Also, while previous studies have employed probabilistic models based on first-order co-occurrences, the work presented in this dissertation provides the first distributional, similarity-based model of logical metonymy interpretation. I will now describe in more detail the class of computational models that were used in this dissertation, that is Distributional Semantic Models, and I will then turn to some final methodological remarks on how to relate data from the computational models and from the behavioral experiments.

2.3.1. Distributional Semantic Models (DSMs)

The roots of Distributional Semantics are grounded in the idea that words occurring in similar contexts are semantically related. This idea was already implied by the structuralist concept of paradigmatic relations (de Saussure, 1915; Sahlgren, 2008), but it was not until much later (Harris, 1954; Firth, 1957; Miller and Charles, 1991) that the idea of linking distributional facts (contexts of occurrence) and semantic meaning of words became popular, in the form of an explicit **Distributional Hypothesis**: two linguistic units are more semantically similar the more similar their context of occurrence are. Or, in more evocative words, "You shall know a word by the company it keeps" (Firth, 1957, p. 11).

Distributional Semantic Models (DSMs, also known as Vector Space Models) build on the Distributional Hypothesis and represent linguistic units (e.g. words) by means of corpus-extracted vectors whose dimensions are (function of) co-occurrence frequencies with other words. See for example the toy vectors for *mouse*, *frog* and *hawk*, Table 2.1: the vector for *frog* is more similar to the vector for *mouse* (they both move on the *land*, they can both *jump*, neither of them *flies* or *pounces*) than to the vector

	land	water	sky	run	jump	fly	pounce	seize
mouse	8	0	0	7	4	0	0	4
frog	5	7	0	0	10	0	0	5
hawk	0	0	10	0	0	10	10	9

Table 2.1.: Toy distributional vectors representing the words *mouse*, *frog* and *hawk*.

for *hawk*. If our toy model does not take syntactic relations or thematic roles into account, then all three may co-occur with *seize*, without any difference regarding who *seizes* and who *is seized*.

DSMs differ with regard to their definition of context, their way of representing the distributional facts, and the linking hypothesis regarding the aspects of meaning they are meant to represent (Sahlgren, 2008; Lenci, 2008; Turney and Pantel, 2010). Context features may be content words within a certain distance from the target word (as in our toy vectors in Table 2.1, see also Schütze, 1992; Lund and Burgess, 1996). Alternatively, context features can also be text regions, as in the Latent Semantic Analysis approach (LSA, Landauer and Dumais, 1997). Semantic similarity can be approximated by vector similarity, choosing from a wide range of similarity metrics (Lee, 1999; Turney and Pantel, 2010).

DSMs offer a very straightforward way to represent meaning and compare representations (similarity computation), which makes them appealing for computational linguistics and NLP applications. A variety of DSMs have been used for several NLP tasks such as thesaurus construction (Lin, 1998), word sense discrimination (Schütze, 1998), topical relatedness estimation (Landauer et al., 1998), ontology learning (Buitelaar et al., 2005), event type classification (Zarcone and Lenci, 2008), and many more (Turney and Pantel, 2010).

Also, DSMs have a "cognitive vocation": the Distributional Hypothesis (at least in its strongest version, see Lenci, 2008) takes the shape of a cognitive hypothesis about semantic representations, going as far as claiming that the distributional behavior of a word in context *is* a direct correlate of its semantic content at the cognitive level and that the context of occurrence of a word provides an insight into the organization of the mental lexicon[4]. LSA is a very popular model among psycholinguists and cognitive

[4]de Saussure (1915) observed that co-occurrence relations are relevant for memory, because the context of co-occurrence of a word is relevant for its retrieval: a word, e.g. the French word *enseigne-*

	run-SUBJ	jump-SUBJ	fly-SUBJ	seize-SUBJ	seize-OBJ
mouse	7	4	0	1	3
frog	0	10	0	0	5
hawk	0	0	10	9	0
sparrow	0	1	9	0	4
crocodile	0	0	0	9	2

Table 2.2.: Toy structured distributional vectors representing the words *mouse*, *frog*, *hawk*, *sparrow* and *crocodile*.

linguists, as it is arguably (Landauer et al., 1998) a plausible model to simulate a variety of cognitive phenomena, and DSMs in general have proven to be very compatible with known features of human cognition (gradedness, context-sensitivity, distributedness), as shown by models of graded category membership (Rosch, 1975), multiple sense activation (Erk et al., 2010), lexical development (Li et al., 2004), category-related deficits (Vigliocco et al., 2004), selectional preferences (Erk, 2007), and more (see Landauer et al., 2007; Baroni and Lenci, 2010, for a review).

DSMs can relate words with other words in its context (*bag-of-words* models Schütze, 1992) or with words in a specific syntactic pattern (structured vector spaces, Lin, 1997; Padó and Lapata, 2007; Baroni and Lenci, 2010). The latter, also called **Structured DSMs**, take into consideration not only binary relationships of co-occurrence between two words (for example, how many times *mouse* and *jump* occurred in the same sentence), but also syntactic relations between the words (for example, how many times *mouse* was the subject of *seize* or object of *seize*). The vectors in Table 2.2 keep track of how often an animal was subject or object of the verbs it co-occurred with: crocodiles are amphibious creatures like frogs, but are predators like hawks (they are the subject of *seize*), and are rarely *seized* (with the exception of when they are being smuggled and are seized by authorities). Sparrows are birds, but unlike hawks they tend to *be seized* more than to *seize*. Thus, vectors in Table 2.1 and 2.2 capture different kinds of similarity (the former is closer to topical association, the latter is more similar to cohyponymy).

Structured DSMs, using syntactically parsed data, allow us to take into considera-

ment 'teaching', "will unconsciously call to mind a host of other words", e.g. *education* 'education,' *apprentissage* 'apprenticeship,' etc. (de Saussure, 1915, p. 123).

tion syntactic dependencies and can thus approximate thematic roles by way of these dependencies (e.g. the role of agent with the subject dependency, the role of patient with the object dependency) and effectively model regular and inverse selectional preferences (Erk et al., 2010). Activation of event knowledge in language processing is sensitive to intra-sentential context and in particular to thematic role fillers, and behavioral evidence has shown that thematic role fillers guide our language comprehension and expectations during processing, because fillers that match the verb's selectional preferences are expected, influencing plausibility ratings and reducing processing costs (see McRae and Matsuki, 2009; Elman, 2011, for a review). For example, if we are talking about an event of *serving*, there will probably be a *waiter* (typical agent) and a *customer* (typical patient) involved, and these typical fillers will receive high plausibility ratings and will be expected during processing. Modeling typical fillers for argument positions is particularly relevant for logical metonymies, as we have seen that the interpretation of a covert event can vary greatly depending on intra-sentential context (contrast *The author began the book* → *writing* vs. *The child began the book* → *reading*). Furthermore, DSM models have the advantage of being unsupervised, as they rely only on a large parsed corpus (no labeled data or training set is required). Our model of logical metonymy can potentially be replicated for languages other than English and German. I will discuss the DSM adopted in our work (Distributional Memory), as well as compositionality in DSMs, in Section 6.3.

The question remains open, whether typicality can be reduced to corpus co-occurrences. As observed by Padó (2007), "infrequent events may be perceived as more informative or interesting and therefore more worthy of being communicated, which may cause them to be discussed disproportionally more often than they are experienced" (pp. 30-31) while "frequent events may be perceived as less newsworthy and therefore be mentioned less often than they occur" (p. 31), leading to an imperfect parallelism between events in the world and events in language. Bruni et al. (2012) found that it is difficult to model stereotypical color adjective information (e.g. *bananas are yellow*) on the basis of corpus-extracted information, arguably for the very same reason that corpus-extracted information is not informative with regard to highly typical information. Nevertheless, results from studies on selectional preferences (Erk et al., 2010) and on composition of verb-argument expectations (Lenci, 2011) show that corpus-extracted information can indeed approximate typicality information

by taking structural relations at the sentence level (agent and patient) into account, mirroring effects of generalized event knowledge.

2.4. Modeling Psycholinguistic Data

A variety of methods has been employed to directly relate computational data to behavioral data. In the field of Natural Language Processing, the aim is usually to evaluate the performance of several models (not only DSMs) and, although common practices are fairly well established, the choice of the evaluation task is usually determined by the availability of an evaluation dataset and by the sort of dataset available (bigger or smaller, with or without reaction times or rating scores, averaged or not). Evaluating different computational models is beyond the scope of this dissertation. Rather, we want to employ an established state-of-the-art structured DSM to further analyze psycholinguistic results and psycholinguistic datasets. For example, if a dataset contrasts two group of items that are supposed to differ only for the experimental manipulation, we can use a computational model of thematic fit to evaluate if there is an unreported thematic fit difference that may be influencing reaction times.

The datasets I took into consideration are those yielded by the methods described in this chapter: rating datasets and materials employed in reading time and reaction time studies. I will now review some common methods, and their applicability to the behavioral tasks described in this chapter and to the data which will be discussed later in this dissertation.

2.4.1. Predicting Continuous Behavioral Data

Correlation with Human Judgments

Computational models of thematic fit and selectional preferences (Resnik, 1996; Keller and Lapata, 2003; Erk, 2007; Erk et al., 2010; Baroni and Lenci, 2010; Ó Séaghdha, 2010) are typically evaluated by correlating the plausibilities estimated by the model with human judgment ratings, expressed on a Likert scale or employing a continuously valued scale (Magnitude Estimation, Stevens, 1975)[5]; for example, such datasets may

[5] Peirsman and Padó (2011) show an almost perfect correlation between the plausibility judgments collected by Brockmann and Lapata (2003) with Magnitude Estimation and the plausibility judgments they collected on the same dataset using a Likert scale.

Verb	Thematic role	Noun	Plausibility Rating
chase	lion	agent	6.6
chase	lion	patient	2.6
chase	mouse	agent	3.1
chase	mouse	patient	5.5

Table 2.3.: Verb - thematic role - noun triples with plausibility judgments on a 7-point Likert scale (McRae et al., 1998).

include human ratings reflecting how plausible it is for a *lion* to *chase* or to *be chased*, or for a *mouse* to *chase* or to *be chased* (see Table 2.3). The correlation between the model estimations and the human ratings is usually tested with Spearman's correlation (ρ, a non-parametric rank-order test). Significant correlations (ρ values significantly different than 0) are considered evidence of the predictive relationship between the model and the speakers' judgments.

Predicting Reading Times

Another continuous dependent variable that is a correlate of semantic processing is reading time. A nonparametric correlation evaluation can be performed, to evaluate if the ranking proposed by the computational model correlates with the ranking in processing costs. Reading times can be measured in different ways, and predicting them via a computational model is not a trivial task, for a number of reasons.

Firstly, there is no such thing as a single straightforward measure of reading time: reading times yielded by self-paced reading tasks also include the time needed to press the button while reading, and eye-tracking studies provide several different measures of processing difficulty (first-fixation duration, first-pass reading time, total reading time, regression path duration and number of regressions). This leads to the question of what eye-tracking measures better correlate with what computational models (see for example McDonald and Shillcock, 2003a,b; Frisson et al., 2005).

Secondly, many different factors are known to affect processing costs, such as word frequency and length, the position of the word in the sentence, verb argument structure and syntactic frame frequency (see Just and Carpenter, 1980; Rayner and Duffy, 1986; Baayen and Milin, 2010). When designing a psycholinguistic experiment

featuring two experimental conditions, the experimenter must try to make sure that the items in the two conditions are matched for variables known to affect processing costs, in order to avoid influencing the reading time means for each condition, but a model performing a point-wise prediction of reading times would have to take these factors into account.

Lastly, added to the several word-specific factors known to affect processing costs, one must also take into account the idiosyncrasies of participants and of experimental items. A psychological experiment like those presented in this dissertation (for instance, with two experimental conditions) usually features a sufficiently high number of participants and items (for instance, 30 participants and 96 items), and then tests the significance of the differences between the average means for the experimental conditions (across items and participants), accounting for within-participant and within-item variability by means of a participant random effect and by an item random effect (Baayen et al., 2008). A point-wise prediction of reading times (per item) would rely only on the average reading time per item (computed only across the participants which have seen each single item, which in case of two lists and 30 participants is 30/2 = 15; or fewer, in case of outlier exclusion), which is usually not sufficient to overcome within-participant and within-item variability.

Although predicting reading times is extremely difficult for the above mentioned reasons[6], a possible solution to overcome at least the first two issues is to predict instead the time deltas between experimental conditions (McRae et al., 1998), or to compute and predict the scaled reading time effect in terms of the percentage of processing difficulty contributed by each region of the sentence, that is by measuring the percentage of reading time effect at each given sentence stage (time delta at the given stage) over the total reading time effect (time delta across all stages, Narayanan and Jurafsky, 2004; Padó, 2007). Due to the small size of the psycholinguistic studies described in this dissertation, we have not not performed a point-wise nonparametric correlation evaluation, which would require a high number of participants per item, but two different methods, described in Sections 2.4.2 and 2.4.3.

[6]Similar issues are encountered when predicting decision latencies in priming studies, but still allow for successful correlations (Padó and Lapata, 2007; Lapesa and Evert, 2013).

2.4.2. Pairwise Comparisons

Psycholinguistic studies often contrast two experimental conditions and look for a (significant) difference in behavioral data between them (for example in terms of eye gazes, reading times, decision latencies). A typical psycholinguistic dataset may then be composed of sentence pairs such as these examples from the experiment reported by Bicknell et al. (2010):

(2.1) a. The journalist checked **the spelling** / the brakes.

b. The farmer loaded **the truck** / the pistol.

c. The player tossed **the frisbee** / the leftovers.

The experiment contrasted congruent (objects in boldface) and incongruent conditions, in order to evaluate the effects of event knowledge on the processing of object argument fillers; Bicknell et al. (2010) showed that the congruent conditions yielded lower processing costs than the incongruent ones. A structured DSMs such as the DM + ECU model (which will be introduced in 6.3.2) is able to compute thematic fit values for the paired contrasts in 2.1.a-c (for example, for the tuple ⟨*journalist check spelling*⟩ vs. the tuple ⟨*journalist check brakes*⟩). Lenci (2011) evaluated the DM + ECU on the Bicknell dataset with simple pairwise comparisons: for each contrast in 2.1.a-c, the model scored a "hit" whenever a sentence in the congruent condition yielded a higher thematic fit score than the incongruent one. The percentage of hits over the total number of sentences was used as a measure of accuracy of the model.

This method is a simple, straightforward way to evaluate a computational model, which is not as sophisticated as an evaluation task that involves predicting reaction times, but has the advantages of (1) doing away with the problem of transforming thematic fit values before correlating them, (2) being applicable to simple psycholinguistic experiments, where the number of participants is not enough for a correlation evaluation of the thematic fit scores with the reading time measures.

Pairwise comparisons were used to evaluate our similarity-based model of covert event interpretation reported in Chapter 6.

2.4.3. Pattern Replication

Another option for correlating thematic fit data and psycholinguistic results is a pattern replication of the significance patterns. For example, when a priming study detects

a significant difference in decision latencies between the related-priming pairs and the unrelated-priming pairs, we can expect a good computational model of semantic relatedness to replicate this pattern by showing a significant difference between the similarity scores of the two groups (see for example McDonald and Brew, 2004; Padó and Lapata, 2007; Herdağdelen et al., 2009; Hare et al., 2009b).

There is a crucial caveat to consider when performing this sort of evaluation: ideally, the computational model should find the same significant differences, main effects and interactions reported by the psycholinguistic studies, but it has been observed that some models (for example LSA) "overemphasize associative relations in priming" (Hare et al., 2009b), showing significant differences in the distributional data that are not strong enough to show in the behavioral data. It is then interesting not only to observe what computational models have a "naturalistic" behavior (somehow mimicking the patterns found in behavioral data), but also what parameters bring models to show more significant differences than behavioral data, and why.

Pattern replication was used in the work reported to evaluate our thematic fit model of logical metonymy in Chapter 7.

3. Accounts of Logical Metonymy

Different approaches to lexical semantics and sentence comprehension have provided very different answers to the trigger problem (*what triggers the logical metonymy?*) and to the source problem (*what is the source of the covert event?*). I will now provide an overview of the most widely adopted approaches to logical metonymy interpretation, which differ greatly with regard to the role played by event knowledge in language understanding.

A full-fledged review of the treatment of logical metonymy and coercion in theoretical linguistics (e.g. Egg, 2005; Asher, 2011) would be out of the scope of this thesis, as my focus is on experimental studies (behavioral and corpus-based). Nevertheless, experimental work on logical metonymy is of course theoretically motivated and has implications for theories of the lexicon and of language processing. I will therefore introduce the theoretical work relevant for the experimental study by grouping the reviewed approaches into two predominant tendencies, depending on the theoretical implications they come with, and for each of the two I will report existing supporting evidence and highlight strengths and weaknesses. I will refer to these two tendencies as the **Lexical Hypothesis** and the **Pragmatic Hypothesis** respectively. I will then sketch a third proposal, the **Words-as-cues Hypothesis**.

3.1. The Lexical Hypothesis

The first and most common definition of logical metonymy (as the combination of an event-selecting verb and an entity-denoting object, resulting in a type clash and in the recovery of a covert event) was proposed within an influential theory of lexical semantics, the Generative Lexicon (GL, Pustejovsky, 1991, 1995, 1998), and was later adopted also by other scholars (Jackendoff, 1997; Blutner, 2002).

The GL focuses on semantic composition: we know that language production requires combining words and phrases in a meaningful way, and that language under-

standing requires recognizing meaningful combinations even if we have never heard them before (for example, we may have never heard the combination *eating turtle soup*, but we are still able to understand what it means). Also, subtle meaning changes may emerge from composition: see for example the adjective *fast* in *fast car*, *fast typist*, *fast waltz*, (Vendler, 1968, pp. 88-90), or *book* as text in *summarize the book* vs. *book* as physical support *dust the book*.

A theory of lexical semantics should therefore be (1) *compositional*, i.e. it should define how simpler semantic elements are combined to form more complex ones, (2) *generative*, i.e. it should explain how a finite number of lexical items can be combined to generate an unbounded number of felicitous contexts as well as account for *creative* uses of language, (3) *constrained*, i.e. it should define the conditions for the composition operations that constrain how words and phrases combine, to avoid overgeneration, and (4) *systematic*, i.e. it should account for regular meaning changes and sense alternations.

In order to fulfill these requirements, the GL resorts to lexical decomposition and proposes a strongly typed semantic system, where lexical entries are not atomistic representations, but information-rich structures, incorporating four levels of representations: the LEXICAL INHERITANCE STRUCTURE (specifying the relations between lexical structures in the type lattice), the ARGUMENT STRUCTURE (specifying the number and type of logical arguments and their syntactic realization), the EVENT STRUCTURE (specifying the event type and subevental structure of the lexical item), and the QUALIA STRUCTURE (specifying our "understanding of an object or a relation in the world", Pustejovsky, 1995, p. 87).

Qualia structures specify a finite set of (four) essential aspects of a word's meaning[1]: its relation with its constituents (CONSTITUTIVE QUALE), its relation with other objects in a larger domain (FORMAL QUALE), its purpose and function (TELIC QUALE) and the factors involved in its origin / creation (AGENTIVE QUALE).

[1] Qualia are inspired by Moravcsik's (1975) interpretation of Aristotle's *aitiae*.

(3.1)
$$\begin{bmatrix} \alpha \\ \text{ARGSTR} = \begin{bmatrix} \text{ARG1} = x \\ \dots \end{bmatrix} \\ \text{EVENTSTR} = \begin{bmatrix} \text{E1} = e_1 \\ \dots \end{bmatrix} \\ \text{QUALIA} = \begin{bmatrix} \text{CONST} = \textbf{what } x \textbf{ is made of} \\ \text{FORMAL} = \textbf{what } x \textbf{ is} \\ \text{TELIC} = \textbf{function of } x \\ \text{AGENTIVE} = \textbf{how } x \textbf{ came into being} \end{bmatrix} \end{bmatrix}$$

Example of a lexical structure (Pustejovsky, 1995)

See for example the qualia in the lexical structure for *book*:

(3.2)
$$\begin{bmatrix} \textbf{book} \\ \text{ARGSTR} = \begin{bmatrix} \text{ARG1} = \textbf{x:information} \\ \text{ARG2} = \textbf{y:phys\_obj} \end{bmatrix} \\ \text{QUALIA} = \begin{bmatrix} \textbf{info} \cdot \textbf{physobj\_lcp} \\ \text{FORMAL} = \textbf{hold(y,x)} \\ \text{TELIC} = \textbf{read(e,w,x.y)} \\ \text{AGENTIVE} = \textbf{hold(e',v,x.y)} \end{bmatrix} \end{bmatrix}$$

Lexical structure for *book* (Pustejovsky, 1995)

Note that the qualia structure for *book* includes the knowledge that books contain information but also have physical supports. Qualia are postulated with the precise purpose of achieving optimal explanatory adequacy within a combinatory and generative semantic system, enriching lexical information while still keeping it concise and systematic. The interpretation of the following examples is licensed by the qualia structure of the object *book*:

(3.3) a. John read the **book** (→ book as information)

b. Mary put the **book** on the shelf (→ book as physical object)

c. This is a **good book** (→ to read)

The ambiguity between the two senses of *book* in 3.3.a-b is solved by picking one or another qualia role, and the covert event in 3.3.c is retrieved from the telic quale.

Crucially, the GL identifies logical metonymy as a prototypical case of (apparent) non-compositionality, as some non-lexical knowledge (the covert event) is required for interpretation. The GL account of logical metonymy is then derived from this strongly typed semantic system, from the above-mentioned lexical structures and from *type coercion* (see below), one of the generative devices or semantic transformations (together with *selective binding* and *co-composition*) which are triggered when arguments, arguments types and qualia types undergo semantic composition.

In the GL, the ARGUMENT STRUCTURE in lexical entries for verbs like *begin* or *finish* includes information regarding the type of the fillers for their object positions (specifically: objects of type event):

(3.4)
$$\begin{bmatrix} \textbf{begin} & \\ \text{ARGSTR} = & \begin{bmatrix} \text{ARG1} = \textbf{x:human} \\ \text{ARG2} = \mathbf{e}_2 \end{bmatrix} \\ \text{QUALIA} = & \begin{bmatrix} \text{FORMAL} = \textbf{P}(\mathbf{e}_2,\mathbf{x}) \\ \text{AGENTIVE} = \textbf{begin\_act}(\mathbf{e}_1,\mathbf{x},\mathbf{e}_2) \end{bmatrix} \end{bmatrix}$$

Lexical structure for *begin* (Pustejovsky, 1995)

If metonymic verbs select for an event-denoting argument, then their combination with entity-denoting objects is a violation of their type restrictions, which gives rise to a type clash that must be somehow resolved for the combination to be felicitous. The way the type clash is solved is then via a **type coercion** (**type shift**) operation, transforming the type entity of the object into a type event:

(3.5) $\rho :<< e, t >, t > \rightarrow < e, t >$
$\rho = \{Q_A(NP), Q_T(NP)\}$
(Pustejovsky, 1993)

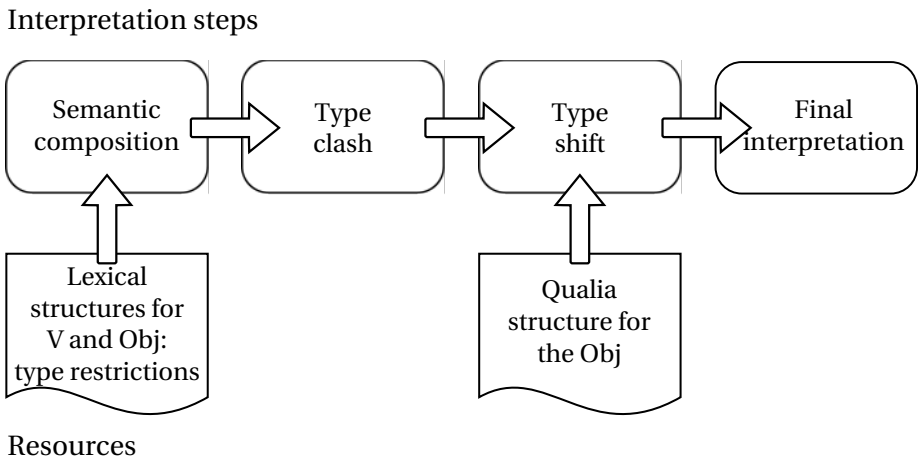

Figure 3.1.: Schematic representation of logical metonymy interpretation for the Lexical Hypothesis.

The functions shifting the type of the NP are Q_A and Q_T: when applied, they return respectively the value of the agentive quale and of the telic quale. This extension of meaning gives the logical metonymy its name, as it denotes a part-for-the-whole transfer of meaning: the logical argument of a semantic type is shifted to denote the semantic type itself (Pustejovsky, 1993):

(3.6) a. John **enjoyed the novel** (→ reading the novel)

b. Mary **finished the novel** (→ reading / writing the novel)

By placing the covert event knowledge in the lexicon, in the form of qualia structures, the GL finds an optimal solution to the problem of the covert event (and to other similar problems of ambiguity and of noncompositionality), complying to the requirements of systematicity and generativity. The **trigger** of the logical metonymy is ascribed to the lexical combinatory properties of metonymic verbs (their type restrictions and the following type clash), and also the **source** of the covert events is ascribed to the lexicon (the qualia structure). For this reason I will label this approach as **the Lexical Hypothesis**.

Let us now sum up what the Lexical Hypothesis consists of with regard to the two problems:

- **the trigger problem:** what triggers the logical metonymy?
  → the (lexical) type restrictions of the metonymic verb determine a type clash with an entity-denoting object;

- **the source problem:** what is the source of the covert event?
  → the covert events are retrieved from the lexicon, were they are stored as events associated with the object in complex lexical entries.

3.1.1. Evidence in Support of the Lexical Hypothesis

Psycholinguistic and Neurolinguistic Studies

A number of psycholinguistic and neurolinguistic studies (see Pylkkänen and McElree, 2006, for a review) searched for possible processing costs for the coercion mechanism. In doing so, they have constructed contrasts which echo the proposal of the Lexical Hypothesis, namely contrasting type clashes with cases where the type clash did not occur.

McElree et al. (2001) and Traxler et al. (2002) contrasted a coercion condition (a semantic type mismatch as defined by the Lexical Hypothesis 3.7.a) with two non-coercive conditions (3.7.b-c):

(3.7) a. ×The author **was starting the book** in his house on the island.

b. ✓ The author **was writing the book** in his house on the island.

c. ✓ The author **was reading the book** in his house on the island.

(McElree et al., 2001; Traxler et al., 2002)

The main verb in the preferred condition (3.7.b) is the verb which would be the most plausible covert event for the coercion condition (given that subject, e.g. *author*: *start the book* → *writing*), the dispreferred condition (3.7.c) features a possible but less plausible covert event as its main verb (*author*: *start the book* → *reading*). Verbs in the preferred and dispreferred conditions were elicited with a fill-in-the-blank (cloze completion) experiment using the metonymic verb as a template (*The author started* _____ *the book*). Predominant answers were chosen for the preferred conditions, whereas less frequent (but still elicited) answers were chosen for the dispreferred conditions, such that the preferred verbs were more than twice as frequent as the

dispreferred verbs. Note that no qualia structure criterion was used to select the interpretations.

In a self-paced reading study (McElree et al., 2001), longer reading times were yielded at the noun (*book*) for both the dispreferred and coerced condition compared to the preferred condition; longer reading times were also yielded at the noun +1 position (*in*) for the coerced condition compared to both the preferred and dispreferred conditions. In an eye-tracking study (Traxler et al., 2002), the coerced condition yielded more first-pass regressions and longer total times at the noun position (*book*) and longer second-pass and total times at the verb position (*starting* / *writing* / *reading*) compared to the preferred condition, with marginal differences between the coerced and the dispreferred condition.

Traxler et al. (2002) also contrasted a coercion condition (3.8.a), where an event-selecting verb is combined with an entity-denoting object, with three conditions without type clash: 3.8.b, where the same verb is combined with an event-denoting object matching its type restrictions, and 3.8.c-d, where a neutral verb (allowing for entity- and event-denoting objects) is combined with both objects, in a 2x2 design study (verb type x object type):

(3.8) a. × The boy **started the puzzle** after school today.

b. ✓ The boy **started the fight** after school today.

c. ✓ The boy **saw the puzzle** after school today.

d. ✓ The boy **saw the fight** after school today.

(Traxler et al., 2002)

The self-paced reading study yielded an interaction between verb type and object type at the noun +1 position (*after*), with the longest reading times for the coercion condition, 3.8.a, and the an eye-tracking study found an interaction between verb type and object type in second-pass and total times at the noun position (*puzzle* / *fight*).

McElree et al. (2001) and (Traxler et al., 2002) interpret these results as evidence that there is indeed a difference between logical metonymies (marked with × in 3.7 and 3.8) and constructions that do not exhibit any type clash (marked with ✓ in 3.7 and 3.8): in short, coercion is costly, and more so than processing dispreferred (but still sensical) combinations. The additional cost is ascribed to the "introduction of additional

semantic structure" (McElree et al., 2001). I will return later to the successive debate about the source for the coercion cost (Chapters 8 and 9).

Additional costs for coercion cases are also reported by Pickering et al. (2005); Traxler et al. (2005); McElree et al. (2006b); Pylkkänen and McElree (2007); Kuperberg et al. (2010); Baggio et al. (2010, 2012). The source problem has largely been ignored by psycholinguistic studies, as none of the studies mentioned here provides direct evidence for or against qualia structures, contrasting a qualia-compatible interpretation with a non-qualia-compatible interpretation.

Corpus Analyses and Computational Modeling

If psycholinguistic and neurolinguistic studies have focused on additional costs for the coercion operation (the trigger problem), neglecting the source problem, quite the opposite has happened in computational linguistics, where corpus analyses and computational modeling (which can not provide data on on-line processing costs) have mostly neglected the trigger problem, usually limiting the scope of their analysis to the source problem.

Corpus studies (Briscoe et al., 1990; Verspoor, 1997a; Zarcone and Rüd, 2012) have shown that qualia cover a majority of logical metonymy interpretations from a corpus, although in some cases less default interpretations do arise (see Chapter 4 for a more detailed analysis of corpus analyses of logical metonymy).

Modeling studies, often with the intended goal of supporting NLP applications (Lapata and Lascarides, 2003; Lapata et al., 2003; Roberts and Harabagiu, 2011; Shutova and Teufel, 2009; Shutova et al., 2013) have mainly focused on the task of generating a range of possible interpretations for a given (potentially metonymic) verb + object construction (for example, generating *enjoy reading the book* and *enjoy writing the book* as possible interpretations of *enjoy the book*)[2]. While most of these studies have not addressed the question of the descriptive power of qualia events, Lapata et al. (2003) have employed qualia to categorize both the interpretations provided by the model and the events elicited by humans, showing that humans tend to agree on

[2] On the other hand, it is worth mentioning that the possible overgeneration of interpretations, such as generating interpretations for verb + object combinations where the object is of type event (which would be ruled out by the Lexical Hypothesis), is actually welcomed by such models, as paraphrases for non-metonymic cases (e.g., *enjoy the lecture* → *enjoy listening to the lecture*) "may be useful for some potential NLP applications" (Lapata and Lascarides, 2003).

the qualia role of the covert event for a logical metonymy, and that the model could successfully predict the qualia category of the covert event.

3.1.2. Strengths and Weaknesses of the Lexical Hypothesis

Despite its great explanatory power, the Lexical Hypothesis has been criticized from various sides. I will now summarize the most interesting points raised by its critics and by its advocates.

The Power of the Lexicon

The Lexical Hypothesis relies heavily on the lexicon and in particular on its generative power: logical metonymies are not a case of "compositionality fail", but rather a case of **enriched composition** (Jackendoff, 1997), a special case of compositionality. The Lexical Hypothesis maintains a traditional distinction between linguistic knowledge and world knowledge, while enriching the lexicon in order to maintain its compositional and generative power, that is its ability of generating a potentially infinite number of sensical sentences with a finite number of lexical items and subsenses. This mechanism allows for productivity in the GL, explaining creative uses of language, while establishing its boundaries, constraining the range of sensical sentences.

This aspect is the greatest strength of this hypothesis, as it is able to account for logical metonymy and many other phenomena within a purely linguistic domain, preserving the systematicity and generativity of the lexicon, by empowering it with rich but still concise and compact representations.

The "Anomaly" of Logical Metonymy?

On the other hand, defining logical metonymies in terms of type clash has the consequence of conceptualizing them as somehow "anomalous". Interestingly, some experimental settings compare logical metonymies with strong violations of type restrictions (e.g. *The journalist began / wrote / astonished the article*, Pylkkänen and McElree, 2007; Kuperberg et al., 2010; Baggio et al., 2010) — which is questionable, since logical metonymies, despite differing from "neutral" contexts for the presence of a covert event, are indeed grammatical and sensical constructions. Comparing metonymic constructions with ungrammatical or nonsensical constructions indicates

at least an implicit prejudicial categorization of logical metonymy as outlandish or marginally acceptable, and ultimately seems to do little for identifying how they are placed within incremental, efficient, fast (and successful) language processing.

A similar instance is often taken with regard to figurative language[3], resulting in a bizarre conundrum: metaphors (e.g. *They are not talkative: these counselors are carps*, Pynte et al., 1996) and metonymies (e.g. *The espresso wanted to pay*, Schumacher and Weiland, 2011) can not be interpreted literally, and should therefore be considered anomalous, but are nevertheless pervasively used in language and are not perceived as anomalous, even more if accompanied by a supporting context (Hahn and Markert, 1997; Wilson and Sperber, 2004).

The Limits of the Qualia Structure

Qualia structures are a formalization of a very reasonable assumption, that is the idea that lexical items referring to entities are associated with some sort of event knowledge, which plays a role in processing and interpretation. On the other hand, the Lexical Hypothesis and in particular qualia structures have been criticized from various sides as rigid and limited (Fodor and Lepore, 1998; Lascarides and Copestake, 1998; Blutner, 2002; Carston, 2002; Egg, 2005; Asher, 2011; Zarcone et al., 2014): the original definitions of telic quale and agentive quale (respectively, the mode of creation or the purpose of an entity) underestimate the range of possible covert events to two at the most, and are not dynamic enough to account for context effects.

Fodor and Lepore (1998) notice that the GL wrongly predicts sentences like 3.9.a-b to be ill-formed, as rocks, according to their complex lexical representation, are not artifacts and therefore lack a telic quale ("not all lexical items carry a value for each qualia role", Pustejovsky, 1995, p. 76):

(3.9) a. **?** Mary **enjoyed that rock**.

b. ✓ The climber **enjoyed that rock**.

Fodor and Lepore (1998) notice that "given a clue" (3.9.b: that is, given sufficient supportive context), the oddness vanishes. In his reply to them, Pustejovsky (1998) argues that "exceptions" such as 3.9.b are possible, but oddly enough he ascribes also

[3] See Gibbs, 1994; Bambini and Resta, 2012 for a discussion of direct vs. indirect access hypotheses regarding figurative language.

this process to the lexicon (namely, to *enjoy* imposing a new telic quale on *rock*), the same lexicon that did not license that interpretation for 3.9.a in the first place. The problem of the restrictiveness of the Lexical Hypothesis is also analyzed by Blutner (2002) and Asher (2011) in similar terms.

Lascarides and Copestake (1998) argue that some pragmatic inference is needed to account for interpretations that go beyond the qualia structure, and point out that the covert event is often determined by the subject filler or requires the integration of wider discourse-derived contextual information. Consider the following examples:

(3.10) a. Jack Kerouac **began the book** around 1949 in New York. →writing

b. I found *On the Road* in a second-hand bookstore. I **began the book** as soon as I got home. →reading

c. The goat **began Jack Kerouac's book**. →eating

d. Jack Kerouac was an amateur wrestler. He always **enjoyed a good match**. → fighting

e. Jack Kerouac's father Leo was a wrestling fan. He always **enjoyed a good match**. → watching

A different covert event is retrieved in 3.10.a-c and 3.10.d-e depending on context (here, the agent). Qualia structures alone do not allow for this flexibility: *eating* and *watching* can hardly be considered the agentive or telic quale of *book* and *match* respectively. Also, *match*, being an event-denoting object, should not trigger a covert event interpretation according to the Lexical Hypothesis, since no type clash occurs with the main verb, and the agentive and telic role are not defined for entities which do not denote artifacts (Zarcone et al., 2014).

The account I am going to propose (Section 3.4) overcomes the rigidity of qualia structure approaches by conceptualizing the covert event interpretation for a logical metonymy as a ranked set of plausible interpretations, whose ranking can be influenced by context.

3.2. The Pragmatic Hypothesis

An alternative account of logical metonymy, which I will refer to as **the Pragmatic Hypothesis**, has been proposed, building on many of the observations in 3.1.2. Las-

carides and Copestake (1998) have argued for pragmatic inferences to generate broader discourse-derived logical metonymy interpretations beyond those determined by the qualia structure. Fodor (1990) and Fodor and Lepore (1998) go as far as claiming that lexical items do not have an internal structure, but rather are atomistic representations that do not encode information beyond their own denotations: "the lexical entry for *dog* says that it refers to 'dogs'; the lexical entry for *boil* says that it refers to 'boiling'; and so forth . . . " (Fodor and Lepore, 1998, p. 54).

A more detailed alternative, stemming from the observations in Fodor and Lepore (1998) and Lascarides and Copestake (1998), is provided by de Almeida (2004) and de Almeida and Dwivedi (2008). According to their theory, logical metonymies are not anomalous, but rather they are underspecified: as we have seen in Chapter 1, many covert event interpretations are possible. Metonymic verbs (de Almeida and Dwivedi, 2008) are supposed to trigger presuppositions, not different than those triggered by other verbs as for example *regret*:

(3.11)
- *regret* → *there is an event that has previously been performed*
- *begin* → *there is an event that the subject begins to perform with the object*
- *enjoy* → *there is an event that the subject enjoys doing*

If the event is not explicitly mentioned, then the construction is underspecified and the event must be recovered from discourse context via general pragmatic principles (for example, those expressed by Relevance Theory, Sperber and Wilson, 1986; Carston, 2002) to interpret an otherwise underspecified construction.

Let us sum up what the Pragmatic Hypothesis consists of with regard to the two problems:

- **the trigger problem:** what triggers the logical metonymy?
  → the underspecification of metonymic expressions triggers post-lexical inferences;
- **the source problem:** what is the source of the covert event?
  → the covert event is recovered post-lexically from discourse context and non-lexical knowledge via general pragmatic principles.

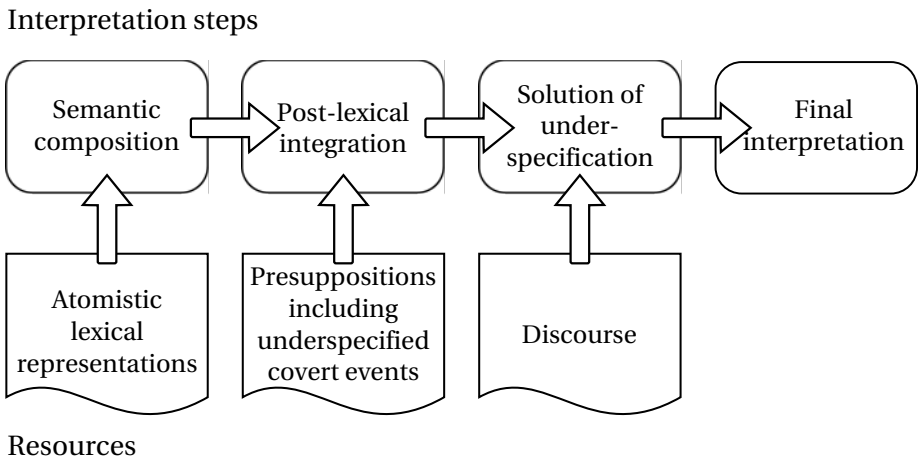

Figure 3.2.: Schematic representation of logical metonymy interpretation for the Pragmatic Hypothesis.

3.2.1. Evidence in Support of the Pragmatic Hypothesis

Psycholinguistic and Neurolinguistic Studies

De Almeida (2004) suggests that the effects reported in McElree et al. (2001) and (Traxler et al., 2002) may also be explained in terms of later (post-lexical) interpretive processes, akin to those assumed by the Pragmatic Hypothesis. Adding context information may have the effect of (1) canceling the type-shifting operation or (2) narrowing the range of covert events, in either case canceling the alleged costs for the coercion operation. If this were not the case, one would then have to assume that the cost is determined by a type-shifting operation.

De Almeida (2004) replicates the experiment in McElree et al. (2001), by adding supporting context to the material sentences:

(3.12) The author was always very busy. His editor asked him to review a book while he was working on his own novel during the summer.

a. ×The author **started the book** in his house on the island.

b. ✓ The author **wrote the book** in his house on the island.

c. ✓ The author **read the book** in his house on the island.

(de Almeida, 2004)[4]

[4]Interestingly, the supporting context for 3.12 makes both 3.12.b and 3.12.c very plausible interpreta-

A significant difference in self-paced reading times is reported only at the object position (*book*) between the preferred condition and the (significantly slower) dispreferred and coerced conditions, and no significant differences are found at the noun +1 position; thus, these results were considered supporting evidence *against* the existence of a coercion operation.

In lack of supporting context, the Pragmatic Hypothesis would predict later reading times or eye fixation correlates, in the light of the traditional distinction between lexical knowledge (activated first) and world knowledge (activated later in processing, see Bornkessel and Schlesewsky, 2006; Warren and McConnell, 2007).

3.2.2. Strengths and Weaknesses of the Pragmatic Hypothesis

A Matter of Communication

Logical metonymies easily comply with the Gricean maxims (Grice, 1975; Lapata and Lascarides, 2003), and in particular with the commandment coming from the Maxim of Manner (*be brief*) of avoiding unnecessary prolixity. There is then no such reason why I should say *I began reading Pride and Prejudice* rather than *I began Pride and Prejudice* if it is clear that I am not Jane Austen and I am in fact *reading* it. Thus, logical metonymies are not an anomaly in the language, but rather a way to make our communication quick and efficient. Placing logical metonymies back into communication is a great merit of the pragmatic approaches.

Open-end Covert Events

The Pragmatic Hypothesis overcomes some of the problems of the Lexical Hypothesis (rigidity of the qualia structure, lack of context-sensitivity), but lacks a concrete characterization of the type of knowledge involved in the retrieval of covert events and of the organization of such knowledge, which appears to be unconstrained. Without any constraint, a sentence like 3.13 should be perfectly acceptable:

(3.13) **?** John **enjoyed that doorstop**.
(Lascarides and Copestake, 1998)

tions of 3.12.a. Since this is not the case for other material sentences in the experiment, I do not believe this was intentional.

whereas the interpretation of 3.13 as *John enjoyed reading that doorstop* is odd even in presence of context where it is clear that a book is used as a doorstop. Also, doing away with the qualia results in a lack of systematicity, as shown in the following examples:

(3.14) a. She was **enjoying the first cigarette** of the day. → smoking

b. She was **enjoying the first coffee** of the day. → drinking

Despite the above-mentioned limits of the qualia structure, assuming such thing as a telic role does have the advantage of capturing some systematicity between 3.14-a and 3.14.b, and such systematicity is lost in the approach of the Pragmatic Hypothesis, resulting in a weaker theory which denies any role to conventionalization and does not provide a (falsifiable) mechanism to distinguish between feasible and unfeasible covert event interpretations for a given context.

3.3. Open Issues

I have provided a detailed overview of the two main (and opposite) hypotheses of logical metonymy interpretation, of their strength and weaknesses, and of supporting evidence for either one or the other. I will now go through two more problematic aspects that seem to be common to both approaches.

Problematic Verb Triggers

A blistering debate[5] has stemmed from experimental reports of results apparently pointing in opposite directions, leaving the reader with an unsatisfactory picture. While supporters of the type clash and type shift solution (Pickering et al., 2005; Traxler et al., 2005) reported effects at the noun +1 position on reading times and interpreted them as evidence for the accommodation of a type-shifting operation, supporters of the Pragmatic Hypothesis (de Almeida, 2004; de Almeida and Dwivedi, 2008) argued for a different interpretation of the same results as evidence for post-lexical inferential processes. Also, de Almeida (2004) reported results supporting the hypothesis that context, by narrowing down the range of covert events and solving the

[5] See de Almeida (2004); Pickering et al. (2005); Traxler et al. (2005); de Almeida and Dwivedi (2008) for the whole debate.

underspecification of logical metonymies, cancels the cost of the (alleged) coercion operation, but these results were criticized and contested by Traxler et al. (2005).

The differences emerging in this debate may not just lay in the interpretation of experimental results, but partially diverging results may be due to differences of experimental design and choice of materials. Regarding experimental materials, it is worth noticing that all the studies mentioned so far use as "metonymic" verbs a mix of aspectual and non-aspectual event-selecting verbs, without much discussion about the criteria of inclusion. The aspectual verbs are a fairly well defined class, sharing semantic commonalities (they refer to the "initiation, continuation or termination of an activity", Levin, 1993, p. 274) as well as similar syntactic behavior, that is sentential complement-taking properties and (for some of them, see the *begin-verbs* in Levin, 1993) a regular syntactic alternation in English (causative alternation):

(3.15) a. Jack Kerouac **began** the journey that would take him back and forth across America.

b. Jack Kerouac's journey **began** less than a few months after the annulment to his wife was final.

c. Jack **began** the sandwich after his discussion with Dean.

d. *Jack's sandwich **began** after his discussion with Dean.

Such shared behavior is among the reasons of interest for logical metonymy in the first place: the causative alternation (despite not being a necessary or sufficient condition for being an aspectual verb) shows that the same restrictions should be imposed by aspectual verbs (which refer to *activities*) both on their subjects in the intransitive construction (3.15.b) and on their objects in the transitive construction (3.15.a), that is they should be something that can be *initiated*, *continued* or *terminated* (events, see also Pustejovsky, 1991). This explains for example why 3.15.d violates a restriction and why 3.15.c may be conceptualized as a type clash.

Unsurprisingly, many of the critiques to the Lexical Hypothesis come from non-aspectual "metonymic verbs", such as *enjoy* or *want* (Fodor and Lepore, 1998), whose inclusion in the "family" of metonymic verbs seem far less obvious or intuitive:

(3.16) a. I **began my sandwich**. → eating

b. Dean and Marylou **wanted some of my sandwiches**.

If 3.16.a is about *beginning* something, and it can be easily argued that *sandwiches* are not events and can thus not be *begun*, it is less intuitive (and seems to be somehow less necessary) to argue that in 3.16.b *sandwiches* can't be *wanted*. The very definition of metonymic verbs is problematic, due to the lack of explicit criteria to define what should be included in this category. I will return to this issue in Section 7.1.

The Emptiness of the Lexicon

A second open issue concerns the old and thorny problem of the boundary between lexicon and world knowledge, which emerges in almost every argument supporting one or the other view on logical metonymy.

Fodor and Lepore (1998) claim that the GL wrongly predicts sentences like 3.17 to be ill-formed:

(3.17) **?** Mary began the rock.

As observed in Section 3.1.2, 3.17 is odd because rocks are not artifacts and therefore lack a telic quale, but Fodor and Lepore (1998) notice that given sufficient supportive context the sentence can be made acceptable. While Pustejovsky (1998) concedes that discourse context can still influence interpretation, Fodor and Lepore (1998) claim that, in a theory that argues for an information-rich lexicon, the conditions for well-formedness should only depend on the lexicon, and not on discourse: if such ill-formed sentences can be understood only with extra lexical information (context or world knowledge), then constraints for semantic well-formedness should not be placed in the lexicon at all. Fodor and Lepore (1998) argue against the necessities for complex meanings in the sense of the GL and for a different type of complex lexical entries, composed of atomistic (purely denotational) lexical meanings and specific composition rules (e.g. *want* → *want to have*), which determine the logical form of the phrases that the lexical entry is contributing to form. Interestingly, the argument here is about two opposite views on the lexicon: If the GL aimed at explaining creativity of use and regular sense alternations by relying on a small set of powerful tools, and loading the lexicon with more information that it was traditionally considered to contain, Fodor and Lepore (1998) strip it back down to atomistic meanings and composition rules, leaving it to the later integration of extra-lexical context (and not to the lexicon) to specify the covert event (ultimately, a matter of discourse or world

knowledge: "a thing about the world, not a thing about the words", Fodor and Lepore, 1998).

Carston (2002), Egg (2005) and Asher (2011) also acknowledge that a good part of the problem is about where to place the systematicity that is typical of logical metonymy, whether in the semantics or in the conceptual knowledge that is within the domain of pragmatic inferential processes. Egg (2005) suspects that much of the information that the GL places in the lexicon should rather be included in generic conceptual knowledge and not modeled as part of the linguistic (lexical) knowledge, while Asher (2011) argues against resorting to "contingent factual information about what the world is actually like" (p. 16), and Carston (2002) concedes that "it may be that there is no general answer and that it varies from case to case" (p. 375).

3.4. The Words-as-cues Hypothesis

I have argued in 1.2 that logical metonymies are a potential challenge to compositionality, that they show at least some degree of regularity and systematicity, and that they are sensitive to discourse context and intra-sentential context. Ideally, a theory of logical metonymy should be able to capture these aspects of the phenomenon.

An extremely problematic point seems to be the type of event knowledge involved in logical metonymy interpretation, and in particular whether it should be ascribed to the lexicon or to world knowledge. I have argued that strictly lexicalist accounts are too weak to explain either the full range of covert events, or their sensitivity to discourse and context. On the other hand, pragmatic approaches, by appealing to general communicative inference underestimate the role of conventional and structured lexical information.

My investigation searches for a third way, inspired by the "words-as-cues" proposal by Elman (2009, 2011) and guided by the hypothesis (**the Words-as-cues Hypothesis**) that covert events can be explained by general mechanisms of event knowledge cued by lexical items during online sentence processing. I will now sketch this third proposal.

3.4.1. Words and Scenarios

Constraint-based and probabilistic approaches to language processing (e.g. Altmann, 1999; Altmann and Kamide, 1999; Elman et al., 2005; MacDonald et al., 1994; McRae et al., 1998; Trueswell et al., 1994) have underlined the role of incrementality in language processing: speakers use syntactic, lexical, semantic and pragmatic information at each point in processing to reach a provisional analysis and consequently build expectations about upcoming linguistic input.

More recently, extensive experimental evidence has been collected (see McRae and Matsuki, 2009; Elman, 2011, for detailed reviews) supporting the hypothesis that speakers, along with these sources of information, also use rich knowledge about common events and their participants, acquired from our first and secondhand experience: we are active event participants (for example *washing a car*, *ordering a meal at a restaurant*, *going to the doctor*...), we see other people do the same (in real life or on television or on the internet), we hear them talk about these events, or we read about these events. These studies argue that this vast amount of information is stored in memory and readily accessible not as a detailed memory of a specific event but rather in a generalized, prototypical form (**generalized event knowledge**, McRae and Matsuki, 2009), similar in spirit to the older idea of event schemata (e.g. Rumelhart, 1975, 1980; Rumelhart et al., 1986), frames (e.g. Minsky, 1975), and scripts (e.g. Schank and Abelson, 1977). For example, we know that *washing hair* typically includes *shampoo* and *a bathroom*; *washing a car* would involve a different scenario, an *outdoor* environment, *a hose* (Matsuki et al., 2011). The generalized event knowledge we share about events and their participants can be cued by linguistic input (Ferretti et al., 2001; McRae et al., 2005; Bicknell et al., 2010; Matsuki et al., 2011), allowing us to effectively anticipate typical upcoming input and to process it more quickly than less typical input:

(3.18) a. **The journalist** / the mechanic checked the spelling of his latest report.

b. The journalist / **the mechanic** checked the brakes on the car.

(Bicknell et al., 2010)

In 3.18.a, participants in a self-paced reading experiment would read *the spelling* faster in a context involving a *journalist* than in a context involving a *mechanic*, while the opposite would happen for *brakes* (3.18.b). Also, words rapidly combine in

sentences to cue specific scenarios and to drive expectations about upcoming input which is relevant to those scenarios (*check* and *journalist* would not cue *spelling* if presented in isolation, but they do when presented in combination) and for syntactic structure (Hare et al., 2009a).

Such anticipatory effects, reflecting the expectations that drive language processing, are not just general "semantic association" effects. Firstly, they are not binary associations, since different cues can be combined (and expectations updated accordingly). Secondly, the elements (*journalist, spelling, mechanic, brakes*) are associated with the two scenarios of *check* (*check spelling* and *check brakes*) in a more narrow sense than what is commonly understood as semantic association: "verbs, event nouns, agents, patients, instruments, and locations used in the experiments presented [...] are indeed associated in the general sense, but these associations are driven by people's knowledge of common events" (McRae and Matsuki, 2009). Lastly, the expectations are mediated by an event template, and structured along the slots provided by its arguments: given an argument slot for a verb, preceding context will build expectations for the filler most fitting with our generalized knowledge of the event.

Generalized event knowledge is shared to a great extent by people from the same linguistic and cultural background: Matsuki et al. (2011) point out that sentences about *cleaning miniatures on the shelf* or *trapping a large goose* are not highly likely to refer to events that the average undergraduate student (who takes part in psycholinguistic experiments) is typically familiar with. If generalized event knowledge is so pervasively employed in language understanding, careful attention must be paid to the choice of materials for reading studies.

3.4.2. Plausibility vs. Typicality

Paying attention to the choice of materials in reading studies involves controlling for *typicality* as well as for *plausibility*. It is very common that the materials in psycholinguistic studies are controlled for plausibility, but typicality is different than high plausibility (for example, both very typical and very plausible sentences can obtain high plausibility ratings). The predictability of a stimulus in sentence contexts is a better correlate of typicality, and a common method to estimate it is the **cloze completion task** (Taylor, 1953), in which speakers are asked to fill in a gap in a sentence or to finish an incomplete sentence and the completions are then used as an estimate

of a word's predictability (the more a word is mentioned as a completion, the more expected / predictable it is).

Matsuki et al. (2011) and Smith and Levy (2013), among others, argued against using cloze completions as a measure of predictability, as they provide only limited information about what is expected in the cloze, and are not a good estimate of small differences in absolute predictability. Event-based production norms (elicitation) were then suggested as a better way to tap into our knowledge of what is highly typical (and thus predictable) and what is less typical, albeit still plausible (Matsuki et al., 2011): we can think of plausibility as a range spanning from *highly anomalous* to *extremely predictable*, where "plausibility ratings nicely capture differences at the lower end: the ratings show substantial differences between anomalous and implausible items" but event-based production norms (such as those employed when constructing our materials in all the psycholinguistic experiments in this dissertation) "sensitively capture differences in the upper part of the continuum where plausibility ratings appear to do so only weakly" (Matsuki et al., 2011, p. 925).

3.4.3. A Words-as-cues Framework

Converging experimental work (reported above) has made the case for a theory on generalized event knowledge and its influence on language processing, which is ultimately a theory on conceptual knowledge. Building on this, but also on expectation-based models of language processing (e.g. Altmann, 1999; McRae et al., 1998; Trueswell et al., 1994), and on dynamical models of cognition (Spivey, 2007; Tabor and Tanenhaus, 2001), Elman (2009, 2011) has suggested a new theory of the lexicon, that we will refer to as the Words-as-cues framework. Elman re-conceptualizes the mental lexicon as a dynamical system where interpretation is modulated incrementally. Words act as cues to meaning and to event knowledge, rather than being semantically meaningful in themselves (see also a similar idea in Rumelhart, 1979).

In the dynamical system proposed, words interact in real time, producing an incremental interpretation and building expectations about upcoming input matching that interpretation. Words are seen not as mental objects residing in a lexicon, but rather as "stimuli that alter mental states". As an example of a fitting computational model for this system, Elman proposes his Simple Recurrent Network (Elman, 1990), a neural network where the internal state (the hidden layer) corresponds to the "mental

state" of the system, varying at each point in time depending on the stimulus (a word) and producing a representation of the next expected word (for example, after the verb *arrest* it might produce something like *the thief*). Interestingly, the computational model proposed by Elman shares some interesting features (projection layers, backpropagation, task of next word prediction) with deep learning approaches in NLP, which have been claimed to be more powerful than count vectors in successfully distinguishing word senses (e.g. *fire projectile* vs. *fire employee*) and have recently gained popularity in distributional semantics (Schwenk, 2007; Schwenk and Koehn, 2008; Mikolov et al., 2013; Schütze, 2013; Yin and Schütze, 2013).

Elman's proposal, along with the work it is inspired by (Hagoort et al., 2004; McRae and Matsuki, 2009), challenges the very distinction between lexicon and world knowledge: the meaning of a word is "rooted in our knowledge of both the material and the social world" (Elman, 2011). Elman acknowledges that "eliminating the lexicon is indeed radical surgery" (Elman, 2011), but strives nevertheless for a way of representing lexical knowledge without a lexicon: not a lexicon in the narrow linguistic sense, but rich and context-sensitive lexical knowledge stored in memory as a dynamic system.

3.4.4. Logical Metonymy in a Words-as-cues Framework

The Words-as-cues framework is built on solid experimental ground, as experimental studies on generalized event knowledge have employed a range of interesting linguistic constructions (predicate-argument composition, Bicknell et al., 2010, verb combination with agent, patient, instruments and locations, Ferretti et al., 2001; McRae et al., 2005; Matsuki et al., 2011, grammatical aspect, Ferretti et al., 2007, causative vs. inchoative constructions, Hare et al., 2009a), but effects of generalized event knowledge on logical metonymy and covert event interpretation had not been investigated before the work reported in this dissertation. Can the Words-as-cues framework provide a satisfactory account of logical metonymy interpretation? This account presents itself as a somehow intermediate position between the Lexical and the Pragmatic Hypothesis: I will now explain what it would entail and what predictions it would lead to.

Like the Lexical Hypothesis, the Words-as-cues Hypothesis links objects to associated events (sometimes leading to overlapping predictions as to what covert event is retrieved). For example, we know that *books* are typically *read*, therefore we associate

them with generalized knowledge about events of *reading*. Nevertheless, like the Pragmatic Hypothesis, the Words-as-cues Hypothesis acknowledges the need for a broader set of covert events, than go beyond those included in the classic Pustejovsky's qualia (1991; 1995; 1998) and are arguably part of our knowledge of typical event scenarios involving objects. For example, we know that *pizzas* are frequently *delivered* and *apples* are *peeled*, but those events would not be comprised in the qualia representations.

Unlike the Pragmatic Hypothesis, the Words-as-cues Hypothesis considers covert events to be determined by rich lexical information about event knowledge[6], and makes this information available for early integration during processing, claiming that the information needed to choose the covert event is not accessed via general communicative inference mechanisms (and therefore is not delayed). Also, the Words-as-cues Hypothesis places logical metonymy in a dynamic system of incremental integration of contextual cues, allowing for contextual influence and for integration of a wide range of information sources, challenging a coarse-grained, radical distinction between linguistic and world knowledge, as well between lexical and pragmatic information.

Another important question concerns the trigger of the logical metonymy. Comparing logical metonymies to type-restriction violations is problematic, because logical metonymies are widely used in communication, so they can not be just an anomaly. Work on selectional preferences (Wilks, 1975; Resnik, 1996; Ferretti et al., 2001) has shown that a verb's selectional behavior is better captured by a graded notion of preference, rather than with binary constraints. Selectional restrictions allow us to say that *eat the chair* is a nonsensical combination, since chairs are not +edible, but what about typical objects of a verb like *arrest*? Assuming a binary restriction on +arrestable objects would not take us very far, whereas our expectation on a possible object of *arrest* would be guided by what we know to be the verb's selectional *preferences*: *thieves* are more likely to be arrested than *policemen*, although the opposite is still possible. Extensive experimental work has shown that generalized event knowledge shapes those preferences and determines what fillers are best (or have a better thematic fit) for a given argument structure (Wilks, 1975; McRae et al., 1998).

Coming back to our trigger problem:

[6]Recall that, despite rejecting a traditional view of the lexicon, the Words-as-cues Hypothesis allows for rich lexical information.

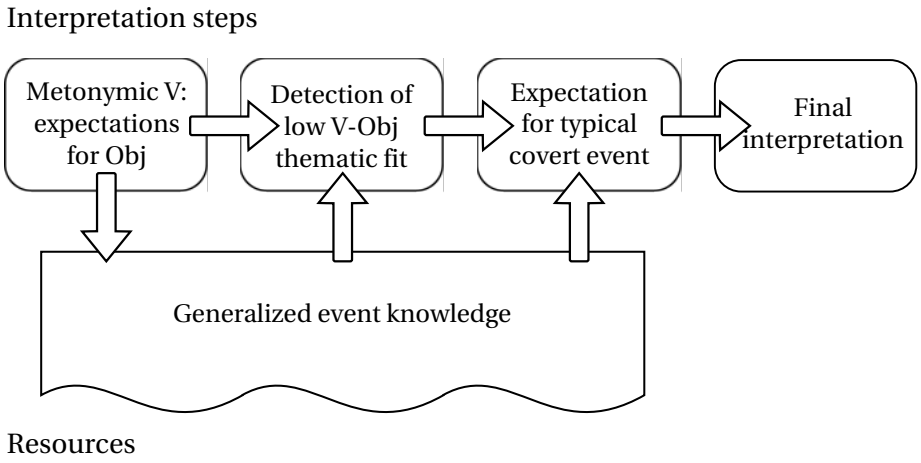

Figure 3.3.: Schematic representation of logical metonymy interpretation for the Words-as-cues Hypothesis, for cases in which the object is presented after the verb. If the object precedes the verb, then the low-thematic fit is detected at the verb region.

(3.19) a. Jack Kerouac **began the journey** that would take him back and forth across America.

b. Jack Kerouac **began the book** around 1949 in New York.

It would not be surprising if *begin* had a *preference* for event-denoting objects such as *journey*. Entity-denoting objects, rather than clashing with the selectional *restriction* of the metonymic verb, would then only be a less ideal match for the verb's selectional *preferences*. Expressing the trigger problem in terms of selectional preferences (informed by thematic fit) has the advantage of (1) using a single mechanism (thematic fit) to account for both problems (the trigger problem and the source problem, and (2) relaxing the strong constraints imposed by the Lexical Hypothesis while maintaining some of its predictive power.

Summing up what the Words-as-cues Hypothesis consists of with regard to the two problems:

- **the trigger problem:** what triggers the logical metonymy?
  → low thematic fit between an event-selecting verb and an entity-denoting object;

- **the source problem:** what is the source of the covert event?
  → the covert event with best thematic fit is recovered from generalized knowledge about events.

3.4.5. Research Questions and Experiment Plan

From these observations, a set of research questions arises, which my investigation will focus on:

1. **the source problem:**
   Different accounts of logical metonymy have either ascribed covert event interpretation to complex lexical entries or to world knowledge and discourse context. I argue for a typicality-based approach, suggesting a ranked set of interpretations (high-thematic-fit events) influenced by context and informed by generalized event knowledge. This raises the following research questions:
   a) Can qualia structures provide a satisfactory account for the range of covert events or do we need a different approach?
   b) Can a model of generalized event knowledge account for the retrieved covert events?
   c) Are covert events part of lexical or non-lexical knowledge?
   d) Do the retrieved covert events for logical metonymies fall in a different range than the explicit event in long forms (e.g. *begin the book* vs. *begin wrapping the book*)?
   e) Can a computational model of thematic fit account for the retrieved covert events?

2. **the trigger problem:**
   Logical metonymies were traditionally defined as the result of a type clash, resulting in a type shift and in a covert event interpretation, and type clash itself was used to distinguish metonymies from non-metonymic constructions (*begin the book* vs. *begin the fight*). I argue that it is thematic fit and not type clash which triggers covert event interpretation, raising the following research questions:

a) Are covert event interpretations only possible for entity-denoting objects or also for event-denoting objects?

b) What happens if the object is sortally ambiguous between an entity- and an event-denoting reading?

c) Is the event-selecting behavior enough to define a class of metonymic verbs?

d) Can a computational model based only on thematic fit distinguish between metonymic contexts and non-metonymic contexts, doing without a notion of type clash?

e) Can thematic fit only determine the cost for the coercion operation, without resorting to type clash?

I have addressed these research questions while trying to keep my feet on as solid experimental ground as possible, following the interdisciplinary approach described in Chapter 2. Elman suggested a neural network similar to his Simple Recurrent Network for a representation of "lexical knowledge without a lexicon" (Elman, 2011), I propose instead a structured DSM, in order to capture the idea that generalized event knowledge is not just plain semantic association, but rather stems from imposing a thematic structure to the event. I have used psycholinguistic experiments that manipulate thematic fit as an experimental condition, in order to investigate whether generalized event knowledge could affect covert event interpretation, and a structured DSM of thematic fit, in order to evaluate if a thematic fit model, lacking any explicit information about type, can successfully predict the correct covert event, and, when tested on known psycholinguistic datasets, can efficiently model coercion effects.

The following studies will be reported in the following chapters:

1. **the source problem** (Part II):
   - a corpus study (Rüd and Zarcone, 2011; Zarcone and Rüd, 2012) looked for instances of metonymic verbs in a corpus in order to address research questions (1a) and (1d) and to study the range of covert event interpretations (Chapter 4, Section 4.1);
   - a crowdsourcing study (Zarcone and Padó, 2010) tackled research questions (1a), (2a) and (2b) by asking participants to elicit covert event inter-

pretations for a set of entity-denoting, event-denoting and event/entity-ambiguous objects and analyzed the range of elicited events (Chapter 4, Section 4.2);

- a psycholinguistic study (one self-paced reading experiment and two probe recognition experiments, Zarcone and Padó, 2011; Zarcone et al., 2012b, 2014) investigated the role of typicality and generalized event knowledge on reading times and reaction times for covert event retrieval, providing an answer to (1b) and offering insights into (1c) (Chapter 5, Sections 5.2, 5.3 and 5.4);
- a similarity-based model (Zarcone et al., 2012d) successfully modeled covert event retrieval using thematic fit information, providing further evidence to our answer to research question (1b) and addressing (1e) (Chapter 6, Section 6.3);

2. **the trigger problem** (Part III):
   - a hybrid computational model of type and thematic fit (Utt et al., 2013) addressed research question (2c) by assessing to what extent some transitive verbs prefer event-denoting objects over entity-denoting objects ("event-hood") and by using eventhood to distinguish between metonymic and non-metonymic verbs (Chapter 7, Section 7.2);
   - another modeling study (Zarcone et al., 2013) analyzed thematic fit values computed from the similarity-based model for test sentences from behavioral studies on coercion, and showed that thematic fit can successfully distinguish the coercion condition from non-coercive conditions (2d); the study suggested that the cost of the coercion operation (2e) may indeed only be ascribed to thematic fit only (Chapter 7, Section 7.4);
   - finally, a last self-paced reading study (Zarcone and Padó, 2013) tackled research question (2e) more directly, showing that not only thematic fit but also type plays a role in determining processing costs for logical metonymies (Chapter 8).

Part II.

The Source of the Covert Event

4. The Range of Covert Events: Usage

An important step in logical metonymy interpretation is the recovery of a covert event (*Jack Kerouack began the book → writing*). Chapter 3 has shown that the recovery of the covert event has been explained either by positing complex lexical entries (**Lexical Hypothesis**, Pustejovsky, 1991, 1995) or through the integration of the covert event via post-lexical inferences (**Pragmatic Hypothesis**, Fodor and Lepore, 1998; de Almeida and Dwivedi, 2008). Both these approaches have advantages and shortcomings when it comes to generating a satisfactory subset of covert events and to accounting for the role played by context.

In this chapter I will present a first pilot data analysis highlighting the limits of a rigid qualia structure hypothesis with two studies: a corpus study, exploring the usage of logical metonymies in naturally-occurring text, and a crowdsourcing study, eliciting covert events for logical metonymies. I will show that the qualia structure, while providing a good generalization for typical events in default-cases, does not account for the wider range of events elicited, both when no extra-sentential context is provided (as in the crowdsourcing study) and in a number of corpus-extracted cases where wider discourse-derived contextual information is required.

4.1. A Corpus Study of Logical Metonymy

The Lexical Hypothesis predicts that the interpretation of logical metonymies in naturally-occurring text should overlap with either the agentive or the telic quale of the object (*begin the book → reading / writing*), whereas non-default interpretations should be marginal and have to be licensed by discourse and supporting context (Pustejovsky, 1998).

The Pragmatic Hypothesis suggests that, as logical metonymies are part of our communication, they should comply with the Gricean Maxim of Manner *be brief* (Grice, 1975; Lapata and Lascarides, 2003), avoiding unnecessary prolixity. Thus,

verb	prog.	inf.	NP	EV-object	EN-object	non-qualia-based	% non-qualia-based
begin	1	57	11	8	3	2	67%
enjoy	6	0	59	21	25	4	16%
finish	8	0	31	8	23	6	26 %
miss	3	0	24	10	13	4	31 %
prefer	4	30	30	10	13	1	8 %
regret	2	0	17	14	0	0	0 %
start	45	28	63	42	21	0	0 %
						17% of EN-objects	

Table 4.1.: Corpus study on the LOB corpus (Briscoe et al., 1990): counts of event-denoting (EV) and entity-denoting (EN) objects, and percentages of non-qualia based interpretations.

when logical metonymies (where the event is implicit) are contrasted with long forms (where the event is explicit), the former should involve more obvious events whereas the latter should involve non-default interpretations (e.g. *begin the book* vs. *begin wrapping the book*).

Corpus studies of logical metonymy in use can show how often type clashes occur, and (if the logical metonymies are annotated with their covert events) what events are involved in the interpretation. Covert events can be analyzed with regard to their overlap with the qualia structure and can be contrasted with events in long forms.

Previous Work

Briscoe et al. (1990) looked for instances of metonymic verbs in the Lancaster - Oslo/ Bergen corpus (LOB, 1 million words, Johansson et al., 1978) and took into consideration seven verbs (*begin, enjoy, finish, miss, prefer, regret, start*, see Table 4.1). They reported the frequencies of their different subcategorization frames (progressive, infinitive or NP), they selected metonymic sentences (that is sentences with entity-denoting object fillers in the NP subcategorization frame)[1], and they observed that in 17 cases (17 %) the covert event was not retrieved from the qualia structure but was

[1]The count of entity- and event-denoting object sentences does not sum up to the total of NP contexts, as the authors were unable to classify some examples.

verb	EN-object	AQ	TQ	AQ + TQ	non-qualia-based	% non-qualia-based
begin	164	65	91	156	8	5%
finish	319	94	211	305	14	4%
begin on	25	4	5	9	16	64%
					7% of EN-objects	

Table 4.2.: Corpus study on the LOB and BNC corpora (Verspoor, 1997a,b).

solved pragmatically, because the immediate context supported the pragmatic inference needed to understand the metonymy. Cases where the event was explicit (the long form, that is progressive and infinitive subcategorization frames) often involved non-default predicates. Briscoe et al. (1990) suggest that a noun's qualia structure should contain typical (frequent) predicates associated with the noun, which are inherited in a lexical taxonomy (e.g. *Burgundy* should inherit the telic role from the hypernym *drink*), whereas encyclopedic (non-lexical) information should be less accessible in the inheritance network. They also argue that both should be recoverable semi-automatically from corpora, suggesting that the difference between the two may be closer to a gradient of accessibility rather than a clear-cut distinction.

Verspoor (1997a,b) carried out a similar analysis on fewer verbs but on a larger corpus, that is the British National corpus (BNC, 100 million words, Burnard, 1995) combined with the LOB. She collected sentences containing the verbs *begin, finish, begin on* followed by a noun phrase, selected metonymic sentences (sentences where the head of the noun phrase was an entity-denoting noun), and annotated them (for *finish*, only a sample of them), indicating if the covert event overlapped with the agentive quale of the noun (AQ), with its telic quale (TQ), or with neither (see Table 4.2). A majority of covert events for *begin* and *finish* were compatible with the qualia roles, while only a minority of cases required larger context information. On the other hand, context-based interpretations were much more frequent for *begin on*, leading to the conclusion that metonymic verbs show some idiosyncratic (lexically-determined) behavior: pragmatics determines the covert events for *begin on*, whereas the covert events for *begin* and *finish* are more likely to be retrieved via conventionalized access to the lexicon.

verb	tot	NP				VP			
		tot	artifacts		LM	tot	artifacts		LF
anfangen	5463	2571	111	2.0%	**112**	2892	446	8.2%	**472**
anfangen mit	4015	3691	337	8.4%	**350**	324	46	1.1%	**51**
aufhören	1223	13	–	–	**–**	1210	97	7.9%	**104**
aufhören mit	1223	1188	46	3.8%	**47**	35	5	0.4%	**5**
beenden	12014	12014	228	1.9%	**231**	–	–	–	**–**
beginnen	41288	30111	242	0.6%	**243**	11177	1058	2.6%	**1110**
beginnen mit	36853	34858	395	1.1%	**406**	1995	94	0.3%	**110**
genießen	20749	20477	1052	5.1%	**1272**	272	31	0.1%	**34**
4547 annotated sentences					**2661**				**1886**

Table 4.3.: Corpus study on the SDEWAC corpus (Rüd and Zarcone, 2011; Zarcone and Rüd, 2012): annotated sentences.

A Corpus Study of German Logical Metonymies

In order to confirm and expand previous work on English, we have performed a similar, more extensive corpus-based analysis of German metonymic verbs (*anfangen (mit) [start (with)], aufhören (mit) [stop (with)], beenden [finish], beginnen (mit) [begin (with)], genießen [enjoy]*, Rüd and Zarcone, 2011; Zarcone and Rüd, 2012) in the SDEWAC corpus (Faaß and Eckart, 2013, 880 million words[2]). We took into consideration both logical metonymies and long forms:

logical metonymies (verb + dependent NP), required semantic annotation:

e.g. *Raucher können mit bestem Blick über die Stadt ihre Zigaretten genießen.* → *rauchen*

"Smokers can enjoy their cigarettes with the best view of the city" → smoke

long forms (verb + dependent VP), where the event was explicitly realized:

e.g. *In dieser Zeit begann er, seine berühmten großformatigen Aquarelle zu malen.*

"In this period he began to paint his famous large-format watercolors"

Logical metonymies and long forms were extracted for each verb (Table 4.3). Only transitive sentences with an animate subject and an artifact-denoting object[3] were

[2]The corpus was parsed with the FSPAR parser (Schiehlen, 2004).

[3]Semi-automatic labeling of subjects and objects was based on GermaNet 5.1 (Kunze and Lemnitzer, 2002): GermaNet categories were grouped into four classes (humans, artifacts, natural entities and

considered for the logical metonymy set, leaving out cases of alternation (e.g. *Der Film begann* — "the movie began") and non-metonymical uses, and the same filter was applied to subjects and objects for the long forms. Note that, since more than one object can occur in a sentence (e.g. *Wir haben Kaffee und Kuchen genossen*, "we have enjoyed coffee and cake" → drinking and eating), the total number of annotated logical metonymies (LM) and long forms (LF) is higher than the number of sentences with artifact NPs.

The extracted sentences were annotated by two expert annotators (ANN1 and ANN2[4], Table 4.4): logical metonymies were annotated with a covert event paraphrase; each covert event and each subordinate event (for long forms) was labelled depending on its overlap with the object's qualia:

AQ: if the event corresponded to the agentive quale;

TQ: if the event corresponded to the telic quale;

OTHER: if the event did not correspond to either of them.

Two more tags were introduced for cases where it was problematic to find an appropriate paraphrase for the logical metonymy:

UNDET: if the agentive and telic quale of the object were unclear;

INSCTXT: if the sentence context was not sufficient to find a paraphrase.

A total of 1886 metonymies and 2661 long forms were annotated, with a substantial inter-annotator agreement both for logical metonymies (5 tags, Krippendorff's $\alpha = 0.6$, Krippendorff, 1980) and for long forms (3 tags, $\alpha = 0.71$). Following Vendler (1968) and Lapata and Lascarides (2003), we considered qualia roles to contain not one single predicate each, but rather prototypical concepts, bundling different predicates together (e.g. as the agentive quale of *Buch* ["book"] we considered also *veröffentlichen* ["publish"] and *herausbringen* ["release"]). The database of metonymies and long forms (German Logical Metonymy Database), labelled by the two annotators and reporting both annotations for each sentence, is publicly available for scientific research purposes[5].

events), and subjects and objects not included in GermaNet were annotated manually.

[4]Stefan Rüd and Niki Hoedoro, both native speakers of German.

[5]The database is available on my website: `http://www.ims.uni-stuttgart.de/institut/mitarbeiter/zarconaa/data/GLMDB.zip`.

		Qualia coverage for covert event interpretation in logical metonymies						
verb	tot		AQ	TQ	AQ + TQ	OTHER	INSCTXT	UNDET
anfangen	112	ANN1	60.7 %	20.5 %	81.3 %	3.6 %	12.5 %	2.7 %
		ANN2	53.6 %	41.1 %	94.6 %	5.4 %	– %	– %
anfangen mit	350	ANN1	18.3 %	34.3 %	52.6 %	22.9 %	22.6 %	2.0 %
		ANN2	25.4 %	62.6 %	88.0 %	9.7 %	2.3 %	– %
aufhören mit	47	ANN1	23.4 %	61.7 %	85.1 %	6.4 %	6.4 %	2.1 %
		ANN2	29.8 %	66.0 %	95.7 %	4.3 %	– %	– %
beenden	231	ANN1	51.1 %	35.9 %	87.0 %	5.2 %	7.8 %	– %
		ANN2	52.8 %	45.9 %	98.7 %	0.9 %	0.4 %	– %
beginnen	243	ANN1	88.1 %	5.8 %	93.8 %	3.3 %	2.9 %	– %
		ANN2	86.4 %	11.5 %	97.9 %	2.1 %	– %	– %
beginnen mit	406	ANN1	35.5 %	31.3 %	66.7 %	19.7 %	13.5 %	– %
		ANN2	39.9 %	49.5 %	89.4 %	9.9 %	0.7 %	0.0 %
genießen	1272	ANN1	– %	90.4 %	90.4 %	2.0 %	1.9 %	5.7 %
		ANN2	– %	96.5 %	96.5 %	2.9 %	0.6 %	– %

		Qualia coverage for events in long forms				
verb	tot		AQ	TQ	AQ + TQ	OTHER
anfangen	472	ANN1	40.0 %	21.4 %	61.4 %	38.6 %
		ANN2	41.3 %	28.8 %	70.1 %	29.9 %
anfangen mit	51	ANN1	– %	13.7 %	13.7 %	86.3 %
		ANN2	– %	43.1 %	43.1 %	56.9 %
aufhören	104	ANN1	27.9 %	37.5 %	65.4 %	34.6 %
		ANN2	26.0 %	48.1 %	74.0 %	26.0 %
aufhören mit	5	ANN1	– %	20.0 %	20.0 %	80.0 %
		ANN2	– %	60.0 %	60.0 %	40.0 %
beginnen	1110	ANN1	45.2 %	19.5 %	64.8 %	35.2 %
		ANN2	42.5 %	34.1 %	76.6 %	23.4 %
beginnen mit	110	ANN1	– %	20.9 %	20.9 %	79.1 %
		ANN2	0.9 %	80.0 %	80.9 %	19.1 %
genießen	34	ANN1	26.5 %	35.3 %	61.8 %	38.2 %
		ANN2	17.6 %	61.8 %	79.4 %	20.6 %

Table 4.4.: Corpus study on the SDEWAC corpus (Rüd and Zarcone, 2011; Zarcone and Rüd, 2012): qualia coverage for events in logical metonymies and long forms.

Similarly to the results in Briscoe et al. (1990) and Verspoor (1997a,b), the majority of logical metonymy interpretations fell into the range of qualia events (80-90% for some verbs), with much lower proportions of qualia events for *anfangen mit* and *beginnen mit* (between 50% and 70%).

We observed lexical differences between verbs (as in Verspoor, 1997a,b), suggesting lexical idiosyncrasies: some verbs are more compatible with qualia interpretations, whereas other tend to require non-qualia interpretations. *Aufhören* and *aufhören mit* had a very strong preference for the telic quale, *beenden* had a tendency for the agentive quale. *Genießen* matched the low number of non-qualia interpretations of *enjoy*, and had a default telic interpretation for logical metonymies, whereas when the interpretation was agentive then it was explicitly formulated in a long form:

(4.1) *Ich habe es wirklich genossen, diesen Film zu drehen wenn man von den Szenen absieht, die ich bis zur Hüfte im Sumpf zubringen musste.*
I really enjoyed making this film apart from the scenes I had to spend up to the hip in the swamp.

Beginnen and *anfangen* showed a strong preference for the agentive quale, whereas the telic quale interpretations (and the number of non-qualia events) increased for the corresponding constructions with *mit* (*anfangen mit* and *beginnen mit*). *Begin* and *begin on* in Verspoor (1997a,b) showed a preference for qualia-interpretations and for context interpretations respectively, and a similar contrast was yielded for *anfangen (mit)*, *beginnen (mit)* and *aufhören (mit)*.

Sweep (2012) carried out a comparative study of Dutch (*beëindigen [finish], beginnen [begin], eindigen [end], genieten van [enjoy]*) and German verbs (*anfangen [start], beend(ig)en [finish], beginnen [begin], enden [end], genießen [enjoy]*), on samples extracted from the DWDS-Kerncorpus and the ANW-corpus respectively (around one million words each). She supported our observation for *anfangen (mit)* and *beginnen (mit)* that prepositional objects yield more context-dependent interpretations, and extended it to Dutch verbs *beginnen (met)* and *beginnen (aan)*, and she observed that for both *genießen* and *genieten van* a telic interpretation is preferred.

While most covert event interpretations for *anfangen*, *aufhören mit*, *beenden*, *beginnen*, *genießen* were qualia events, their long form counterparts, where the event was explicit, yielded higher percentages of non-qualia interpretations, confirming the intuition in Briscoe et al. (1990) that logical metonymy is strongly related to Grice's conversational maxims (1975): qualia capture a basic/default interpretation, that we tend to omit in a logical metonymy; whereas if the event is a less typical one (or if the event does not correspond to the verb's lexical preferences as in example 4.1), we need to express it explicitly.

4.2. A Crowdsourcing Study of Logical Metonymy

The Lexical Hypothesis predicts that, given a logical metonymy, one or two default covert event interpretations should be elicited (agentive and telic quale), and non-default interpretations should not emerge in out of context conditions. Also, such covert events should be triggered by a type clash, and only be retrieved when a metonymic verb is combined with an entity-denoting object. I have started approaching the question of whether qualia structures can provide a satisfactory account for the range of covert events, and corpus studies have provided interesting first insights into the use of logical metonymies, but covert events are by definition not attested in the corpus.

We carried out a crowdsourcing experiment (Zarcone and Padó, 2010), to more directly address the question of the range of the elicited covert events, as well as two more research questions:

- Are covert event interpretations only possible for entity-denoting objects in logical metonymy constructions, or also for event-denoting objects?
- What happens if the object is sortally ambiguous between an entity- and an event-denoting reading?

Studies on logical metonymy have often included offline norming studies, either to estimate the plausibilities for given covert event interpretations in a metonymical construction (Lapata and Lascarides, 2003) or to elicit a covert event in a cloze completion task (McElree et al., 2001; Lapata et al., 2003), but did not explore differences between metonymic and non-metonymic interpretations and limited the range of elicitations to only one event. In our crowdsourcing study participants were presented with out-of-context sentences (e.g. *Kate began the newspaper*), combining metonymic and non-metonymic verbs and event-denoting, entity-denoting and entity/event ambiguous nouns and were asked whether a cover event interpretation was necessary for a logical metonymy; if a covert event was necessary, they were asked to provide an appropriate covert event for the metonymy; also, more than one elicitation was possible.

Materials and Design 10 triplets of event-denoting, entity-denoting and entity/event ambiguous English nouns (EV, EN, EN/EV) were combined with a metonymic verb

	EN		EN/EV		EV	
	meton.	non-meton.	meton.	non-meton.	meton.	non-meton.
% CE	63%	11%	39%	6%	18%	6%
% no-CE	37%	89%	61%	94%	82%	94%

Table 4.5.: Covert event (CE) and non-covert event (no-CE) answers in the crowdsourcing study.

(e.g. *begin*) and a non-metonymic verb (e.g. *spot*) to form 60 sentences (see Appendix A.1 for the complete list of sentences). The metonymic verbs were chosen among a set of verbs which have been argued to give rise to logical metonymy and have been used in experimental studies. The nouns in each triplet were matched for length, frequency (estimated from the Brown Corpus, Kučera and Francis, 1967) and for their co-occurrence frequency with the metonymic and the non-metonymic verb (from the UKWAC corpus, Baroni et al., 2009) used as an estimation of plausibility.

The 30 nouns (10 triplets x 3 nouns) were selected from a list of 100 nouns after a threefold **expert annotation study**, where three linguists were asked to annotate the nouns as EV-denoting, EN-denoting or EN/EV ambiguous. α for the selected noun triples was 0.71 (good agreement). Weighted α (Krippendorff, 1980), which incorporates the idea that EN vs. EV is a stronger disagreement than the disagreement between either one of the types vs. the ambiguous EN/EV type[6], was 0.79. A **non-expert annotation study** was performed to confirm the results from the expert annotation and to verify that the annotation did not change in the sentence context provided. 14 annotators from the US took part in the non-expert annotation study, which was delivered on the Crowdflower web platform, and yielded reasonably good agreement for a crowdsourcing experiment (weighted $\alpha = 0.52$).

Participants and Procedure The elicitation experiment was delivered using the crowdsourcing paradigm on the Crowdflower web platform. 15 participants from the US were asked to choose between a metonymic 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.

[6]A weight of 1 was assigned to the EN-EV disagreement and a weight of 0.5 to the EN-EN/EV disagreement and to the EV-EN/EV disagreement.

no additional activity, from now on CE — covert event — vs. no-CE), and, when answering *additional activity*, participants were asked to provide instances of possible activities. The instructions included two examples, but did not mention entity or event interpretations of nouns.

Results and Discussion Agreement among participants for CE/no-CE answers was rather low ($\alpha = 0.35$). Weighted α was 0.36 when excluding EN/EV ambiguous nouns, showing that the low agreement was not due to their presence. A possible explanation for low agreement may be the degree of conventionalization of some logical metonymies: for a highly conventionalized metonymy (e.g. *begin the newspaper*) we may not expect participants to feel that an additional activity is required, and our instructions (asking for additional activities) may not have prompted covert event interpretations for such cases. Nevertheless, our data show that participants do consistently provide covert events for a majority of cases with combinations such as *begin the newspaper* or *begin the breakfast*. Rather, the low agreement seems to be determined by combinations such as *enjoy the automobile* or *remember the brandy*, whose metonymic behavior is arguably less straightforward (see Table 4.6 and item-wise analysis below).

We also computed the majority vote for CE/no-CE answers and compared it with the predictions coming from the Lexical Hypothesis (CE answers for EN-denoting objects with metonymic verbs, no-CE answers for the other conditions), obtaining good agreement ($\alpha = 0.60$). A binomial logistic regression on the CE/no-CE answers[7] yielded a significant effect of the object type ($p < 0.001$), and of the verb type ($z = -8.322$; $p < 0.001$), with interaction ($p < 0.001$).

These results seem to confirm the prediction from Lexical Hypothesis that the object type determines a metonymic interpretation, but consider Table 4.5: 37% of metonymic-verb/EN-noun combinations did not elicit CEs, while 18% of metonymic-verb/EV-noun combinations did. EN nouns generally *tended* to show a strong majority of CE answers with metonymic verbs; EV nouns showed a strong majority of no-CE answers with metonymic and non-metonymic verbs, but exceptions emerged in an item-wise analysis (Table 4.6): for example, *enjoy the conference*, despite featuring an EV-object, did show a 24% of CE elicitations. Not all the non-metonymic verbs blocked CE interpretations (e.g. *remember the brandy*), and the behavior of EN/EV

[7] answer ~ obj_type * verb_type

V-N pair	condition	% CE	% no-CE
begin the newspaper	meton,EN	89%	11%
begin the breakfast	meton,EN/EV	81%	19%
enjoy the automobile	meton,EN	50%	50%
enjoy the translation	meton,EN/EV	39%	61%
remember the brandy	non-meton,EN	34%	66%
enjoy the conference	meton,EV	24%	76%
remember the revolt	non-meton,EV	10%	90%
remember the shower	non-meton,EN/EV	8%	92%
endure the shower	meton,EN/EV	7%	93%
endure the revolt	meton,EV	3%	97%
approve the automobile	non-meton,EN	–	100%
organize the breakfast	non-meton,EN/EV	–	100%
organize the afternoon	non-meton,EV	–	100%

Table 4.6.: Covert event (CE) and non-covert event (no-CE) answers for single items in the crowdsourcing study.

ambiguous nouns appeared to be highly lexically determined (contrast for example *begin the breakfast, enjoy the translation* and *endure the shower*).

As to the range of covert events, each participant elicited on average 1.4 covert events (range 1-6) per each verb-object combination. The number of events to be elicited was not limited by the interface, and eliciting only one event was rather common. Nevertheless, even when participants elicited not more than one event, a variety of different events per verb-object combination was still provided (average 3.2, range 1-7), often enough covering a broader set than the one given by the telic and agentive quale.

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

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	agentive quale	telic quale	non qualia-based
elicited CEs	542	132	162	248
(tokens)	%	24%	30%	46%
elicited CEs	205	31	25	149
(types)	%	15%	12%	73%

Table 4.7.: Covert events accounted for by a qualia-based theory vs. other covert events in the crowdsourcing study.

The set of elicited events form semantically coherent verb classes of plausible events (e.g. *portrait*: *paint, draw, sketch*). Among the elicited events there were also events which would be difficult to account for with the qualia structure of *portrait* even within a theory of extended qualia (Busa et al., 2001): *hang, model, review*. EV objects (e.g. *conference*) also elicited covert events (*enjoy attending/holding a conference*), and for EN/EV ambiguous objects like *harvest* both readings often gave rise to elicited events, including not only light verbs (*performing a translation*), which would be semantically largely transparent, but also full verbs (*reading / completing a translation*).

Table 4.7 reports on the amount of covert events which can be accounted for by a qualia-based theory. The annotation was performed by assigning an agentive quale and a telic quale to each noun and comparing them with the elicited covert events. We considered qualia as classes of meaning, in order to cover also synonyms of the assigned qualia, thus interpreting the set of events encompassed by the qualia structure in a fairly generous way. Almost half of the elicited CEs did not fall in either the agentive quale category or in the telic quale category.

The Lexical Hypothesis thus seems to capture a tendency in the data rather than predicting the participants' answers in every case: metonymic and non-metonymic interpretations emerged as a continuum of behaviors, rather than clear-cut separate categories, and covert events are elicited also for EV and EN/EV nouns. Ascribing the range of covert events to the qualia structure of the noun and limiting it to one or two events seems to be an unsatisfying solution, at least if the qualia are understood as specific verbs, rather than concepts or semantically coherent verb classes.

4.3. Beyond Qualia Roles

The picture emerging from these pilot analyses (the corpus study and the crowd-sourcing study) shows that qualia are a good model of covert events, but have a few shortcomings, mainly undergeneration and lack of context-dependence.

We have defined qualia roles in a rather broad sense, as we did not consider each qualia role as corresponding to a single predicate but to a bundle of semantically-related predicates. Nevertheless, even within a theory of extended qualia, there is a portion of covert events that can not be accounted for by the qualia structure. These constitute up to 20% of the interpretations for some verbs in the corpus study, where the annotators had to provide one (often context-determined) covert event interpretation per context, and almost 50% in the elicitation study, where sentences were presented out of context and participants were left free to come up with more than one or two possible covert event interpretations per context (and often did).

Also, metonymic verbs differ not only with regard to their tendency to give rise to a metonymic interpretation but also with regard to their qualia-based interpretations. Lastly, beyond a verb's lexical behavior in isolation, a mechanism to select one or another covert event depending on context is lacking. This aspect will be tackled in the next chapter, which reports on the psycholinguistic experiments addressing the source problem.

5. The Source of the Covert Event: Psycholinguistic Evidence

In this chapter I will address the problem of the **source** of the covert events and of their context-sensitivity. The corpus study and the crowdsourcing study in Chapter 4 have shown how the qualia structure, while providing a good generalization, suffers from a number of shortcomings. I will suggest that (as predicted by the **Words-as-cues Hypothesis**) covert events are better understood as part of generalized knowledge about events involving the subject and the object of the logical metonymy, and that when we understand logical metonymies we resort to our generalized event knowledge to predict typical covert events compatible with the preceding context.

I will then present three psycholinguistic experiments (a self-paced reading experiment — Experiment 1 — and two probe recognition experiments — Experiments 2 and 2b), providing evidence that intra-sentential elements (agent + patient) cue a covert event that matches the generalized event knowledge of the scenario that typically involves them, and thus that generalized event knowledge plays an important role in determining the covert event. The experiments also show that generalized knowledge about covert events is integrated early during processing and therefore is unlikely to be post-lexical.

5.1. Words as Cues to Covert Event Interpretation

It is reasonable to assume that we associate some typical event knowledge with entity-denoting lexical items, stored in our mental lexicon in a form that may be similar to a qualia structure (as predicted by the Lexical Hypothesis). However, the corpus study and the elicitation study (Chapter 4) have shown that a strictly lexicalist approach does not cover the whole range of covert event interpretations and does not account for the

integration of wider discourse-derived or intra-sentential contextual information. On the other hand, the Pragmatic Hypothesis, by appealing to general communicative inference and to world knowledge, underestimates the role of lexical information, as it does not explicitly predict what the range of the retrieved covert events should be.

Solid experimental evidence has shown how speakers make extensive use of rich knowledge about generalized events and their typical participants to predict plausible upcoming input (Becker, 1980; Altmann and Kamide, 1999; Kamide et al., 2003; van der Meer et al., 2005; McRae and Matsuki, 2009). The Words-as-cues Hypothesis (Elman, 2009, 2011) claims that this knowledge is easily accessible, and can be activated by (combinations of) linguistic cues, determining our expectations about the input and allowing us to process it more quickly when expectations are met. Experimental work shows that such cues are indeed integrated early:

(5.1) a. Donna used **the shampoo** to wash her filthy hair.

b. Donna used **the hose** to wash her filthy hair.

(Matsuki et al., 2011)

In Matsuki et al. (2011), for example, the instruments (*shampoo, hose*) narrow down the range of possible washing scenarios (and of the patients involved in those scenarios: *shampoo* + *wash* → *hair, hose* + *wash* → *car*), leading to a facilitation effect for the expected patient (*hair*) in 5.1.a compared to 5.1.b.

I suggest a Words-as-cues Hypothesis of covert event interpretation: covert events are retrieved via generalized event knowledge, which is activated by contextual cues such as the subject and object of a logical metonymy (e.g. *the writer began the book*): we use this knowledge (e.g. we know that writers *write* books) to infer the covert event. Similar to the Lexical Hypothesis, this account also assigns events to nouns but, rather than assigning a fixed set of events to each noun, it claims that contextual cues can activate knowledge from a wide range of information sources, often generating events that go beyond the classical qualia. For example, *car* can be associated with *fix*, but arguably this event does not refer to either the coming into being of cars (*build*) or their purpose (*drive*). Covert event interpretation is context-sensitive: different generalized event knowledge scenarios are activated if the context varies.

The Pragmatic Hypothesis, claiming that covert event interpretation is a post-lexical, pragmatic phenomenon, would predict later effects on reading times or eye fixations,

for example after the object region, for logical metonymies (assuming that lexical knowledge is activated first, while world knowledge intervenes later in processing, see for example Bornkessel and Schlesewsky, 2006; Warren and McConnell, 2007). Interestingly, the Lexical Hypothesis leads to overlapping predictions, ascribing late effects to the integration of an event structure in the sentence meaning. The opposite prediction follows from the Words-as-cues Hypothesis: if covert events are part of an enriched lexicon, informed by generalized event knowledge, then this knowledge should be integrated early, driving expectations about covert events and giving rise to early facilitation effects during processing of expected events.

5.2. Experiment 1

Experiment 1 (Zarcone and Padó, 2011; Zarcone et al., 2014) verifies the viability of a Words-as-cues Hypothesis of covert event interpretation. Previous studies employing self-paced reading (see Pylkkänen and McElree, 2006, for a review) have addressed the trigger problem, contrasting metonymic and non-metonymic sentences (*The journalist began / wrote / astonished the article*). Our study focuses on the source problem, investigating the role of sentential cues to determine the covert event of the object.

The Words-as-cues Hypothesis claims that sentential context (namely, the subject and the object of the metonymic verb) taps into the associated generalized event knowledge scenario, producing expectations for a covert event. These expectations are not measurable on a logical metonymy, where the event is implicit. We thus tested this hypothesis in a self-paced reading study, capitalizing on German word order and presenting the event explicitly in the final position:

(5.2) a. Der Konditor hörte mit der Glasur auf. (logical metonymy)
The baker finished the icing.

b. Der Konditor hörte auf, die Glasur **aufzutragen**. (long form)
The baker finished the icing **to spread**.

Our design is based on the observation that the covert event interpretation in a logical metonymy is guided by the same principles that determine expectations on the subordinate event in the corresponding paraphrase (long forms, see also Lapata et al., 2003; Zarcone and Padó, 2010).

The event (e.g. *auftragen* [*spread*], the same in both conditions) is in final position, crucially after inter-sentential cues that match the generalized event knowledge required to retrieve it (e.g. *baker* + *the icing*). The cues build up to the event in a high-typicality condition, whereas in the low-typicality condition they are chosen in such a way that the event is a possible but less typical event for the agent-patient pair (*child* + *icing*):

(5.3) Der Konditor / das Kind hörte auf, die Glasur **aufzutragen**.
The baker / the child finished the icing **to spread**.

Note that the high-typicality and the low-typicality condition are both well-formed and plausible, and only differ with regard to the typicality of the event given the agent and patient. We do not predict an inhibitory effect on low-typicality events, as they are not less plausible, but rather an **early facilitation effect** (at the target verb, similar to the effect reported in Matsuki et al., 2011), because generalized event knowledge about events and their typical participants in the high-typicality context generates **expectations** for the subordinate event in the high-typicality condition, making the target verb easier to read.

5.2.1. Method

Materials The materials for Experiment 1 were prepared using norming studies inspired by the procedure in Matsuki et al. (2011).

Norming Study 1 Thematic-based event generation norms were collected for 50 patient nouns on Amazon Mechanical Turk (AMT), asking participants to "list typical actions performed with these objects" (e.g. *what do you do with icing?* → *eat, spread, lick off*...). For each item, space was provided for 10 responses, and no time limit was imposed. Each item was presented to an average of 20 German participants. Participants were very productive, eliciting on average 7.8 events per item per participant. We chose four events for each item, from those named early by many participants (weighting method from Matsuki et al., 2011), ensuring that they all referred to different scenarios. We thus obtained 200 patient-event pairs (50 patients x 4 events).

Norming Study 2 In order to select agents that could cue the events, we collected thematic-based agent generation norms for the 200 patient-event pairs obtained after Norming Study 1. Participants on AMT were asked to "list who typically performs these actions" (e.g. *who spreads the icing?* → *the baker, the confectioner, the cook* ...). For each item, space was provided for 10 responses, ant no time limit was imposed. Each item was presented to an average of 10 German participants. Again participants were very productive (on average, 7 agents per item and per participant). For each patient-event pair, we selected four agents from those named early by many participants.

From the resulting 800 agent-event-patient triplets obtained after Norming Study 2 (50 patients x 4 events x 4 agents), we selected 24 patients (about half of the ones chosen initially) with 2 events each (half of the elicited ones), and per each event we selected one of the best agents, obtaining 48 high-typicality agent-event-patient triplets (24 patients x 2 events x 1 agent). 48 low-typicality triplets were obtained by crossing agents between the two events selected for each patient, as shown in Table 5.1. When selecting the triplets, we ensured that the agents assigned to the events in the low-typicality triplets were never elicited for that event-patient pair (for example, *Konditor* [*baker*] was not elicited for *Glasur* + *essen* [*icing* + *eat*]).

96 sentences were constructed from the 48+48 triplets by embedding them as verb-final subordinate sentences under metonymic main verbs. Similar to Lapata et al. (2003), we used German verbs equivalent to the metonymic verbs most commonly included in theoretical and experimental literature on logical metonymy for English (see Appendix A.2.1 for the complete list of sentences). The sentences continued after the subordinate event, in order to check for possible effects at a later region (up to three words after the target verb). The metonymic verb was the same among the four sentences that featured the same patient, and sentences sharing the same patient only differed in the agent and the subordinate event:

High-typicality:

(5.4) a. Der Konditor hörte auf, die Glasur **aufzutragen**, und fing mit den Pralinen an.
The baker finished the icing **to spread**, and began with the pralines.

b. Das Kind hörte auf, die Glasur **zu essen**, und fing mit den Pralinen an.
The child finished the icing **to eat**, and began with the pralines.

	agent	patient	event
high-typicality triplet	Konditor *baker*	Glasur *icing*	auftragen *spread*
	Kind *child*	Glasur *icing*	essen *eat*
low-typicality triplet	Kind *child*	Glasur *icing*	auftragen *spread*
	Konditor *baker*	Glasur *icing*	essen *eat*

Table 5.1.: Triplets for *Glasur* [*icing*].

Low-typicality:

(5.5) a. Das Kind hörte auf, die Glasur **aufzutragen**, und fing mit den Pralinen an.
The child finished the icing **to spread**, and began with the pralines.

b. Der Konditor hörte auf, die Glasur **zu essen**, und fing mit den Pralinen an.
The baker finished the icing **to eat**, and began with the pralines.

Norming Study 3 In order to check that the low-typicality triplets were, although not highly typical, still sensible and plausible and that they did not violate any selectional restriction, we collected plausibility ratings on AMT for our materials on a five-point Likert scale (no time limit was imposed). Participants (on average, 10 German participants per sentence) were presented with the 96 high- and low-typicality sentences (48 + 48), both in their paraphrased version (e.g. *Der Gast begann, das Schwein zu essen*, The guest began eating the pork) and as non-metonymic base sentences (e.g. *Der Gast aß das Schwein*, The guest ate the pork), along with 52 sentences with selectional restriction violations (nonsensical fillers: e.g. *Der Fisch fährt Fahrrad*, The fish rides the bicycle). The order of presentation was randomized.

The ratings yielded high agreement (Krippendorff's α for ordinal data = 0.73); high-typicality sentences yielded a mean rating of 4.12 (SD = 1.05) in their metonymic form and of 4.71 (SD = 0.72) in the non-metonymic base form; sentences in the low-typicality condition yielded a mean rating of 2.85 (SD = 2.93) in the metonymic form and of 2.61 (SD = 0.93) in their base form. Finally, nonsensical fillers yielded a mean

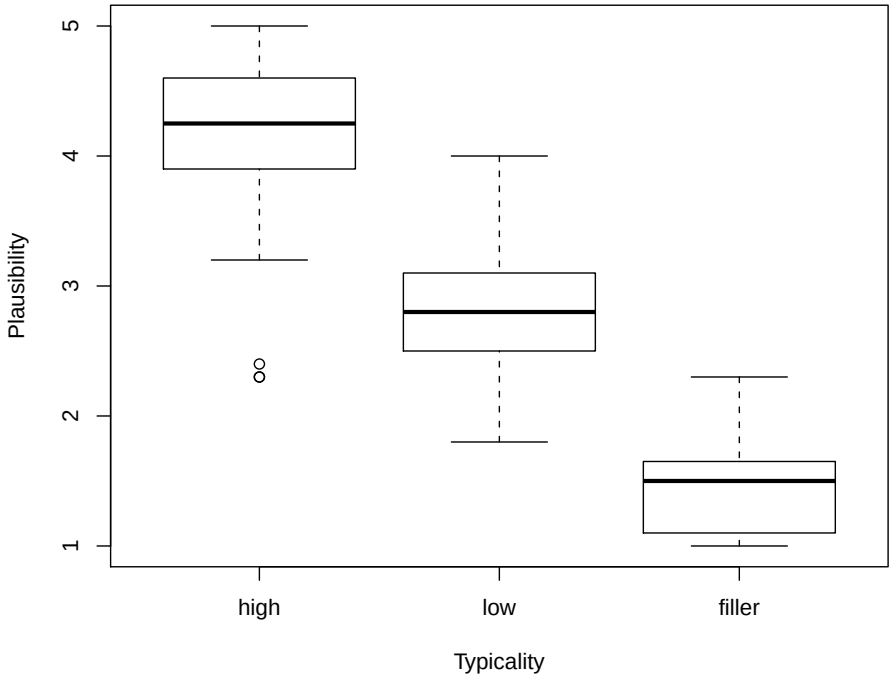

Figure 5.1.: Norming Study 3: Comparing plausibility ratings for high- and low-typicality test sentences and nonsensical fillers in Experiment 1.

rating of 1.44 (SD = 0.63). The plausibility scores for the low-typicality sentences were significantly higher than those for nonsensical fillers (Wilcoxon rank sum test: $W = 198448.5, p < 0.001$ for the base sentences and $W = 210052, p < 0.001$ for the metonymic sentences) and significantly lower than those for the high-typicality sentences ($W = 39767, p < 0.001$ for the base sentences and $W = 40981, p < 0.001$ for the metonymic sentences, see the box plot in Figure 5.1). Furthermore, the correlation between the ratings of metonymic and base sentences was significant (Spearman's $\rho = 0.8, p < 0.01$).

These results support our claims that (a) the low-typicality sentences do indeed differ in plausibility from the high-typicality ones as well as from the nonsensical fillers (they still make sense), and that (b) there is a strong link between the plausibility of base and metonymic sentences. We can then rule out the possibility that the typicality effect in Experiment 1 may be an effect of semantic anomaly of the low-typicality

condition. Note that both very typical and very plausible sentences can obtain high plausibility ratings, whereas the use of production norms allowed us to tap into the speakers' knowledge of what is highly typical (and thus predictable) and what is less typical, albeit still plausible (Matsuki et al., 2011, see also discussion in Section 3.4.2).

Procedure Two lists of 120 sentences each (24 high-typicality, 24 low-typicality, 72 filler sentences) were created to ensure that the same participant would not see the same agent-event-patient triple twice: for each group of four sentences sharing the same patient, the two high-typicality sentences were put in one list (to which half of the participants was assigned) and the two low-typicality ones in the other list (to which the other half of the participants was assigned). The sentences were presented to the participants with a one-word-at-a-time moving-window self-paced reading paradigm. Each trial began with strings of dashes on the screen, each dash replacing a non-space character of the sentence. Participants pressed a button to reveal the next word, simultaneously reverting the previous to dashes. After each sentence, participants were required to answer a yes/no comprehension question. Participants were allowed to take two breaks during the experiment, after the first and second thirds of the sentences.

Participants Thirty students of Universität Stuttgart (age range 19-31, mean 24; 21 females; 2 self-reportedly left-handed participants were assigned to different groups), all native speakers of German with normal or corrected-to-normal vision, volunteered to participate in the experiment and were paid for their participation.

5.2.2. Results and Discussion

All participants answered more than 77% of the comprehension questions correctly (M = 94%, SD = 0.07). Reading times were analyzed one word before the target verb and three after. Items from sentences that received incorrect answers as well as outliers were excluded from the analysis (8% of data points). We chose a threshold (reading time per word above 100 ms and below 3000 ms) so that no more than 10% of items were removed.

Reading times in each region were analyzed through a generalized mixed effect regression model, as suggested by recent critiques to the use of ANOVA in psycholin-

Position		patient	target V	V+1	V+2	V+3
	Examples	*Glasur*	*aufzutragen*	*und*	*fing*	*mit*
		icing	to spread	and	began	with
Latency (ms)	low-typicality	441	591	485	426	422
	high-typicality	442	539	477	420	435
Difference (ms)		-1	52	8	6	-13
Mixed-Effect	t	< 1	2.24	1.21	< 1	< 1
Regression	p	0.33	0.03	0.23	0.41	0.26

Table 5.2.: Experiment 1: Reading latencies (in ms) and mixed-effect regressions.

guistics (Baayen et al., 2008). Mixed-effect models have been shown to be a more powerful tool to analyze reading times, as they on the one hand allow for separating random effects for item and for participant, and on the other hand they are able to take into account trial-to-trial longitudinal dependencies between observations (for example, by including reading times or response latencies at preceding trials in time as covariates). Following the procedure in Baayen et al. (2008), we used an empirical procedure to decide what factors to include in the model, ruling out factors that did not significantly contribute to the model's goodness of fit, determined by a likelihood ratio test. The model's covariates which contributed to the goodness of fit and were thus included were the reading times at the previous word and the order of presentation of each trial (rank-order of a sentence in its experimental sequence).

Table 5.2 shows mean reading times and the associated mixed-effect regressions. No significant differences were found between high- and low-typicality conditions at the patient noun region (*Glasur*), which was not surprising as the sentences were identical in both conditions up to this region. A main effect of typicality was found at the target verb region (*aufzutragen*). As shown in Figure 5.2, events were read 52 ms faster when cued by the agent-patient combination (*Konditor* + *Glasur* → *aufzutragen*) than when not cued (*Kind* + *Glasur* → *aufzutragen*). No difference after the target verb region (*und fing mit* ...) reached significance.

Experiment 1 showed that events cued by a highly typical agent-patient combination were read faster. Under our assumption that the expectations for sentence-final events in long forms are analogous to expectations for covert events in logical

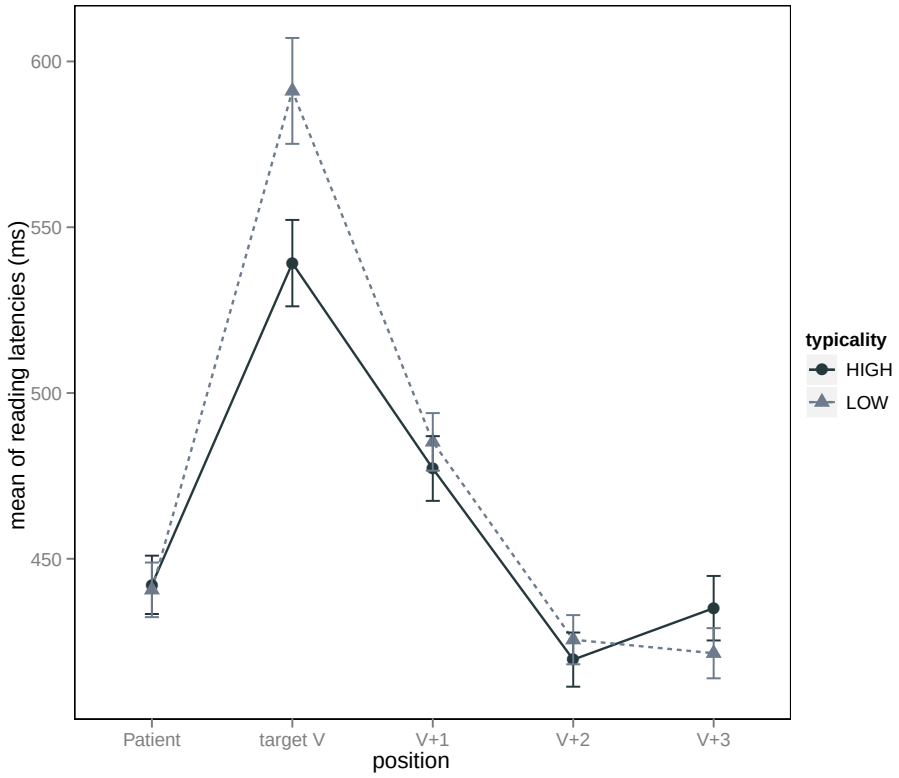

Figure 5.2.: Experiment 1: Comparing reading latencies (in ms) for each position and for each condition.

metonymies, the results from Experiment 1 can be interpreted as evidence that generalized event knowledge is involved in guiding expectations about covert events and in determining their retrieval in logical metonymies.

5.3. Experiment 2

Experiment 1, while providing evidence for the influence of generalized event knowledge in covert event interpretation, relied on the assumption that the same cognitive resources come into play both when interpreting the covert event in a logical metonymy and when predicting sentence-final events in long forms. Nevertheless, there are reasons to question this assumption, as the corpus studies reported in Chapter 4 showed crucial differences between logical metonymies and logical forms, supporting a Gricean account of logical metonymy: while covert events are implicit

and thus more likely to refer to a default / obvious predicate, if the predicate is not obvious then it needs to be mentioned explicitly in the long form.

Experiment 2 (Zarcone et al., 2012b, 2014) aims at strengthening the case for generalized event knowledge in covert event retrieval by providing further experimental evidence, this time avoiding the assumption Experiment 1 relied upon, and investigating the processing of logical metonymies, without resorting to their long form paraphrases. We used a probe recognition task: participants were presented with a metonymic sentence followed by a possible covert event presented as a probe (e.g. *The baker finished the icing* → SPREAD).

Previous work suggests that, when activated by linguistic cues, typical elements from a scenario are difficult to suppress. For example, after viewing a scene (e.g. a farm scene), participants may incorrectly respond that a typical object (e.g. a tractor) was present when the object was not there, as it is difficult to suppress "the interpretations of visual arrays that comprise scenes" (Biederman et al., 1988). Similarly, when participants were presented with the word referring to a typical object (e.g. TRACTOR) in a probe recognition study (Gernsbacher and Faust, 1991) and asked if it was present in a scene presented previously (e.g. a farm scene or a kitchen scene), their decision latencies were delayed, by an effect of interference in the typical scene (the farm scene) compared to scenes where the object is not typically found (the kitchen scene). The delay was found for skilled readers at a short (50 ms) inter-stimulus interval (ISI) but not at long ISI (1000 ms).

Our hypothesis for Experiment 2 is that, when reading a logical metonymy, participants will cue the covert event (presented as a probe) in a high-typicality condition (*baker* + *the icing* → SPREAD), but not in the low-typicality condition (*child* + *the icing* → SPREAD). Thus, when asked to decide if the event was mentioned in the sentence or not (in a word decision task), the correct response to both groups of test sentences will always be "no" (neither SPREAD or EAT were explicitly mentioned in the sentence). We predict that, if the event has been inferred, it will be active in memory and participants will take longer to reject it, and will **delay decision latencies** (when it is presented after a high-typicality sentence for the probe), which instead will not be delayed when the probe is presented after a low-typicality sentence.

Decision latencies at short (100 ms) and long (900 ms) ISI are contrasted. The Words-as-cues Hypothesis predicts that covert events are indeed activated early, yielding a difference in decision latencies **as early as possible** (here, at the short ISI), whereas

the Pragmatic Hypothesis would predict this difference to appear later on (at the long ISI).

5.3.1. Method

Materials 96 metonymic sentences were constructed from the triplets used in Experiment 1, and the covert events from Experiment 1 were used as probes for the probe recognition experiment, appearing once after a high-typicality sentence and once after a low-typicality sentence (see Appendix A.2.2 for the complete list of sentences):

High-typicality:

(5.6) a. Der Konditor hörte mit der Glasur auf. → **AUFTRAGEN**
The baker finished the icing. → **SPREAD**

b. Das Kind hörte mit der Glasur auf. → **ESSEN**
The child finished the icing. → **EAT**

Low-typicality:

(5.7) a. Das Kind hörte mit der Glasur auf. → **AUFTRAGEN**
The child finished the icing. → **SPREAD**

b. Der Konditor hörte mit der Glasur auf. → **ESSEN**
The baker finished the icing. → **EAT**

Probes were on average 8 characters long (min 5, max 14, SD 2); average log frequency in the CELEX word frequency list for German (Baayen et al., 1993) was 1.32 (min 0, max 2.5, SD 0.86).

Procedure Two lists of 120 sentences each (24 high-typicality, 24 low-typicality, 72 filler sentences) were created to ensure that the same participant would not see the same agent-event-patient triple twice: for each group of four sentences sharing the same patient, the two high-typicality sentences were put in one list (to which half of participants were assigned) and the two low-typicality ones in the other list (to which the other half of the participants was assigned). The fillers were the same for both lists, and (as the answer was "no" for all 48 metonymic test sentences — where the covert events were not mentioned), 60 of the fillers included the probe (and thus required

		ISI	
		100 ms	900 ms
Error rates (%)	high-typicality	0.2%	0.9%
	low-typicality	0.9%	0.5%
		ISI	
Latency (ms)	high-typicality	906	854
	low-typicality	853	835
Difference (ms)		53	19
Mixed-Effect Regression	t	-3.10	-0.77
	p	0.002	0.45

Table 5.3.: Experiment 2: Error rates, decision latencies (in ms) and mixed-effect regressions for 100 ms and 900 ms ISI.

a "yes" answer) and 12 did not, for a total of 60 "yes" and 60 (12 + 48) "no" answers in each list. Among the 60 fillers requiring a "yes" answer, 42 had a metonymic verb as main verb, in order to avoid the association between metonymic verbs and "no" answers.

The experiment employed a $2x2$ mixed factorial design: ISI (short / long, that is 100 / 900 ms) was varied between subject; typicality (high / low) was varied within subjects. Each trial began with a 500-ms fixation cross in the middle of the screen, followed by a sentence. Participants pressed a button after reading the sentence, eliciting the presentation of the probe word after a short (100 ms) or long (900 ms) ISI, and were instructed to decide as quickly and as accurately as possible whether or not the probe had been mentioned in the sentence, and to respond accordingly by pressing one of two designated keys (the "no" answers were always given with the non-dominant hand). Participants were allowed to take two breaks during the experiment, after the first and second thirds of the sentences.

Participants Thirty-six students of Universität Stuttgart (age range 18-40, mean 25; 25 females; 3 self-reportedly left-handed were distributed among groups), all native speakers of German with normal or corrected-to-normal vision, volunteered to participate in the experiment and were paid for their participation.

5.3.2. Results and Discussion

All participants scored less than 5% wrong answers (M = 1%, SD = 0.01) in the probe recognition task and average error rates per condition were all below 1%, and thus too small to permit a statistical test (descriptive statistics are reported in Table 5.3). Items that received incorrect answers and decision latency outliers (> 2.5 SDs from the mean) were excluded from the analysis (3% of the data points).

Decision latencies were analyzed to test for an effect of ISI and typicality, via a generalized mixed effect regression using the order of presentation (rank-order of a trial in its experimental sequence), the reading time at the sentence preceding the probe and the decision latency at the preceding probe as covariates (again following the procedure in Baayen et al., 2008). The mixed-effect regression yielded a main effect of typicality ($t = -2.22$; $p = 0.03$), but no effect of ISI. As shown in Figure 5.3, decision latencies for covert event probes (e.g. AUFTRAGEN) at short ISI were 53 ms slower when the covert event was cued by the agent-patient combination (*Konditor-Glasur*) than when it was not (*Kind-Glasur*), while the difference at long ISI was smaller (10 ms). A pair-wise mixed-effect regression at both ISI showed that the 53 ms difference at the short ISI was significant ($t = -3.10$; $p = 0.002$), whereas the 19 ms difference at long ISI was not (see Table 5.3).

Participants in the short ISI group took longer to reject event probes cued by the combinations of objects with high-typicality agents. Experiment 2 showed that when the generalized event knowledge associated with the sentential context cues the covert event, participants are slowed down when deciding whether or not the event was explicitly mentioned in a logical metonymy, confirming the partial conclusion from Experiment 1 that generalized event knowledge determines the covert event in logical metonymies.

5.4. Experiment 2b

While Experiment 2 confirmed the results of Experiment 1, some critiques can be raised with regard to the design and the choice of materials.

A first potential critique pertains to the choice of metonymic verbs used in Experiment 1 and 2, that is a rather heterogeneous set of German verbs, roughly equivalent to the English metonymic verbs commonly used in experiments on logical metonymy,

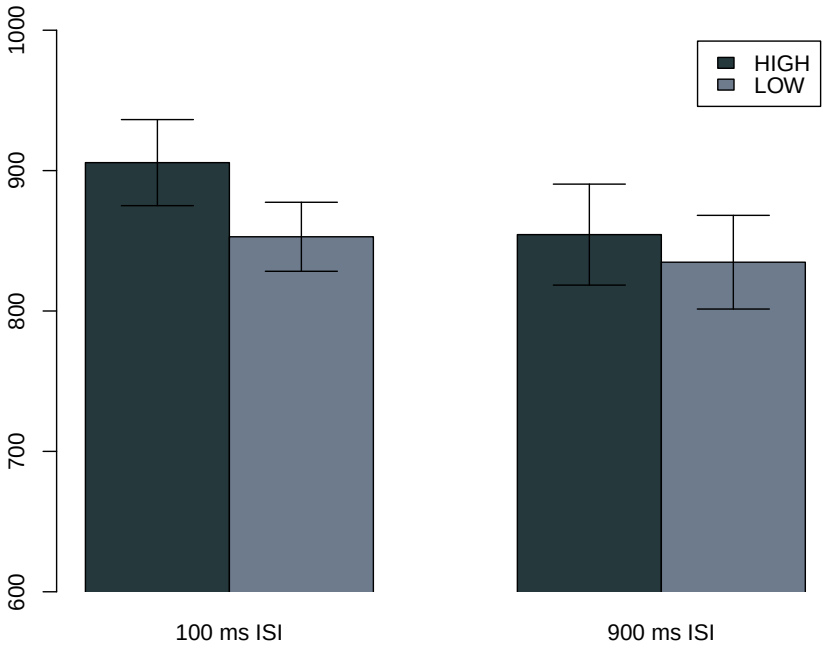

Figure 5.3.: Experiment 2: Comparing decision latencies (in ms) for each position and for each condition.

including aspectual verbs (e.g. *anfangen [start], aufhören [finish]*) as well as psychological verbs (e.g. *genießen [enjoy], hassen [hate]*) and other (less coherent) classes of verbs. The crowdsourcing study in Chapter 4 has shown that verbs may differ with regard to their tendency to give rise to a metonymic interpretation, and Katsika et al. (2012) have reported eye-tracking results showing that not all metonymic verbs are equal, arguing that aspectual verbs trigger processing costs due to the coercion operation, while psychological predicates (e.g. *enjoy*) do not[1].

Secondly, a semantic priming effect between the agent and the event may be an alternative explanation for the difference between the high- and the low-typicality conditions in Experiments 1 and 2, and we would ideally want to exclude that the events were primed by the agents only and not (as claimed) cued by the (highly typical)

[1]For more on this aspect, including lists of metonymic verbs used in different psycholinguistic studies, see Part III, and in particular the computational study in Utt et al., 2013 (Chapter 7).

agent + patient combination. We are inclined to exclude it because, as observed by Rayner et al. (2004); Murray (2006); Matsuki et al. (2011), a potential priming effect should be strongly diminished due to intervening words.

Lastly, one could argue that the short ISI condition in Experiment 2 was actually not short enough to argue for early integration of generalized event knowledge, because we did not control for the presentation time of the sentences and thus for the real time interval between the end of the sentence reading step (which was self-paced) and the presentation of the probe.

In order to address these issues, we designed Experiment 2b, that is a repetition of Experiment 2 with minimal changes: (1) to ensure that our results were not affected by the heterogeneity of the verbs used, only aspectual verbs were used as metonymic verbs for Experiment 2b; (2) a semantic association study (Norming Study 4) on the materials of Experiment 2 was carried out, singling out which items needed to be replaced; (3) the probe recognition task in Experiment 2b was cross-modal, that is the sentences were not read but were presented as audio stimuli, thus allowing for a better control of the inter-stimulus interval; (4) three ISI (rather than two) were contrasted (0, 500 and 1000 ms), in order to have a more fine-grained picture of the time course of covert event activation.

5.4.1. Method

Materials The same 96 sentence-probe pairs from Experiments 1 and 2 were used, with some changes introduced for Experiment 2b that affected 18 of the 24 groups of four sentence-probe pairs. The first change was made on the metonymic verbs: sentences containing non-aspectual metonymic verbs were modified, ensuring that only metonymic verbs that referred to the temporal structure of the event were used (*anfangen [start], aufhören [finish], beginnen [begin], vertagen [postpone], weitermachen [continue]*). The second change was made on the covert events after Norming Study 4.

Norming Study 4 Similarly to Matsuki et al. (2011), in order to exclude that the typicality effects yielded by Experiments 1 and 2 were due to strong associations between the agent and the event, we collected semantic association norms on AMT for

the agents used in Experiments 1 and 2 following the procedure in Nelson et al. (1998)[2]. For each agent we asked 30 German participants to write 1 to 3 words that came spontaneously to mind (e.g. *snow* → *winter, cold, sleigh*). 16 out of 32 covert events which were mentioned among the elicited answers were replaced with a synonym that had not been elicited, to obscure the association.

As in Experiment 2, the probes appeared once after a high-typicality sentence and once after a low-typicality sentence (see Appendix A.2.2 for the complete list of sentences). Probes were on average 8 characters long (min 5, max 16, SD 2); average log frequency in the CELEX word frequency list for German (Baayen et al., 1993) was 1.15 (min 0, max 2.8, SD 0.85).

Procedure The lists and fillers were created in the same way as in Experiment 2. The experiment employed a $3x2$ mixed factorial design: ISI (0 / 500 / 1000 ms) was varied between subject; typicality (high / low) was varied within subjects. The experiment was cross-modal: each trial began with a fixation cross in the middle of the screen, participants pressed a button to hear a sentence (audio stimulus), which was followed by the probe word (written stimulus) after a 0, 500 or 1000 ms ISI, and were instructed to decide as quickly and as accurately as possible whether or not the probe had been mentioned in the sentence, and to respond accordingly by pressing one of two designated keys (the "no" answers were always given with the non-dominant hand). Participants were allowed to take two breaks during the experiment, after the first and second thirds of the sentences.

Participants Thirty-six students of Universität Stuttgart and Hochschule der Medien in Stuttgart (age range 18-31, mean 22; 15 females; 3 self-reportedly left-handed were distributed among groups), all native speakers of German with normal or corrected-to-normal vision, volunteered to participate in the experiment and were paid for their participation.

[2] I would like to thank Sabine Schulte im Walde for including Norming Study 4 in her batch of association norm studies.

		ISI		
		0 ms	500 ms	1000 ms
Error rates (%)	high-typicality	1.74%	1%	1.74%
	low-typicality	0.35%	1%	1.74%
		ISI		
Latency (ms)	high-typicality	881	964	824
	low-typicality	843	961	829
Difference (ms)		38	3	-5
Mixed-Effect Regression	t	-2.43	-0.46	0.39
	p	0.02	0.6	0.7

Table 5.4.: Experiment 2b: Error rates, decision latencies (in ms) and mixed-effect regressions for 0 ms, 500 ms and 1000 ms ISI.

5.4.2. Results and Discussion

All participants scored less than 9% wrong answers (M = 1%, SD = 1.8) in the probe recognition task and average error rates per condition were all below 2%, and thus too small to permit a statistical test (descriptive statistics are reported in Table 5.4). Items that received incorrect answers and decision latency outliers (> 2.5 SDs from the mean) were excluded from the analysis (4% of the data points).

Decision latencies were analyzed to test for an effect of ISI and typicality, via a generalized mixed effect regression using the order of presentation (rank-order of a trial in its experimental sequence) and the decision latency at the preceding probe as covariates. The mixed-effect regression yielded a main effect of typicality ($t = -2.42$; $p = 0.02$), no effect of ISI and a significant interaction between typicality and ISI (0 ms compared to 1000 ms: $t = 1.96$; $p = 0.05$). As shown in Figure 5.4, decision latencies for covert event probes (e.g. AUFTRAGEN) at the 0 ms ISI were 38 ms slower when the covert event was cued by the agent-patient combination (*Konditor-Glasur*) than when it was not (*Kind-Glasur*), and different pair-wise mixed-effect regressions at 0 ms, 500 ms and 100 ms ISI showed that the 38 ms difference at the 0 ms ISI was significant ($t = -2.43$; $p = 0.02$), whereas the differences at 500 and 100 ms were not.

The 500 ms group yielded apparently longer decision latencies than the 0 ms and

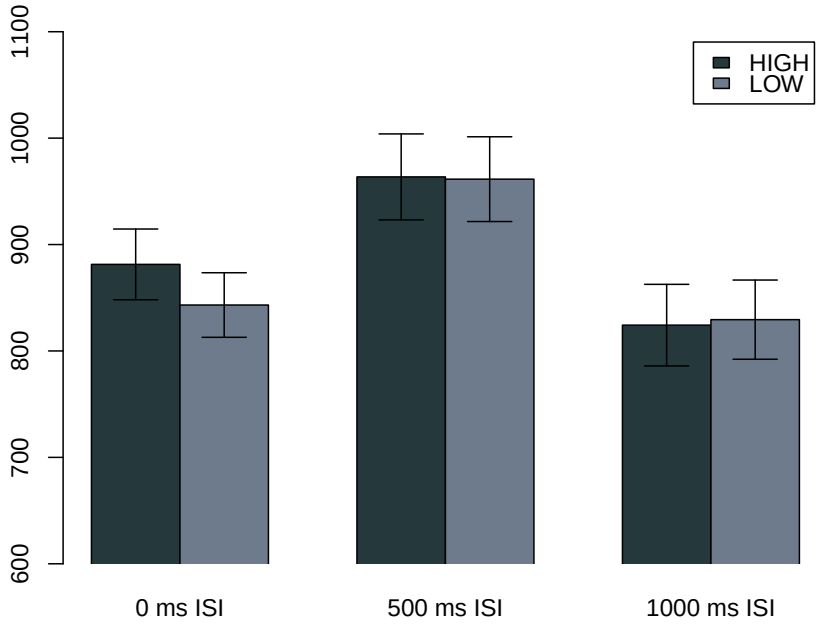

Figure 5.4.: Experiment 2b: Comparing decision latencies (in ms) for each position and for each condition.

the 1000 ms group respectively, but neither difference was significant, neither in the general model nor in pair-wise comparisons (with the only exception of the difference between decision latencies at 500 and 100 ms for high-typicality items: $t = 2.03$, $p = 0.04$). In a similar analysis on a subset of the data points, considering only the 0 and 100 ms ISI groups and disregarding 500 ms group, a main effect of typicality was yielded ($t = -2.39$; $p = 0.02$), without any effect of ISI and with a significant interaction between typicality and ISI (0 ms compared to 1000 ms: $t = 1.93$; $p = 0.05$).

Even after making some changes to the materials and the design, the results from Experiment 2 were confirmed: participants were slower in rejecting event probes cued by high-typicality sentential context. As in Experiment 2, decision latencies were only delayed at short ISI (at 0 ms, but not at 500 ms and 1000 ms ISI): the effect of generalised event knowledge on covert event interpretation is clearly an early effect. This is in accordance with the results obtained in Experiment 1, with previous

work using probe recognition (Gernsbacher and Faust, 1991), as well as with priming studies on generalised event knowledge (McRae et al., 2005), where short stimulus onset asynchrony is used as a window into the generation of implicit expectations guiding normal sentence processing and effects at long stimulus onset asynchrony are instead considered to be determined by strategic processing (Becker, 1980). At short ISI, the high-typicality interpretation is cued by the preceding sentence and rejecting it becomes costly. At longer ISIs, both the high-typicality and low-typicality interpretation are activated as plausible interpretations and are equally as difficult to reject.

5.5. General Discussion

The experiments in this chapter were designed to evaluate the role played by generalized event knowledge in the interpretation of covert events in logical metonymy. I have shown that covert events matching our typical knowledge of common events and their participants are activated by contextual cues. Generalized event knowledge is activated immediately, providing a source of expectation for upcoming input, and thus leading to a facilitation effect during processing (Experiment 1), and is quickly integrated into the sentence meaning, leading to delayed decision latencies when participants have to rule out that the event was mentioned in the sentence (Experiments 2 and 2b).

Generalized event knowledge, being quickly and dynamically activated and updated, overcomes the rigidity of the qualia structure, as it provides a model of covert event interpretation which can account for the influence of intra-sentential context. The facilitation effect reported for Experiment 1 and the delay in decision latencies reported for Experiments 2 and 2b can not be straightforwardly explained in terms of qualia roles, because they do not contrast two different covert events (namely a qualia-related covert event and a non-qualia-related one), but rather they evaluate differences in processing costs on a same target event given different typicality conditions (e.g. *baker / child + icing → SPREAD*). The Lexical Hypothesis would not be able to account for such typicality effects, as it lacks a dynamic mechanism to account for context integration.

Frisson and McElree (2008) partially addressed the problem of the retrieval of the

covert event, but they did not commit to a specific hypothesis regarding the range and the source of the covert event, assuming only that "coerced senses are computed from a broader range of properties than the Qualia structure of the complement noun" (p. 2), and did not address the question of context integration. Also, unlike Frisson and McElree (2008), we use the same patient noun with two typical events, which appear in two different contexts each (a highly typical one and a less typical one): our design thus restricts variability due to item idiosyncrasies. Also, both contexts are metonymic, while Frisson and McElree (2008) aim at contrasting a metonymic and a non-metonymic condition.

The influence of sentential cues could be accounted for by the Pragmatic Hypothesis by generating different post-lexical inferences for different agents. However, this would predict a late (delayed) effect, not at the target verb region, but one or two words later in the self-paced reading experiment, and at long ISI in the probe recognition experiments. The results I reported (early effects in self-paced reading and at short ISI in probe recognition) are in contrast with this prediction, and speak in favor of early integration of generalized event knowledge in the interpretation of covert events in logical metonymy, in line with the predictions of the Words-as-cues Hypothesis.

6. Computational Models of Covert Event Interpretation

I have argued for a primary role played by intra-sentential cues (agent and patient) and their thematic fit with the covert event in selecting a suitable interpretation for logical metonymies, and my claim was supported by the behavioral experiments in Chapter 5.

Previous computational models of logical metonymy focused on the range of the covert events (as applications can benefit from making the implicit information explicit) and acknowledged the role of intra-sentential cues, providing a context-based mechanism to identify the correct covert event by ranking a list of possible interpretations. In this chapter I will review the most prominent among these approaches (the Probabilistic Model in Lapata and Lascarides, 2003 and Lapata et al., 2003), and I will contrast it with a distributional, Similarity-based Model of covert event interpretation. The Similarity-based Model shares the context-sensitivity of the Probabilistic Model but also exploits the notions of thematic fit and similarity, which are central to a Words-as-cues model of covert event interpretation. Both models are evaluated on psycholinguistic datasets from Chapter 5: the Similarity-based Model can account for the role played by intra-sentential cues while outperforming the coverage of the Probabilistic Model.

I will focus on the theoretical contributions that a computational model can make in the study of logical metonymy. As discussed in Chapter 2, if the architecture of a model (in this case, compositional, probabilistic, similarity-based) contributes to its success in replicating behavioral results, then the model can inform the theory with architectural constraints to be set on a model of covert event interpretation.

6.1. Modeling Covert Event Interpretation

The Words-as-cues Hypothesis claims that generalized event knowledge determines covert event interpretation, and the psycholinguistic experiments in Chapter 5 have shown that covert event interpretation is sensitive to intra-sentential cues (agent and patient). Also, the crowdsourcing study in Chapter 4 has shown that covert events are better conceptualized as ranked sets of (often semantically-related) predicates.

More specifically, the Words-as-cues Hypothesis claims that typical arguments guide expectations for events, and that the predicted covert event is the event with the best thematic fit given the intra-sentential cues (agent and patient). A computational model of covert event interpretation in accordance with the Words-as-cues Hypothesis should thus be able to integrate contextual cues (it should be **compositional**), predicting the most fitting covert event for a given logical metonymy (it should incorporate a notion of **thematic fit**). Also, it should assign sensibly high plausibility scores to plausible alternative events. For example, for the metonymy *the aunt finished the tea*, the model should predict *drinking*, but should also be able to assign a high plausibility score to a semantically similar verb like *sipping*. Thus, it should ideally incorporate a notion of **semantic similarity**.

I will contrast two models of covert event interpretation: a Probabilistic Model (Lapata and Lascarides, 2003; Lapata et al., 2003) and a Similarity-based Model (Zarcone et al., 2012d). The models were evaluated on psycholinguistic datasets from Chapter 5: if a model is able to predict covert events, and to produce results which are comparable with the psycholinguistic experiments, then we can conclude that (a) the information exploited by the model does indeed play an important role in covert event interpretation and that (b) the computational model can provide an alternative solution to qualia, thus contributing specific architectural elements (thematic-fit dependency, similarity) to a theoretical model of logical metonymy which speaks in favor of the Words-as-cues Hypothesis.

6.2. A Probabilistic Model

Computational work on logical metonymy (mainly NLP-oriented) aimed at finding one unambiguous interpretation for a logical metonymy (e.g. *the author began the book* → *reading, writing*). The first and most prominent computational model was the

Probabilistic Model of logical metonymy presented in Lapata and Lascarides (2003) and Lapata et al. (2003).

The Probabilistic Model of logical metonymy is actually a model of covert event retrieval, because it avoids the trigger problem by making no distinction between entity- and event-denoting objects: interpretations are computed for both, and covert events for potential metonymies including event-denoting objects are not ruled out because they are considered beneficial for NLP applications[1]. The authors claim that (as observed by Vendler, 1968 for telic adjectives) not a single verb, but a "family of verbs" is needed in order to account for the interpretations of a logical metonymy, and that a corpus-based computational model can successfully predict a **ranked set of covert event interpretations**.

They acknowledge the role played by intra-sentential cues: their approach models the covert event interpretation of a logical metonymy (e.g. *The student enjoyed the book*) as the joint distribution $P(s,v,o,e)$ of the variables s (the subject, e.g. *student*), v (the metonymic verb, e.g. *enjoy*), o (the object, e.g. *book*), e (the covert event, *reading*). The selected covert event $\hat{e}$ for a given context is the event which maximizes $P(s,v,o,e)$. They present two models with different independence assumptions (the Simplified Model and the Full Model), but I have only taken the best performing one into consideration (the Simplified Model). In order to distinguish it from the other models presented in this chapter, it will be referred to as the ***SOV_p* model**.

The ***SOV_p* model** assumes a generative process which first generates the covert event e and then generates all other variables based on the choice of e:

$$\arg\max_e P(s,v,o,e) = \arg\max_e P(e)\ P(o|e)\ P(v|e)\ P(s|e)$$

These distributions are estimated as follows:

$$\hat{P}(e) = \frac{f(e)}{N}, \quad \hat{P}(o|e) = \frac{f(e \overset{o}{\leftarrow} o)}{f(e \overset{o}{\leftarrow} \cdot)},$$
$$\hat{P}(v|e) = \frac{f(v \overset{c}{\leftarrow} e)}{f(\cdot \overset{c}{\leftarrow} e)}, \quad \hat{P}(s|e) = \frac{f(e \overset{s}{\leftarrow} s)}{f(e \overset{s}{\leftarrow} \cdot)}$$

[1] Interestingly, the NLP focus brings the authors to consider potential extensions of the model to other phenomena including covert events, such as telic adjectives: *difficult language → to learn, speak write; good cook → at cooking; good soup → to eat.*

N is the number of occurrences of verbs in the corpus (excluding modals and auxiliaries); $f(e)$ is the frequency of the verb e; $f(e \overset{o}{\leftarrow} \cdot)$ and $f(e \overset{s}{\leftarrow} \cdot)$ are the frequencies of e with a direct object and subject, respectively; and $f(e \overset{c}{\leftarrow} \cdot)$ is the number of times e is the complement of another verb.

Lapata and Lascarides (2003) and Lapata et al. (2003) embrace a Gricean view of metonymy, acknowledging the difference between covert events (default interpretations) and explicit events in long forms (less default interpretations). However, explicit events in long forms are attested in the corpus and can be exploited to estimate what the most likely covert event will be, thus they are taken into consideration ($P(s|e)$ is estimated from long forms), but without any tight correspondence between a long form and its corresponding logical metonymy: $\hat{P}(s|e)$ is estimated without taking the subject and the object into consideration; in other words, in the example *The student enjoyed the book* → *reading*, what counts is how frequently *students* are said to *read*, *books* to *be read*, and *reading* activities to be *begun*, not how frequently *students* are explicitly said to *begin reading books*.

Since my evaluation was be carried out on datasets from Chapter 5, where the metonymic verb v was kept constant for each group of sentences and the covert event e was a function of subject s and object o, I introduce also a second version of the model (the ***SO*$_p$ model**), which does not take v into consideration:

$$P(s, v, o, e) \approx P(s, o, e) \approx P(e)\ P(o|e)\ P(s|e)$$

$$\arg\max_e P(s, o, e) = \arg\max_e P(e)\ P(o|e)\ P(s|e)$$

The **compositionality** requirement is met by the Probabilistic Model: the model provides a straightforward way to account for the influence of agent and patient (approximated as subject and object) as random variables.

The Probabilistic Model does not model **thematic fit** in a straightforward way as it is based on conditional probability of co-occurrence and not on a measure of typicality, although co-occurrence and typicality are arguably correlated.

As to the **similarity** requirement, the Probabilistic Model returns a ranked set of verbs, but does not address the problem of the semantic relations between them (e.g. synonymy, hyponymy, cohyponymy). For example, if *drinking* is highly ranked in the set of covert events for *the aunt finished the tea*, but *sipping* is not attested in the corpus in combination with the other predictors (e.g. ⟨*aunt sips*⟩, ⟨*sip tea*⟩, ⟨*finish*

sipping⟩), the model is not going to predict *sipping* as a plausible covert event.

Related work improved the set of covert events modeled by the Probabilistic Models[2]: Shutova (2009); Shutova and Teufel (2009); Shutova et al. (2013) clustered together verb senses using WordNet synsets and obtained ranked lists of *senses* rather than verbs; Roberts and Harabagiu (2011) enhanced a non subject-dependent model, estimating $P(v,e,o)$, with information on the verb's selectional restrictions[3], to rule out event-invoking covert events.

I will now introduce a distributional, similarity-based model of covert event interpretation, which on the one hand maintains the integration of intra-sentential context of the Probabilistic Model and on the other hand introduces the notions of thematic fit, prototype and similarity, predicting event concepts (not verbs) which match our generalized event knowledge of the expected scenario.

6.3. A Similarity-based Model

The Similarity-based Model (Zarcone et al., 2012d) is based on Distributional Memory (Baroni and Lenci, 2010) and on its ECU extension (Lenci, 2011) and is the first distributional model of logical metonymy. In contrast to most experimental studies on the phenomenon (with the exception of Lapata and Lascarides, 2003), it does not deal with English data but focuses on German.

6.3.1. Distributional Memory (DM)

DSMs have been used for several semantic tasks: attributional (Grefenstette, 1994; Lund and Burgess, 1996; Padó and Lapata, 2007) and relational (Turney, 2006) similarity, property extraction (Cimiano and Wenderoth, 2007), selectional preferences (Erk, 2007), event types (Zarcone and Lenci, 2008), and many more. Baroni and Lenci (2010) observe that this contrasts with the multi-purpose nature of semantic memory, and argue for a "one distributional model, multiple semantic tasks" approach. They call the model **Distributional Memory** (DM, Baroni and Lenci, 2010), reinforcing the

[2] Another group of studies induced qualia structures from web corpora using lexical patterns (Cimiano and Wenderoth, 2007; Baroni and Lenci, 2010), without addressing the question of the feasibility of a qualia-based account of covert event retrieval.

[3] Interestingly, Roberts and Harabagiu (2011) argue that selectional restrictions can also model the trigger problem. I will suggest a similar account based on thematic fit in Part III.

idea already in de Saussure (1915) that distributional facts are part of the tissue that makes up our mental lexicon.

DM is a structured DSM, composed of two parts:

- an **offline part** (3-way tensor), made of corpus-extracted co-occurrences;
- an **online part** (2-way matrices), built ad hoc depending on the required semantic task.

The **offline part** is a "storehouse", built from the corpus once and for all and represented in terms of a third-order tensor of weighted *word-link-word* tuples extracted from a very large corpus; *word-link-word* tuples are mapped onto a weight by a function σ: $\langle w_1\ l\ w_2\rangle \rightarrow \mathbb{R}^+$. For example, ⟨*book obj read*⟩ is a *word-link-word* tuple (*obj* is the link, indicating that *book* is the object of *read*), which has a higher weight than ⟨*label obj read*⟩, and both have higher weights than ⟨*elephant obj read*⟩, because books are more typically encountered as objects or *read* than labels, and *elephant* is never encountered as object of *read*. The set of links can be defined in different ways, thus instantiating different DM models: DepDM (mainly syntactic links, e.g. *subj_tr*), LexDM (strongly lexicalized links, e.g., *such_as*), or TypeDM (syntactic and lexicalized links).

While the offline part (all-purpose, stored) is extracted once and for all, the **online part** can be generated on demand, whenever a task is selected, as a dedicated 2-way semantic space: a *word by link-word* space ($W_1 \times LW_2$), a *word-word by link* space ($W_1W_2 \times L$), a *word-link by word* space ($W_1L \times W_2$), a *link by word-word* space ($L \times W_1W_2$). Each space specializes in modeling different aspects of meaning and performs as well as or better than state-of-the-art DSMs in its dedicated tasks (Baroni and Lenci, 2010).

Due to its versatility, effectiveness and availability, DM was adopted as the distributional model of reference for this dissertation. In particular, for the purposes of this dissertation I will only refer to TypeDM, which is presented by Baroni and Lenci (2010) as the best of the three (henceforth, DM = TypeDM). The scoring function σ is the *Local Mutual Information* (LMI, Evert, 2005) on link type frequency (negative LMI values are raised to 0):

$$\text{LMI} = O_{ijk} \log \frac{O_{ijk}}{E_{ijk}}$$

w_1	l	w_2	σ	w_1	l	w_2	σ
marine	own	bomb	40.0	sergeant	use	gun	51.9
marine	use	bomb	82.1	sergeant	own	book	8.0
marine	own	gun	85.3	sergeant	use	book	10.1
marine	use	gun	44.8	teacher	own	bomb	5.2
marine	own	book	3.2	teacher	use	bomb	7.0
sergeant	own	bomb	16.7	teacher	use	gun	4.7
sergeant	use	bomb	69.5	teacher	own	book	48.4
sergeant	own	gun	73.4	teacher	use	book	53.6

	j=1:own	*j=2:use*	*j=1:own*	*j=2:use*	*j=1:own*	*j=2:use*
	k=1:bomb		*k=2:gun*		*k=3:book*	
i=1:marine	40.0	82.1	85.3	44.8	3.2	3.3
i=2:sergeant	16.7	69.5	73.4	51.9	8.0	10.1
i=3:teacher	5.2	7.0	9.3	4.7	48.4	53.6

Table 6.1.: A toy weighted tuple structure and a labeled tensor from Baroni and Lenci (2010).

	1:	*2:*	*3:*	*4:*	*5:*	*6:*
	⟨*own,bomb*⟩	⟨*use,bomb*⟩	⟨*own,gun*⟩	⟨*use,gun*⟩	⟨*own,book*⟩	⟨*use,book*⟩
1:marine	40.0	82.1	85.3	44.8	3.2	3.3
2:sergeant	16.7	69.5	73.4	51.9	8.0	10.1
3:teacher	5.2	7.0	9.3	4.7	48.4	53.6

Table 6.2.: A labeled matricization of the tensor in Table 6.1 (Baroni and Lenci, 2010).

where O_{ijk} is the observed frequency and E_{ijk} is the expected frequency of a triple $\langle w_i\ l_j\ w_k \rangle$.

The *word by link-word* space ($W_1 \times LW_2$) in DM is the most apt at modeling selectional preferences and event knowledge and is therefore the space we are interested in for the purposes of this dissertation. From the $W_1 \times LW_2$ space we can easily retrieve the most typical fillers for the object position of a given verb (e.g. *read*) as the top n objects (patients) most highly associated with it via an object link. The thematic roles of agent and patient are approximated with the syntactic relations of subject and object respectively. Given a verb and an object (e.g. *read* + *label*), we can also compute

a **thematic fit** score for the object, that is the plausibility / typicality of the object as a filler of the argument position (as an object / patient of *read*) and compare it with the plausibility of another object (e.g. *book*). This can be done with a method adopted by Baroni and Lenci (2010) for DM[4]: after the top *n* most highly associated fillers are extracted (that is, those with the highest thematic fit), a prototype vector is computed as the centroid of the top *n* fillers (note that this vector may not correspond to any lexical item but is rather an abstract distributional representation of the prototypical filler of that argument position); then the thematic fit of a given filler (e.g. *label* as an object of *read*) is computed as its **similarity** (cosine) with the prototypical filler for that argument position. This method is of particular interest because it allows to compute typicality scores for unseen verb-argument pairs: if the combination *eat topinambur* (Baroni and Lenci, 2010) is not attested in the corpus, but the model knows enough about topinamburs to estimate that they can be considered similar to other vegetables, then the thematic fit for *eat topinambur* will be higher than, for example, *eat sympathy*.

The study on the English datasets (Chapter 7) was conducted on the English TypeDM, built by Baroni and Lenci (2010) from a concatenation of the Web-derived UKWAC corpus (Baroni et al., 2009, c.a. 1.9 billion words, publicly available from `http://clic.cimec.unitn.it/dm`), a mid-2009 dump of the English Wikipedia (about 820 million words) and the British National Corpus (100 million words). The German TypeDM used in this Chapter was developed and released by Padó and Utt (2012) from the SDEWAC web corpus (Faaß and Eckart, 2013, 880 million words) parsed with the MATE German dependency parser (Bohnet, 2010). Both were employed with the assistance and collaboration of Jason Utt for the purposes of this study.

6.3.2. DM and Compositionality: ECU

The thematic fit of an argument filler is not only determined by one contextual element (for example, the verb: *read*), but also by other argument fillers. Consider the following examples from Bicknell et al. (2010):

(6.1) a. The journalist checked **the spelling** / the brakes.

b. The mechanic checked the spelling / **the brakes**.

[4] A similar method (extracting all fillers, not just the top *n* fillers) was previously described by Erk (2007) and Padó et al. (2007).

If the agent is *journalist*, then the object with the highest thematic fit will be **spelling**; conversely, if *mechanic* is the agent, the patient with the highest thematic fit will be **brakes**. Such cases are problematic for the models just described, as DSMs are traditionally non-compositional. Thus, for the purposes of this dissertation, I have adopted a simple yet effective module to implement at least some degree of compositionality within DM, the **Expectation Composition and Update** module or **ECU** (Lenci, 2011), which is able to model how a given agent determines the semantic expectation for possible patients of a given verb (cf. examples 6.1.a vs. 6.1.b) by combining the expectations for a typical filler coming from both the agent and the verb and by assigning different thematic fit scores to the same patient with different agent-verb combinations. Given an agent and a verb (e.g. *journalist* and *check*), ECU extracts two different sets of expected patient fillers (toy example in Figure 6.1)[5]:

- $EX_{PA}(v)$ is the set of the expected patients of a verb v (in this case, things that are *checked*), and contains the weighted TypeDM tuples $\langle\langle v\ obj^{-1}\ n_i\rangle, \sigma_i\rangle$, that is the expected nouns n linked to v by an inverse *object* link (obj^{-1} is the link connecting the transitive verb and its object[6]), and their weights;

- $EX(n_{AG})$ is the set of the expected patients acted upon by the agent noun n_{AG} (in this case, the most typical things *the journalist* acts upon), and contains the weighted TypeDM tuples $\langle\langle n_{AG}\ verb\ n_j\rangle, \sigma_j\rangle$, that is the expected nouns n linked to n_{AG} by a *verb* link (*verb* is the link connecting the subject and the object of a transitive verb), and their weights.

The expectations for the patient from the verb are then composed with those from the agent by a function f for expectation composition and update, modeled as either sum or product:

$$EX_{PA}(\langle n_{AG}\rangle) = f(EX(n_{AG}), EX_{PA}(v))$$

Then the same generalization step described in 6.3.1 is applied: (1) from the updated set of expectations, the top n ($n = 20$) are selected, (2) a prototypical (expected) patient

[5] Recall that the thematic roles are approximated by way of syntactic dependencies (e.g. the role of agent with the subject dependency, the role of patient with the object dependency).

[6] l^{-1} is used to denote the inverse link of l (i.e., exchanging the positions of w_1 and w_2).

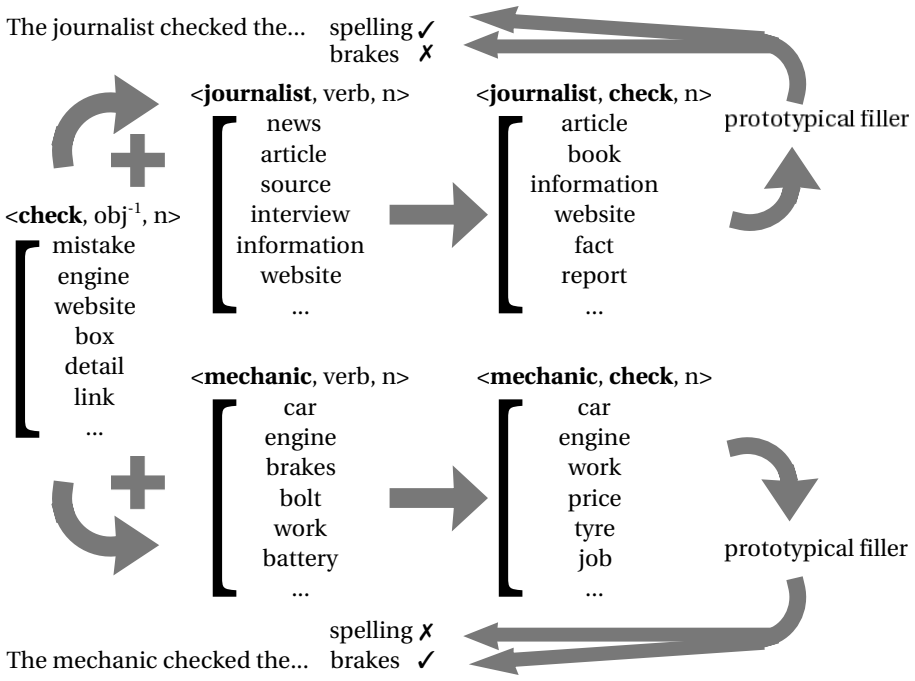

Figure 6.1.: Toy example for ECU: typical patients for ⟨*journalist, check*⟩ and ⟨*mechanic, check*⟩. The expected objects for the verb and for the subject are retrieved from DM in a scored and ranked list, then the expectations are composed to generate the expected patients for the agent + verb combination.

filler is obtained (centroid), (3) the thematic fit scores of possible fillers (*spelling* vs. *brakes*) are computed (cosine) and (4) the one with the best fit is chosen.

As distributional semantics has been recently extended to account for compositional phenomena, several models have been proposed, using different functions for component-wise vector combination and exploring ways to combine functional application à la Montague with vector-based representations (e.g. Erk and Padó, 2008; Mitchell and Lapata, 2010; Baroni and Zamparelli, 2010). As is common practice, in this and the following chapter I have focused on the two most typically used composition functions (sum and product, used also in ECU Lenci, 2011), testing empirically the differences between them and commenting on how they differently affect the results obtained from the model.

ECU was evaluated against the dataset in Bicknell et al. (2010), where patients (e.g. *spelling*) were matched with a high-typicality agent-verb combination (*journalist-*

check) and with a low-typicality one (*mechanic-check*). Since the materials were completely counterbalanced, and the low-typicality agent was the high-typicality agent for another patient (e.g. *mechanic-check-brakes*), the evaluation task kept the context constant, associating it with two patients (*spelling* vs. *brakes*). ECU was able to compute different thematic fit values for both patients in context and to correctly choose the most typical one over the less typical one.

6.3.3. A Similarity-based Model of Covert Event Interpretation

We modified ECU to obtain a Similarity-based Model of covert event interpretation (Zarcone et al., 2012d). The sentential cues and the target were modified: the target is the covert event and the cues are the agent (e.g. *The student* / *the brewer finished the beer*), the patient (e.g. *The student finished the beer* / *the essay*) and the metonymic verb (e.g. *The student finished* / *enjoyed the essay*). Given a logical metonymy (e.g. *The student finished the beer*), we first compute the expectations for the covert event e given the agent n_{AG}, the patient n_{PA} and the metonymic verb v individually, derived again from direct dependency relations in TypeDM:

- $EX(n_{AG})$ is the set of the expected events for the agent noun n_{AG} (in this case, the most typical things *students* do), and contains the weighted TypeDM tuples $\langle\langle n_{AG}\ subj\ e_i\rangle, \sigma_i\rangle$, that is the expected events e linked to n_{AG} by a *subject* link (*subj* is the link connecting the subject and its verb), and their weights;

- $EX(n_{PA})$ is the set of the expected events for the patient noun n_{PA} (in this case, what is typically done with *beer*), and contains the weighted TypeDM tuples $\langle\langle n_{PA}\ obj\ e_j\rangle, \sigma_j\rangle$, that is the expected events e linked to n_{PA} by an *object* link (*obj* is the link connecting the object and its verb), and their weights;

- $EX_e(v)$ is the set of the expected events for a metonymic verb v (in this case, things that are *finished*), containing the weighted TypeDM tuples $\langle\langle v\ comp^{-1}\ e_k\rangle$, $\sigma_k\rangle$, that is the expected events n linked to v by an inverse *complement* link ($comp^{-1}$ is the link connecting the verb and its complement), and their weights.

Then we combine and update these basic expectations to compute the covert event e for a logical metonymy. Depending on the tensor updating function used,

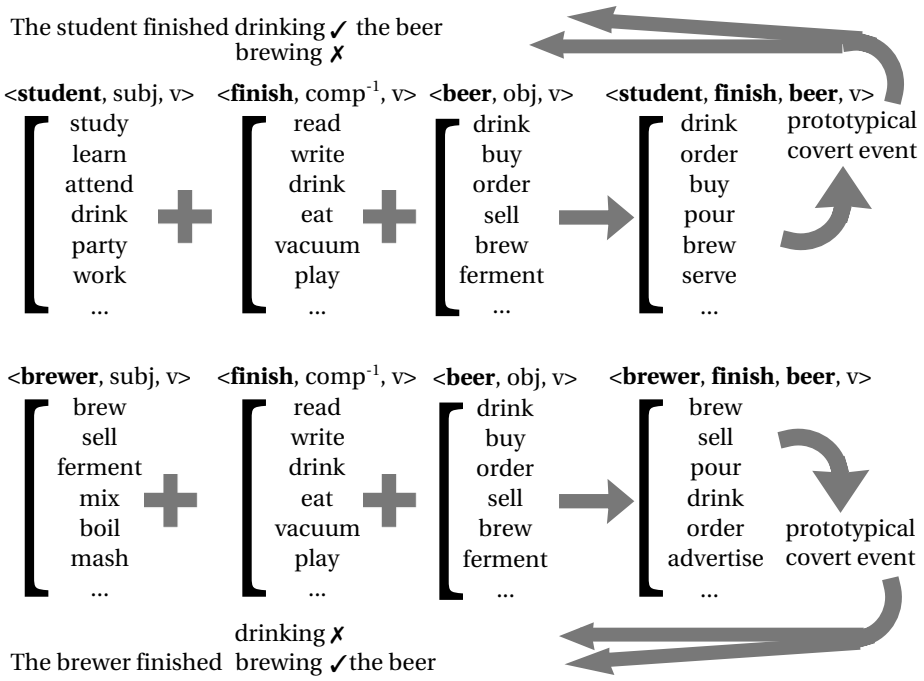

Figure 6.2.: Toy example for the SOV models: typical covert events for ⟨*student, finish, beer*⟩ and ⟨*brewer, finish, beer*⟩. The expected events for the subject, the metonymic verb and the object are retrieved from DM in a scored and ranked list, then the expectations are composed to generate the expected covert events for the logical metonymy.

and depending on the number of predictors used, we define four variations of the Similarity-based Model (toy example in Figure 6.2):

- The $\boldsymbol{SOV_{\Sigma}}$ **model** generates expected covert events for the agent, the metonymic verb and the patient, and then composes all three expectations using *sum* as the composition function;

- The $\boldsymbol{SOV_{\Pi}}$ **model** also generates expected covert events for the agent, the metonymic verb and the patient, and then composes all three expectations using *product* as the composition function;

- The $\boldsymbol{SO_{\Sigma}}$ **model** abstracts away from the metonymic verb, generating expected covert events for the agent and the patient, and composes the expectations using *sum* as the composition function;

- The SO_{Π} **model** also abstracts away from the metonymic verb, generating expected covert events for the agent and the patient, and composes the expectations using *product* as the composition function.

The product composition favors events which are strongly preferred by all sentential cues considered, but only retains events which appear in all expectation sets (those which do not are assigned a zero weight). The sum composition does not give such a high advantage to events which are strongly preferred by all sentential cues, but it does not strike out events which do not appear in all expectation sets either (they may just receive lower weights). Both functions are symmetrical, so the order of composition of the sentential cues is not relevant. What is relevant is that the model is provided with all the cues before generating a list of plausible covert events (similarly, the participants in the psycholinguistic experiments in Chapter 5 were presented with all sentential cues before reading the event).

After the update, the prototype computation proceeds as defined in the original ECU: (1) from the updated set of expectations, the top n ($n = 20$) are selected, (2) a prototypical (expected) covert event is obtained (centroid), (3) the thematic fit scores of possible events (*drinking* vs. *brewing*) are computed (cosine) and (4) the one with the best fit is chosen.

Distributional Semantic Models (DSMs) are traditionally non-compositional: distributional semantics has focused on the representation of words in isolation (for example, to compute similarities between them), but not in combination, and the integration of context into the vector computation is still largely an open question. Only recently did a number of promising studies investigate ways of combining different distributional representations of simpler units into a representation of the meaning of a more complex linguistic unit (Mitchell and Lapata, 2010; Guevara, 2011; Reddy et al., 2011) and, perhaps more ambitiously, of bringing together distributional semantics and the traditional realm of compositionality, formal semantics (Baroni et al., 2012; Erk, 2012). The Similarity-based Model presented here achieves **compositionality** by relying on ECU.

In the Similarity-based Model, covert events (without any a priori limitation on their range) are sampled from distributional knowledge about typical predicate-argument structures (Padó and Lapata, 2007; Erk et al., 2010), exploiting **thematic fit** information. The integration of cues is performed in a way which is analogous to prototype-based

concept combination (see Rosch, 1975; Smith et al., 1988; Hampton, 1991; Kamp and Partee, 1995): event prototype(s) are evoked by agent and patient, and then combined into a new prototype event.

Assigning plausibility scores to covert events semantically similar to the predicted event was problematic for the Probabilistic Model[7]: if *sipping* is not found in the corpus in combination with *aunt* and *tea*, then the Model is not able to score *sipping* as a covert event for *the aunt finished the tea*. The Similarity-based Model achieves this by **similarity-based** generalization: for each logical metonymy, a prototypical event (with high-thematic fit with the arguments) will be expected, and other events will be more or less expected depending on their semantic similarity to the prototype. For example, *sipping*, being semantically similar to *drinking*, may receive a high plausibility score as covert event for *the aunt finished the tea*, even if it was never encountered in combination with *aunt* and *tea*.

6.4. Evaluation

6.4.1. Task and Dataset

Two experimental conditions (high vs. low typicality) were contrasted in the psycholinguistic experiments in Chapter 5, which yielded a significant effect of typicality on reading and decision latencies. We expect a computational model of covert event interpretation to be equally sensitive to typicality effects determined by generalized event knowledge, and to be able to capture the same differences by choosing the covert event matching the generalized event knowledge cued by the sentential context (agent and patient)[8].

The dataset used for the evaluation was the same for Experiment 1: 96 sentences, built from 24 sets of four $\langle s, v, o, e \rangle$ tuples (two high-typicality ones and two low-typicality ones), where s is the object, v the metonymic verb, o the object and e the covert event (e.g. $\langle$ *baker, finish, icing, spread* $\rangle$). We also replicated the evaluation for the dataset from Experiment 2b (again 96 sentences, build from 24 sets of four

[7] It is problematic unless some sort of smoothing is provided (see General Discussion at the end of this chapter).

[8] The probe recognition experiments showed that the ISI played a role as well, as the effect of typicality emerged early (at short ISI), but this aspect can not be modeled by ECU, as the weights assigned to the covert events in the model do not decay over time.

	covert event	
	high typicality	low typicality
Der *Konditor* hörte mit der *Glasur* auf The baker finished the icing	**auftragen** **spread**	**essen** **eat**
Der *Kind* hörte mit der *Glasur* auf The baker finished the icing	**essen** **eat**	**auftragen** **spread**

Table 6.3.: Example materials for the experiments in Chapter 5.

$\langle s, v, o, e \rangle$ tuples), as its materials introduced some changes compared to the other two psycholinguistic experiments.

The items were completely counterbalanced (similar to the Bicknell et al., 2010 dataset used by Lenci, 2011), so it was possible to obtain 48 agent-verb-patient combinations, each paired with the two covert events assigned to it in the dataset (high vs, low. typicality, see Table 6.3). We compared the Probabilistic Models with the Similarity-based Models on a pairwise comparison task similar to the one in Lenci (2011): given a $\langle s, v, o \rangle$ tuple and a pair of covert events e, e' (e.g. $\langle$ *baker, finish, icing* $\rangle$ $\rightarrow$ *spread*, *eat*), the task was to pick the high-typicality covert event for the given triple. Due to the size of the psycholinguistic experiments in Chapter 5, which did not allow for a point-wise prediction of experimental measurements, we employed the pairwise comparison as a simple and straightforward evaluation strategy (see the discussion on evaluation methods in Chapter 2).

For the Probabilistic Models, we compare the probabilities $P(s, v, o, e)$ and $P(s, v, o, e')$ – a model scores a "hit" if the $P(s, v, o, e)$ for the high-typicality e is higher than the $P(s, v, o, e')$ for the low-typicality e'. Analogously, for the Similarity-based Models, we computed the thematic fit of e and e' as the similarity of their vectors the prototype vectors for the expected covert event and predict the one with higher similarity: a model scores a "hit" if the prototypical event vector for $\langle s, v, o \rangle$ has a higher thematic fit /similarity with the high-typicality e than with the low-typicality e'.

The models were evaluated for coverage (as the percentage of data points where a prediction can be made[9]) and accuracy (on the covered contexts, the ratio of correct

[9] A prediction can be made if at least one of the two scores ($P(s, v, o, e)$ and $P(s, v, o, e')$ for the Probabilistic Models and the thematic fit scores of e and e' for the Similarity-based Models) is higher than zero.

predictions to the number of predictions made by the model), but also on a measure that combines the two, *Back-off Accuracy*:

$$coverage \times accuracy + ((1 - coverage) \times 0.5)$$

This last measure emulates a back-off procedure (when a prediction is not available, the model assumes baseline performance, that is 50%) and tends to degrade towards baseline performance for low-coverage models.

We evaluated the significance of differences between models with a χ^2 test, applied to a 2×2 contingency matrix containing the number of correct and incorrect answers, where data points outside a model's coverage count half for each category, which corresponds exactly to the definition of Back-off Accuracy.

6.4.2. Baselines

Following Lapata and Lascarides (2003), we evaluated the Probabilistic Models against a baseline ($\boldsymbol{B_p}$). Given a $\langle s, v, o \rangle$ tuple and two covert events e and e', the Probabilistic Baseline scores a "hit" if the $P(e|o)$ for the high-typicality e is higher than the $P(e'|o)$ for the low-typicality e'.

The Similarity-based Models were evaluated against a Similarity Baseline ($\boldsymbol{B_s}$) which, given an $\langle s, v, o \rangle$ tuple and two covert events e and e', scores a "hit" if the prototypical event vector for o has a higher thematic fit /similarity with the high-typicality e than with the low-typicality e'.

Interestingly, what both baseline models do (namely, associating one event to the patient, depending on corpus-based probability / typicality and regardless of other cues) corresponds to a simple notion of probability-based qualia roles, which associates each noun with at most two qualia events (in our case one). For example, the baselines will probably associate *Bier* (*beer*) with *trinken* (*drink*) and *Auto* (*car*) with *fahren*, (*drive*).

Our dataset is counterbalanced — that is, each covert event (e.g. *fahren*, drive) appears once as the high-typicality event for a given patient (e.g. *Chauffeur* + *Auto*, driver + car) and once as the low-typicality event (with a different agent, e.g. *Mechaniker* + *Auto*, mechanic + car). Without taking the agents into consideration, the baselines will make a correct prediction for exactly 50% of the agent-patient combinations: for example, it will score a "hit" for $\langle$*Chauffeur, vermeiden, Auto*$\rangle \rightarrow$ *fahren, reparieren*

and a "miss" for ⟨*Mechaniker, vermeiden, Auto*⟩ → *fahren, reparieren*. Note, however, that this is not a *random* baseline: the choice of the covert event is made deterministically by the baseline models on the basis of the input parameters (o, e and e'), and it achieves 50% accuracy only because of the balanced design of the dataset.

6.4.3. Results

We contrasted results from both classes of models (Probabilistic and Similarity-based), and from the baselines. Recall that both the Probabilistic and the Similarity-based models are presented in two versions, one including the metonymic verb (*SOV*) and one abstracting away from it (*SO*)[10], and that the Similarity-based Models employ two different composition functions, sum and product.

Probabilistic Models

The Probabilistic Models yielded lower coverage than the Similarity-based Models (see Table 6.4), with SO_p scoring better coverage than SOV_p (Experiment 1 dataset: 75% vs. 44%; Experiment 2b dataset: 65% vs. 35%). The SOV_p model was unable to make a prediction for more than half of all contexts, because many $\langle o, v \rangle$ combinations were simply not attested in the corpus. Even on the covered items, the SO_p model was still more reliable than the SOV_p model (Experiment 1 dataset: 75% vs. 62% accuracy; Experiment 2b dataset: 71% vs. 59% accuracy). The metonymic verb did not systematically help to predict the covert event, but rather introduced noisy estimates.

Similarity-based Models

The Similarity-based Models did not have major coverage issues, the only problematic item being ⟨ *Pizzabote, Pizza* ⟩ (i.e. ⟨ *pizza delivery man, pizza* ⟩) which was paired with the covert events *liefern (deliver)* and *backen (bake)*: the Similarity-based Models required transitive constructions for *Pizzabote*, which were not attested in the corpus (or were not found by the parser). The difference between the *SOV* and *SO* models for the Experiment 1 dataset was not clear-cut and was dependent on the choice of

[10] Note that the comparison between *SOV* and *SO* models was not carried out by Lapata and Lascarides (2003) and Lapata et al. (2003), as only the former (the SOV_p model, corresponding to their Simplified Model) is implemented by them, while the *SO* model was introduced by us for better comparison with the Similarity-based Models we considered.

	Dataset Experiment 1							
	Probabilistic Models			Similarity-based Models				
	B_p	SOV_p	SO_p	B_s	SOV_Σ	SOV_Π	SO_Σ	SO_Π
Accuracy	50%	62%	75%	50%	66%	53%	66%	68%
Coverage	100%	44%	75%	100%	98%	94%	98%	98%
Back-off Accuracy	50%	55%	**69%**	50%	66%	53%	66%	**68%**

	Dataset Experiment 2b							
	Probabilistic Models			Similarity-based Models				
	B_p	SOV_p	SO_p	B_s	SOV_Σ	SOV_Π	SO_Σ	SO_Π
Accuracy	50%	59%	71%	50%	57%	45%	64%	62%
Coverage	100%	35%	65%	100%	98%	83%	98%	98%
Back-off Accuracy	50%	53%	**64%**	50%	57%	46%	**64%**	62%

Table 6.4.: Results for all Probabilistic and Similarity-based models on datasets from Experiments 1 and 2b. Due to a small error in the implementation in Zarcone et al. (2012d), the results reported here are slightly different than those reported the article. The overall analysis though remains unchanged.

composition operation. As sum is more robust, for sum models the inclusion of the metonymic verb (SOV_Σ vs. SO_Σ) did not make a big difference. On the other hand, a major difference was found between the two product models SOV_Π and SO_Π, the former being the worst model at near-baseline performance, and the latter being the best one. Again, the metonymic verb introduced noisy expectations which disrupted the update process, which was a problem in particular for the product model, because with the product any factor (also one noisy variable such as the metonymic verb) had a greater influence on the final result.

A bigger difference between the *SOV* and *SO* Similarity-based models was yielded for the Experiment 2b dataset. Recall that the metonymic verbs in Experiment 2b were only aspectual verbs, and thus semantically rather "empty". It is then reasonable to expect them to bring even more noise to the update process, again more evidently in the product model.

Best Models and Baseline Comparison

The back-off accuracy scores for both best models for the Experiment 1 dataset significantly outperformed the deadline (Baseline vs. SO_p model: $\chi^2 = 5.506, p = 0.02$, Baseline vs. SO_Π model: $\chi^2 = 5.506, p = 0.02$) but did not significantly differ from each other in the χ^2 test (not surprisingly, given the small size of our dataset). The Experiment 2b dataset was somewhat harder (highly associated covert events in Experiment 1 were replaced by less frequent synonyms), and the χ^2 tests did not reach significancy.

The best models from the two families (SO_p and SO_Π for Experiment 1 and SO_p and SO_Σ for Experiment 2b) were only informed by the subject and the object, showing that the subject contributed to the models with a significant improvement. While the accuracy of the best Probabilistic Model was higher than the accuracy of the best Similarity-based Model (75% vs. 68%), its coverage was much lower (only 75% of the contexts), while the distributional model SO_Π covered all items but one (98%). A similar picture emerged for Experiment 2b, albeit with slightly lower accuracy scores.

Qualitative Analysis

Let us now look at some examples. The Probabilistic Models, relying on first-order co-occurrence, had coverage problems. For example, neither of them could assign a probability to ⟨ *Dieb, schmuggeln* / *schleifen, Diamant* ⟩ and ⟨ *Juwelier, schmuggeln* / *schleifen, Diamant* ⟩ (⟨ *thief, smuggle* / *cut, diamond* ⟩ and ⟨ *jeweler, smuggle* / *cut, diamond* ⟩), which appeared in both datasets, because neither subject occurred with either of the verbs in corpus, even though *Diamant* did occur as the object of both. In contrast, the Similarity-based Models were able to compute expectations for these triples from second-order co-occurrences, by exploiting other verbs co-occurring with *Diamant*, and were not punished by the extra context, as both *Dieb* and *Diamant* were associated with the verbs: *stehlen* (steal), *rauben* (thieve), *holen* (*get*), *entwenden* (*purloin*), *erbeuten* (*snatch*), *verkaufen* (*sell*), *nehmen* (*take*). All these events are typical events for a thief, which fits the intuition that *Dieb* is more predictive of the event than *Diamant*.

Both Probabilistic Models predicted *fahren* (drive) for ⟨ *Chauffeur, Auto* ⟩ (⟨ *driver, car* ⟩) and for ⟨ *Mechaniker, Auto* ⟩ (⟨ *mechanic, car* ⟩), due to the high overall frequency of *fahren*. The Similarity-based Models, however, took the mutual information into

$EX_{SO}(\langle Chauffeur, Auto \rangle)$		$EX_{SO}(\langle Mechaniker, Auto \rangle)$	
fahren	*(drive)*	*bauen*	*(build)*
bringen	*(bring)*	*lassen*	*(let/leave)*
lassen	*(let/leave)*	*sehen*	*(see)*
parken	*(park)*	*reparieren*	*(repair)*
sehen	*(see)*	*brauchen*	*(need)*
machen	*(make)*	*besitzen*	*(own)*
halten	*(keep/hold)*	*machen*	*(make)*
steuern	*(steer)*	*stellen*	*(put)*

Table 6.5.: Updated covert event expectations in SO_{Π} for *Chauffeur* (*chauffeur*) and for *Mechaniker* (*mechanic*) combined with the expectations for *Auto* (*car*).

account and were thus able to determine events that were more strongly associated with *Mechaniker* (e.g. *bauen, reparieren*, etc.) while at the same time discounting the uninformative verb *fahren*.

There were, however, items that all models have difficulty with. Some were due to a frequency disparity between the high- and low-typicality event. *Schreiben*, for example, occurred more frequently than *benoten*, leading to incorrect predictions for ⟨ *Lehrerin, benoten / schreiben, Klausur* ⟩ (⟨ *teacher, grade / take, exam* ⟩). In the case of ⟨ *Schüler, lernen / schreiben, Geschichte* ⟩ (⟨ *student, learn / write, story* ⟩), none of the models correctly chose *lernen* as the preferred event. The Probabilistic Models were influenced by the very frequent *Geschichte schreiben* which is part of an idiomatic expression (*to write history*). On the other hand, the Similarity-based Models choose the *story* and *history* sense of the word given the following most informative events: *erzählen* (*tell*), *schreiben* (*write*), *lesen* (*read*), *hören* (*hear*), and *studieren* (*study*).

Neither the Probabilistic Models or the Similarity-based Models were able to correctly choose *auspacken* (*unwrap*) over *einpacken* (*wrap*) for ⟨ *Geburtstagskind, Geschenk* ⟩ (⟨ *birthday-boy/girl, present* ⟩): the Probabilistic Models were not able to make a prediction due to the models' coverage problems, whereas for the Similarity-based Models, while both *auspacken* and *verpacken* (*wrap*) were highly associated with *Geschenk*, the most strongly associated actions of *Geburtstagskind* were many and highly diverse: e.g. *bekommen* (*receive*), *sagen* (*say*), *auffuttern* (*eat up*), *herumkommandieren* (*boss around*), *ausblasen* (*blow out*). This diversity of scenarios made it difficult for the Similarity-based Models to identify a clear event prototype.

6.5. General Discussion

I have presented a contrastive study of two classes of computational models of covert event interpretation: Probabilistic Models and Similarity-based Models. Both classes of models were able to integrate contextual cues (agent and patient), outperforming the baselines which only took into account information coming from the patient. Furthermore, Similarity-based Models rivaled (in accuracy) and outperformed (in coverage) Probabilistic Models.

The coverage issues with simple probabilistic models based on maximum likelihood estimate, such as those described in this chapter, are not likely to be solved with a larger amount of data. The model is already relying on data from a web-sized corpus of German (which is a considerably-sized corpus and definitely bigger than the corpora available for many other languages not as well represented on the web). These models only use first-order co-occurrence information, which suffers from sparsity issues even in large web corpora, as both unfrequent (but possible) and dispreferred combinations are assigned zero probability. This does not exclude that other probabilistic models, resorting to some sort of smoothing (e.g. generative models introducing latent variables to model clusters based on higher-order co-occurrences, Prescher et al., 2000), may achieve coverage and accuracy values comparable to those of the Similarity-based Models.

The Similarity-based Models on the other hand did not have coverage problems, because they have a way of smoothing the estimated scores for unseen items by taking advantage of higher-order co-occurrences (Dagan et al., 1999). First, a prototype event is computed for the given contextual cues, and then, given a new event (possibly unseen with such cues due to data sparsity), its thematic fit given those contextual cues is computed as the similarity to the prototype. Thematic fit / similarity scores can thus also be computed for covert events with low co-occurrence frequencies with the agent and patient, allowing Similarity-based Models to achieve higher coverage than the Probabilistic Models while maintaining their accuracy.

The main attractiveness of Similarity-based Models is that they constitute a model of meaning representation in the lexicon, as they are strictly connected with a cognitive hypothesis about semantic representations (the Distributional Hypothesis). On the other hand, while the Probabilistic Models clearly represent the combination of contextual cues as joint probabilities, the Similarity-based Models do not have a

straightforward way to take compositionality and context-dependence into account. As we have shown, both functions used here (sum and product) have their limitations (the sum models suffer less from noise, but neither of them performs better than the SO_{Π} model for the Experiment 1 dataset) and the model might profit from a different way of implementing the combination of intra-sentential cues. For instance, the metonymic verb may not simply be a source of noise, rather it may only require being weighted differently than the agent cue and the patient cue.

The scope of Similarity-based Models is also somewhat limited: if on the one hand the idea of composing and updating expectations in the vector-space seems compatible with the integration of a growing amount of cues (the metonymic verb, but also, for example, extra-sentential context), on the other hand models such as ECU are not scalable in a straightforward way, but are bound to the cues which can be retrieved via syntactic dependency (intra-sentential cues: in this case, subjects and objects). It is then more challenging to integrate different sources of knowledge (as, for example, extra-sentential knowledge as represented in a topic model) in a Similarity-based Model such as ECU. On the other hand, Probabilistic Models are easier to scale to extra-sentential knowledge, which can be represented as another variable (*Ctxt*) in a joint distribution (e.g. $P(s, v, o, e, Ctxt)$, see also the Surprisal model of processing difficulty in Levy, 2008).

A limitation of both groups of models is that none has a way to account for differences between event activation at different ISI intervals. Neither the probability estimates from the Probability Models or the thematic fit scores from the Similarity-based Models are a function of the time elapsed after the presentation of the stimulus. Nevertheless, the fact that in the pairwise comparison task the models successfully replicate the results from the psycholinguistic experiments suggests that the distributional information the models rely on is a correlate of generalized event knowledge activation at the same early stages where it intervened in the psycholinguistic experiments.

6.6. A Thematic-fit Model of Covert Events

I have shown that a thematic-fit model (such as the Similarity-based Models) can successfully account for fine-grained (high- vs. low- typicality) differences between covert events for a given metonymy which I have ascribed to generalized event knowledge.

The model was able to distinguish between covert events by relying on thematic fit only, without encoding any information on qualia roles, suggesting that thematic fit can be a convincing computational counterpart of typicality determined by generalized event knowledge.

The model has an effective (albeit not perfect) way of combining contextual cues, which is defined as their role in reshaping the expectation for a covert event given a logical metonymy. Elman (2009, 2011) had sketched a model similar to his simple recurrent network (Elman, 1990) to account for the incremental combination of contextual cues in the Words-as-cues framework, but did not provide a concrete implementation and evaluation of the model in the light of this hypothesis. Following Lenci (2011), we have adapted the ECU model to account for covert event interpretation, thus providing the Words-as-cues Hypothesis with a model of how contextual cues may be combined.

The model exploits a similarity-based generalization step (e.g. computing the thematic fit for the rare *sipping* as covert event for *the aunt finished the tea*), providing a way of conceptualizing covert event *senses*: after computing the prototype, the model can compute a thematic fit score for several possible covert events, even if they do not occur with the agent and the patient of the logical metonymy. Interestingly, this generalization is the factor that helps our model rival the Probabilistic Model in Lapata and Lascarides, 2003 and Lapata et al., 2003, by smoothing estimated scores for unseen events, and thus achieving better coverage. Similarity-based generalization also allows us to conceptualize covert events not in terms of single predicates, but as event senses, with more prototypical and less prototypical representatives. The similarity (distance) of an event to the prototype determines how expected the event is: given certain contextual cues, the model is directed towards an event scenario (e.g. scenarios involving *aunts* and *tea*) with a prototypical event (*drinking*), which will cause similar events also to be expected (e.g. *sipping*). Different contextual cues will direct the model towards a different scenario, with different expected events (e.g. *company*, *tea* → *export*).

The convergence of results from the psycholinguistic experiments and from the computational modeling supports a thematic-fit model of covert event interpretation. I will now move on to the problem of the trigger of the logical metonymy, which will be discussed in Part III, and I will explore the hypothesis that thematic fit can account for the trigger problem as well as for the source / range problem.

Part III.

The Trigger of the Logical Metonymy

7. The Trigger of the Logical Metonymy: Computational Models

In the previous part, the main focus was the covert event: I presented converging evidence from psycholinguistic experiments and from computational models that thematic fit (informed by generalized event knowledge) guides covert event interpretation: given a logical metonymy, the event with the highest thematic fit with agent and patient is chosen. I will now address the trigger problem: what **triggers** the metonymic interpretation, distinguishing constructions which require covert events from those which do not?

Metonymic verbs are a key element in accounting for the trigger of the logical metonymy, and the two main approaches differ with regard to their characterization of these verbs: for the **Lexical Hypothesis**, metonymic verbs are event-selecting verbs whose argument-selecting behavior determines when they are combined with an entity-denoting object (e.g. *begin the book*), triggering a logical metonymy interpretation, whereas the **Pragmatic Hypothesis** simply claims that metonymic verbs trigger presuppositions which are not different than those triggered by other verbs, and that such presuppositions arise from the underspecification of logical metonymies (e.g. *begin the book* can be interpreted as *reading, writing, translating the book*...). The crowdsourcing study in Chapter 4 showed that verbs differ in their tendency to trigger metonymic interpretations, which can be depicted as a continuum spanning from more metonymic verbs to less metonymic ones. I will now address the issue of defining what a metonymic verb is, and I will then evaluate the prediction from the **Words-as-cues Hypothesis** that the logical metonymy is triggered by low thematic fit between the event-selecting metonymic verb and its non-event-denoting argument.

This chapter will be devoted to two computational studies: the first study presents a model characterizing verbs in terms of their eventhood, that is their preference for event-denoting objects, which is applied to sets of metonymic and non-metonymic

verbs used in two psycholinguistic studies; the second study models verb-object thematic fit scores for well-known psycholinguistic datasets used in studies on coercion. Both studies suggest that thematic fit can successfully account for the distinction between coercive and non-coercive sentences, suggesting (as the Words-as-cues Hypothesis predicts) that it may be the sole factor responsible for the trigger of the logical metonymy.

7.1. What is a Metonymic Verb?

Studies of logical metonymy have employed an incredibly diverse set of verbs which arguably select for event-denoting objects and trigger metonymic interpretations (see for example the English verbs in Table 7.1). Besides aspectual verbs (e.g. *begin, finish*, which form a fairly well defined class, Levin, 1993), the set includes a mix of non-aspectual verbs, whose criteria of inclusion are not discussed, such as psychological verbs (e.g. *enjoy, endure, savor*), or others that do not seem to belong to the same verb class (e.g. *attempt*, *expect*, *survive*, *try*). Consider the following examples, from a well-known psycholinguistic study on logical metonymy:

(7.1) a. The victim **endured** / reported the robbery / **the driver**.

b. The banker **expected** / remembered the audit / **the check**.

(Traxler et al., 2002)

The first verb is used in the metonymic condition and the second one in the non-metonymic condition (both verbs are then combined with an event-denoting object and an entity-denoting object), but it is not clear why *endure* and *expect* should be considered event-selecting verbs, imposing restrictions on its object arguments, while *report* and *remember* should not. Recall that Traxler et al. (2002) reported higher processing costs for the metonymic condition (in boldface in 7.1.a-b: metonymic verb + entity-denoting object), which were ascribed to the coercion operation. They define metonymic verbs as the verbs triggering a logical metonymy, and the logical metonymy as a shift triggered by metonymic verbs: if the argument-selecting characteristics of metonymic verbs are the trigger of the logical metonymy, then it is important to justify the choice of verbs in experimental studies (and even more in the studies searching

McElree et al. (2001),de Almeida (2004) Pickering et al. (2005)	*attempt, begin, endure, enjoy, expect, finish, master, prefer, resist, savor, start, survive, try*
Traxler et al. (2002)	*attempt, begin, complete, end, endure, enjoy expect, finish, master, prefer, resist, start, try*
Lapata and Lascarides (2003)	*attempt, begin, enjoy, finish, postpone, prefer, resist, start, survive, try, want*
Traxler et al. (2005)	*begin, complete, enjoy, finish, master, resist, start, try*
Pylkkänen and McElree (2007)	*attempt, begin, complete, endure, enjoy, finish, master, start, try*
Frisson and McElree (2008)	*attempt, begin, complete, continue, endure, enjoy, finish, prefer, resist, start, try*
Baggio et al. (2010)	*attempt, begin, complete, endure, enjoy, finish, master, manage, resist, start, try*
Zarcone and Padó (2010)	*begin, continue, endure, end, enjoy, finish, prefer, savor, start, try*

Table 7.1.: English metonymic verbs used in studies on coercion.

for processing costs of coercion) using criteria which are different than the mere assumption that these verbs trigger metonymic shifts, in order to avoid circularity.

Recent work by Katsika et al. (2012) has questioned a homogeneous notion of "metonymic verbs", arguing that "the hypothesis that eventive inferences must be attributed to the same mechanism of building meaning (coercion + type-shifting) [for all metonymic verbs] is too strong" (Katsika et al., 2012, p. 60). Their study contrasts an aspectual verb condition (7.2.a), where the verb selects for an event-denoting object, a psychological verb condition, where the verb is compatible with entity- and event-denoting objects (7.2.b), and a control condition with a clearly entity-selecting verb (7.2.c)[1]:

[1]The aspectual and most of the psychological verbs used in Katsika et al. (2012) were selected from those used in previous experimental studies.

(7.2) The new interns, Alexandra and John, loved to read novels

a. Alexandra **was completing a sci-fi book** when the secretary announced the meeting. (aspectual)

b. Alexandra **was enjoying a sci-fi book** when the secretary announced the meeting. (psychological)

c. Alexandra **was shelving a sci-fi book** when the secretary announced the meeting. (control)

(Katsika et al., 2012)

An eye-tracking experiment showed that participants spent significantly more time re-reading the verb (second-pass time) in the aspectual condition than in the other two, and the objects of aspectual verbs also evoked more first-pass regressions and longer second-pass times than those of psychological and control verbs. Two regions after the verb, no significant differences were reported between aspectual and psychological conditions, whereas the aspectual condition yielded significantly more regressions than the control condition. These results support the hypothesis that not all metonymic verbs are equal: the authors claim that aspectual verbs trigger compositional and inferential processes (the compositional processes correspond to the coercion operation, whereas the inferential processes, that is the retrieval of the covert event, that they do not consider costly), and that psychological predicates only trigger inferential processes.

This difference in behavioral patterns has already determined a repetition of our Experiment 2 (Experiment 2b), which excluded non-aspectual metonymic verbs and substantially confirmed the results from Experiment 1 and 2. I will now present a corpus-based approach to verb classification, which characterizes verbs in terms of their behavior at the syntax-semantics interface, namely in terms of their preference for event-denoting objects. If a verb's selectional behavior (as modeled from corpus-extracted data) can effectively distinguish between aspectual metonymic, psychological metonymic and non-metonymic verbs, this result would speak against the claim from the Pragmatic Hypothesis that metonymic verbs are not different from other verbs that trigger post-lexical presuppositions, and would suggest that such difference is indeed grounded in a verb's fit with its arguments.

7.2. A Computational Model of Eventhood

We built a corpus-based model of verb classification based on a measure of **eventhood** (Utt et al., 2013), that is a quantitative measure of a verb's argument selecting behavior in terms of its preference for event-denoting objects, in order to ground the distinction between metonymic and non-metonymic verbs on a more objective footing than the experimenter's intuitions and assumptions.

The model computed eventhood scores on metonymic and non-metonymic verbs used in existing psycholinguistic datasets (which are based on this distinction), allowing us to perform post-hoc analyses on these datasets. If metonymic verbs are indeed more associated with event-denoting objects, then we should expect the eventhood measure to distinguish them from non-metonymic verbs. We also expect metonymic aspectual verbs to yield higher eventhood scores than metonymic non-aspectual verbs. This more fine-grained distinction would further bolster the proposal in Katsika et al. (2012), that we should carefully control for the type of metonymic verbs used in experimental studies on logical metonymy.

7.2.1. Measuring the Event Expectations of Verbs

Eventhood is defined in terms of expectations, as the degree to which verbs expect event-denoting rather than entity-denoting objects. The computational model of eventhood presented here is hybrid: it is thematic-fit driven (similar to the distributional models presented in Chapter 6), as it estimates the verb's expectations for typical objects from Distributional Memory (DM, Baroni and Lenci, 2010, see Chapter 6[2]), but it also incorporates a type-based component, which relies on the WordNet lexical hierarchy (Fellbaum, 1998)[3] to discover whether a noun has an event sense. The typical objects for a verb are mapped into event and non-event "types", and the typed expectations are then exploited to estimate a verb's preference for event-denoting objects.

[2]As the evaluation datasets were in English, we used the English TypeDM, which is available at `http://clic.cimec.unitn.it/dm/`.

[3]We used version 3 of WordNet.

WordNet node	Count	Examples
EVENT	11248	*training, splat, Alamo, suicide, hyperalimentation*
ACT/DEED/HUMAN ACTION/ HUMAN ACTIVITY, ACTION, ACTIVITY	9845	*banditry, dissolution, beanball, messaging, banishment*
PROCESS/PHYSICAL PROCESS	2590	*ultracentrifugation, desensitization, extinction, superconductivity*
PROCESS/COGNITIVE PROCESS/ MENTAL PROCESS/OPERATION/ COGNITIVE OPERATION	998	*reminiscence, breakdown, score, analogy, inference*
ORGANIC PROCESS/ BIOLOGICAL PROCESS	878	*recuperation, emission, autoregulation, drinking, blossoming*
all	14143	

Table 7.2.: High-level event-denoting nodes in WordNet (Utt et al., 2013).

Estimating Object Expectations

The verbs' selectional preferences have been successfully modeled in previous work by exploiting the (typical) fillers of a verb's argument slot within a particular subcategorizing frame, relying on a lexical hierarchy (Resnik, 1996) and / or on distributional information (Rooth et al., 1999; Erk et al., 2010; Schulte im Walde et al., 2009). In the case of metonymic verbs, we are interested in the object slot and in particular in learning how event-like their most expected (most associated) objects are. The most expected objects for a verb (that is the most typical fillers for the verb's object position) can be easily estimated as the top k object fillers most highly associated with it via an object link in the *word by link-word* space in DM, where $\langle w_1\ l\ w_2 \rangle$ triples are weighted with regard to their Local Mutual Information (Evert, 2005), indicating how strongly w_1 is associated with w_2 (see Chapter 6). Thus, we simply extracted fillers linked to the target verbs by an *obj* link (e.g. for *postpone* we may select ⟨*meeting obj postpone*⟩ and ⟨*breakfast obj postpone*⟩).

It is common to select the top k most associated (prototypical / expected) fillers as a reliable method to characterize a verb's selectional preferences (Baroni and Lenci, 2010; Lenci, 2011, see also the model of covert event interpretation in Chapter 6). We fix k at 100, in order to eliminate the issue of using words from DM which are not covered in WordNet (as it may be the case for less frequent objects).

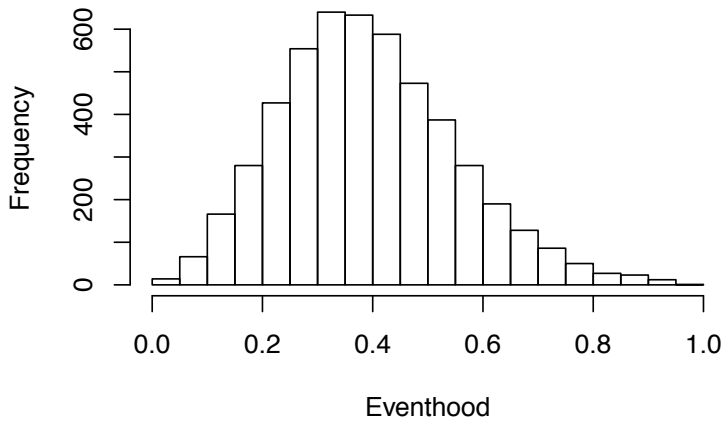

Figure 7.1.: Histogram of eventhood across verbs in DM (Utt et al., 2013).

Discovering Event Senses

We distinguished event-denoting objects from non-event-denoting objects using WordNet, defining event-denoting nouns as nouns with at least one synset that is dominated in the hierarchy by an event-denoting top node (see a list of event-denoting nodes in Table 7.2). This is simply an approximation and does not take into account distributional information when estimating the degree to which the noun is used in an event-denoting sense. Thus, we obtained a set of 14K event nouns (out of WordNet's 170K nouns).

Measuring Eventhood

We then selected the top k (100) object fillers $obj_k(v)$ for each target verb v and defined the eventhood ϵ_k of a verb's object slot (in short, the verb's eventhood) as the proportion of fillers with an event sense (EV):

$$\epsilon_k(v) = \frac{|EV \cap obj_k(v)|}{k}$$

We can then rank the verbs in DM according to their eventhood (see the distribution in Figure 7.1): interestingly, the leftmost bar in the histogram ($\epsilon < 0.05$, low-eventhood verbs) contains verbs which are typically combined with people as patients (e.g. *marry, behead*) and which would be rather uncommon with event-denoting objects; the other side of the spectrum ($\epsilon > 0.9$, high-eventhood verbs) contains verbs concerning the

temporal unfolding of an event (e.g. *commence, cease, halt, delay*). The most frequent range accepts both entity- and event-denoting objects and is semantically rather diverse (e.g. *prance, emaciate, exorcise, downsize*), which is not surprising, as we are only considering one aspect of the verbs' meaning, that is their event-selecting behavior.

7.2.2. Evaluation on Psycholinguistic Datasets

The eventhood model was tested on two experimental datasets from the above mentioned psycholinguistic studies on logical metonymy:

Traxler et al. (2002) Dataset: this dataset is composed of 24 verbs used in Experiment 2 and 3 in Traxler et al. (2002), divided into *metonymic* and *non-metonymic* verbs[4].

Katsika et al. (2012) Dataset: this dataset is composed of 38 verbs used in Katsika et al. (2012) and taken mostly from previous psycholinguistic experiments on coercion. The authors distinguished three sets of verbs: *metonymic aspectual*, *metonymic psychological* and *non-metonymic* verbs[5].

Recall that both studies contrasted experimental conditions where verbs were grouped with respect to their event-selecting behavior, expecting higher processing costs when "metonymic verbs" were combined with non-event-denoting objects than when "non-metonymic verbs" were. If it is indeed the verbs' event-selecting behavior that determines the higher processing costs, then we expect our eventhood measure to successfully distinguish between the classes used in the psycholinguistic studies.

We evaluated the eventhood measure in two ways:

1. The Wilcoxon rank sum test (a non-parametric alternative to the Student's t-test) was used to check for differences in eventhood between verb classes in both datasets.

[4] *Event verbs* and *neutral verbs*, according to the terminology of the study.
[5] *Aspectual*, *psychological* and *entity-selecting*, according to the terminology of the study. We excluded non-transitive verbs (*subscribe to, work on*).

Traxler et al.'s (2002) dataset	
metonymic	*begin, complete, end, endure, enjoy, expect, finish, prefer, start*
non-metonymic	*approve, curse, describe, forget, ignore, observe, outline, praise, prepare, recall, recollect, remember, report, see, watch*
Katsika et al.'s (2012) dataset	
metonymic aspectual	*begin, complete, continue, finish, start*
metonymic psychological	*enjoy, face, favor, prefer, resist, stomach, tolerate*
non-metonymic	*access, auction, buy, conduct, contribute, deliver, destroy, drop, fax, find, inspect, misplace, open, peruse, purchase, photocopy, rent, sell, send, shelve, shred, submit, trash, unearth, unpack, write*

Table 7.3.: Datasets from Traxler et al. (2002) and Katsika et al. (2012).

2. In Traxler et al.'s (2002) dataset, each sentence appears once with a metonymic verb and once with a non-metonymic verb, which gives us a list of verb pairs. This list allows us to compute the number of times the metonymic verb yielded a higher eventhood score than the corresponding non-metonymic verb in the same sentence pair.

7.2.3. Results and Discussion

On the Traxler et al. (2002) dataset, the difference between metonymic and non-metonymic verbs was close to significance, with p just above 0.05 ($W = 100.5, p < 0.053$), and reached significance when we removed the four non-aspectual metonymic verbs in the dataset (*endure, enjoy, expect, prefer*: $W = 67.5, p < 0.01$). On the Katsika et al. (2012) dataset, metonymic aspectual verbs yielded higher eventhood scores than metonymic psychological verbs and non-metonymic verbs, and all pairwise comparisons were significant: metonymic aspectual vs. metonymic psychological verbs ($W = 30, p < 0.05$); metonymic aspectual vs. non-metonymic verbs ($W = 125, p < 0.01$); metonymic psychological vs. non-metonymic verbs ($W = 18.5, p < 0.01$). See also the

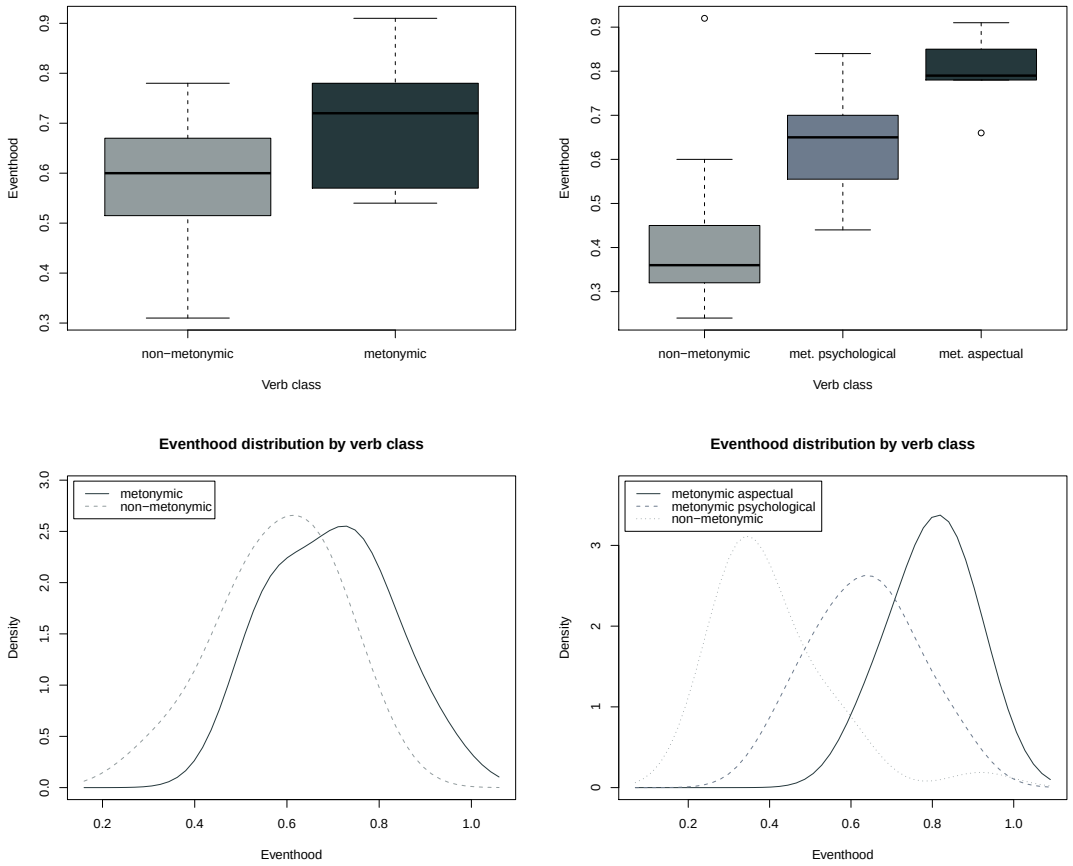

Figure 7.2.: Comparing eventhood distributions for verb classes in the Traxler et al. (2002) dataset (left) and in the Katsika et al. (2012) dataset (right), with box plots and density plots (Utt et al., 2013).

the eventhood distributions for the verb classes in both datasets (box plots and density plots in Figure 7.2), which show that the more homogeneous three-class distinction in Katsika et al. (2012) clearly identifies three different argument-selection behaviors (aspectual, psychological, non-metonymic), whereas the two classes in Traxler et al. (2002) overlap substantially.

The model yielded 72% accuracy in the pairwise comparisons on the Traxler et al. (2002) dataset (23/32 metonymic verbs received higher eventhood scores than the non-metonymic verbs in the same sentence pair), and errors occurred more often for metonymic psychological verbs than for metonymic aspectual verbs, as some of them (*recall, report*) preferred events to a higher degree than some metonymic verbs (*enjoy,*

metonymic		non-metonymic		prediction
verb	eventhood	verb	eventhood	correct?
begin	0.91	praise	0.55	y
complete	0.79	recall	0.67	y
start	0.78	see	0.51	y
endure	0.73	report	0.78	n
end	0.72	outline	0.64	y
finish	0.66	prepare	0.41	y
enjoy	0.57	watch	0.60	n
enjoy	0.57	curse	0.31	y
prefer	0.54	praise	0.55	n

Table 7.4.: Eventhood values for some verb pairs from Traxler et al. (2002) and model prediction (Utt et al., 2013).

prefer, examples in Table 7.4).

We showed that a corpus-based measure of eventhood can not only distinguish metonymic from non-metonymic verbs, but can also account for a more fine-grained distinction between aspectual metonymic verbs, non-aspectual metonymic verbs, and non-metonymic verbs. Our results support those in Katsika et al. (2012), as well as their observation that the set of metonymic verbs typically used in experimental studies is too diverse. With Katsika et al. (2012), we argue that metonymic verbs should not include verbs which are less event-selecting than aspectual verbs: the differences emerging between aspectual and non-aspectual metonymic verbs both in the psycholinguistic study in Katsika et al. (2012) and in our corpus-based study provide empirical evidence against the use of non-aspectual metonymic verbs in studies of metonymy, or at the very least call for careful consideration of the verb classes employed.

Our results raise the question of the relationship between metonymy and a verb's event-selecting behavior. Verbs at the extreme end of the spectrum (e.g. *undergo, protest, conduct, spearhead, facilitate, undertake*, cf. Figure 7.1) have the strongest event-selecting tendency and simply do not occur with entity-denoting objects. As a result, they disprefer metonymic constructions, which arguably require an entity-denoting object. We may wonder what happens if these verbs are combined with

entity-denoting objects: the following examples (from American discussion forums on the web) show that when occasionally combined with entity-denoting objects, they seem to imply some sort of covert event and to show some occasional (productive) metonymic behavior:

(7.3) a. There's a huge connection between Prematurity and GBS morbidities and mortalities and I too would be more then willing to **undergo the antibiotics** if such a risk factor was involved.
→ taking the antibiotics

b. *[The Adventures of Tom Sawyer]* is called the first real work of the American Literature movement, which in general spawned **the Hemingways and Faulkners I would later undertake**.
→ reading the Hemingways and Faulkners

c. Taking an IPD approach, we collaborated with Zeemac using 3D modeling known as "real time design" to **facilitate the floor plan**.
→ designing the floor plan

Thus, even though eventhood serves as a good indicator of "metonymicity", it does not seem to be the case that metonymic verbs are at the extreme high end of the eventhood spectrum. Generally, we expect metonymic verbs to be placed at the high end of the eventhood spectrum, but not at the extreme. If, as I have observed in Chapter 1, metonymic combinations are not anomalous, then metonymic verbs should also allow for entity-denoting objects, albeit less frequently. For instance, *begin*, arguably a "true" metonymic verb (metonymic aspectual, $\epsilon = 0.91$), does indeed occur with entity-denoting objects in the corpus (that is, in logical metonymies). High eventhood can then be considered a necessary but not sufficient indicator of metonymic behavior, as it seems to have a strong (but not perfect) correlation with metonymicity.

As a fundamentally graded measure, eventhood does not predict a clear-cut binary distinction between metonymic and non-metonymic verbs, rather it shows a spectrum of verb classes: at the high-eventhood end of the spectrum, **strongly event-selecting verbs** (e.g. *undergo, facilitate*), which disprefer entity-denoting objects but may still be combined with them in a creative and productive way (giving rise to metonymic interpretation, as in 7.3.a); **metonymic aspectual verbs** (e.g. *begin, finish*), which

have a strong preference for event-denoting objects but are not uncommonly (albeit less frequently) combined with entity-denoting objects; **psychological verbs** (e.g. *enjoy, prefer*), showing a strong bias for event-denoting objects but yielding different behavioral patterns compared with aspectual verbs (Katsika et al., 2012); a wide range of non-metonymic verbs which are **neutral** with regards to the type of the object; and finally, at the other end of the spectrum, non-metonymic verbs which are **strongly biased against event-denoting arguments** (e.g. *marry* or *behead*).

Such a graded continuum, based on eventhood, is rather different than the binary distinction between event-selecting verbs and entity-selecting verbs suggested by the Lexical Hypothesis. Also, it confirms the picture emerging from the corpus analyses and the crowdsourcing study reported in Chapter 4: our corpus analysis for German (as well as previous work for English) showed that metonymic verbs are indeed attested in combination with entity-denoting objects, and the crowdsourcing study showed that verb + object constructions could be placed on a continuum depicting their tendency to elicit covert events: on the one end of the spectrum aspectual verbs, when combined with entity-denoting objects, elicited covert event interpretations in the majority of cases (e.g. *begin the newspaper*: 89% covert event interpretations), on the other end entity-selecting verbs (e.g. *approve the automobile*) did not elicit any covert event interpretation, and in between non-aspectual "metonymic" verbs had a mixed behavior (e.g. *enjoy the automobile*: 50% covert event interpretations).

I thus suggest that the type clash hypothesis has to be reformulated in terms of a mismatch between preference (expectation) for an object and the encountered object. This hypothesis will be further explored by the next computational model, a model of metonymy trigger based on thematic fit.

7.3. Type Clash or Thematic Fit?

The Pragmatic Hypothesis claims that the underspecification of a logical metonymy (e.g. *begin the book* → *reading, writing, translating...*) triggers post-lexical presuppositions about the covert event, which are not different for aspectual verbs (e.g. *begin*) and for other verbs such as *enjoy* or *regret*. On the other hand, the psycholinguistic study in Katsika et al. (2012) and the computational study reported in this Chapter (Section 7.2) provided convincing evidence that non aspectual "metonymic" verbs

do indeed differ from aspectual verbs both in their thematic fit with event-denoting objects and in the processes they trigger.

In particular, Katsika et al. (2012) argue that, while non-aspectual metonymic verbs trigger inferential processes, at least for aspectual verbs the prediction from the Lexical Hypothesis should hold that logical metonymies are triggered by a **type clash** caused by the type restrictions of the metonymic verb when this is combined with an entity-denoting object. Experimental studies (e.g. McElree et al., 2001; Traxler et al., 2002) have supported this hypothesis, reporting higher processing costs for metonymic conditions compared to non-metonymic ones, ascribed to the type clash which triggers the coercion mechanism (compatibly with the Lexical Hypothesis).

However, the computational model of eventhood in Section 7.2 represented the event-selecting behavior of metonymic verbs as a graded continuum. This is in principle compatible with the Words-as-cues Hypothesis, which proposes an alternative account of the trigger problem in terms of **thematic fit** between an event-selecting verb and an entity-denoting object: entity-denoting objects simply have a low fit as objects of event-selecting verbs.

The idea of conceptualizing type clashes in terms of thematic fit is not dissimilar from the idea of **selectional preferences** (as opposed to **selectional restrictions**, Hanks, 2007). The theory of selectional restrictions (Katz and Fodor, 1963; Chomsky, 1965) claimed that verbs impose strict restrictions on what fillers they can take for their argument positions, and these binary selectional restrictions distinguish well-formed combinations from nonsensical ones (for example, the verb *weigh* selects for a subject of the type PHYSICAL OBJECT, thus making *The pain weighs three pounds* anomalous, Katz and Fodor, 1963). Wilks (1975, 1978) instead adopted a graded notion of selectional *preferences* (for example, the preferred subject of *drink* is ANIMATE, but we are nevertheless able to understand *My car drinks gasoline*, Wilks, 1978). Selectional preferences allow us to model the expectations of a lemma for highly associated (typical) fillers in its argument slots (with a high thematic fit with the predicate), expectations which have been known to facilitate processing (for example, *crook* is a more fitting object for *arrest* than *cop*, and will be primed or processed faster, McRae et al., 1998; Ferretti et al., 2001; Matsuki et al., 2011).

Similarly, I maintain the observation from the Lexical Hypothesis that metonymic verbs disprefer entity-denoting objects (*contra* the Pragmatic Hypothesis, which does not acknowledge such distinction), but I account for this observation not with a

type clash but via a graded model of selectional preferences and thematic fit. The model proposed by Generative Lexicon (Pustejovsky, 1991, 1995), coherently with the Lexical Hypothesis, is a strongly typed semantic system which does not account for graded preferences. The hybrid model in Section 7.2 incorporates a graded notion of thematic fit, built on typed expectations for typical objects (given a verb and its most associated objects, we have labeled the objects with their type using WordNet). I will now introduce a model of the trigger problem which does not encode any information about type but is fully thematic fit driven (as it is based on DM + ECU).

A thematic-fit explanation of the trigger problem has the advantage of being more economical, as it does not need to postulate a separate type-clash mechanism but simply relies on the same mechanism (thematic fit) which was singled out as responsible for the recovery of the covert event (Chapter 5) and which was employed in the computational modeling of covert event retrieval (Chapter 6). The model (compared to the model in Section 7.2) distinguishes not between metonymic and non-metonymic verbs but between metonymic and non-metonymic subject-verb-object constructions and was evaluated on datasets from psycholinguistic experiments on logical metonymy.

7.4. A Thematic Fit Model of Metonymy Trigger

7.4.1. Measuring Thematic Fit

We used three DM + ECU models (Baroni and Lenci, 2010; Lenci, 2011), based on the English TypeDM, which I have described in Chapter 6 and in Section 7.2:

- The **verb-only model** computes thematic fit values for a verb-object pair: first, it selects the TypeDM weighted tuples $\langle\langle v\ obj^{-1}\ n_i\rangle, \sigma\rangle$, that is the filler nouns n linked to v by an inverse *object* link and their weights σ; then it selects the top 20 most informative (highest-scoring) object fillers and (following a method similar to Erk, 2010 and Lenci, 2011, see Chapter 6) it computes their prototype vector (centroid). The thematic fit value for an object o given a verb v is then computed as the similarity (cosine) between the context vector of o and the prototype object vector for v.

- The **sum model** is simply the ECU model described by Lenci (2011) and introduced in Chapter 6, which computes thematic fit values for a subject-verb-object

combination: first, it selects the object fillers for the verb v and those for the subject s in DM, then it composes the two sets using sum as the composition function and then it computes a prototype vector as the centroid of the (updated) top 20 object fillers. The thematic fit value for an object o given a verb v and a subject s is then computed as the similarity (cosine) between the context vector of o and the prototype object vector.

- The **product model** also computes thematic fit values for a subject-verb-object combination, the only difference being the composition function (product).

7.4.2. Evaluation on Psycholinguistic Datasets

None of the above mentioned models encodes any information about type. In order to test whether a computational model based only on thematic fit can distinguish metonymic sentences from non-metonymic ones, doing without a notion of type clash, we computed thematic fit scores from the models in 7.4.1 for two datasets used in two self-paced reading experiments (McElree et al., 2001; Traxler et al., 2002)[6] and one eye-tracking experiment (Frisson and McElree, 2008), and we evaluated the models on their ability to replicate all significant experimental findings from the psycholinguistic studies (Zarcone et al., 2012c, 2013). The two datasets contained metonymic and non-metonymic sentences, which were contrasted in the experimental design:

McElree et al. (2001) dataset: this dataset is composed of 31 triplets of *metonymic*, *high-typicality* and *low-typicality* sentences, differing with regard to the verb[7] (e.g. *the writer finished / wrote / read the novel*). The self-paced reading study reports a main effect of verb type on reading times one region after the object; pairwise comparisons yielded (a) longer reading times for the metonymic condition and (b) no significant differences between the high- and low-typicality condition. Longer reading times for the metonymic condition are ascribed to the coercion operation.

[6] The same datasets with minimal changes were also used in eye-tracking studies, and the results mirrored those of the self-paced reading study. McElree et al. (2001); Traxler et al. (2002) and Frisson and McElree (2008) report in detail how the datasets were built.

[7] *Type-shifting*, *preferred* and *non-preferred*, according to the terminology of the study. We excluded two triplets from the original dataset due to problems of coverage.

Traxler et al. (2002) dataset: this dataset is composed of 30 sentence quadruplets crossing two factors (a) *verb type*, metonymic vs. non-metonymic[8] and (b) *object type*, event- vs. entity-denoting (EV vs. EN; e.g. *The boy [started / saw] [the fight / puzzle]*). The self-paced reading study reports a main effect of object type on reading times one region after the object and a significant verb∗object interaction, with higher processing costs for the metonymic condition (for sentences with metonymic verbs + entity-denoting objects compared to sentences with non-metonymic verbs + entity-denoting objects). The higher processing costs for the metonymic condition are ascribed to the coercion operation.

Frisson and McElree (2008) dataset: this dataset is composed of 25 sentence quadruplets crossing two factors (a) *verb type*, metonymic vs. non-metonymic[9] and (b) *preference*, strongly- vs. weakly-preferred. The strongly-preferred sentences differed from the weakly-preferred ones because the former had one strongly preferred covert event interpretation (e.g. *The teenager began the novel* → *reading*), whereas the latter had multiple plausible covert event interpretations (e.g. *The waitress started the coffee* → *drinking, serving, ...*). A main effect of verb type on eye movements is reported, with higher processing costs for the metonymic conditions compared to the non-metonymic conditions both for strongly-preferred covert events and for weakly-preferred covert events (first-pass time at the object and the object+1 region, second-pass time at the object region and on total reading time at the object region and the object+1 region). An effect of preference was yielded for second-pass time at the object region and for total reading time at the object region and the object+1 region, without interaction. The metonymic condition appears to be costly both when there is one strongly preferred covert event and when more interpretations are possible: based on these results, the authors argue that the cost of coercion is unrelated to the retrieval of one or many plausible interpretations. For better comparison with the previous studies, I will refer to the total reading times in the spillover region (object+1 region).

[8] *Event verb* and *neutral verb*, according to the terminology of the study. We excluded one quadruplet from the original dataset due to problems of coverage.

[9] *Coerced* and *control*, according to the terminology of the study (e.g. *The teenager [began / read] the novel; The waitress [started / served] the coffee*). We excluded seven quadruplets from the original dataset due to problems of coverage.

triplets from McElree et al. (2001)				
		high-typicality	low-typicality	metonymic
reading times at the obj.+1 position		360	361	385
1 – thematic fit	verb-only	0.484	0.571	0.763
	sum	0.483	0.569	0.754
	product	0.584	0.625	0.714

quadruplets from Traxler et al. (2002)					
		metonymic verb		non-metonymic verb	
		EN obj.	EV obj.	EN obj.	EV obj.
reading times at the obj.+1 position		512	427	467	455
1 – thematic fit	verb-only	0.770	0.664	0.717	0.718
	sum	0.767	0.661	0.714	0.712
	product	0.727	0.658	0.724	0.681

quadruplets from Frisson and McElree (2008)					
		strongly preferred		weakly preferred	
		meton. verb	non-meton. verb	meton. verb	non-meton. verb
total reading times at the obj.+1 position		391	353	404	374
1 – thematic fit	verb-only	0.808	0.461	0.778	0.644
	sum	0.809	0.462	0.776	0.643
	product	0.776	0.708	0.744	0.685

Table 7.5.: Reading time data (self-paced reading) from McElree et al. (2001) and Traxler et al. (2002) and total reading time data (eye tracking) from Frisson and McElree (2008); thematic fit data from the computational models.

We used the three models to compute verb-object or subject-verb-object thematic fit for the combinations in the dataset. We expected processing load to be inversely related to thematic fit, thus we assumed that processing cost (reading time) corresponds to 1 – thematic fit. We manipulated thematic fit as a dependent variable, and we searched for main effects of factors (object type, verb type) on thematic fit with linear regression analyses and as well as Wilcoxon rank sum task (to test the significance of pairwise thematic fit differences between conditions). We then verified that the same main effects and significant pairwise differences were yielded by the psycholinguistic

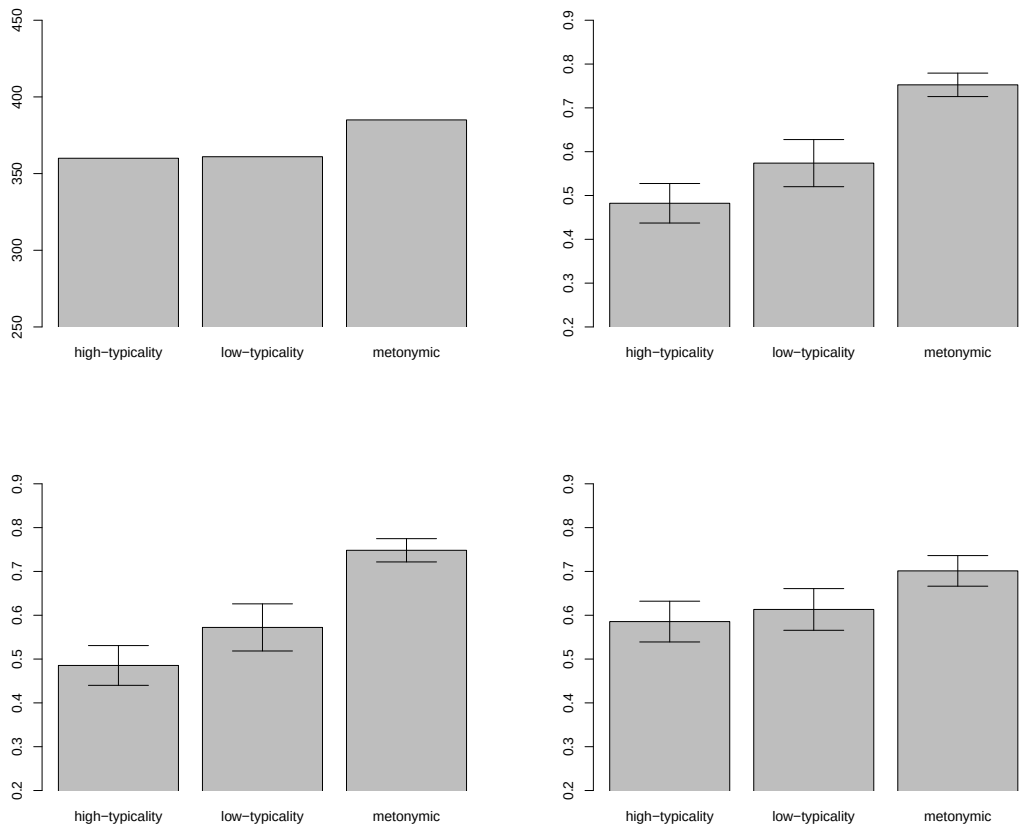

Figure 7.3.: Comparing reading times (top left, in ms) in McElree et al. (2001) with scores (1 – thematic fit) from the verb-only model (top right), from the sum model (bottom left) and from the product model (bottom right).

models and by the computational model. We employed a pairwise comparison strategy, as only the mean reading times per condition are reported (also because of the size of the experiments, which does not allow for a point-wise correlation — see the discussion on evaluation methods in Chapter 2).

7.4.3. Results and Discussion

McElree et al. (2001) Dataset: all models successfully yielded lower thematic fit scores (higher 1 – thematic fit scores) for the metonymic sentences (see Table 7.5) and mirrored the self-paced reading study (see Figure 7.3), yielding a main effect of

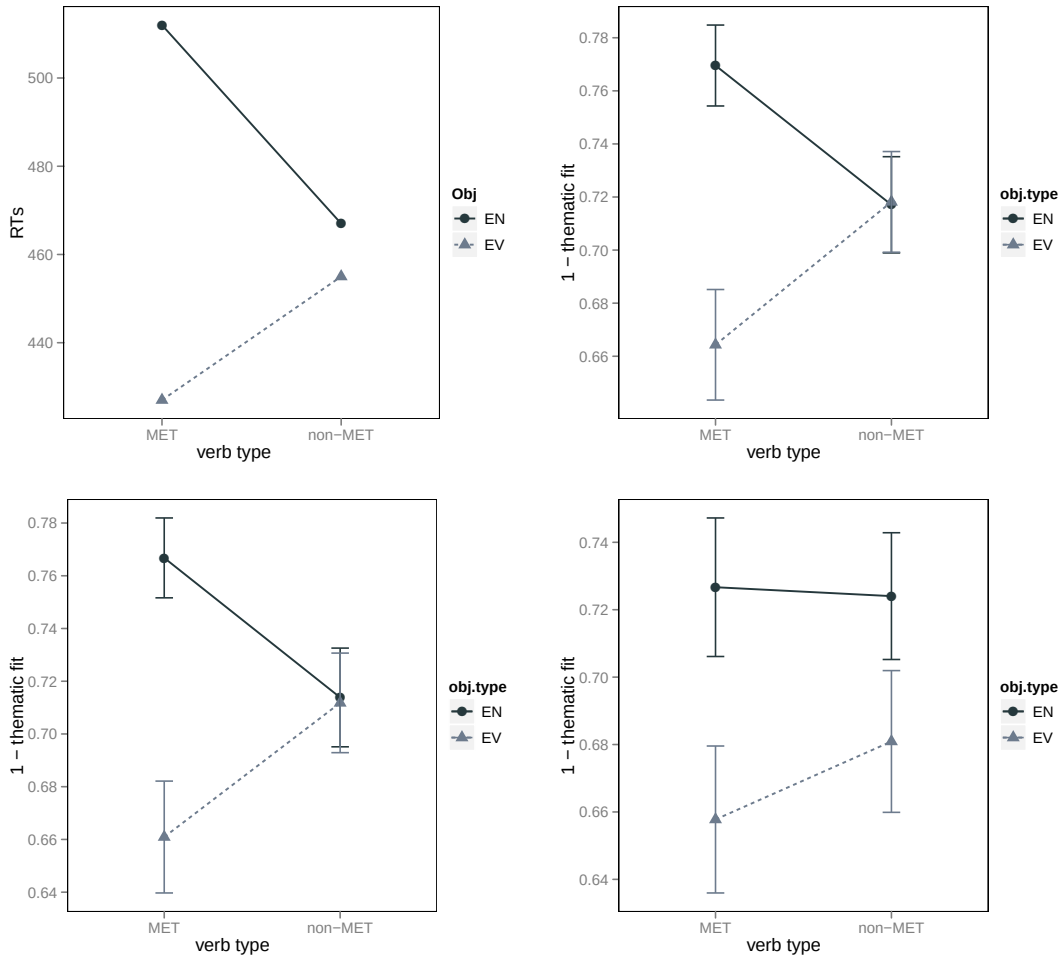

Figure 7.4.: Comparing reading times (top left, in ms) in Traxler et al. (2002) with scores (1−thematic fit) from the verb-only model (top right), from the sum model (bottom left) and from the product model (bottom right).

verb type (verb-only model: $F = 20.247, p < 0.001$; sum model: $F = 19.738, p < 0.001$; product model: $F = 4.5847, p < 0.05$), significant differences between the metonymic condition and both high-typicality (verb-only model: $W = 877, p < 0.001$; sum model: $W = 870, p < 0.001$; product model: $W = 689, p < 0.01$) and low-typicality conditions (verb-only model: $W = 740, p < 0.001$; sum model: $W = 740, p < 0.001$; product model: $W = 617, p = 0.055$), and no significant difference between the high- and low-typicality conditions (verb-only model: $W = 595, p > 0.05$; sum model: $W = 591, p > 0.05$; product model: $W = 552, p > 0.05$).

Traxler et al. (2002) Dataset: all models successfully yielded lower thematic fit scores (higher 1 − thematic fit scores) for the metonymic condition (metonymic verb + entity-denoting object, see Table 7.5) and mirrored the main effect of object type (verb-only model: $F = 8.0039, p < 0.01$; sum model: $F = 8.3997, p < 0.01$; product model: $F = 7.4133, p < 0.01$) reported by the self-paced reading study, and all but the product model yielded a significant verb∗object interaction (verb-only model: $F = 8.3455, p < 0.01$; sum model: $F = 7.7712, p < 0.01$; product model: $F = 0.3927, p = 0.53$). Figure 7.4 shows the close correspondence between experimental results from self-paced reading and thematic-fit modeling results. All but the product model also yielded the same pair-wise differences reported by the self-paced reading study: within the sentences with entity-denoting objects, metonymic verbs yield lower thematic fit compared to non-metonymic sentences (verb-only model: $W = 300, p < 0.05$; sum model: $W = 318, p = 0.052$; product model: $W = 300, p = 0.75$).

Frisson and McElree (2008) Dataset: all models successfully yielded lower thematic fit scores (higher 1 − thematic fit scores) for the metonymic conditions (see Table 7.5) and mirrored the main effect of verb type (verb-only model: $F = 56.703, p < 0.001$; sum model: $F = 56.2965, p < 0.001$; product model: $F = 4.5499, p < 0.05$) reported by the eye-tracking study, as well as the significant pairwise comparisons between metonymic and non-metonymic sentences, both for the strongly-preferred condition (verb-only model: $W = 60, p < 0.001$; sum model: $W = 59, p < 0.001$; product model: $W = 212, p = 0.052$) and (for all but the product model) the weakly-preferred condition (verb-only model: $W = 155, p < 0.01$; sum model: $W = 158, p < 0.01$; product model: $W = 255, p = 0.27$). Additionally, all but the product model yielded a significant effect of preference (verb-only model: $F = 5.664, p < 0.05$; sum model: $F = 5.3236, p < 0.05$; product model: $F = 0.8585, p = 0.36$) and a significant verb∗preference interaction (verb-only model: $F = 11.171, p < 0.01$; sum model: $F = 11.1160, p < 0.01$; product model: $F = 0.0244, p = 0.88$, see also Figure 7.5).

Our models (models of thematic fit without any explicit type information) were able to replicate the experimental findings from three psycholinguistic studies (McElree et al., 2001; Traxler et al., 2002; Frisson and McElree, 2008).

On the McElree et al. (2001) dataset, the models yielded a main effect of verb type on thematic fit and assigned the lowest thematic fit scores to the metonymic condition (where the psycholinguistic study had found the longest reading times).

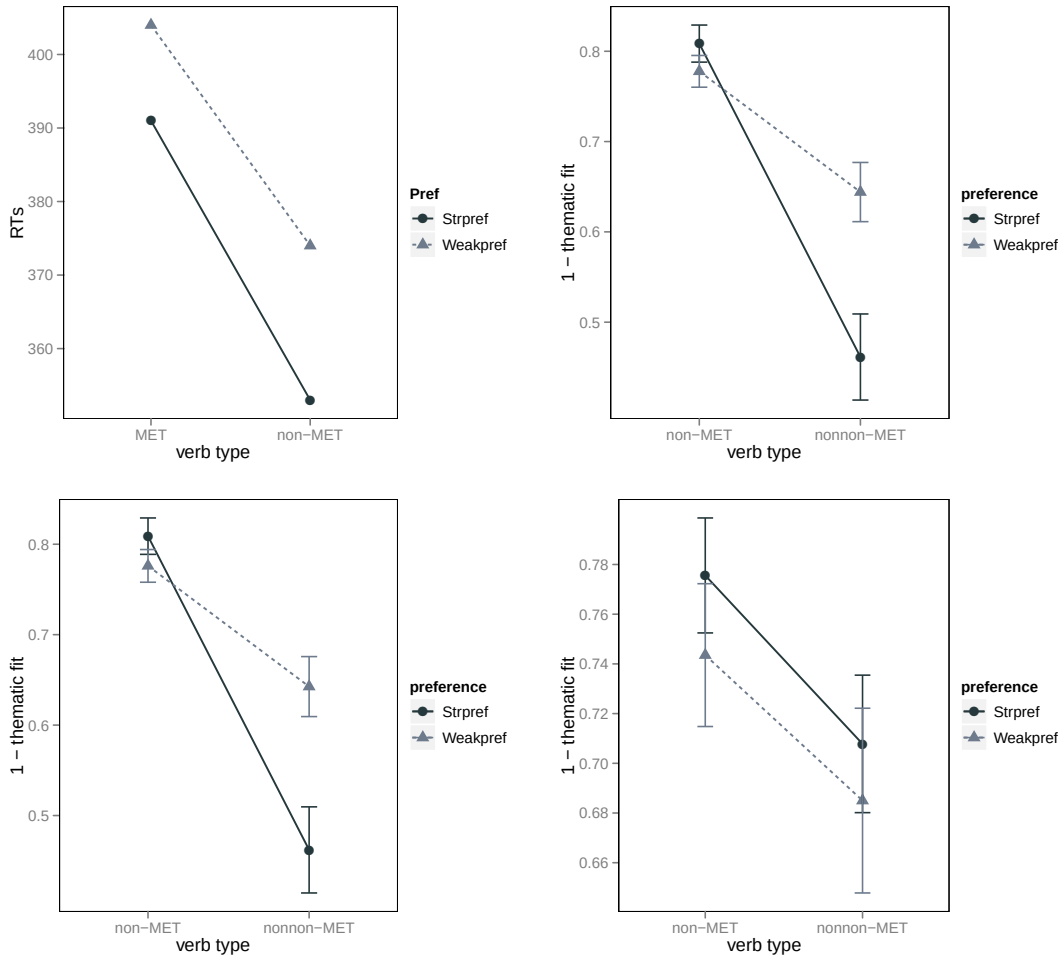

Figure 7.5.: Comparing reading times (top left, total reading time for the object +1 region in ms) in Frisson and McElree (2008) with scores (1 – thematic fit) from the verb-only model (top right), from the sum model (bottom left) and from the product model (bottom right).

The significant and non-significant differences yielded by the models in the pairwise comparisons also matched those in the experimental study: metonymic sentences had lower thematic fit values than both the high- and low-typicality conditions, whereas no difference was found between the high- and low-typicality conditions. Interestingly, this was the case for all the computational models, both for the verb-only model (which does not take the subject into account) and for the sum and product models (which incorporate information about the subject). Since the authors claim that the high- or low-typicality of the non-metonymic conditions (*preferred* and *non-*

preferred in the original terminology) depend on the subject (e.g. *write* is the preferred verb for ⟨*author began book*⟩, whereas *read* is non-preferred), this calls into question whether the balancing of the materials truly reflected a predominance of the preferred condition over the non-preferred condition. For example, in 7.4 one could argue that it is as untypical for *surfers* to *wear* as it is to *rent tuxedos*:

(7.4) The surfer **endured** / wore / rented **the tuxedo** but felt very uncomfortable. (McElree et al., 2001)

As to the Traxler et al. (2002) dataset, the models yielded 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. Again, the significant differences detected by the model matched those in the experimental studies: for the entity-denoting objects, metonymic verbs yielded lower thematic fit than non-metonymic verbs.

Also the effect of coercion in the Frisson and McElree (2008) dataset was successfully replicated by the thematic fit models. The authors argue that the lack of a strong effect of preference on eye movements, as well as the lack of interaction with verb type, are convincing evidence for the predominance of the coercion effect (which should be encountered no matter how underspecified the covert event interpretation is) over effects of plausibility / typicality (preference). Nevertheless, if this was the case, I would expect to find an interaction between preference and verb type (with longer reading times for metonymic sentences, and different reading times for strongly- and weakly-preferred interpretations, shorter for the former and longer for the latter), but the interaction failed to reach significance in the eye-tracking study (see Figure 7.5). Interestingly, this is the picture that most of the computational models for this dataset provide. The computational model, without resorting to any object type information, offers a picture (possibly even more) pursuant to the claim in Frisson and McElree (2008), that logical metonymy is not dependent on the range and ranking of covert event interpretation available.

7.5. General Discussion

The results from the computational models in this chapter show that a computational model based on thematic fit only can distinguish metonymic contexts from non-

metonymic contexts, doing without a notion of type clash. Furthermore, all models mirrored the results from psycholinguistic studies which were designed to investigate (and which ultimately yielded) extra processing costs for logical metonymies. The only model which at times did not reach significance was the product model (recall from Chapter 6 that sum was more robust, whereas product introduced noise to the expectation update process). The results from the verb-only model and from the sum model were rather similar.

As these psycholinguistic findings were interpreted as evidence for type-clash induced coercion, the results from the model support an alternative interpretation of these results (coherent with the Words-as-cues Hypothesis), suggesting that thematic fit may account for both the retrieval of the covert event and the trigger of the metonymic interpretation, and that thematic fit on its own may determine the cost of the coercion operation, without resorting to type clash. A similar proposal is also presented by Roberts and Harabagiu (2011), who address the problem of coercion detection (to determine if a verb-object pair is a coercion or not) with a scoring method based on a Latent Dirichlet Allocation selectional preference model, but they do not make any claim about the cognitive plausibility of their model.

An alternative explanation may be that there is indeed a cost for the coercion operation, but the experimental materials taken into consideration in this chapter are simply not balanced enough in terms of thematic fit. For example, some subjects in the sentences used (Traxler et al., 2002) seem to have a bias for the event-denoting objects used (*cheating husband → affair, soprano → concert, pastor → funeral*) compared to the entity-denoting objects:

(7.5) a. The cheating husband **began** / recalled the affair / **the letter**.

b. The soprano **began** / praised the concert / **the letter**.

c. The pastor **finished** / prepared the funeral / **the sandwich**.

(Traxler et al., 2002)

This may have introduced a lexical-semantic bias that influenced the results, overshadowing the costs of the coercion.

Structured distributional models can be an interesting tool for a critical reanalysis of psycholinguistic datasets and of the interpretation of the results obtained with these datasets, showing that thematic fit is indeed a key factor to consider. It is worth

mentioning, however, that the effects reported are also yielded by the verb-only model, suggesting that they are indeed determined by the verb-object thematic fit rather than by the influence of the subject. On the other hand, this does not necessarily rule out a possible effect of type clash and type shift: the distributional fact, even without encoding any information about type, may still correlate with type information, which may in turn be reflected in corpus distributions. Before discarding the notion of type in logical metonymy interpretation and in order to evaluate if thematic fit can truly account for the trigger of the logical metonymy, in the next chapter I will discuss the results of a second self-paced reading experiment that manipulates type and thematic fit.

8. The Trigger of the Logical Metonymy: Psycholinguistic Evidence

In the previous chapter I presented the results from two computational models as evidence in favor of the **Words-as-cues Hypothesis**, and in particular in support of the proposal that the trigger of the logical metonymy can be ascribed to the low thematic fit between the event-selecting metonymic verb and its non-event-denoting argument. Previous work in psycholinguistics (McElree et al., 2001; Traxler et al., 2002; Frisson and McElree, 2008) reported results in favor of the **Lexical Hypothesis**, showing higher processing costs for logical metonymies compared to non-metonymic constructions and interpreting them as evidence for a type-clash and coercion operation. Nevertheless, a computational model of thematic fit (Chapter 7) successfully replicated the results from the psycholinguistic studies, doing away with a notion of type and suggesting that these results may be reinterpreted as an effect of thematic fit only.

However, an alternative explanation is still possible: logical metonymies may indeed be more costly than non-metonymic constructions, and it may simply be the case that the experimental materials taken into consideration in the psycholinguistic studies were not balanced enough in terms of thematic fit. I will now present a self-paced reading experiment manipulating type and thematic fit aimed at evaluating the role played by each of them separately and at investigating whether thematic fit only could be responsible for the costs of the logical metonymy.

8.1. Previous Work

One of the most relevant experimental studies on the trigger of logical metonymy is the work of Traxler et al. (2002). Recall that the design of their study crossed two factors (verb type and object type), contrasting metonymic constructions (metonymic verb + entity-denoting object) with constructions containing non-metonymic verbs and event-denoting objects (see 8.1 for some examples, metonymic combinations in boldface):

(8.1) a. The boy **started** / saw the fight / **the puzzle**.

b. The victim **endured** / reported the robbery / **the driver**.

c. The banker **expected** / remembered the audit / **the check**.

d. The cheating husband **began** / recalled the affair / **the letter**.

e. The soprano **began** / praised the concert / **the letter**.

f. The pastor **finished** / prepared the funeral / **the sandwich**.

(Traxler et al., 2002)

Traxler et al. (2002) reported higher processing costs for the metonymic condition, both in an eye-tracking study (Experiment 2) and in a self-paced reading study (Experiment 3). More specifically, they reported an effect of object type on second-pass time in the object region, and a significant verb∗object interaction on second-pass time and total time in the object region, again with the longest reading times for the metonymic condition. The self-paced reading study yielded an effect of object type on reading times one region after the object (object + 1) and a significant verb∗object interaction, with higher processing costs for the metonymic condition (metonymic verbs + entity-denoting objects) compared to conditions with non-metonymic verbs and with entity-denoting objects. They also reported an effect of object type on first-pass time and regression-path time at the object region, but in a different direction than the other effects (event-denoting objects take longer than entity-denoting objects), which was ascribed to "some differential difficulty processing the NPs", as the event-denoting nouns were reportedly on average 0.2 characters longer than entity-denoting nouns and numerically less frequent (by 15 per million occurrences).

I have already highlighted some potential issues in the design and materials in Traxler et al. (2002), such as their selection of metonymic verbs and the problem of

thematic fit balancing. I will now refer back to them and mention one more issue, regarding the differences between entity- and event-denoting verbs.

Aspectual vs. Non-aspectual Metonymic Verbs

As discussed in Chapter 7, Experiments 2 and 3 in Traxler et al. (2002) employ a diverse array of metonymic verbs, including aspectual verbs (*begin*, *finish*) but also psychological verbs (*endure*, *enjoy*) and other verbs whose criteria of inclusion are more obscure (*attempt*, *resist*). For example, in 8.1.b-c it is not clear why *endure* and *expect* should be more "event-selecting" than *report* and *remember* and thus why the former two should be employed in the "metonymic" conditions whereas the latter two should not. Converging evidence from an eye-tracking study (Katsika et al., 2012) as well as from the computational model of eventhood in Section 7.2 argued for a separation of aspectual metonymic verbs and non-aspectual metonymic verbs in experimental studies, or at least for careful consideration of the potential differences between these two classes when preparing experimental items.

Thematic Fit Balancing

Traxler et al. (2002) did run a plausibility norming study, to exclude the possibility that one of their conditions might be significantly more or significantly less plausible from the others, and reported that no difference reached significance. Nevertheless, some of their items seem to have a bias towards their event-denoting objects (e.g. 8.1.d: *cheating husband* → *affair*, 8.1.e: *soprano* → *concert*, 8.1.f *pastor* → *funeral*) compared to their entity-denoting objects (*cheating husband* → *letter, soprano* → *letter, pastor* → *sandwich*). The thematic fit model in Section 7.4 introduced the suspicion that the higher processing costs reported for sentences with entity-denoting objects and interpreted as costs for the logical metonymy might rather be accounted for in terms of thematic fit.

As observed in Section 3.4.2, plausibility does not correspond to typicality, and it could then be the case that, while the participants in the plausibility norming study in Traxler et al. (2002) have judged the *plausibility* of the test sentences (it is perfectly plausible for *cheating husbands* and *sopranos* to *begin letters*, and for *pastors* to *finish sandwiches*), a plausibility rating study may not reflect the fact that *begin letters* is not a *typical* activity performed by *cheating husbands* (or at least that it is not as typical

for them to *begin letters* as it is to *begin affairs*). Thus, the materials in Traxler et al. (2002), while balanced for plausibility, need to be balanced for *typicality* as well, in order to exclude that differences in typicality (quantified as thematic fit) conceal the cost of the logical metonymy and in order to verify if these can be ascribed to typicality / thematic fit only, thus doing away with a notion of type clash.

Entity- vs. Event-denoting Nouns

The entity-denoting objects in Traxler et al. (2002) were matched for frequency and length (both known to affect processing costs), in order to exclude that the differences in processing costs found at the object and object + 1 region can be ascribed to anything other than the type clash responsible for the metonymy. Nevertheless, using different words in the target region introduces inter-item variance, leading to a problem known as the Language-as-Fixed-Effect Fallacy (Coleman, 1964; Clark, 1973): even matching them for frequency and length, some items will be easier (and read faster) than others, thus item variance should be considered a random effect (Raaijmakers et al., 1999).

Furthermore, there are reasons to doubt that entity-denoting and event-denoting nouns are processed in a similar way: it has been argued that event nouns, unlike entity-denoting nouns, have an argument structure, just like the verbs they derive from (Zubizarreta, 1987; Grimshaw, 1990; Alexiadou, 2001), and neurolinguistic studies have reported differences between entity- and event-denoting nouns in the performance of agrammatic patients during picture naming (Collina et al., 2001; Tabossi et al., 2010) and have shown different hemodynamic responses in brain imaging for entity- and event-denoting nouns, as well as analogies between verbs and event nouns (Garbin et al., 2012; Bedny et al., 2013). The longer reading times reported by Traxler et al. (2002) for event-denoting nouns may therefore be determined by differences in the argument structure of those nouns.

Also, the experiments in Traxler et al. (2002) were carried out in English, a language with zero-derivation verb to noun word-formation processes (e.g. *to call* → *a call*, *to jump* → *a jump*, Plag, 2003): many event-nouns were zero-derivation deverbals, looking identical in their (base) form to the verbs they are derived from, and were morphologically ambiguous between two categories (noun and verb). Psycholinguistic studies (Farmer et al., 2006; Hohenstein and Kliegl, 2013) suggest that readers make use

of information on the visual form of words when reading (for example, if a word has a form that is either typical or atypical for a noun or for a verb), which reduces processing costs for words whose form matches the typical form for its morphological category according to a measure of phonological typicality[1]. Thus, using zero-derivation event nouns might have introduced a confounding factor.

In the light of such considerations, a design aimed at measuring differences at the target object (or in the spillover region) between the entity-denoting and event-denoting does not appear very robust, even when the items are matched for frequency, length and plausibility.

8.2. Experiment 3

Traxler et al. (2002) interpreted their results as evidence for type-clash induced logical metonymy (supporting the Lexical Hypothesis). The Words-as-cues Hypothesis suggests instead that the thematic fit between the event-selecting metonymic verb and its non-event-denoting argument should be the main (and perhaps only) factor to trigger logical metonymy interpretation. Experiment 3 (Zarcone and Padó, 2013) introduces some key changes to the design in Traxler et al. (2002), which are aimed at overcoming the potential issues which I raised about the design and materials in Traxler et al. (2002), as well as at evaluating the Words-as-cues Hypothesis on the trigger problem, disentangling the role played by type and thematic fit as triggers of the metonymic interpretation.

The first change was that the experiment was conducted in German. This had two advantages: (a) the German deverbals used in the materials are not formed by zero derivation, and (b) German nouns are capitalized (capitalization is arguably exploited by German readers to facilitate processing, Hohenstein and Kliegl, 2013); thus, the event-denoting nouns do not look identical to the verbs they are derived from, but are clearly marked in different ways, overcoming potential problems of English zero-derivation deverbals.

Secondly, we exploited the participle-final word order in German, as in the following example:

[1] For example, *amuse* is a verb-like verb, *ignore* is a noun-like verb, *insect* is a verb-like noun, *marble* is a noun-like noun (Farmer et al., 2006).

(8.2) a. Das Geburtstagskind hat mit den Geschenken **angefangen**.
The birthday boy has with the presents **begun**.

b. Das Geburtstagskind hat mit der Feier **angefangen**.
The birthday boy has with the party **begun**.

This word order allows us to measure reading times at the same target word (in our case, the metonymic verb ***angefangen***, ***begun***) in different contexts (for example after an entity-denoting object, *Geschenken*, *presents*, as well as after an event-denoting object, *Feier*, *party*). Note that the type clash hypothesis can still be evaluated with a different word order: if a type clash occurs between an event-selecting verb and an entity-denoting object, this should happen as soon as the two are combined, regardless of which of the two is presented first. We can thus verify if the reading times at the target region are influenced by the choice of object, namely if the entity-denoting object causes a type clash with the metonymic verb and if such a clash influences reading times. This allows for a more robust design than the one in Traxler et al. (2002), as in all conditions we can measure the same word (i.e. the metonymic verb).

Lastly, we made a crucial distinction in the design between two factors: **type** and **thematic fit**. This is done by introducing thematic fit as a second factor (the other factor is type: entity vs. event) and by using both high-thematic-fit objects (one entity-denoting, one event-denoting, obtained with an elicitation study, see example 8.2) and low-thematic fit objects (one entity-denoting, one event-denoting):

(8.3) a. Das Geburtstagskind hat mit der Suppe **angefangen**.
The birthday boy has with the soup **begun**.

b. Das Geburtstagskind hat mit der Schicht **angefangen**.
The birthday boy has with the shift **begun**.

This design allows us to better control for the thematic fit of the objects and to disentangle its contribution to reading times from the role played by type, distinguishing the predictions of the Lexical Hypothesis and of the Words-as-cues Hypothesis: a type clash account, such as the Lexical Hypothesis, would predict an effect of object type (longer reading times for sentences with entity-denoting objects compared to those with event-denoting objects), whereas a pure thematic fit based account would predict an effect of thematic fit, regardless of object type.

8.2.1. Method

Materials The materials for Experiment 3 were prepared using an elicitation study, similar to those used for the experiments in Chapter 5. This is crucial to obtain objects for the high-typicality conditions which are actually typical and not just very plausible for the sentences used.

Norming Study 5 Thematic-based patient generation norms were collected for 25 sentence templates (e.g. *Der Student hat mit dem / der _____ angefangen.* The student has begun with the _____.) on Amazon Mechanical Turk (AMT), asking participants to "provide words that could plausibly fill the blank". All sentence templates contained an aspectual verb from those used for Experiment 2b. For each sentence, space was provided for 5 responses, and no time limit was imposed. Each sentence was presented to an average of 6 German participants. Participants were very productive, eliciting on average 3 patients per sentence per participant. We chose 4 patients for 21 sentences, from those named early by many participants (using the weighting method from Matsuki et al., 2011), selecting the sentences where at least two entity-denoting and two event-denoting nouns were elicited: e.g. *Der Student hat mit dem / der _____ angefangen* → *Aufsatz* (essay), *Buch* (book), *Studium* (study), *Prüfung* (exam). We thus obtained 84 patient-sentence pairs (4 patients x 21 sentences).

Expert Annotation Study 40 of the 84 patient nouns (2 patients x 20 sentences) were selected after a threefold expert annotation study, where three linguists (native speakers of German) were asked to annotate the nouns as event-denoting (EV), entity-denoting (EN) or EN/EV ambiguous. The 40 patient nouns selected were the ones with the highest agreement scores for both nouns elicited for the sentence: Weighted Krippendorff's α (Krippendorff, 1980) for the selected 40 nouns was 0.56. Weighted α for the selected 40 nouns, incorporating the idea that EN vs. EV is a stronger disagreement than the disagreement between either one of the types vs. the ambiguous EN/EV type[2], was 0.71 (good agreement).

40 high-typicality sentences were constructed from the 40 patient-sentence pairs (one entity-denoting patient and one event-denoting patient per sentence template).

[2]A weight of 1 was assigned to the EN-EV disagreement and a weight of 0.5 to the EN-EN/EV disagreement and to the EV-EN/EV disagreement.

An adverb was inserted between the object and the participle, as a buffer to exclude possible spillover effects from the object to interfere with the verb region, and a continuation was included after the participle.

High-typicality:

(8.4) a. [entity] Das Geburtstagskind hat **mit den Geschenken** sofort **angefangen**, obwohl seine Mutter nicht da war.
The birthday boy has **with the presents** at once **started**, although his mother wasn't there.

b. [event] Das Geburtstagskind hat **mit der Feier** sofort **angefangen**, obwohl seine Mutter nicht da war.
The birthday boy has **with the party** at once **started**, although his mother wasn't there.

Low-typicality:

(8.5) a. [entity] Das Geburtstagskind hat **mit der Suppe** sofort **angefangen**, obwohl seine Mutter nicht da war.
The birthday boy has **with the soup** at once **started**, although his mother wasn't there.

b. [event] Das Geburtstagskind hat **mit der Schicht** sofort **angefangen**, obwohl seine Mutter nicht da war.
The birthday boy has **with the shift** at once **started**, although his mother wasn't there.

40 low-typicality sentences were obtained by crossing patients with another sentence template using the same metonymic verb (in the case of 8.4, the sentence template with *Kellnerin*, *waitress* as subject), which in turn obtained the high-typicality objects from the first template as its low-typicality objects. When crossing the materials to obtain the low-typicality sentences, we ensured that the objects assigned to the subjects in the low-typicality sentences were never elicited for that subject-object pair (for example, *Suppe* [*soup*] and *Schicht* [*shift*] were never elicited for *Das Geburtstagskind hat mit dem / der _____ angefangen*).

As all objects occurred once in the high-typicality group and once in the low-typicality group, both groups were balanced with regard to length and frequency.

Entity- and event-denoting objects did not differ with regard to frequency (average log frequency in the CELEX word frequency list for German (Baayen et al., 1993) was 1.49 for entity-denoting objects and 1.38 for event-denoting objects; Wilcoxon rank sum test: $W = 881, p < 0.438$), but they did differ with regard to length (entity-denoting objects were on average 13 characters long; event-denoting objects 15; $W = 489, p = 0.003$). This is practically unavoidable, because both groups of objects were obtained with an elicitation task (thus we could not match them for length) and because, as mentioned above, event-denoting objects in German are formed by suffixation and not by zero derivation, often resulting in longer words.

Norming Study 6 In order to check that the low-thematic-fit triplets were, although not highly typical, still sensible and plausible and that they did not violate any selectional restriction, we collected plausibility ratings on AMT for our materials on a five-point Likert scale (no time limit was imposed). Participants (on average, 16 German participants per sentence) were presented with the 80 high- and low-thematic-fit sentences (40 + 40), along with 64 sentences with selectional restriction violations (nonsensical fillers: e.g. *Die Gitarre ging ins Kino*, The guitar went to the cinema). The order of presentation was randomized. The ratings yielded high agreement (Krippendorff's α for ordinal data = 0.72); sentences in the high-thematic-fit condition yielded a mean rating of 4.41 (SD = 0.58), sentences in the low-thematic-fit condition yielded a mean rating of 2.92 (SD = 0.56), nonsensical fillers yielded a mean rating of 1.58 (SD = 0.56). The plausibility scores for the low-thematic-fit sentences were significantly higher than those for nonsensical fillers (Wilcoxon rank sum test: $W = 113.5, p < 0.001$) and significantly lower than those for the high-thematic-fit sentences ($W = 1522, p < 0.001$, see the box plot in Figure 8.1). Sentences with entity-denoting objects yielded a mean rating of 3.48 (SD = 0.96), sentences with event-denoting objects yielded a mean rating of 3.84 (SD = 0.89). The plausibility scores for the sentences with entity- and event-denoting objects were not significantly different ($W = 618, p > 0.05$). These results support our claims that (a) the low-thematic-fit sentences do indeed differ in plausibility from the high-thematic-fit ones as well as from the nonsensical fillers (they still make sense), and that (b) the sentences with event- and entity- denoting objects did not differ with regard to plausibility. We can therefore rule out the possibility that effects of thematic fit or effects of type in Experiment 3 may be caused by a semantic anomaly of the low-thematic-fit condition or by the

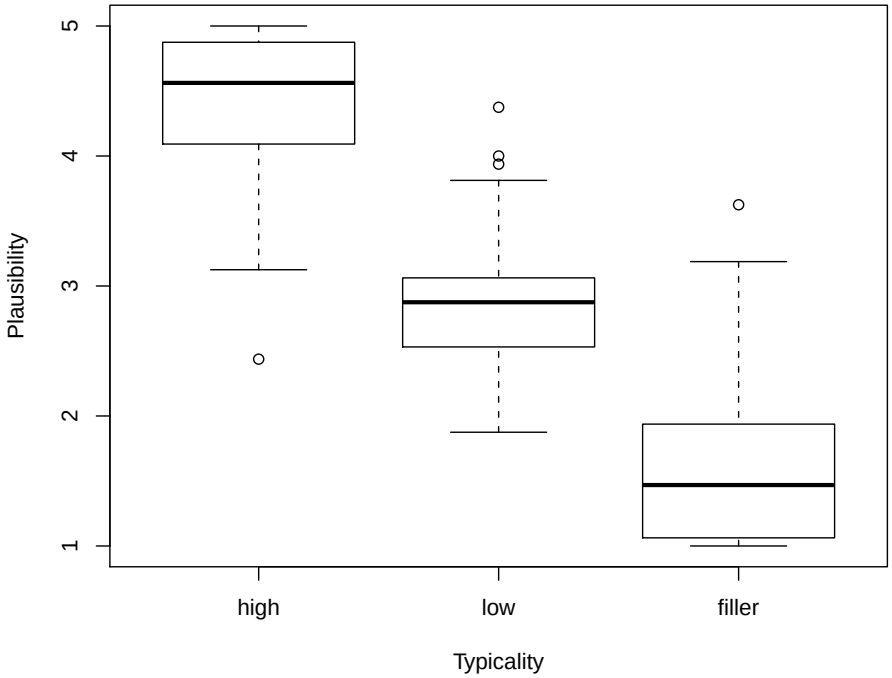

Figure 8.1.: Norming Study 6: Comparing plausibility ratings for high- and low-typicality test sentences and nonsensical fillers in Experiment 3.

lower plausibility of sentences with event- or entity- denoting objects.

Procedure Two lists of 92 sentences each (5 high-typicality/EN, 5 high-typicality/EV, 5 low-typicality/EN, 5 low-typicality/EV, 72 filler sentences) were created to ensure that the same participant would not see the same agent-patient combination twice: for each group of four sentences sharing the same agent, each sentence was put in a different list (to which a quarter of the participants was assigned). Participants were presented with sentences with a one-word-at-a-time moving-window self-paced reading paradigm. Each trial began with strings of dashes on the screen, each dash replacing a non-space character of the sentence. Participants pressed a button to reveal the next word, and revert the previous to dashes. After each sentence, participants were required to answer a yes/no comprehension question. Participants were allowed to take two breaks during the experiment, after the first and second thirds of

Position			patient	adverb	target V	V+1
	Examples		*mit der Feier*	*sofort*	*angefangen,*	*obwohl*
			with the party	at once	started,	although
Latency (ms)	high-fit EN		642	656	819	508
	high-fit EV		655	644	736	473
	low-fit EN		667	693	802	520
	low-fit EV		710	682	806	505
Mixed-Effect Regression	Type:	t	−2.2	−1.26	−2.5	−3.32
		p	0.03	0.21	0.01	0.001
	Thematic fit:	t	2.28	2.19	−0.42	2.83
		p	0.02	0.03	0.68	0.01
	Interaction:	t	–	–	2.04	–
		p	–	–	0.04	–

Table 8.1.: Experiment 3: Reading latencies (in ms) and mixed-effect regressions.

the sentences.

Participants Forty-eight students of Universität Stuttgart (age range 18-32, mean 23; 20 females; 3 self-reportedly left-handed participants were assigned to different groups), all native speakers of German with normal or corrected-to-normal vision, volunteered to participate in the experiment and were paid for their participation.

8.2.2. Results and Discussion

All participants answered more than 85% of the comprehension questions correctly (M = 95%, SD = 0.05). Items that received incorrect answers and decision latency outliers (> 2.5 SDs from the mean per region) were excluded from the analysis (10% of the data points).

Reading times in each region were analyzed through a generalized mixed effect regression model, separating random effects for item and for participant and taking into account trial-to-trial longitudinal dependencies between observations (as for Experiment 1). Following Baayen et al. (2008), we used an empirical procedure to decide what factors to include in the model, ruling out factors that did not significantly

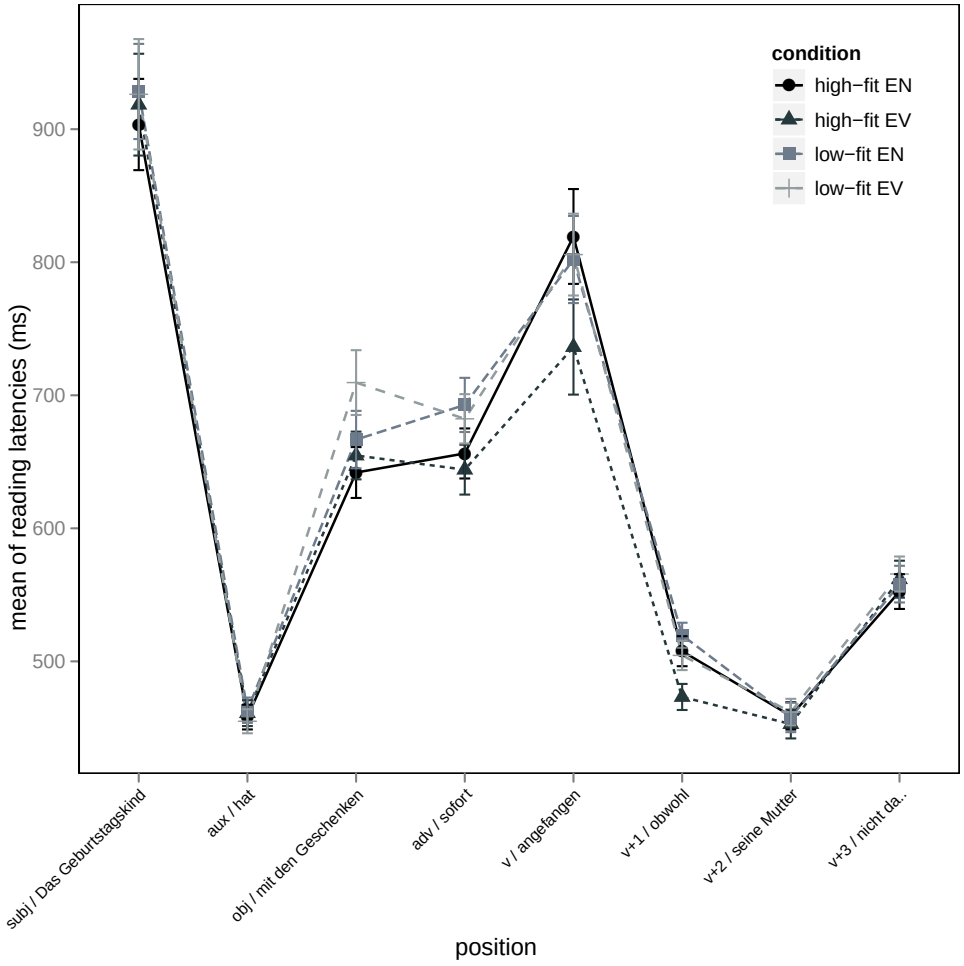

Figure 8.2.: Experiment 3: Comparing reading latencies (in ms) for each position and for each condition.

contribute to the model's goodness of fit, determined by a likelihood ratio test. The model's covariates which contributed to the goodness of fit and were thus included were the reading times at the previous word and the order of presentation of each trial (rank-order of a sentence in its experimental sequence). Table 8.1 shows mean reading times per condition. No significant effect was found before the object region, which was not surprising as the sentences were identical in both conditions up to this region, and after the verb + 1 region.

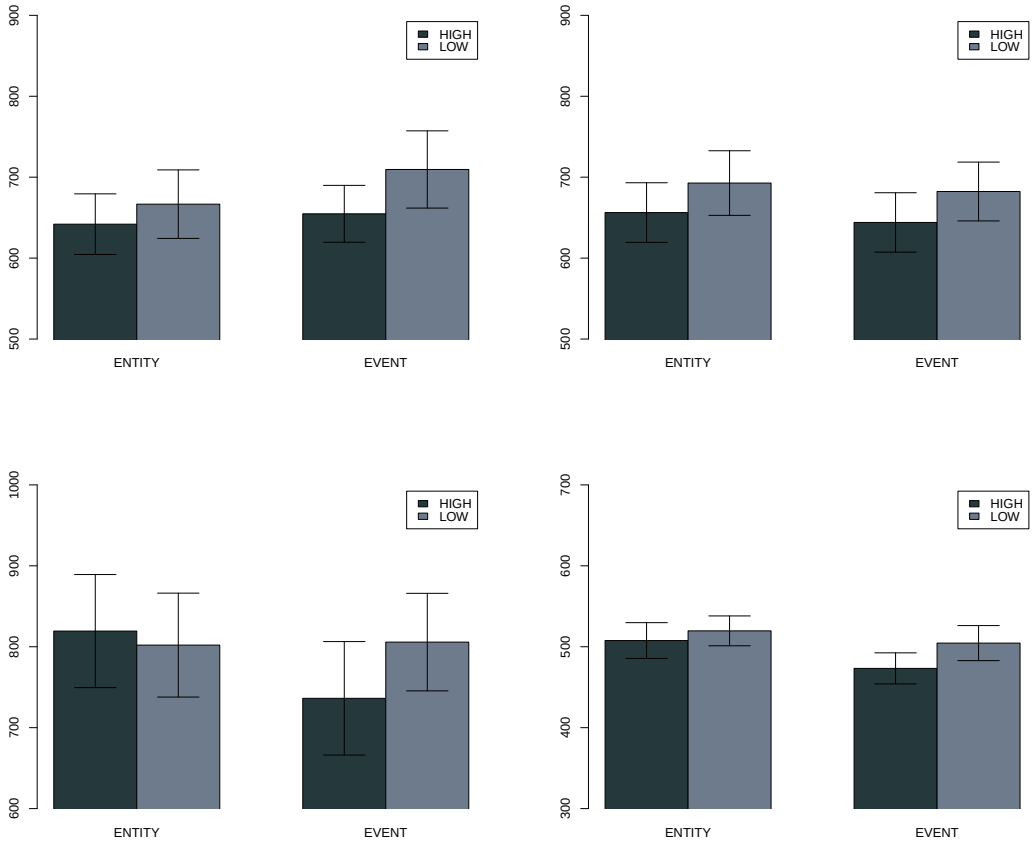

Figure 8.3.: Experiment 3: Comparing reading latencies (in ms) for each condition at the object region (top left), at the adverb region (top right), at the verb region (bottom left) and at the verb + 1 region (bottom right).

Object Region (*mit der Feier*): Entity-denoting objects were read faster than event-denoting objects ($t = -2.2$; $p = 0.03$), not surprisingly because event-denoting objects arguably have argument structures and because our event-denoting objects were longer than our entity-denoting objects. An effect of typicality / thematic fit was also found at the object region ($t = 2.28$; $p = 0.02$): high-thematic-fit objects, matching the reader's expectations, were read faster.

Object + 1 Region (*sofort*): The effect of thematic fit lingered at the object + 1 region (the adverb region: $t = 2.19$; $p = 0.03$). No effect of type was yielded.

Verb Region (*angefangen*): At the verb region, which is the region where the type

clash is expected to take place, we observed a main effect of object type ($t = -2.5$; $p = 0.01$) which was qualified by an interaction of object type and thematic fit ($t = 2.04$; $p = 0.04$).

Verb + 1 Region (*obwohl*): A main effect of object type ($t = -3.32$; $p = 0.001$) and of thematic fit ($t = 2.83$; $p = 0.01$) was yielded at the verb + 1 region. Both at the verb and at the verb + 1 region, sentences with high-typicality event-denoting objects yielded the shortest reading times (see Figure 8.3). While sentences with event-denoting objects yielded longer reading times up to the verb region, here the opposite tendency occurred: metonymic verbs were read faster after event-denoting objects compared to entity-denoting objects (provided that the thematic fit of the latter was high enough).

The word order exploited in Experiment 3 allowed us to separate the two factors, showing that thematic fit information (cued by contextual elements) and type are exploited early: Experiment 3 showed an effect of both thematic fit and type clash on reading times for logical metonymies: metonymic verbs were read more quickly after an event-denoting object, but only if the object matched the generalized event knowledge cued by the subject of the sentence (e.g. *Das Geburtstagskind hat mit der Feier angefangen*, *The birthday boy has with the party begun*).

High-typicality event-denoting nouns had a 70 ms advantage over low-typicality event-denoting nouns at the verb region, whereas the opposite was found for entity-denoting nouns (a 17 ms disadvantage for the high-typicality condition over the low-typicality condition). This can be easily explained in terms of expectation-based processing: after a high-typicality entity-denoting object, a metonymic verb is an unexpected continuation, whereas a content-loaded action verb would probably be more expected (e.g. *presents → open* vs. *presents → start)*.

The effect of thematic fit appeared early (at the object region), and was maintained through the spillover (v + 1) region. The type effect appeared at the region where the type clash was expected to take place (at the verb region, with an interaction with thematic fit) and continued into the spillover region. A late effect of type was also yielded at the spillover region.

In sum, both type and thematic fit influence processing costs for logical metonymies: thematic fit has an effect on reading times (the costs for the logical metonymy were modulated by varying the thematic fit), but it was the interaction of the two factors to be decisive (the shortest reading times were yielded by sentences with high-thematic-

fit event-denoting objects). This suggests that the Lexical Hypothesis does indeed provide the appropriate trigger for the logical metonymy (type), despite not being context-sensitive enough to account for the effects of thematic fit.

Part II and Part III have addressed the source problem and the trigger problem respectively, evaluating the predictions from the Words-as-cues-Hypothesis (and, conversely, of the Lexical Hypothesis and the Pragmatic Hypothesis) regarding those two problems. While Part II provided evidence from psycholinguistic experiments and from computational modeling that thematic fit (informed by generalized event knowledge) constitutes a valuable alternative to qualia-based covert event interpretation for the source problem (as predicted by the Words-as-cues Hypothesis), specifically by providing a context-sensitive and dynamic mechanism to generate a ranked list of possible covert events, Part III has shown – somewhat surprisingly – that the picture may be different than that predicted by a thematic-fit-only account: while the computational model of thematic fit in Chapter 7 had suggested that thematic fit may play a decisive (and perhaps exclusive) role in distinguishing logical metonymies from non-metonymic constructions, Experiment 3 has shown that thematic fit does not provide a sufficient answer for the trigger problem. This result is in principle not incompatible with a Words-as-cues account: experiments in this framework have consistently shown that syntactic cues (such as selectional restrictions on type) are rapidly combined with thematic fit information, influencing the expectation-building process during online language comprehension (see for example Trueswell et al., 1993, 1994; McRae et al., 1998; Hare et al., 2003, 2009a; Matsuki, 2013). In the last part of this dissertation I will address the problem of type, comparing the results obtained for logical metonymy with experimental work on standard metonymy and metaphor and discussing the role of type in a Words-as-cues account of logical metonymy.

Part IV.

The Words-as-cues Hypothesis Revisited

9. The Words-as-cues Hypothesis Revisited

The psycholinguistic experiments in Chapter 5 and the computational models in Chapter 6 have provided evidence consistent with the hypothesis that generalized knowledge about typical events and their participants is the source of the covert event, but Experiment 3 in Chapter 8 has shown that both thematic fit and type clash influence reading times, suggesting that the type of the object is indeed a necessary factor when accounting for differences between metonymic and non-metonymic constructions, and that both type and thematic fit play a role in logical metonymy interpretation.

I will now shift from logical metonymy, to consider other non-compositional phenomena which, as logical metonymy, involve a transfers of meaning (that is, standard metonymy and metaphor), in order to discuss how the results reported in this dissertation compare with related work investigating their behavioral correlates and what shared processes and resources might be involved. Results from Experiment 3 and from related work call for (a) a consideration of the role of context in non-compositional phenomena of meaning transfer, (b) a reconsideration of the notion of type, which is mostly taken for granted in type-shift-based accounts, and whose relation with thematic fit or generalized event knowledge has not been explored, and (c) a revision of the Words-as-cues Hypothesis, which can not prescind from a notion of type. I will then propose a context-sensitive model of logical metonymy interpretation that exploits an information-rich lexicon, but also includes a notion of type, which is reconciled with the notion of thematic fit.

9.1. The Cost of Meaning Transfers

The psycholinguistic experiments reported in this dissertation show that context helps readers build expectations about typical event scenarios, but at the same time the preference of the metonymic verb for entity-denoting objects also affects expectations (and is reflected in reading times): it appears that, while the verb's lexical properties impose structural constraints on the type of its arguments (a metonymic verb requires a complement of type event), the generalized event knowledge activated by contextual cues generates expectations about typical objects. Two factors seem to be at work (and to interact) here: the verb's lexical (often idiosyncratic) argument-selecting properties and context-sensitive generalized event knowledge. In Experiment 3 the facilitating context (cueing high-typicality objects) did interact with the verb's lexical properties (selecting event-denoting objects), ultimately resulting in reduced processing costs for (high-typicality) event-denoting objects compared to entity-denoting objects.

The facilitating role of context and its interaction with type mismatches in phenomena of meaning transfers has been explored by a number of experimental studies on constructions which defy a simple compositional interpretation, such as standard metonymy and metaphor (Nunberg, 1995). I will now review the main findings from studies on standard metonymy and metaphor, in order to sketch a more complete picture of context integration in non strictly compositional phenomena.

9.1.1. Standard metonymy

Standard metonymy (Stern, 1931; Nunberg, 1979) requires a (part-for-the-whole) transfer of meaning, and more specifically an entity-to-entity transfer (*reading Jack Kerouac* → *reading the books of Jack Kerouac*). McElree et al. (2006a) argued that (a) standard metonymy is not costly per se (and, if supported by preceding context, is as easy to process as non-metonymic constructions), and that (b) logical metonymy is computationally more complex than standard metonymy (and consequently more costly), because the former includes building an event template, whereas the latter does not require an extra event meaning.

The first claim (that standard metonymy is not costly per se) is supported by an eye-tracking study investigating the role of supporting context in standard metonymies. Frisson and Pickering (2007) contrasted familiar and unfamiliar [producer-for-product]

metonymies (e.g. *My great-grandmother often read Dickens / Needham*) with non-metonymic controls (e.g. *My great-grandmother often met Dickens / Needham*). The metonymies were presented either with supporting context (e.g. preceding sentences making clear that *Needham is a writer*) or without. The study did not yield any particular difficulty with unfamiliar metonymies supported by context[1]: both unfamiliar metonymies and familiar ones were processed efficiently (as efficiently as the non-metonymic controls), provided that the preceding context introduced some noteworthy features necessary for the metonymic shift (e.g. *Needham is a writer* → we are probably talking about *his books*, Nunberg, 2004; McElree et al., 2006a; Frisson and Pickering, 2007).

The second claim (that logical metonymy is more costly than standard metonymy) is supported by another eye-tracking study (McElree et al., 2006a), contrasting standard metonymies (e.g. *The editor published Rushdie before the death threats were issued*) as well as logical metonymies (*completed Rushdie*) with non-metonymic controls (*invited Rushdie*). They report that logical metonymy was more difficult to process than the standard metonymy condition (first-pass regressions at the object region) and the control (first-pass regressions at the object region and total time), and interpret this result as evidence in favor of the higher complexity of the logical metonymy shift compared to the standard metonymy shift. At this point, it seems reasonable to conclude that standard metonymies are easier to process than logical metonymies.

However, there are reasons to question these two claims. Recent work (Schumacher and Weiland, 2011; Schumacher, 2011, 2013) provided evidence that standard metonymy does have an inherent processing cost which can not be cancelled by providing supporting context. These studies investigated standard metonymies by examining event-related brain potentials (ERPs), that is the modulations of electrical activity in the brain in response to a cognitive stimulus, time-locked to the stimulus presentation (Kutas and Van Petten, 1994). Recording and investigating ERPs as a correlate of cognitive processes allows for a more fine-grained temporal resolution (in the order of milliseconds) than reading time data. Schumacher and colleagues identified a biphasic pattern typical of metonymic processes in ERP components: the elicited signatures are the N400 response (a negative deflection in the ERP waveform peaking around 300 ms after the stimulus onset), which is associated with expectation-driven

[1] See also Frisson and Pickering (1999) for similar results [place-for-institution] and [place-for-event] metonymies.

parsing guided by the fit of the stimulus with prior context (Kutas and Federmeier, 2000; Schumacher, 2009), and the Late Positivity window (a positive deflection in the ERP waveform peaking around 550 ms after the stimulus onset), which is associated with an update of the discourse representation structure (Bornkessel and Schlesewsky, 2006; Schumacher, 2009).

Familiar metonymies (e.g. *Tim's uncle once read / met Brecht*) were contrasted with non-metonymic controls and elicited a typical N400 / Late Positivity biphasic pattern (Schumacher and Weiland, 2011). Unfamiliar metonymies (e.g. *The espresso wanted to pay*) were contrasted with non-metonymic controls (e.g. *The espresso was out of stock*), either with supporting context (e.g. *The waitress asks the barkeeper who wanted to pay. The espresso wanted to pay*, Schumacher, 2011) or without supporting context (e.g. *Kristen asks Geoff who wanted to pay. The espresso wanted to pay*, Schumacher, 2013), in order tease apart the role played by familiarity and context. In the absence of supporting context, a larger N400 response as well as a more pronounced Late Positivity effect were reported for the unfamiliar metonymies compared to the non-metonymic controls. When supporting context was provided, no differences were yielded in the N400 response between the unfamiliar metonymies and the non-metonymic controls, whereas a more pronounced Late Positivity effect was still reported for the unfamiliar metonymies. In conclusion, the claim that standard metonymy is as costly as non-metonymic interpretation may not be supported by ERP evidence, which shows a Late Positivity response for standard metonymies also in the presence of context supporting a metonymic interpretation.

Also, note that possible differences may exist between the experimental conditions in McElree et al. (2006a) with regard to typicality. Among the 24 writers whose names were employed in the test sentences, 23 (e.g. Sartre, Tolstoy) passed away before 1980[2]: one could then argue that, while it may be considered equally *plausible* that *the student welcomed Sartre* or *the student read Sartre* (but we would have to place that sentence in a far away past), the latter seems to match our knowledge of what students *typically* do (*read books*) much more than the former (*meet famous writers*). Provided that there might still be a difference in terms of computational complexity between standard and logical metonymy, the conclusions that standard metonymies are not more costly than their non-metonymic counterparts may be overlooking or masking

[2]Also, the only one living (Rushdie) famously lived under police protection because of a fatwa until 1999.

possible differences with regard to typicality between between standard metonymies and the non-metonymic controls.

As to the second claim (that logical metonymy should be computationally more complex than standard metonymy), note that the logical metonymies employed in McElree et al. (2006a) involved *two* metonymic steps: a [producer-for-product] standard metonymy (*Rushdie* → *Rushdie's books*) and the logical metonymy proper (*Rushdie's books* → *reading Rushdie's books*). The difference may then lie in the number of type shifts rather than in their complexity: a more balanced contrast would be between *reading Rushdie* and *beginning the book* rather than between *reading Rushdie* and *beginning Rushdie*.

In sum, a dissociation emerged in ERP results on standard metonymy between (context-based) expectation effects on the one hand and enrichment processes on the other hand (Schumacher, 2013): contextual licensing was shown to reduce context-induced N400 responses in the ERPs, but did not prevent processing costs arising from the type shift, which in turn resulted in a later signature of metonymic processes, that is a Late Positivity, which distinguishes even context-supported standard metonymies from non-metonymic controls (the former being more costly). Also, the choice of experimental stimuli in McElree et al. (2006a) casts some doubt on the conclusion that standard metonymies are not more costly than non-metonymic constructions and that logical metonymies in turn are more costly than both.

9.1.2. Metaphor

The ERP methodology was also applied to the study of metaphors (see Bambini and Resta, 2012 for a review), in particular to investigate whether (a) metaphors are first interpreted literally, and are reinterpreted metaphorically only at a second stage, upon failure of the literal interpretation (literal-first hypothesis, Janus and Bever, 1985), or whether (b) the metaphoric interpretation is accessed straightforwardly (direct-access hypothesis, Gibbs and Gerrig, 1989). According to this latter view, contextually-relevant metaphors should be as easy to process as literal meanings (see also the Graded Salience Hypothesis, Giora, 2003, for an intermediate position).

Pynte et al. (1996) contrasted short familiar metaphors (e.g. *Those fighters are lions*), unfamiliar metaphors (e.g. *Those apprentices are lions*) and controls (e.g. *Those animals are lions*). A larger N400 response was elicited for metaphors compared to the

non-metaphorical controls, which was reduced when the metaphor was supported by context (e.g. *They are not naive: those fighters are lions*). A larger Late Positivity response was elicited for unfamiliar metaphors compared with the controls, both when they were supported by context and when they were not, but not for familiar metaphors.

Pynte et al. (1996) claims that the N400 response is modulated by context effects, and considers this as evidence in favor of a context-dependent (direct-access) account of metaphor comprehension, arguing that the literal meaning is accessed only when the metaphor is not supported by preceding context, whereas others (for example, De Grauwe et al., 2010) interpreted these as well as similar results as evidence in favor of the literal-first hypothesis, arguing that this is the only hypothesis consistent with Late Positivity effects.

Despite the lack of consensus in the interpretation, a N400 / Late Positivity biphasic pattern (see also Coulson and Van Petten, 2002 for a similar pattern) did consistently emerge as the typical signature of metaphor, similar to that encountered for standard metonymy.

9.2. The Cost of the Logical Metonymy

Studies on standard metonymy and on metaphor have shown that, if on the one hand supporting context can significantly ease the processing of these constructions, on the other hand these are still more costly than the non-metonymic or non-metaphoric controls. I will now review the main results from related work on logical metonymy, in particular regarding the problem of the processing cost of the logical metonymy.

9.2.1. Lexical Hypothesis vs. Pragmatic Hypothesis

Supporters of the Lexical Hypothesis (and in particular of the type-shift solution) argued that logical metonymies are more costly than their non-metonymic counterparts. Their experimental results showed effects on reading times at the obj+1 position for logical metonymies (e.g. *the boy started the puzzle* /*the fight* ***after school*** *today*, Traxler et al., 2002), which were interpreted as evidence for the accommodation of a type-shifting operation. On the other hand, supporters of the Pragmatic Hypothesis (de Almeida, 2004; de Almeida and Dwivedi, 2008; de Almeida et al., 2009) argued

that such results are determined by post-lexical inferential processes, and that they should disappear when the range of covert events is narrowed down and the need for an inference to solve the underspecification of the logical metonymy is cancelled.

In their responses to de Almeida and colleagues, the group of researchers supporting the type-shift solution (Pickering et al., 2005; Traxler et al., 2005; McElree et al., 2006a; Frisson and McElree, 2008) identified four possible sources of the cost of the logical metonymy:

1. the meaning shift in itself;
2. the retrieval of the covert event;
3. the conflict between different covert event interpretations;
4. the construction of the event sense.

They then ruled out that the meaning shift itself (1) is costly, according to experimental results in McElree et al. (2006a) where logical metonymy is considered more costly than standard metonymy; they also ruled out that the retrieval of the covert event (2) is costly, because Traxler et al. (2005) reported extra costs for logical metonymies also when the event is explicitly mentioned in preceding context, and that conflicting event interpretations (3) are the cause of extra costs, as Frisson and McElree (2008) reported no differences between logical metonymies with one covert event interpretation and others which are ambiguous with regard to the possible covert events. This has led them to suggest that the extra processing costs for logical metonymies should be ascribed to a fourth cause, that is (4) "the **construction of an event sense** for a complement that is of a different semantic type" (Frisson and McElree, 2008, p. 7), or in other words the resolution of the metonymic shift resulting in the construction of an argument structure including the covert event.

This conclusion though, reached at the exclusion of other hypotheses, is not entirely convincing: in Section 9.1.1 I questioned the conclusion that standard metonymy is not costly, whereas logical metonymy requires more computational effort. Also, the debate between the supporters of two different views on the trigger problem as well as previous work by Katsika et al. (2012) suggest that diverging results between the supporters of the Lexical Hypothesis and the supporters of the Pragmatic Hypothesis may be due to differences of experimental design and choice of materials. Lastly,

our own results (the modeling study in Chapter 7, as well as the methodological considerations in Chapter 8) have argued for the presence of a confounding factor (thematic fit) which was not controlled for in previous experimental studies on logical metonymy, calling for more controlled experimental design and materials, in order to disentangle possible effects of type and thematic fit. Similar considerations were proposed by work on surprisal effects in logical metonymy interpretation, which I will now review.

9.2.2. Logical Metonymy as Surprisal

Logical Metonymy has recently been investigated with ERPs methodologies as well. Baggio et al. (2010) and Kuperberg et al. (2010) contrasted logical metonymies with anomalous sentences (violating animacy constraints) and non-metonymic controls (*The journalist began / astonished / wrote the article*)[3]. Both Baggio et al. (2010) and Kuperberg et al. (2010) report an N400 effect evoked by the entity-denoting object in the logical metonymy condition (*began the article*) and in the anomalous condition (*astonished the article*) compared to the control[4]. In light of the studies on standard metonymy and metaphor reviewed in this chapter, the N400 may be a correlate of the (low) predictability of the entity-denoting object. Baggio et al., 2010 and Kuperberg et al., 2010, however, did not further investigate this aspect (for example, by modulating context in the design of a second study). Rather, they observed that the N400 effect could also be ascribed to the detection of a semantic mismatch and the consequent integration of the meaning of the entity-denoting object in the sentence meaning.

Delogu et al. (2013) further investigated the nature of this N400 effect, moving from the hypothesis that coercion effects in logical metonymy may be explained in terms of **surprisal**. According to Surprisal Theory (Hale, 2001; Levy, 2008; Smith and Levy, 2013), the predicted processing difficulty of a word given its preceding context equals to its surprisal, which is proportional to the logarithm of its conditional probability of

[3] As I have already pointed out, comparing logical metonymies with semantic anomalies may be misleading, as logical metonymies are not perceived as anomalous. Also, note that both Baggio et al. (2010) and Kuperberg et al. (2010) employed a varied array of "metonymic verbs", including not only aspectual verbs but also verbs like *master, endure, manage, resist.*

[4] Baggio et al. (2010) also report a sustained late negativity (700-1000 ms after the stimulus onset) for the metonymy compared to the other two conditions, whereas Kuperberg et al. (2010) report a P600 effect (a positive-going component in the ERP signal 500-900 ms after the stimulus onset) for the anomalous condition compared to the other two.

appearing (increased surprisal → increased reading times):

$$\text{Surprisal}(w) = -\theta \log P(\text{word}|\text{context})$$

Note that this approach is quite comparable with the expectation-based approach adopted in this dissertation (the Words-as-cues Hypothesis). Surprisal and thematic fit offer two different measures of predictability: the former is breadth-based, as it is defined in terms of a word's conditional probability of appearing, computed on the base of co-occurrence with (potentially every) word occurring in preceding context (as in a Language Model), whereas the latter is depth-based, in that it is based on syntactic relations between words (e.g. subject, object). Nevertheless, it is not excluded that P could be computed differently, for example by taking into account syntactic dependencies, thematic roles and local mutual information as in our model of thematic fit. Delogu et al. (2013) also observed that in previous psycholinguistic work on logical metonymy, the surprisal scores for the complement noun were lower in the coercion conditions compared to the control conditions, and that even small differences in cloze probability between conditions in previous work may have lead to an (underestimated) effect of predictability on reading times, which in turn could have been potentially misunderstood as an effect of logical metonymy.

In order to tease apart coercion and surprisal, Delogu et al. (2013) contrasted a logical metonymy with two non metonymic conditions, a neutral condition and a preferred condition (*Peter began / bought / read the book*[5]). The neutral condition matched the surprisal score of the metonymic condition, whereas the preferred condition had a higher surprisal score for the entity-denoting object compared with the other two. Both the neutral and the metonymic condition evoked larger N400 responses for the entity-denoting object compared to the preferred condition. Recall that Baggio et al. (2010) and Kuperberg et al. (2010) had observed that the N400 effect in logical metonymy interpretation could be ascribed either to the low predictability of the entity-denoting object or to the detection of a semantic mismatch (and the consequent integration of the event meaning). Interestingly, Delogu et al. (2013) by controlling for surprisal were able to rule out the hypothesis that the N400 reflects the detection of a semantic mismatch, and to ascribe the N400 effect to the predictability of the entity-denoting object in the metonymic condition.

[5]Note the absence of a semantically anomalous condition.

Also, in an eye-tracking experiment they showed that surprisal can also account for early effects in logical metonymy interpretation (first-pass regressions and regression path time in the spillover region), but (crucially) not for total reading time at the object region. These late effects seem indeed to be characteristic of logical metonymy interpretation, and possibly coherent with the effect of type reported for Experiment 3 in this dissertation.

In sum, it seems that predictability (Surprisal or thematic fit) does indeed influence the processing of logical metonymies in early measures (in a similar way as for standard compositional processing), but then in later stages of processing type-shifting operations intervene, which are typical of the logical metonymy itself.

9.3. Type revisited

Related work on non-compositional phenomena involving transfers of meaning (standard metonymy and metaphor) has investigated the interaction of context-dependent predictability on the one hand (reflected in early behavioral measures, such as the N400 component in ERP studies) and of other more structural aspects, related to the subcategorization frame of the verb and its selectional behavior (type clashes) on the other hand (reflected in late behavioral measures, e.g. a Late Positivity for standard metaphor and metonymy or total time in eye tracking studies on logical metonymy). This picture is compatible with the results of Experiment 3: predictability (modeled as surprisal or as thematic fit) influences processing costs and is likely to be the factor determining the covert event for a logical metonymy, but is not sufficient to determine the cost of a logical metonymy, as high-thematic-fit entity-denoting objects do not facilitate the processing of the metonymic verb as much as high-thematic-fit event-denoting objects do: another factor (type) needs to be taken into account (in its interaction with predictability) in logical metonymy interpretation, which can not be reduced to predictability only.

Previous work on logical metonymy, while ascribing the additional processing costs for logical metonymy to a type-clash and type-shift mechanism, considered types to be somewhat "given", without explaining what semantic types are or what cognitive reality they have. I will now start addressing the role of semantic types, and in particular how they can be compatible with expectation-based accounts of language understanding such as the Words-as-cues Hypothesis.

9.3.1. Type and Verb Bias in a Words-as-cues Framework

The evidence presented in this dissertation, coming from different experimental methodologies (corpus analysis, web experiments, psycholinguistic studies and computational modeling), has shown that logical metonymy is mainly determined by a verb's preference for entity-denoting objects, and in particular by the argument-selecting properties of a very specific class of verbs (aspectual verbs), but also that verbs can vary greatly with regard to their idiosyncratic event-selecting behavior along an "eventhood" axis (see for example the Eventhood Model in Chapter 7).

The idea that verbs have a lexically-determined bias for certain structures and for certain argument types is not in principle incompatible with the Words-as-cues framework: a number of studies have explored the interaction between a verb's structural biases and context-based typicality information in shaping people's expectations about upcoming linguistic input. Trueswell et al. (1993) showed that lexically specific constraints determined by verbs are one of the main information sources to guide parsing, and have an early influence on processing. For example, the verbs *forget* and *hope* in English differ with regard to their subcategorization bias, which prefers an NP complement (for the former) or a sentence complement (for the latter)[6]:

(9.1) a. The student forgot the solution was in the back of the book.

b. The student hoped the solution was in the back of the book.

(Trueswell et al., 1993)

9.1.a-b are syntactically ambiguous at *the solution*, as this can be either the direct object of the main verb or the subject of the sentence complement. The early use of subcategorization information affects the syntactic analysis of the ambiguous NP, causing readers to slow down when encountering a sentence complement for *forget* (due to its NP-bias) and when encountering an NP complement for *hope* (due to its S-bias).

Trueswell et al. (1994) also showed that subcategorization biases interact with thematic fit. In the following examples, the verb *examine* has a bias for an animate agent and an inanimate patient:

[6] Subcategorization biases were estimated with a sentence completion study.

(9.2) a. The defendant examined by the lawyer turned out to be unreliable.

b. The evidence examined by the lawyer turned out to be unreliable.

(Trueswell et al., 1994)

Also 9.2.a-b are syntactically ambiguous at *examined*, as this can be either a reduced relative or a main verb in the simple past form. The early use of thematic fit information affects the syntactic analysis of the ambiguous region: readers expect *examined* to be the main verb when it is introduced by a good agent (the animate noun), and thus slow down when reading the reduced relative (disambiguated at *by the lawyer*) compared to the condition when the reduced relative was introduced by a good patient (the inanimate noun).

Similarly, the interpretation of verbs showing an alternation between a transitive (causative) and an intransitive (inchoative) structure (e.g. *chill*) was shown in Hare et al. (2009a) to be influenced by thematic fit:

(9.3) a. The wind chilled the spectators who stood drinking wine on the terrace.

b. The wine chilled the spectators who stood drinking it on the terrace.

c. The wine chilled through the afternoon but they never bothered to open it.

d. The wind chilled through the afternoon but they never bothered to put sweaters on.

(Hare et al., 2009a)

The wind is a good-cause subject for *chill*, favoring the transitive (causative) construction (9.3.a), whereas *the wine* is a good-theme subject, favoring the intransitive (inchoative) construction (9.3.c). As a result, at the region after the verb (where the causative-inchoative ambiguity is resolved), participants were faster to read the transitive continuation for the condition with the good-cause subject (9.3.a) compared to the transitive condition with the good-theme subject (9.3.b); whereas they were faster to read the intransitive continuation for the condition with the good-theme subject (9.3.a) compared to the intransitive condition with the good-cause subject (9.3.a). A similar interaction between generalized event knowledge / thematic fit and subcategorization information was also found by many other experimental studies, see for example Hare et al. (2003); McRae et al. (1998); Matsuki (2013).

There is thus ample evidence that thematic fit interacts with a verb's idiosyncratic behavior (and its structural biases) and that they both contribute to the generation of expectations about upcoming linguistic input. Similarly, in the light of these studies, the results from Experiment 3 may be interpreted as an interaction between a verb's expectations for a certain (expected) *meaning structure* (in this case, for a *complement of type event*) and the expectations for typical (high-thematic fit) objects (with the constraint that these have to be event-denoting objects).

Type is essentially a way of imposing a structure on the generalized event knowledge activated by a verb: a verb like *begin* will trigger expectations for high-thematic fit objects, but only for those which match a specific type (trivially, *beginnable* objects are objects of *type event*). Type then determines the way generalized event knowledge is accessed, playing a somewhat similar role to that of grammatical aspect: it has been shown that events prime typical locations (*was skating* → *arena*) in the imperfective aspect form, but not when presented as completed (perfect aspect: *had skated* → *arena*, Ferretti et al., 2007).

As I mentioned in Chapter 8, different behavioral correlates were yielded by entity-denoting and event-denoting nouns (Collina et al., 2001; Tabossi et al., 2010; Garbin et al., 2012; Bedny et al., 2013). Although the psychological reality of semantic types is probably unquestionable, the way they are represented has not been investigated, for example, whether they should simply be represented as properties, as sets with binary membership functions (+*event*/−*event*) or, in accordance with prototype theories of concepts (Rosch, 1975; Osherson and Smith, 1982; Smith et al., 1988; Garbin et al., 2012; Bedny et al., 2013), they could be more similar graded membership sets, with identifiable prototypes and with typicality defined as the distance from the prototype (some items are more event-like than others) and determined by the shared properties with the prototype and the other members of the set (Erk, 2010).

Another interesting challenge is posed by the problem of their granularity: if types are a way to represent expectations at a more abstract level, then experiments such as those presented and reviewed in this dissertation (aimed at disentangling generalized event knowledge and abstractions over this knowledge) are a helpful tool to investigate which levels of abstractions are cognitively relevant (in our case, the entity vs. event distinction), which are stored and which created on demand. Considering the importance of semantic type and their interaction with generalized event knowledge for lexical semantics and their relevance for language processing, answering

these questions is a necessary step to gain insight into the organization of our mental lexicon.

9.3.2. Type in a Computational Model

Two-stage models of parsing assume that syntax has a primary role in guiding the reader / hearer's processing and that lexical information and world knowledge intervene only at a second stage, to revise the initial parse (see for example the Garden Path Model, Frazier and Rayner, 1982). Their predictions, though, are at odds with psycholinguistic evidence which shows that, in the sequential processing of linguistic input, people exploit information coming from an extremely diverse range of sources, and that in particular subcategorization information is available very early during processing. Constraint-based models (MacDonald et al., 1994; McRae et al., 1998, see for McRae and Matsuki, 2013 for a review) implement the interplay of such constraints in guiding parsing and in solving local ambiguities, making more plausible claims about the way these constraints are integrated to predict upcoming input which are in accordance both with the predictions from the Words-as-cues approach and with the experimental results corroborating it. In a constraint-based model, semantic type may act as yet another constraint contributing to the expectation building process.

The Eventhood Model in Section 7.2 provides a unified model encoding both thematic fit and type, in the spirit of previous work modeling selectional preferences as distributions over WordNet synsets (see for example Resnik, 1996; Abe and Li, 1996; Clark and Weir, 2001; Schulte im Walde, 2006). The latter models also represent corpus-based selectional preferences for conceptual classes as distributions over classes of fillers (for example, mapping them onto WordNet synsets), and then generalize over these classes[7] to characterize a verb's selectional preferences.

These unified models rely on a given taxonomy (e.g. WordNet, GermaNet), but do not tackle the problem of how this taxonomy should be implemented or how types should be represented. An interesting answer to this problem comes from another model in accordance with the Words-as-cues approach, the Simple Recurrent Network (Elman, 1990), which is proposed as an example of a dynamical system, where words interact in real time during processing and act as "stimuli that alter mental

[7] Their generalizations involve lower levels of abstractions than our Eventhood Model, with the exception of Schulte im Walde (2006), who uses GermaNet top nodes.

states" (Elman, 2009, 2011). Interestingly, as it appears from an analysis of the learned average internal states of the network, the model is able to create its own lexical representations. From a clustering of these representations, a category structure clearly emerges, where syntactic and semantic categories can be distinguished (e.g. noun vs. verb, animate vs. inanimate). A semantic type may then be an emergent category, which is not encoded in the model but emerges from observed activation patterns.

Recent approaches in compositional distributional semantics have strived to find an optimal way to interface concepts of formal logic with distributional representations, mapping predicates in the logical forms with distributional representations. A common strategy (Baroni et al., 2012; Erk, 2012, 2013) is to map lexical entries with semantic types, which determine the algebraic type of the distributional representation (vector, matrix, tensor...). Nevertheless, models which associate each lexical entry to one ("given") semantic type (and to one shape of distributional representation) may be problematic when dealing with event nouns (would they be shaped in the same way as verbs, as they arguably have an argument structure, or as non-entity-denoting nouns?) or in general with ambiguous words (what is the representation of an event-/entity- ambiguous noun, e.g. *breakfast*) and when accounting for non strictly compositional aspects of sentence meaning such as type shifts (how does the distributional representation for an entity turn into the distributional representation of a covert event involving the entity?).

A different strategy to incorporate a notion of semantic type in distributional semantic models is to exploit the correlation between semantic type and patterns of corpus occurrence (similarly to Elman's Simple Recurrent Network), relying on behavior patterns (in this case, distributional patterns) to successfully distinguish between types without encoding any explicit information regarding semantic types. For instance, the DM + ECU model employed in this dissertation (Baroni and Lenci, 2010; Lenci, 2011), which did not encode type information, was able to mirror experimental results whose interpretation was based on the existence of semantic types. Other models have exploited a small set of seed words and corpus-extracted distributional patterns to infer semantic category associations via bootstrapping (Ravichandran and Hovy, 2002; Thelen and Riloff, 2002). It could then be the case that semantic types do not need to be explicitly included in such models because they emerge from corpora distributions.

Compared to compositional distributional models which rely on a strong type theory background, such a treatment of semantic types as emerging patterns of behavior has the advantage of relying on minimal assumptions regarding the granularity of the type ontology, as well as of allowing for large-scale multi-purpose modeling of linguistic phenomena, as these models are unsupervised or minimally supervised. This proposal is particularly intriguing, as pattern recognition is a key aspect of human cognition (see for example Rumelhart and McClelland, 1987; Saffran et al., 1996; Marcus et al., 1999; Tomasello, 2009) and is thus plausible to assume that types emerge from distributional patterns.

9.4. The Words-as-cues Hypothesis Revisited

In the first part of this dissertation I have sketched the main approaches to logical metonymy (the Lexical Hypothesis and the Pragmatic Hypothesis) and I have proposed a third approach, the Words-as-cues Hypothesis, inspired by work on expectation-based linguistic processing.

Let us look back at the predictions for the trigger problem and for the source problem as I had formulated them:

- **the trigger problem:** what triggers the logical metonymy?
  → low thematic fit between an event-selecting verb and an entity-denoting object;

- **the source problem:** what is the source of the covert event?
  → the covert event with best thematic fit is recovered from generalized knowledge about events.

The psycholinguistic experiments and the computational models presented in this dissertation supported the predictions from the Words-as-cues Hypothesis about the source problem: the events cued by a highly typical agent-patient combination were read faster and were correctly predicted by a similarity-based model based on thematic fit. Covert event interpretation is thus determined by the same source of information that drives people's expectations about upcoming input (generalized event knowledge, quantified as thematic fit, McRae and Matsuki, 2009; Elman, 2011) and which is then identified as relevant for implicit linguistic content as well.

On the other hand, thematic fit on its own did not suffice to account for the trigger problem. The last psycholinguistic experiment presented in this dissertation suggests that thematic fit and type interact, both acting as constraints determining which constructions are more costly and which are not: metonymic verbs generate expectations for event-denoting objects; logical metonymies (combining metonymic verbs and entity-denoting objects) were more costly than similar constructions with event-denoting objects, but only if these met the expectations built by previous context (high-typicality event-denoting objects). This is not incompatible with other phenomena observed within the Words-as-cues framework, characterized by the interaction between generalized event knowledge and a verb's selectional behavior: both expectations about type and expectations about generalized event knowledge are seamlessly integrated during language processing.

The Words-as-cues Hypothesis thus needs to be reformulated, to account for a notion of type as well as thematic fit (see Figure 9.1):

- **the trigger problem:** what triggers the logical metonymy?
  → the (lexical) type restrictions of the metonymic verb drive expectations for event-denoting objects, determining a type clash when an entity-denoting object is encountered;

- **the source problem:** what is the source of the covert event?
  → the covert event with best thematic fit is recovered from generalized knowledge about events.

The revisited Words-as-cues Hypothesis incorporates a notion of type, which structures generalized event knowledge information by determining the way it is accessed. Upon encountering a metonymic verb, which triggers expectations for an event-denoting object, high-typicality event-denoting object fillers are cued; if an entity-denoting object is processed instead, the expectations are not met and an operation of enrichment (similar to those required for standard metonymy, Schumacher, 2013) is required to obtain an event slot, whose typical fillers are then retrieved from our generalized event knowledge, contributing (and often anticipating) the final interpretation.

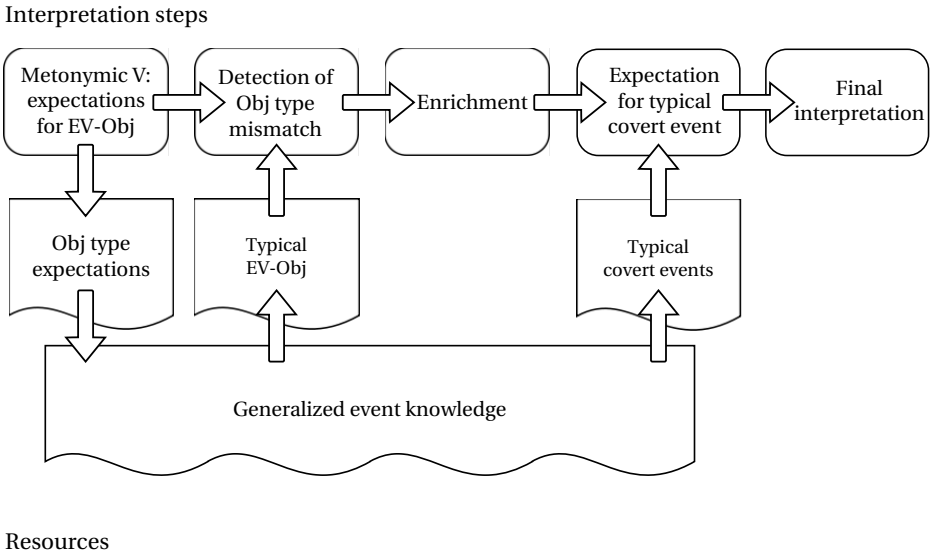

Figure 9.1.: Schematic representation of logical metonymy interpretation for the Revisited Words-as-cues Hypothesis, for cases in which the object is presented after the verb. If the object precedes the verb, then the low-thematic fit is detected at the verb region.

In the next and final chapter I will summarize the contributions of a Words-as-cues approach to the study of logical metonymy interpretation and draw some final conclusions.

10. Conclusions

This dissertation has employed a range of methodologies to carry out a critical analysis of various hypotheses of logical metonymy interpretation and to evaluate a new proposal based on the Words-as-cues framework, exploring the hypotheses that (a) generalized knowledge about typical events determines the covert event interpretation of a logical metonymy and that (b) thematic fit, informed by generalized event knowledge, is ultimately responsible for the trigger of the metonymic interpretation and of the recovery of the covert event, determining extra processing costs; I eventually revised this second hypothesis to incorporate a notion of type.

I will now provide a brief overview of the results reported in the previous chapters, summarizing the main contributions of this dissertation and in general of a Words-as-cues approach to the study of logical metonymy interpretation and I will draw some conclusions on lexical knowledge and world knowledge in models of language understanding and in particular in the Words-as-cues framework.

10.1. Models of Logical Metonymy Interpretation

The Lexical Hypothesis and the Pragmatic Hypothesis

The lexicalist theories of logical metonymy interpretation, based on the Lexical Hypothesis, has focused on the systematicity and regularity of logical metonymy, treating it as a special case of compositionality (*enriched composition*, Pustejovsky, 1991, 1995; Jackendoff, 1997) which is solved by relying on generative devices and lexical representations. This approach on the one hand maintains a clear distinction between lexicon and world knowledge, while the other hand it expands the domain of lexical knowledge, as it relies on a type-based system, on complex lexical entries and on generative devices to elegantly account for metonymic shifts within the lexical domain (both for the trigger of the metonymy — a type clash — and for the range of the

covert event, retrieved from the qualia structure of the lexical entry for the object). Some kind of event knowledge (e.g. *books* are *read* and *written*) is then included in the lexicon (in the form of qualia structures), albeit fairly restricted[1]. The web elicitation studies and the corpus analyses reported in this dissertation (Chapter 4) have shown that qualia structures are too narrow to account for the wide range of events that we associate with objects in our mental lexicon (and that are arguably relevant for the interpretation of covert events in logical metonymies). Also, intra-sentential context plays an important role in determining the covert event for a logical metonymy, but qualia are probably not context-sensitive and dynamic enough to account for fast and efficient expectation-based linguistic processing.

The Pragmatic Hypothesis arose from criticism of the idea that lexical items have an internal structure, and claimed that logical metonymies are challenging only because they are underspecified with regard to the covert event interpretation, which needs to be integrated (from world knowledge) via post-lexical inferences (Fodor and Lepore, 1998; de Almeida and Dwivedi, 2008). This approach, while bringing logical metonymy closer to "normal" compositional phenomena in our communication and allowing for an open-end set of possible covert events depending on context, looses the systematicity brought by the Lexical Hypothesis: the event knowledge necessary to logical metonymy interpretation is not structured, and ultimately does not make a clear prediction about what covert events are retrieved. Work by Katsika et al. (2012) as well as a computational model of Eventhood (Chapter 7 of this dissertation) have shown that, even though many metonymic verbs employed in experimental work on logical metonymy can be said to trigger post-lexical inferential processes, not all metonymic verbs are equal: aspectual metonymic verbs (e.g. *begin*, *finish*) do indeed have a stronger preference for event-denoting objects, and should thus give raise to different interpretation processes compared to non-aspectual "metonymic" verbs (e.g. *prefer*, *enjoy*).

The Words-as-cues Hypothesis

I have suggested an approach to logical metonymy interpretation based on a third hypothesis (the Words-as-cues Hypothesis), claiming that the type of event knowledge

[1] Not surprisingly, Pustejovsky's latest work significantly expands the idea of qualia into *habitats*, turning them into frames capturing salient aspects of a situation and its affordances (Pustejovsky, 2012, 2013).

involved in logical metonymy interpretation is the same generalized event knowledge (cued by lexical items during online sentence processing and computed in terms of thematic fit) which has been shown to be exploited during linguistic processing and which allows people to effectively anticipate typical upcoming input (McRae and Matsuki, 2009; Elman, 2011). This hypothesis was supported by the results of the psycholinguistic experiments in Chapter 5, which showed that people, when understanding a logical metonymy, resort to our generalized event knowledge to predict typical covert events compatible with the preceding context. Covert events are thus better understood as part of generalized knowledge about events involving the subject and the object of the logical metonymy. The Similarity-based Model in Chapter 6 was able to successfully predict the correct covert event for a logical metonymy relying on thematic fit information about the subject and the object of the metonymy and to effectively account for the role played by intra-sentential cues in covert event interpretation. A Words-as-cues approach to logical metonymy can achieve a much greater context sensitivity than qualia-based approaches (here intra-sentential context, but similar considerations can be made for a broader discourse context), while still making clear predictions about covert event interpretations for a given metonymy.

Thematic fit and generalized event knowledge were shown to provide a valuable (context-sensitive and dynamic) extension for the qualia structure. I have then explored the hypothesis (suggested by the computational models in Chapter 7) that thematic fit alone may be enough to account for the trigger of the covert event interpretation and to distinguish between metonymic and non-metonymic constructions. The psycholinguistic experiment in Chapter 8 then contradicted this hypothesis, showing that a verb's preference for event-denoting objects is indeed a necessary trigger of logical metonymy interpretation. Both the object type and its thematic fit influence processing costs for the logical metonymy, interacting early on: the event-denoting objects with high thematic fit are those that facilitate processing the most. This calls for a revision of my previous formulation of the Words-as-cues Hypothesis for logical metonymy interpretation (Chapter 9), which needs to take into account the interaction between type and thematic fit.

Revising the Words-as-cues Hypothesis to accommodate for the interaction of type and generalized event knowledge is not incompatible with other experimental results within the Words-as-cues framework, which have shown that a verb's selectional be-

havior can interact early with the activation of generalized event knowledge and that these can both be considered constraints intervening in efficient (expectation-based) language processing. This treatment of logical metonymy places the phenomenon within a broader framework of expectation-based interpretation rather than considering it an "anomaly" of language processing.

A Note on Cross-linguistic Comparisons

Previous psycholinguistic and computational work on logical metonymy was carried out mainly for English (see Lapata and Lascarides, 2003, for an exception). The choice of German for the psycholinguistic experiments reported on in this dissertation was determined by practical reasons (availability of German participants for lab studies) and by the experimental design (Experiments 1 and 3), which capitalized on German word order. The computational modelling studies reported in this dissertation were carried out for German (when the aim was to model experimental studies for German) and English (when the aim was to model previous experimental studies for English).

The corpus studies reported in Chapter 4 show that very similar considerations hold for logical metonymies in English, German and Dutch (Sweep, 2012), thus supporting the assumption that our considerations from experimental work on German can be extended to languages sharing strong structural similarities with it, but also exhibiting typological differences (e.g. English does not have verb-final word order for subordinate sentences). It would not be possible to replicate Experiments 1 and 3 for English with the same experimental design, as they both rely on verb-final word order. On the other hand, Experiment 3 is a more solid replication of previous work on English (Traxler et al., 2002), and the results from Experiment 1 are supported by additional experimental evidence from Experiments 2 and 2b, which present the covert event as a probe rather than as embedded in a subordinate clause. This corroborates the assumption that the expectations for sentence-final events in long forms are analogous to expectations for covert events in logical metonymies (and that the same cognitive resources come into play when interpreting the covert event also for languages without verb-final word order).

The DM+ECU model can potentially be extended for different languages, as they do not require labeled data but only a large parsed corpus. Also, both functions used (*sum* and *product*) are symmetrical, so applying the distributional model to languages with different word orders would not significantly change the ECU model.

10.2. Lexical Meaning and World Knowledge

The main difference between theories of logical metonymy interpretation is the role they attribute to event knowledge and its position in the cognitive architecture (in the lexicon or as part of world knowledge), and a study of logical metonymy cannot help but touch the raw nerve of the distinction between lexical meaning and world knowledge (see for example discussions in Carston, 2002; Egg, 2005; Asher, 2011). Traditionally, linguistic (lexical) knowledge has been depicted as systematic and compositional, amenable to generalization, and as a more feasible object of analysis. Conversely, world knowledge has been considered to include situated, culture-dependent knowledge, which seems to elude a systematic characterization and analysis. Jackendoff (2002) observes that the motivation for such distinction may not be based on solid empirical ground but rather on "lurking fear that general-purpose knowledge and belief are a bottomless pit, and that in order to make the enterprise of semantics manageable it must somehow be restricted. And therefore some distinction must be made so we can stop before drowning in endless detail" (p. 283). Hobbs (2009) shares a similar suspect that "the most common argument in linguistics and related fields for drawing a strict boundary between lexicon and world is a kind of despair that a scientific study of world knowledge is possible" (p. 758). Work on generalized event knowledge (e.g. Ferretti et al., 2001; McRae et al., 2005; McRae and Matsuki, 2009; Bicknell et al., 2010; Matsuki et al., 2011) has had the merit of showing that it is indeed possible to make predictions and verify hypotheses regarding world knowledge and its role in linguistic processing.

From the architectural distinction (or from lack thereof) between lexicon and world knowledge, different predictions about processing follow. Theories of sentence comprehension that separate lexicon and world knowledge have usually predicted that the former is accessed immediately, whereas the latter is delayed (e.g. Katz, 1972; Chomsky, 1975; Fodor, 1983; Sperber and Wilson, 1986; Bornkessel and Schlesewsky, 2006; Warren and McConnell, 2007); other approaches have questioned the existence of a sharp distinction, as both lexical and world knowledge intervene and interact early in processing (e.g. Münte et al., 1998; Federmeier and Kutas, 1999; Hagoort et al., 2004; Nieuwland and van Berkum, 2006; McRae and Matsuki, 2009); others have proposed single-step models of language interpretation (without a priority of semantics over world knowledge, e.g. Bates and MacWhinney, 1989; MacDonald et al., 1994;

Trueswell et al., 1994) or a parallel architecture (e.g. Jackendoff, 2002), which is in principle compatible with a separation of the two domains while still accommodating for the experimental results which support an early interaction.

Note though that what a linguist or a computational linguist calls lexical knowledge may not overlap with what the psycholinguist or the cognitive scientist calls lexical knowledge (and the same goes for world knowledge). For example, in the Generative Lexicon (Pustejovsky, 1991, 1995), what to include in the lexicon seems to be determined by a gain in the generative / explanatory power obtained by enriching the lexicon rather than by experimental considerations regarding when this (richer) information kicks in during processing: Pustejovsky includes some world knowledge into the lexicon, by observing patterns and regularities and capturing them with systematic representational structures, ultimately in order to impose structure on a domain which was considered too elusive for rigorous analysis. This is even more apparent in the most recent revisions of the Generative Lexicon (Pustejovsky, 2012, 2013), where qualia structures transition into *habitats*: habitats are frames depicting generalizations about a situation which arise from world knowledge and on which compositional process can operate, reaching to a domain closer to affordances and to perceptual and motor capacities, which the cognitive scientist would hesitate to call lexical, if anything because it is relevant to other cognitive processes besides language.

10.3. The Richness of the Lexicon

The accessibility of world knowledge to non-linguistic cognitive processes (e.g. reasoning, planning, see Jackendoff, 2002) seems to be a stronger argument for distinguishing between conceptual knowledge (event knowledge) and lexical knowledge. In this respect, Elman (2009; 2011) considered three possible configurations regarding linguistic-specific knowledge and non-linguistic specific knowledge:

1. a parallel architecture (as proposed by Jackendoff, 2002), where only systematic information about words which can be generalized across classes of words (non idiosyncratic) is part of a lexical representation and where non-lexical representation should interact fully and bidirectionally with language-specific modules;

2. an information-rich lexicon, where all information that was shown to be exploited during processing (facilitating processing and generating expectations for upcoming input) should be included (generalized event knowledge about events and their participants, preference for subcategorization frames, thematic role filler information, see for example Langacker, 1987; Kamide et al., 2003; van Berkum et al., 2005);

3. an empty lexicon (Fodor and Lepore, 1998), stripped from all the information mentioned in (2) which has not traditionally been considered part of the lexicon.

Elman does not consider (1) a feasible option, due to the early interaction of allegedly lexical and non-lexical knowledge during processing, and he argues that the reason for separating the domains should be empirical (rather than meeting any requirement of "architectural tidiness"). Option (2) is presented as the most reasonable, but it does pose an interesting problem: if the lexicon is enriched with information which was traditionally considered to be world knowledge, then this information should be reduplicated outside the lexicon in order to be accessible also to non-linguistic cognitive processes, or rather should be placed outside the lexicon (option 3), which again is not desirable as we know this information strongly affects language processing. In Elman's words (2009; 2011, p. 568), "can we take the world out of language and put language in the world"? Elman's solution, a "lexical knowledge without a lexicon" does take information-rich lexical knowledge outside the lexicon (doing away with an old metaphor of the mental lexicon as a dictionary) and claims that words are cues to this information-rich knowledge, which can also be accessed by other (non-linguistic) cognitive processes.

Elman's solution arises from the need for a unified account of information-rich lexical knowledge, accessible both to linguistic and non-linguistic cognitive processes. However, an interesting criticism is voiced by the proponents of the Dual Coding Theory (DCT, Paivio, 2010; Paivio and Sadoski, 2011). Paivio and Sadoski acknowledge that the Words-as-cues framework has the merit of reconceptualizing lexical knowledge by putting strong emphasis on context-dependence and on event knowledge. As they argue, it is in fact a very compatible model to DCT: in DCT, lexical representations (*logogens*) are not meaningful in themselves, but are cues to meaning. On the other hand, they argue that we should not strive for a unified lexicon and that Elman's proposal does not account for multimodality (or at least it remains vague with regard

to this aspect) and thus is at odds with evidence for modality-specific lexical forms (e.g. visual, auditory, non-verbal, see Caramazza, 1997; Coltheart, 2004).

Addressing multimodality is thus a crucial challenge for a Words-as-cues approach to language understanding. Work on Grounded Cognition (Barsalou, 1999; see Barsalou, 2008 and Pecher and Zwaan, 2005 for a review) has also argued against amodal representations of knowledge and experience, showing evidence in support of the hypothesis that multimodal representations (acquired during sensorimotor experience) are reactivated in the form of simulations during cognitive processes (including language). The Language and Situated Simulation theory (LASS, Barsalou et al., 2008) argues that a linguistic system and a multimodal simulation system are both activated during language processing (the former slightly earlier than the latter). Interestingly, multimodality and grounded representations are also a recent challenge for distributional semantics, and some recent approaches have been proposed which exploit visual information extracted from images to build distributional and perceptually grounded models of word meaning (e.g. Feng and Lapata, 2010; Bergsma and Goebel, 2011; Bruni et al., 2012).

In conclusion, an information-rich lexicon (or, to use Elman's words, information-rich *lexical knowledge without a lexicon*) is necessary to account for people's predictive capabilities when processing language, both regarding explicit and implicit content (as for example for covert events). Also, it seems to be the case that language-specific aspects interact early and quickly with rich event knowledge, and an architecture different than that described by Elman (2009, 2011) might have to be assumed in order to implement concepts from formal semantics and linguistic theory (if they prove to be relevant for language processing, as semantic types for logical metonymy), and to account for modality-specific phenomena.

Part V.

Appendix

A. Stimuli for the Experiments

A.1. Stimuli for the Crowdsourcing Study

Each triplet was composed by three objects: an entity/event ambiguous noun (EN/EV), an entity noun (EN) and an event noun (EV). Each object was matched with two verbs (one event-selecting, or *begin-verb*, one not, or *spot-verb*).

1. **EN/EV** Walter enjoyed the translation on the premises of the company.
   Daniel approved the translation on the premises of the company.

   EN Charlie enjoyed the automobile on the premises of the company.
   Brian approved the automobile on the premises of the company.

   EV Keith enjoyed the conference on the premises of the company.
   Edward approved the conference on the premises of the company.

2. **EN/EV** Anne preferred the collection from the museum in the presence of her professor.
   Sarah discussed the collection from the museum in the presence of her professor.

   EN Martha preferred the instrument from the museum in the presence of her professor.
   Sophie discussed the instrument from the museum in the presence of her professor.

   EV Jane preferred the expedition from the museum in the presence of her professor.
   Helen discussed the expedition from the museum in the presence of her professor.

3. **EN/EV** James ended the conquest from the camp on the hill.
   Matt spotted the conquest from the camp on the hill.

   EN Steve ended the magazine from the camp on the hill.
   Tim spotted the magazine from the camp on the hill.

   EV Scott ended the ceremony from the camp on the hill.
   Nick spotted the ceremony from the camp on the hill.

4. **EN/EV** Robert started the blessing on the campus of the university.
   Thomas disdained the blessing on the campus of the university.

   EN Richard started the portrait on the campus of the university.
   Bruce disdained the portrait on the campus of the university.

EV Joseph started the semester on the campus of the university.
Jason disdained the semester on the campus of the university.

5. **EN/EV** Jack savored the praise with his wife over breakfast.
Luke considered the praise with his wife over breakfast.

   EN John savored the butter with his wife over breakfast.
Frank considered the butter with his wife over breakfast.

   EV Paul savored the debate with his wife over breakfast.
Greg considered the debate with his wife over breakfast.

6. **EN/EV** Rose began the breakfast on the patio after her long sickness.
Linda organized the breakfast on the patio after her long sickness.

   EN Kate began the newspaper on the patio after her long sickness.
Susan organized the newspaper on the patio after her long sickness.

   EV Mary began the afternoon on the patio after her long sickness.
Lisa organized the afternoon on the patio after her long sickness.

7. **EN/EV** Bernard finished the harvest for the autumn with his family.
Martin prepared the harvest for the autumn with his family.

   EN Michael finished the package for the autumn with his family.
Albert prepared the package for the autumn with his family.

   EV Andrew finished the holiday for the autumn with his family.
Philip prepared the holiday for the autumn with his family.

8. **EN/EV** Bart tried the bath in the park by the waterfall.
Colin recalled the bath in the park by the waterfall.

   EN David tried the tent in the park by the waterfall.
Peter recalled the tent in the park by the waterfall.

   EV George tried the swim in the park by the waterfall.
Chris recalled the swim in the park by the waterfall.

9. **EN/EV** Claire continued the dinner with the team during the training session.
Nancy reviewed the dinner with the team during the training session.

   EN Louise continued the letter with the team during the training session.
Karen reviewed the letter with the team during the training session.

   EV Emily continued the season with the team during the training session.
Laura reviewed the season with the team during the training session.

10. **EN/EV** Bill endured the shower on the island at the sunset.
Tom remembered the shower on the island at the sunset.

    EN Ben endured the brandy on the island at the sunset.
Sam remembered the brandy on the island at the sunset.

    EV Mark endured the revolt on the island at the sunset.
Alex remembered the revolt on the island at the sunset.

A.2. Stimuli for the Psycholinguistic Experiments

A.2.1. Stimuli for Experiment 1

The first agent in each sentence was matched with the first event in the high-typicality condition, and with the second one in the low-typicality condition. The second agent was matched with the first event in the low-typicality condition, and with the second one in the high-typicality condition.

1. Der Chauffeur / der Mechaniker vermied es, das Auto zu fahren / zu reparieren, weil er sehr müde war.
2. Der Bäcker / die Bäuerin fing an, die Äpfel zu schälen / zu pflücken, nachdem er / sie den Hund gefüttert hatte.
3. Der Bergsteiger / der Künstler versuchte, den Berg zu erklimmen / zu malen, aber es war schon zu dunkel.
4. Der Braumeister / der Student fing an, das Bier zu brauen / zu trinken, und goss ein bisschen auf seine Hand.
5. Der Kunstsammler / der Zeichner probierte, das Bild zu kaufen / malen, aber er hatte nicht genug Geld.
6. Der Dieb / der Juwelier genoss es, den Diamanten zu schmuggeln / zu schleifen, weil er so edel war.
7. Der Handwerker / die Hausfrau ertrug es, das Fenster einzubauen / zu putzen, obwohl er / sie keine Lust hatte.
8. Der Journalist / der Regisseur genoss es, den Film zu drehen / zu kritisieren, weil es eine sehr interessante Geschichte war.
9. Das Geburtstagskind / die Verkäuferin fing an, das Geschenk auszupacken / einzupacken, bevor es / sie mit seiner / ihrer Arbeit fertig war.
10. Der Autor / der Schüler begann, die Geschichte zu schreiben / zu lernen, nachdem er mit der Übersetzung fertig war.
11. Das Kind / der Konditor hörte auf, die Glasur zu essen / aufzutragen, und fing mit den Pralinen an.
12. Der Maurer / die Maklerin versuchte, das Haus zu bauen / zu verkaufen, aber das Grundstück war viel zu teuer.
13. Der Abiturient / die Lehrerin hasste es, die Klausur zu schreiben / zu benoten, weil er / sie lieber auf der Party gewesen wäre.
14. Der Pianist / der Transporteur probierte, das Klavier zu spielen / zu transportieren, aber seine Hände taten weh.

15. Das Kind / die Kellnerin verschob es, den Nachtisch zu essen / zu servieren, bis die Mutter mit dem Käse fertig war.
16. Der Koch / der Pizzabote hasste es, die Pizza zu backen / zu liefern, weil es so warm war.
17. Das Baby / der Ober hörte auf, den Saft zu trinken / einzugießen, weil er übergelaufen war.
18. Der Gast / der Metzger begann, das Schwein zu essen / zu schlachten, nachdem er mit dem Huhn fertig war.
19. Der Möbelpacker / die Putzfrau ertrug es, das Sofa zu tragen / abzusaugen, obwohl er sie sehr müde war.
20. Der Informatiker / der Junge verschob es, das Videospiel zu programmieren / spielen, bis der neue Computer angekommen war.
21. Der Professor / die Studentin hörte auf, die Vorlesung vorzubereiten / zu besuchen, weil er / sie zu beschäftigt war.
22. Der Bauarbeiter / der Maler hasste es, die Wand zu einreißen / zu streichen, weil sein Gehalt nicht hoch genug war.
23. Der Patient / der Redakteur vermied es, die Zeitschrift durchzublättern / zu schreiben, weil er schon ein Buch zu lesen hatte.
24. Der Verleger / der Zeitungsjunge probierte, die Zeitung zu drucken / zu verteilen, aber er war krank und konnte nicht arbeiten

A.2.2. Stimuli for Experiment 2

The first agent in each sentence was matched with the first probe in the high-typicality condition, and with the second one in the low-typicality condition. The second agent was matched with the first probe in the low-typicality condition, and with the second one in the high-typicality condition.

1. Der Chauffeur / der Mechaniker fing mit dem Auto an. (FAHREN / REPARIEREN)
2. Der Bäcker / die Bäuerin fing mit den Äpfeln an. (SCHÄLEN / PFLÜCKEN)
3. Der Bergsteiger / der Künstler versuchte es mit dem Berg. (ERKLIMMEN / MALEN)
4. Der Braumeister / der Student vermied das Bier. (BRAUEN / TRINKEN)
5. Der Kunstsammler / der Zeichner probierte das Bild. (KAUFEN / MALEN)
6. Der Dieb / der Juwelier begann mit den Diamanten. (SCHMUGGELN / SCHLEIFEN)
7. Der Handwerker / die Hausfrau probierte es mit dem Fenster. (EINBAUEN / PUTZEN)

8. Der Journalist / der Regisseur genoss den Film. (KRITISIEREN / DREHEN)
9. Das Geburtstagskind / die Verkäuferin fing mit dem Geschenk an. (AUSPACKEN / EINPACKEN)
10. Der Autor / der Schüler begann mit der Geschichte. (SCHREIBEN / LERNEN)
11. Das Kind / der Konditor hörte mit der Glasur auf. (ESSEN / AUFTRAGEN)
12. Der Maurer / die Maklerin versuchte es mit dem Haus. (BAUEN / VERKAUFEN)
13. Der Abiturient / die Lehrerin hasste die Klausur. (SCHREIBEN / BENOTEN)
14. Der Pianist / der Transporteur probierte es mit dem Klavier. (SPIELEN / TRANSPORTIEREN)
15. Das Kind / die Kellnerin verschob den Nachtisch. (ESSEN / SERVIEREN)
16. Der Koch / der Pizzabote ertrug die Pizza. (BACKEN / LIEFERN)
17. Das Baby / der Ober hörte mit dem Saft auf. (TRINKEN / EINGIESSEN)
18. Der Gast / der Metzger begann mit dem Schwein. (ESSEN / SCHLACHTEN)
19. Der Möbelpacker / die Putzfrau versuchte es mit dem Sofa. (TRAGEN / ABSAUGEN)
20. Der Informatiker / die Junge hasste das Videospiel. (PROGRAMMIEREN / SPIELEN)
21. Der Professor / die Studentin hasste die Vorlesung. (VORBEREITEN / BESUCHEN)
22. Der Bauarbeiter / der Maler verschob die Wand. (EINREISSEN / STREICHEN)
23. Der Patient / der Redakteur ertrug die Zeitschrift. (DURCHBLÄTTERN / SCHREIBEN)
24. Der Verleger / der Zeitungsjunge hörte mit der Zeitung auf. (DRUCKEN / VERTEILEN)

A.2.3. Stimuli for Experiment 2b

The first object in each sentence was matched with the first probe in the high-typicality condition, and with the second one in the low-typicality condition. The second agent was matched with the first probe in the low-typicality condition, and with the second one in the high-typicality condition. Only the modified sentences are reported, the others remained unchanged from Experiment 2 (see A.2.2).

1. Der Chauffeur / der Mechaniker fing mit dem Auto an. (RASEN / REPARIEREN)
3. Der Bergsteiger / der Künstler vertagte den Berg. (ERKLIMMEN / SKIZZIEREN)
4. Der Braumeister / der Student fing mit dem Bier an. (BRAUEN / TRINKEN)

5. Der Kunstsammler / der Zeichner vertagte das Bild. (KAUFEN / ENTWERFEN)

7. Der Handwerker / die Hausfrau vertagte das Fenster. (EINBAUEN / REINIGEN)

8. Der Journalist / der Regisseur hörte mit dem Film auf. (KRITISIEREN / DREHEN)

10. Der Autor / der Schüler begann mit der Geschichte. (LESEN / AUSDENKEN)

12. Der Maurer / die Maklerin vertagte das Haus. (VERPUTZEN / VERMITTELN)

13. Der Abiturient / die Lehrerin hörte mit der Klausur auf. (SCHREIBEN / BENOTEN)

14. Der Pianist probierte / der Transporteur machte mit dem Klavier weiter. (SPIELEN / TRAGEN)

15. Das Kind / die Kellnerin machte mit dem Nachtisch weiter. (PROBIEREN / BRINGEN)

16. Der Koch / der Pizzabote machte mit der Pizza weiter. (BACKEN / AUSFAHREN)

18. Der Gast / der Metzger begann mit dem Schwein. (ESSEN / ZERLEGEN)

19. Der Möbelpacker / die Putzfrau machte mit dem Sofa weiter. (ABTRANSPORTIEREN / ABSAUGEN)

20. Der Informatiker / die Junge begann mit dem Videospiel. (ENTWERFEN / SPIELEN)

21. Der Professor / die Studentin vertagte die Vorlesung. (VORBEREITEN / BESUCHEN)

22. Der Bauarbeiter / der Maler machte mit der Wand weiter. (EINREISSEN / VERZIEREN)

23. Der Patient / der Redakteur fing mit der Zeitschrift an. (DURCHBLÄTTERN / SCHREIBEN)

A.2.4. Stimuli for Experiment 3

Each agent was matched with two high thematic fit objects (one entity-denoting, one event-denoting object) and with two low thematic fit objects. The high thematic fit objects for each odd-numbered item were used as low thematic fit objects for the following even-numbered item, and the high thematic fit objects for each even-numbered item were used as low thematic fit objects for the preceding even-numbered item.

1. [high] Das Kind hat mit dem Spielzeug / mit der Schlägerei ohne zu klagen aufgehört, weil es sehr müde war.
   [low] Das Kind hat mit dem Medikament / mit der Therapie ohne zu klagen aufgehört, weil es sehr müde war.

2. [high] Der Patient hat mit dem Medikament / mit der Therapie ohne zu klagen aufgehört, weil er Kopfschmerzen hatte.
   [low] Der Patient hat mit dem Spielzeug / mit der Schlägerei ohne zu klagen aufgehört, weil er Kopfschmerzen hatte.

3. [high] Der Autor hat das Buch / die Buchvorstellung endlich begonnen, um nicht in Verzug zu geraten.
   [low] Der Autor hat das Bier / die Säuberung endlich begonnen, um nicht in Verzug zu geraten.

4. [high] Der Braumeister hat mit dem Bier / mit der Säuberung endlich angefangen. nachdem er mit seinem Chef gesprochen hatte.
   [low] Der Braumeister hat mit dem Buch / mit der Buchvorstellung endlich angefangen, nachdem er mit seinem Chef gesprochen hatte.

5. [high] Der Metzger hat die Würste / die Wurstherstellung vorsichtig begonnen, nachdem er den Hund gefüttert hatte.
   [low] Der Metzger hat den Schrank / den Umzug vorsichtig begonnen, nachdem er den Hund gefüttert hatte.

6. [high] Der Möbelpacker hat mit dem Schrank / mit dem Umzug vorsichtig weitergemacht, nachdem er eine Zigarette geraucht hatte.
   [low] Der Möbelpacker hat mit den Würsten / mit der Wurstherstellung vorsichtig weitergemacht, nachdem er eine Zigarette geraucht hatte.

7. [high] Der Regisseur hat mit dem Drehbuch / mit dem Casting pünktlich aufgehört, um seinen Zug nicht zu verpassen.
   [low] Der Regisseur hat mit dem Aufsatz / mit dem Studium pünktlich aufgehört, um seinen Zug nicht zu verpassen.

8. [high] Der Student hat mit dem Aufsatz / mit dem Studium pünktlich angefangen, um vor seinem Geburtstag fertig zu sein.
   [low] Der Student hat mit dem Drehbuch / mit dem Casting pünktlich angefangen, um vor seinem Geburtstag fertig zu sein.

9. [high] Das Geburtstagskind hat mit den Geschenken / mit der Feier sofort angefangen, obwohl seine Mutter nicht da war.
   [low] Das Geburtstagskind hat mit der Suppe / mit der Schicht sofort angefangen, obwohl seine Mutter nicht da war.

10. [high] Die Kellnerin hat mit der Suppe / mit der Schicht sofort angefangen, obwohl sie keine Lust hatte.
    [low] Die Kellnerin hat mit den Geschenken / mit der Feier sofort angefangen, obwohl sie keine Lust hatte.

11. [high] Der Informatiker hat den Code / die Fehlersuche umgehend begonnen, nachdem er eine große Tasse Kaffee getrunken hatte.

[low] Der Informatiker hat den Motor / die Reparatur umgehend begonnen, nachdem er eine große Tasse Kaffee getrunken hatte.

12. [high] Der Mechaniker hat mit dem Motor / mit der Reparatur umgehend aufgehört, weil er mit anderen Dingen zu beschäftigt war.
[low] Der Mechaniker hat mit dem Code / mit der Fehlersuche umgehend aufgehört, weil er mit anderen Dingen zu beschäftigt war.

13. [high] Der Journalist hat mit dem Artikel / mit der Recherche ohne Überzeugung weitergemacht, weil es schon sehr spät war.
[low] Der Journalist hat mit dem Kuchen / mit dem Verkauf ohne Überzeugung weitergemacht, weil es schon sehr spät war.

14. [high] Der Konditor hat den Kuchen / den Verkauf ohne Überzeugung vertagt, weil er zuerst die Weihnachtsplätzchen backen wollte.
[low] Der Konditor hat den Artikel / die Recherche ohne Überzeugung vertagt, weil er zuerst die Weihnachtsplätzchen backen wollte.

15. [high] Der Professor hat mit dem Beispiel / mit der Vorlesung ohne zu zögern weitergemacht, weil er es eilig hatte.
[low] Der Professor hat mit dem Haus / mit dem Verkauf ohne zu zögern weitergemacht, weil er es eilig hatte.

16. [high] Die Maklerin hat das Haus / den Verkauf ohne zu zögern vertagt, weil es einfach nicht der richtige Moment war.
[low] Die Maklerin hat das Beispiel / die Vorlesung ohne zu zögern vertagt, weil es einfach nicht der richtige Moment war.

17. [high] Der Gast hat den Kuchen/ das Gespräch mit Freude begonnen, weil er sich schon lange darauf gefreut hatte.
[low] Der Gast hat die Mauer / den Aufbau mit Freude begonnen, weil er sich schon lange darauf gefreut hatte.

18. [high] Der Bauarbeiter hat mit der Mauer / mit dem Aufbau mit Freude weitergemacht, weil er einen guten Tag hatte.
[low] Der Bauarbeiter hat mit dem Kuchen / mit dem Gespräch mit Freude weitergemacht, weil er einen guten Tag hatte.

19. [high] Der Redakteur hat den Artikel / die Besprechung aus gutem Grund begonnen, obwohl er nicht wirklich daran interessiert war.
[low] Der Redakteur hat das Pilzragout / den Frühjahrsputz aus gutem Grund begonnen, obwohl er nicht wirklich daran interessiert war.

20. [high] Die Hausfrau hat das Pilzragout / den Frühjahrsputz aus gutem Grund vertagt, sie war nämlich krank.
[low] Die Hausfrau hat den Artikel / die Besprechung aus gutem Grund vertagt, sie war nämlich krank.

Bibliography

Abe, N. and Li, H. (1996). Learning word association norms using tree cut pair models. arXiv preprint cmp-lg/9605029.

Alexiadou, A. (2001). *Functional structure in nominals: nominalization and ergativity*. John Benjamins Publishing, Amsterdam, The Netherlands.

Altmann, G. (1999). Thematic role assignment in context. *Journal of Memory and Language*, 41:124–145.

Altmann, G. and Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, 73:247–264.

Apresjan, J. D. (1974). Regular polysemy. *Linguistics*, 142:5–32.

Asher, N. (2011). *Lexical meaning in context: A web of words*. Cambridge University Press, Cambridge, UK.

Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59:390–412.

Baayen, R. H. and Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, 3(2):12–28.

Baayen, R. H., Piepenbrock, R., and van H, R. (1993). *The CELEX lexical database on CD-ROM*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Baggio, G., Choma, T., van Lambalgen, M., and Hagoort, P. (2010). Coercion and compositionality. *Journal of Cognitive Neuroscience*, 22(9):2131–2140.

Baggio, G., van Lambalgen, M., and Hagoort, P. (2012). The processing consequences of compositionality. In Werning, M., Hinzen, W., and Machery, E., editors, *The Oxford Handbook of Compositionality*, pages 657–674. Oxford University Press, Oxford, UK.

Bambini, V. and Resta, D. (2012). Metaphor and experimental pragmatics: when theory meets empirical investigation. *Humana Mente*, 23:37–60.

Baroni, M., Bernardi, R., and Zamparelli, R. (2012). Frege in space: a program for compositional distributional semantics. *Linguistic Issues in Language Technologies*. To appear.

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43:209–226.

Baroni, M. and Lenci, A. (2010). Distributional Memory: a general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

Baroni, M. and Zamparelli, R. (2010). Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference of Empirical Methods in Natural Language Processing*, pages 1183–1193, MIT, Massachussetts.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22:577–660.

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, 59:617–645.

Barsalou, L. W., Santos, A., and Simmons, W. K. (2008). Language and simulation in conceptual processing. In Vega, M. D., Glenberg, A. M., and Graesser, A. C., editors, *Symbols, Embodiment, and Meaning*, pages 245–283. Oxford University Press, Oxford, UK.

Bates, E. and MacWhinney, B. (1989). Functionalism and the competition model. In Bates, E. and MacWhinney, B., editors, *The cross-linguistic study of sentence processing*, pages 3–73. Cambridge University Press, Cambridge, UK.

Becker, C. A. (1980). Semantic context effects in visual word recognition: an analysis of semantic strategies. *Memory & Cognition*, 8(6):493–512.

Bedny, M., Dravida, S., and Saxe, R. (2013). Shindigs, brunches, and rodeos: The neural basis of event words. *Cognitive, Affective, & Behavioral Neuroscience*, pages 1–11.

Bergsma, S. and Goebel, R. (2011). Using visual information to predict lexical preference. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 399–405, Hissar, Bulgaria.

Bicknell, K., Elman, J. L., Hare, M., McRae, K., and Kutas, M. (2010). Effects of event knowledge in processing verbal arguments. *Journal of Memory and Language*, 63:489–505.

Biederman, I., Blickle, T. W., Teitelbaum, R. C., and Klatsky, G. J. (1988). Object search in nonscene displays. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14:456–467.

Blutner, R. (2002). Lexical semantics and pragmatics. *Linguistische Berichte*, 10:27–58.

Bohnet, B. (2010). Very high accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 89–97, Beijing, China.

Boland, J. E. (2004). Linking eye movements to sentence comprehension in reading and listening. In Carreiras, M. and Clifton, C. E., editors, *The on-line study of sentence comprehension: Eyetracking, ERP, and beyond*, chapter 4, pages 51–76. Psychology Press, New York, NY.

Bornkessel, I. and Schlesewsky, M. (2006). The extended argument dependency model: a neurocognitive approach to sentence comprehension across languages. *Psychological Review*, 113(4):787–821.

Briscoe, T., Copestake, A., and Boguraev, B. (1990). Enjoy the paper: Lexical semantics via lexicology. In *Proceedings of the 13th International Conference on Computational Linguistics*, pages 42–47, Helsinki, Finland.

Brockmann, C. and Lapata, M. (2003). Evaluating and combining approaches to selectional preference acquisition. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, pages 27–34, Budapest, Hungary.

Bruni, E., Boleda, G., Baroni, M., and Tran, N.-K. (2012). Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 136–145, Jeju Island, South Korea.

Buitelaar, P., Cimiano, P., and Magnini, B. (2005). *Ontology learning from text: methods, evaluation and applications*. IOS press, Amsterdam, The Netherlands.

Burnard, L. (1995). *Users reference guide for the British National Corpus*. Oxford University Computing Services, Oxford, UK.

Busa, F., Calzolari, N., and Lenci, A. (2001). Generative Lexicon and the SIMPLE Model: Developing Semantic Resources for NLP. In *The language of word meaning*, pages 333–349. Cambridge University Press, Cambridge, UK.

Callison-Burch, C. (2009). Fast, cheap, and creative: evaluating translation quality using Amazon's Mechanical Turk. In *Proceedings of the 14th Conference of Empirical Methods in Natural Language Processing*, pages 286–295, Singapore.

Caramazza, A. (1997). How many levels of processing are there in lexical access? *Cognitive Neuropsychology*, 14(1):177–208.

Carreiras, M. and Clifton, C. E. (2004). *The on-line study of sentence comprehension: Eyetracking, ERP and beyond*. Psychology Press, New York, NY.

Carston, R. (2002). *Thoughts and utterances*. Blackwell, Oxford, UK.

Casati, R. and Varzi, A. (2010). Events. *The Stanford Encyclopedia of Philosophy*. `http://plato.stanford.edu/archives/spr2010/entries/events/`.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA.

Chomsky, N. (1975). *Reflections on Language*. Pantheon, New York, NY.

Cimiano, P. and Wenderoth, J. (2007). Automatic acquisition of ranked qualia structures from the web. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 888–895, Prague, Czech Republic.

Clark, H. H. (1973). The Language-as-Fixed-Effect Fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, 12:335–359.

Clark, S. and Weir, D. (2001). Class-based probability estimation using a semantic hierarchy. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 1–8, Pittsburgh, PA.

Coleman, E. B. (1964). Generalizing to a language population. *Psychological Reports*, 14:219–226.

Collina, S., Marangolo, P., and Tabossi, P. (2001). The role of argument structure in the production of nouns and verbs. *Neuropsychologia*, 39:1125–1137.

Coltheart, M. (2004). Are there lexicons? *The Quarterly Journal of Experimental Psychology*, 57A:1153–1171.

Coulson, S. and Van Petten, C. (2002). Conceptual integration and metaphor: An event-related potential study. *Memory & Cognition*, 30(6):958–968.

Crocker, M. W. (2010). Computational Psycholinguistics. In Clark, A., Fox, C., and Lappin, S., editors, *The Handbook of Computational Linguistics and Natural Language Processing*. Wiley Blackwell, London, UK.

Dagan, I., Lee, L., and Pereira, F. C. (1999). Similarity-based models of word cooccurrence probabilities. *Machine Learning*, 34:43–69.

de Almeida, R. G. (2004). The effect of context on the processing of type-shifting verbs. *Brain and Language*, 90:249–261.

de Almeida, R. G. and Dwivedi, V. D. (2008). Coercion without lexical decomposition: Type-shifting effects revisited. *The Canadian Journal of Linguistics / La revue Canadienne de Linguistique*, 53(2/3):301–326.

de Almeida, R. G., Riven, L., Manouilidou, C., Lungu, O., Dwivedi, V., Jarema, G., and B., G. (2009). Coercion effects are pragmatic: fMRI and behavioral evidence. Poster presented at the 15th Conference on Architectures and Mechanisms for Language Processing, Barcelona, Spain.

De Grauwe, S., Swain, A., Holcomb, P. J., Ditman, T., and Kuperberg, G. R. (2010). Electrophysiological insights into the processing of nominal metaphors. *Neuropsychologia*, 48(7):1965–1984.

de Saussure, F. (1915). Cour de Linguistique Générale (Course in General Linguistics), trans. Wade Baskin, 1959 edition.

Delogu, F., Drenhaus, H., and Crocker, M. (2013). Teasing apart coercion and surprisal: Evidence from ERPs and eye-movements. Poster presented at the 26th CUNY Conference.

Dijkstra, A. and de Smedt, K. (1996). Computer models in psycholinguistics: An introduction. In Dijkstra, A. and de Smedt, K., editors, *Computational psycholinguistics: AI and connectionist models of human language processing*, pages 3–23. Taylor & Francis, London, UK.

Egg, M. (2005). *Flexible semantics for reinterpretation phenomena*. CSLI Publications, Stanford, CA.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14:179–211.

Elman, J. L. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive Science*, 33(4):547–582.

Elman, J. L. (2011). Lexical knowledge without a lexicon? *The Mental Lexicon*, 6(1):1–33.

Elman, J. L., Hare, M., and McRae, K. (2005). Cues, constraints, and competition in sentence processing. In Tomasello, M. and Slobin, D., editors, *Beyond nature-nurture: Essays in honor of Elizabeth Bates*, pages 111–138. Lawrence Erlbaum Associates, Mahwah, NJ.

Erk, K. (2007). A simple, similarity-based model for selectional preferences. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 216–223, Prague, Czech Republic.

Erk, K. (2010). What is word meaning, really? (and how can distributional models help us describe it?). In *Proceedings of the 2010 Workshop on Geometrical Models of Natural Language Semantics*, pages 17–26, Uppsala, Sweden.

Erk, K. (2012). Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, 6(10):635–653.

Erk, K. (2013). Towards a semantics for distributional representations. In *Proceedings of the IWCS Workshop "Towards A Formal Distributional Semantics"*, Potsdam, Germany.

Erk, K. and Padó, S. (2008). A structured vector space model for word meaning in context. In *Proceedings of the 2008 Conference of Empirical Methods in Natural Language Processing*, pages 897–906,, Honolulu, Hawaii.

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

Evert, S. (2005). *The statistics of word cooccurrences*. PhD thesis, Universität Stuttgart.

Faaß, G. and Eckart, K. (2013). SdeWaC – A Corpus of Parsable Sentences from the Web. In Gurevych, I., Biemann, C., and Zesch, T., editors, *Language Processing and Knowledge in the Web*, volume 8105 of *Lecture Notes in Computer Science*, pages 61–68. Springer Berlin Heidelberg.

Fabrigar, L. R., Krosnick, J. A., and MacDougall, B. L. (2005). Attitude measurement: Techniques for measuring the unobservable. In Brock, T. and Green, M., editors, *Persuasion: Psychological insights and perspectives*, pages 17–40. Sage, Thousand Oaks, CA.

Farmer, T. A., Christiansen, M. H., and Monaghan, P. (2006). Phonological typicality influences on-line sentence comprehension. *Proceedings of the National Academy of Sciences of the United States of America*, 103(32):12203–12208.

Federmeier, K. D. and Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language*, 41(4):469–495.

Fellbaum, C. (1998). *WordNet. An electronic lexical database*. MIT Press, Cambridge, MA.

Feng, Y. and Lapata, M. (2010). Visual information in semantic representation. In *Proceedings of the 11th Meeting of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 91–99, Los Angeles, CA.

Ferretti, T. R., Kutas, M., and McRae, K. (2007). Verb aspect and the activation of event knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(1):182–196.

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

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. In *Studies in Linguistic Analysis*, pages 1–32. Blackwell, Oxford, UK.

Fodor, J. (1983). *The Modularity of Mind*. MIT Press, Cambridge, MA.

Fodor, J. A. (1990). *A theory of content and other essays*. MIT Press, Cambridge, MA.

Fodor, J. A. and Lepore, E. (1998). The emptiness of the lexicon: reflections on James Pustejovsky's The Generative Lexicon. *Linguistic Inquiry*, 29(2):269–288.

Fort, K., Adda, G., and Cohen, K. B. (2011). Amazon Mechanical Turk: Gold mine or coal mine? *Computational Linguistics*, 37(2):413–420.

Frazier, L. and Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, 14(2):178–210.

Frege, G. (1892). Über Sinn und Bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, 100:25–50.

Frisson, S. and McElree, B. (2008). Complement coercion is not modulated by competition: Evidence from eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(1):1–11.

Frisson, S. and Pickering, M. J. (1999). The processing of metonymy: Evidence from eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25(6):1366–1383.

Frisson, S. and Pickering, M. J. (2007). The processing of familiar and novel senses of a word: Why reading Dickens is easy but reading Needham can be hard. *Language and Cognitive Processes*, 22(4):595–613.

Frisson, S., Rayner, K., and Pickering, M. J. (2005). Effects of contextual predictability and transitional probability on eye movements during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(5):862–877.

Garbin, G., Collina, S., and Tabossi, P. (2012). Argument structure and morphological factors in noun and verb processing: An fMRI study. *PloS ONE*, 7(9):e45091.

Gernsbacher, M. A. and Faust, M. E. (1991). The mechanism of suppression: A component of general comprehension skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 17(2):245–262.

Gibbs, R. W. (1994). *The poetics of mind: Figurative thought, language, and understanding*. Cambridge University Press, Cambridge, UK.

Gibbs, R. W. and Gerrig, R. J. (1989). How context makes metaphor comprehension seem 'special'. *Metaphor and Symbol*, 4(3):145–158.

Giora, R. (2003). *On our mind: Salience, context, and figurative language*. Oxford University Press, New York, NY.

Grefenstette, G. (1994). *Explorations in automatic thesaurus discovery*. Kluwer Academic Publishers, Norwell, MA.

Grice, H. P. (1975). Logic and conversation. In Cole, P. and Morgan, J. L., editors, *Syntax and semantics 3: Speech Acts*, pages 41–58. Academic Press, New York, NY.

Grimshaw, J. (1990). *Argument Structure*. MIT Press, Cambridge, MA.

Guevara, E. (2011). Computing semantic compositionality in distributional semantics. In *Proceedings of the 9th International Conference on Computational Semantics*, pages 135–144, Oxford, UK.

Hagoort, P., Hald, L., Bastiaansen, M., and Petersson, K. M. (2004). Integration of word meaning and world knowledge in language comprehension. *Science*, 304:438–441.

Hahn, U. and Markert, K. (1997). In support of the equal rights movement for literal and figurative language: A parallel search and preferential choice model. In *Proceedings of the 19th Annual Conference of the Cognitive Science Society*, pages 1215–1220, Palo Alto, CA.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 1–8, Pittsburgh, PA.

Hampton, J. A. (1991). The combination of prototype concepts. In Schwanenflugel, P., editor, *The Psychology of Word Meanings*, pages 91–116. Lawrence Erlbaum Associates, Hillsdale, NJ.

Hanks, P. (2007). Preference syntagmatics. In Ahmad, K., Brewster, C., and Stevenson, M., editors, *Words and Intelligence II. Essays in Honor of Yorick Wilks*, pages 119–135. Springer, Berlin, Germany.

Hare, M., Elman, J. L., Tabaczynski, T., and McRae, K. (2009a). The wind chilled the spectators, but the wine just chilled: Sense, structure, and sentence comprehension. *Cognitive Science*, 33(4):610–628.

Hare, M., Jones, M., Thomson, C., Kelly, S., and McRae, K. (2009b). Activating event knowledge. *Cognition*, 111(2):151–167.

Hare, M., McRae, K., and Elman, J. L. (2003). Sense and structure: Meaning as a determinant of verb subcategorization preferences. *Journal of Memory and Language*, 48(2):281–303.

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

Herdağdelen, A., Erk, K., and Baroni, M. (2009). Measuring semantic relatedness with vector space models and random walks. In *Proceedings of the ACL-IJCNLP Workshop on Graph-based Methods for Natural Language Processing*, pages 50–53, Singapore.

Hobbs, J. R. (2009). Word meaning and world knowledge. In Maienborn, C., von Heusinger, K., Portner, P., and van Leusen, N., editors, *Semantics: An International Handbook of Natural Language Meaning*, pages 740–761. Mouton de Gruyte, The Hague.

Hodges, W. (2012). Formalizing the relationship between meaning and syntax. In Werning, M., Hinzen, W., and Machery, E., editors, *The Oxford Handbook of Compositionality*, pages 245–261. Oxford University Press, Oxford, UK.

Hohenstein, S. and Kliegl, R. (2013). Eye movements reveal interplay between noun capitalization and word class during reading. In *Proceedings of the 35th Annual Meeting of the Cognitive Science Society*, pages 2554–2559, Berlin, Germany.

Jackendoff, R. (1997). *The architecture of the language faculty*. MIT Press, Cambridge, MA.

Jackendoff, R. (2002). *Foundations of language: Brain, meaning, grammar, evolution*. Oxford University Press, Oxford, UK.

Janus, R. A. and Bever, T. G. (1985). Processing of metaphoric language: An investigation of the three-stage model of metaphor comprehension. *Journal of Psycholinguistic Research*, 14(5):473–487.

Johansson, S., Leech, G. N., and Goodluck, H. (1978). Manual of information to accompany the Lancaster-Oslo/Bergen Corpus of British English, for use with digital computer. Unpublished manuscript.

Just, M., Carpenter, P., and Woolley, J. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, 111(2):228–238.

Just, M. A. and Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87:329–354.

Kamide, Y., Altmann, G., and Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*, 49:133–156.

Kamp, H. and Partee, B. (1995). Prototype theory and compositionality. *Cognition*, 57:129–191.

Katsika, A., Braze, D., Deo, A., and Piñango, M. (2012). Complement coercion: Distinguishing between type-shifting and pragmatic inferencing. *The Mental Lexicon*, 7(1):58–76.

Katz, J. J. (1972). *Semantic theory*. Harper & Row, New York, NY.

Katz, J. J. and Fodor, J. A. (1963). The structure of a semantic theory. *Language*, 39(2):170–210.

Keller, F., Gunasekharan, S., Mayo, N., and Corley, M. (2009). Timing accuracy of web experiments: A case study using the webexp software package. *Behavior Research Methods*, 41(1):1–12.

Keller, F. and Lapata, M. (2003). Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics*, 29(3):459–484.

Kerouac, J. (1957). *On the Road*. Viking Press, New York, NY.

Krippendorff, K. (1980). *Content Analysis*. Sage Publications, Beverly Hills, CA.

Kučera, H. and Francis, W. N. (1967). *Computational analysis of present-day American English*. Brown University Press, Providence, RI.

Kunze, C. and Lemnitzer, L. (2002). GermaNet – representation, visualization, application. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, pages 1485–1491, Las Palmas , Spain.

Kuperberg, G. R., Choi, A., Cohn, N., Paczynski, M., and Jackendoff, R. (2010). Electrophysiological correlates of complement coercion. *Journal of Cognitive Neuroscience*, 22(12):2685–2701.

Kutas, M. and Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences*, 4(12):463–470.

Kutas, M. and Van Petten, C. (1994). Psycholinguistics electrified. In Gernsbacher, M. A., editor, *Handbook of Psycholinguistics*, pages 83–143. Academic Press, San Diego, CA.

Landauer, T. and Dumais, S. (1997). A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104(2):211–240.

Landauer, T. K., Foltz, P. W., and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, 25(2-3):259–284.

Landauer, T. K., McNamara, D. S., Dennis, S. E., and Kintsch, W. E. (2007). *Handbook of Latent Semantic Analysis*. Lawrence Erlbaum Associates, Mahwah, NJ.

Langacker, R. W. (1987). *Foundations of Cognitive Grammar*, volume 1. Stanford University Press, Stanford, CA.

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

Lapata, M. and Lascarides, A. (2003). A probabilistic account of logical metonymy. *Computational Linguistics*, 29(2):261–315.

Lapesa, G. and Evert, S. (2013). Evaluating neighbor rank and distance measures as predictors of semantic priming. In *Proceedings of the 4th Workshop on Cognitive Modeling and Computational Linguistics*, pages 66–74, Sofia, Bulgaria.

Lascarides, A. and Copestake, A. (1998). Pragmatics and word meaning. *Journal of Linguistics*, 34:387–414.

Lee, L. (1999). Measures of distributional similarity. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 25–32, College Park, MD.

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Italian Journal of Linguistics*, 20(1):1–31.

Lenci, A. (2011). Composing and updating verb argument expectations: A distributional semantic model. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 58–66, Portland, OR.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. University of Chicago Press, Chicago, IL.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

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

Lin, D. (1997). Using syntactic dependency as local context to resolve word sense ambiguity. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pages 64–71, Madrid, Spain.

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of the joint Annual Meeting of the Association for Computational Linguistics and International Conference on Computational Linguistics*, pages 768–774, Montréal, Canada.

Lund, K. and Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28(2):203–208.

MacDonald, M. C., Pearlmutter, N. J., and Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review*, 101(4):676–703.

Marcus, G. F., Vijayan, S., Rao, S. B., and Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science*, 283(5398):77–80.

Markert, K. and Hahn, U. (2002). Understanding metonymies in discourse. *Artificial Intelligence*, 135(1–2):145–198.

Matsuki, K. (2013). *The Roles of Thematic Knowledge in Sentence Comprehension*. PhD thesis, University of Western Ontario, London, Ontario, Electronic Thesis and Dissertation Repository. Paper 1661. `http://ir.lib.uwo.ca/etd/1661`.

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

McCarthy, D. and Navigli, R. (2009). The English lexical substitution task. *Language Resources and Evaluation*, 43:139–159.

McConkie, G. W. and Rayner, K. (1975). The span of the effective stimulus during a fixation in reading. *Perception & Psychophysics*, 17(6):578–586.

McDonald, S. and Brew, C. (2004). A distributional model of semantic context effects in lexical processing. In *Proceedings of the 42th Annual Meeting of the Association for Computational Linguistics*, pages 17–24, Barcelona, Spain.

McDonald, S. A. and Shillcock, R. C. (2003a). Eye movements reveal the on-line computation of lexical probabilities during reading. *Psychological Science*, 14(6):648–652.

McDonald, S. A. and Shillcock, R. C. (2003b). Low-level predictive inference in reading: The influence of transitional probabilities on eye movements. *Vision Research*, 43(16):1735–1751.

McElree, B., Frisson, S., and Pickering, M. J. (2006a). Deferred interpretations: Why starting Dickens is taxing but reading Dickens isn't. *Cognitive Science*, 30(1):181–192.

McElree, B., Pylkkänen, L., Pickering, M. J., and Traxler, M. J. (2006b). A time course analysis of enriched composition. *Psychonomic Bulletin & Review*, 13(1):53–59.

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

McEnery, T. and Hardie, A. (2012). *Corpus linguistics: method, theory and practice*. Cambridge University Press, Cambridge, UK.

McKoon, G. and Ratcliff, R. (1986). Inferences about predictable events. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 12:82–91.

McRae, K., Hare, M., Elman, J. L., and Ferretti, T. (2005). A basis for generating expectancies for verbs from nouns. *Memory & Cognition*, 33(7):1174–1184.

McRae, K. and Matsuki, K. (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. and Matsuki, K. (2013). Constraint-based models of sentence processing. In Van Gompel, R. P. G., editor, *Sentence Processing*, chapter 3, pages 51–77. Psychology Press, New York, NY.

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

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119, Lake Tahoe, NV.

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

Minsky, M. (1975). A framework for representing knowledge. In Winston, P. H., editor, *The Psychology of Computer Vision*, pages 211–277. McGraw-Hill, New York.

Mitchell, J. and Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1429.

Moravcsik, J. M. (1975). Aitia as generative factor in Aristotle's philosophy. *Dialogue*, 14(04):622–638.

Munro, R., Bethard, S., Kuperman, V., Lai, V. T., Melnick, R., Potts, C., Schnoebelen, T., and Tily, H. (2010). Crowdsourcing and language studies: the new generation of linguistic data. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 122–130, Los Angeles, CA.

Münte, T. F., Schiltz, K., and Kutas, M. (1998). When temporal terms belie conceptual order. *Nature*, 395(6697):71–73.

Murray, W. S. (2006). The nature and time course of pragmatic plausibility effects. *Journal of Psycholinguistic Research*, 35(1):79–99.

Narayanan, S. and Jurafsky, D. (2004). A Bayesian model of human sentence processing. Unpublished manuscript, `http://www.icsi.berkeley.edu/~snarayan/newcog.pdf`.

Nelson, D., McEvoy, C., and Schreiber, T. (1998). The University of South Florida word association, rhyme, and word fragment norms. `http://www.usf.edu/freeassociation/`.

Nieuwland, M. S. and van Berkum, J. J. (2006). When peanuts fall in love: N400 evidence for the power of discourse. *Journal of Cognitive Neuroscience*, 18(7):1098–1111.

Nunberg, G. (1978). *The pragmatics of reference*. Indiana University Linguistics Club, Bloomington, Indiana.

Nunberg, G. (1979). The non-uniqueness of semantic solutions: Polysemy. *Linguistics and Philosophy*, 3(2):143–184.

Nunberg, G. (1995). Transfers of meaning. *Journal of Semantics*, 12(2):109–132.

Nunberg, G. (2004). The pragmatics of deferred interpretation. In Horn, L. and Ward, G., editors, *Handbook of pragmatics*, pages 344–364. Blackwell, Oxford, UK.

Ó Séaghdha, D. O. (2010). Latent variable models of selectional preference. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 435–444, Uppsala, Sweden.

Osherson, D. N. and Smith, E. E. (1982). Gradedness and conceptual combination. *Cognition*, 12(3):299–318.

Padó, S. and Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

Padó, S., Padó, U., and Erk, K. (2007). Flexible, corpus-based modelling of human plausibility judgements. In *Proceedings of the joint Conference of Empirical Methods in Natural Language Processing and Conference on Natural Language Learning*, pages 400–409, Prague, Czech Republic.

Padó, S. and Utt, J. (2012). A Distributional Memory for German. In *Proceedings of the 11th KONVENS*, pages 462–470, Vienna, Austria.

Padó, U. (2007). *The integration of syntax and semantic plausibility in a wide-coverage model of human sentence processing*. PhD thesis, Universität des Saarlandes, Saarbrücken.

Paivio, A. (2010). Dual coding theory and the mental lexicon. *The Mental Lexicon*, 5(2):205–230.

Paivio, A. and Sadoski, M. (2011). Lexicons, contexts, events, and images: Commentary on Elman (2009) from the perspective of dual coding theory. *Cognitive Science*, 35(1):198–209.

Partee, B. H., ter Meulen, A. G., and Wall, R. E. (1990). *Mathematical methods in linguistics*. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Pecher, D. and Zwaan, R. A. (2005). *Grounding cognition: The role of perception and action in memory, language, and thinking*. Cambridge University Press, Cambridge, UK.

Peirsman, Y. and Padó, S. (2011). Semantic relations in bilingual lexicons. *ACM Transactions in Speech and Language Processing*, 8(2):3:1–3:21.

Pickering, M. J., Frisson, S., McElree, B., and Traxler, M. J. (2004). Eye movements and semantic composition. In Carreiras, M. and Clifton, C. E., editors, *The on-line study of sentence comprehension: Eyetracking, ERP, and beyond*, pages 33–50. Psychology Press, New York, NY.

Pickering, M. J., McElree, B., and Traxler, M. J. (2005). The difficulty of coercion: A response to de Almeida. *Brain and Language*, 93:1–9.

Plag, I. (2003). *Word-formation in English*, chapter 5, Derivation without Affixation, pages 107–131. Cambridge University Press, Cambridge, UK.

Potts, C. (2012). Goal-driven answers in the Cards dialogue corpus. In *Proceedings of the 30th West Coast Conference on Formal Linguistics*, pages 1–20, Santa Cruz, CA.

Prescher, D., Riezler, S., and Rooth, M. (2000). Using a probabilistic class-based lexicon for lexical ambiguity resolution. In *Proceedings of the 18th International Conference on Computational Linguistics*, pages 649–655, Saarbrücken, Germany.

Pustejovsky, J. (1991). The Generative Lexicon. *Computational Linguistics*, 17(4):409–441.

Pustejovsky, J. (1993). Type coercion and lexical selection. In Pustejovsky, J., editor, *Semantics and the Lexicon*, pages 73–94. Kluwer Academic Publishers, Dordrecht, The Netherlands.

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

Pustejovsky, J. (1998). Generativity and explanation in semantics: A reply to Fodor and Lepore. *Linguistic Inquiry*, 29(2):289–311.

Pustejovsky, J. (2012). The semantics of functional spaces. In Schalley, A., editor, *Practical Theories and Empirical Practice: Facets of a Complex Interaction*. John Benjamins Publishing, Amsterdam, The Netherlands.

Pustejovsky, J. (2013). Dynamic event structure and habitat theory. In *Proceedings of the 6th International Conference on Generative Approaches to the Lexicon (GL2013)*, pages 1–10, Pisa, Italy.

Pustejovsky, J. and Bouillon, P. (1995). Aspectual coercion and logical polysemy. *Journal of Semantics*, 12(2):133–162.

Pylkkänen, L. and McElree, B. (2006). The syntax-semantics interface: On-line composition of sentence meaning. In Traxler, M. and Gernsbacher, M. A., editors, *Handbook of Psycholinguistics*, pages 539–579. Elsevier, Amsterdam, The Netherlands, 2nd edition.

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

Pynte, J., Besson, M., Robichon, F.-H., and Poli, J. (1996). The time-course of metaphor comprehension: An event-related potential study. *Brain and Language*, 55(3):293–316.

Raaijmakers, J. G., Schrijnemakers, J., and Gremmen, F. (1999). How to deal with "The Language-as-Fixed-Effect Fallacy": Common misconceptions and alternative solutions. *Journal of Memory and Language*, 41:416–426.

Ravichandran, D. and Hovy, E. H. (2002). Learning surface text patterns for a Question Answering System. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 41–47, Philadelphia, PA.

Rayner, K. and Duffy, S. A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition*, 14(3):191–201.

Rayner, K., Warren, T., Juhasz, B. J., and Liversedge, S. P. (2004). The effect of plausibility on eye movements in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(6):1290–1301.

Reddy, S., McCarthy, D., and Manandhar, S. (2011). An empirical study on compositionality in compound nouns. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 210–218, Chiang Mai, Thailand.

Resnik, P. (1996). Selectional constraints: An information-theoretic model and its computational realization. *Cognition*, 61:127–159.

Roberts, K. and Harabagiu, S. M. (2011). Unsupervised learning of selectional restrictions and detection of argument coercions. In *Proceedings of the 16th Conference of Empirical Methods in Natural Language Processing*, pages 980–990, Edinburgh, UK.

Roller, S., Schulte im Walde, S., and Scheible, S. (2013). The (un)expected effects of applying standard cleansing models to human ratings on compositionality. In *Proceedings of the 9th Workshop on Multiword Expressions*, pages 32–31, Atlanta, GA.

Rooth, M., Riezler, S., Prescher, D., Carroll, G., and Beil, F. (1999). Inducing a semantically annotated lexicon via EM-based clustering. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 104–111, College Park, MD.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, 104(3):192–233.

Rothstein, S. (2008). *Structuring events: A study in the semantics of aspect*. Blackwell, Oxford, UK.

Rüd, S. and Zarcone, A. (2011). Covert events and qualia structures for German verbs. In *Proceedings of the Metonymy 2011 Workshop*, pages 17–22, Stuttgart, Germany.

Rumelhart, D., Smolensky, P., McClelland, J., and Hinton, G. (1986). Schemata and sequential thought processes in PDP models. In *Parallel distributed processing: Explorations in the microstructure of cognition*, volume 2, pages 7–57. MIT Press, Cambridge, MA.

Rumelhart, D. E. (1975). Notes on a schema for stories. In *Representation and understanding: Studies in cognitive science*. Academic Press, New York.

Rumelhart, D. E. (1979). Some problems with the notion of literal meanings. In Ortony, A., editor, *Metaphor and thought*. Cambridge University Press, Cambridge, UK.

Rumelhart, D. E. (1980). On evaluating story grammars. *Cognitive Science*, 4:313–316.

Rumelhart, D. E. and McClelland, J. L. (1987). Learning the past tenses of English verbs. Implicit rules or parallel distributed processing. In *Mechanisms of language acquisition*, pages 249–308. Lawrence Erlbaum Associates, Hillsdale, NJ.

Saffran, J. R., Aslin, R. N., and Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928.

Sahlgren, M. (2008). The distributional hypothesis. *Italian Journal of Linguistics*, 20(1):33–54.

Schank, R. C. and Abelson, R. P. (1977). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Schiehlen, M. (2004). Annotation strategies for probabilistic parsing in German. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 390–396, Geneva, Switzerland.

Schulte im Walde, S. (2006). Experiments on the automatic induction of German semantic verb classes. *Computational Linguistics*, 32(2):159–194.

Schulte im Walde, S., Hying, C., Scheible, C., and Schmid, H. (2009). Combining EM training and the MDL principle for an automatic verb classification incorporating selectional preferences. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies*, pages 496–504, Singapore.

Schulte im Walde, S., Melinger, A., Roth, M., and Weber, A. (2008). An empirical characterisation of response types in German association norms. *Research on Language and Computation*, 6(2):205–238.

Schumacher, P. B. (2009). Definiteness marking shows late effects during discourse processing: Evidence from ERPs. In *Anaphora Processing and Applications. Lecture Notes in Computer Science*, pages 91–106. Springer, Berlin, Germany.

Schumacher, P. B. (2011). The hepatitis called... electrophysiological evidence for enriched composition. In *Experimental Pragmatics / Semantics*, pages 199–219. John Benjamins Publishing, Amsterdam, The Netherlands.

Schumacher, P. B. (2013). Content and context in incremental processing: "the ham sandwich" revisited. *Philosophical Studies*, pages 1–15.

Schumacher, P. B. and Weiland, H. (2011). Reading Brecht and talking to the espresso: Electrophysiological investigations of conventional and novel metonymy. In *Proceedings of the Metonymy 2011 Workshop*, pages 23–28, Stuttgart, Germany.

Schütze, H. (1992). Dimensions of meaning. In *Proceedings of Supercomputing '92*, pages 787–796, Minneapolis, MN.

Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

Schütze, H. (2013). Plenary talk. Dagstuhl Seminar on Computational Models of Language Meaning in Context.

Schwenk, H. (2007). Continuous space language models. *Computer Speech & Language*, 21:492–518.

Schwenk, H. and Koehn, P. (2008). Large and diverse language models for statistical machine translation. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing*, pages 661–666, Hyderabad, India.

Shutova, E. (2009). Sense-based interpretation of logical metonymy using a statistical method. In *Proceedings of the ACL-IJCNLP Student Research Workshop*, pages 1–9, Singapore.

Shutova, E., Kaplan, J., Teufel, S., and Korhonen, A. (2013). A computational model of logical metonymy. *ACM Transactions on Speech and Language Processing*, 10(3):11:1–11:28.

Shutova, E. and Teufel, S. (2009). Logical metonymy: Discovering classes of meanings. In *Proceedings of the CogSci Workshop on Semantic Space Models*, pages 29–34.

Smith, E. E., Osherson, D. N., Rips, L. J., and Keane, M. (1988). Combining prototypes: A selective modification model. *Cognitive Science*, 12:485–527.

Smith, N. J. and Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y. (2008). Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference of Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii.

Sperber, D. and Wilson, D. (1986). *Relevance*. Harvard University Press, Cambridge, MA.

Spivey, M. (2007). *The continuity of mind*. Oxford University Press, New York, NY.

Staub, A. and Rayner, K. (2007). Eye movements and on-line comprehension processes. In Gaskell, G., editor, *The Oxford Handbook of Psycholinguistics*, pages 327–342. Oxford University Press, Oxford, UK.

Stern, G. (1931). *Meaning and change of meaning; with special reference to the English language*. Wettergren & Kerbers, Göteborg.

Sternberg, S. (1966). High-speed scanning in human memory. *Science*, 153(3736):652–654.

Sternberg, S. (1975). Memory scanning: New findings and current controversies. *The Quarterly Journal of Experimental Psychology*, 27:1–32.

Stevens, S. S. (1975). *Psychophysics: Introduction to its Perceptual*. John Wiley, New York, NY.

Sweep, J. (2012). Logical Metonymy in Dutch and German: Equivalents of Begin, Finish, and Enjoy. *International Journal of Lexicography*, 25(2):117–151.

Tabor, W. and Tanenhaus, M. K. (2001). Dynamical systems for sentence processing. In Christiansen, M. and Chater, N., editors, *Connectionist Psycholinguistics*, pages 177–211. Ablex Publishing, Westport, CO.

Tabossi, P., Collina, S., Caporali, A., Pizzioli, F., and Basso, A. (2010). Speaking of events: The case of CM. *Cognitive Neuropsychology*, 27(2):152–180.

Taylor, W. L. (1953). "Cloze procedure": A new tool for measuring readability. *Journalism Quarterly*, 30:415–433.

Thelen, M. and Riloff, E. (2002). A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In *Proceedings of the 2002 Conference of Empirical Methods in Natural Language Processing*, pages 214–221, Philadelphia, PA.

Tomasello, M. (2009). *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press, Cambridge, MA.

Traxler, M. J., McElree, B., Williams, R. S., and Pickering, M. J. (2005). Context effects in coercion: Evidence from eye movements. *Journal of Memory and Language*, 53:1–25.

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

Trueswell, J., Tanenhaus, M., and Garnsey, S. (1994). Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language*, 33:285–318.

Trueswell, J. C., Tanenhaus, M. K., and Kello, C. (1993). Verb-specific constraints in sentence processing: separating effects of lexical preference from garden-paths. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19(3):528.

Turney, P. D. (2006). Similarity of semantic relations. *Computational Linguistics*, 32(3):379–416.

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

Utt, J., Lenci, A., Padó, S., and Zarcone, A. (2013). The curious case of metonymic verbs: A distributional characterization. In *Proceedings of the IWCS Workshop "Towards A Formal Distributional Semantics"*, Potsdam, Germany.

van Berkum, J. J., Brown, C. M., Zwitserlood, P., Kooijman, V., and Hagoort, P. (2005). Anticipating upcoming words in discourse: evidence from erps and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(3):443.

van der Meer, E., Krüger, F., and Nuthmann, A. (2005). The influence of temporal order information in general event knowledge on language comprehension. *Zeitschrift für Psychologie/Journal of Psychology*, 213(3):142–151.

Vendler, Z. (1968). *Adjectives and nominalizations*. Mouton, The Hague.

Verspoor, C. (1997a). Conventionality-governed logical metonymy. In *Proceedings of the 2nd International Workshop on Computational Semantics*, pages 300–312, Tilburg, The Netherlands.

Verspoor, C. M. (1997b). *Contextually-dependent lexical semantics*. PhD thesis, University of Edinburgh. College of Science and Engineering. School of Informatics.

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

Warren, T. and McConnell, K. (2007). Investigating effects of selectional restriction violations and plausibility violation severity on eye-movements in reading. *Psychonomic Bulletin & Review*, 14(4):770–775.

Weinreich, U. (1966). Explorations in semantic theory. In Sebeok, T., editor, *Current Trends in Linguistics*. Mouton, The Hague.

Werning, M., Hinzen, W., and Machery, E. (2012). *The Oxford Handbook of Compositionality*. Oxford University Press, Oxford, UK.

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

Wilks, Y. (1978). Making preferences more active. *Artificial Intelligence*, 11(3):197–223.

Wilson, D. and Sperber, D. (2004). Relevance theory. In Horn, L. R. and Ward, G., editors, *Handbook of Pragmatics*. Blackwell, Oxford, UK.

Wurm, L. H. and Cano, A. (2010). Stimulus norming: It is too soon to close down brick-and-mortar labs. *The Mental Lexicon*, 5(3):358–370.

Yin, W. and Schütze, H. (2013). Deep learning embeddings for discontinuous linguistic units. arXiv preprint arXiv:1312.5129.

Zarcone, A. and Lenci, A. (2008). Computational models of event type classification in context. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, pages 1232–1238, Marrakech, Morocco.

Zarcone, A., Lenci, A., Padó, S., and Utt, J. (2013). Fitting, not clashing! a distributional semantic model of logical metonymy. In *Proceedings of the 10th International Conference on Computational Semantics*, Potsdam, Germany.

Zarcone, A., Lipenkova, J., and Michelbacher, L. (2012a). Easy / difficult constructions as triggers of implicit content: comparing covert event elicitations and events extracted from a very large corpus. Poster presented at Linguistic Evidence 2012.

Zarcone, A. and Padó, S. (2010). "I like work: I can sit and look at it for hours" - Type clash vs. plausibility in covert event recovery. In *Proceedings of Verb 2010 - Interdisciplinary Workshop on Verbs*, pages 209–214, Pisa, Italy.

Zarcone, A. and Padó, S. (2011). Generalized event knowledge in logical metonymy resolution. In *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*, pages 944–949, Boston, MA.

Zarcone, A. and Padó, S. (2013). Logical metonymy: Disentangling object type and thematic fit. Poster presented at the 19th Conference on Architectures and Mechanisms for Language Processing, Marseille, France.

Zarcone, A., Padó, S., and Lenci, A. (2012b). Inferring covert events in logical metonymies: a probe recognition experiment. In *Proceedings of the 34th Annual Meeting of the Cognitive Science Society*, pages 1215–1220, Sapporo, Japan.

Zarcone, A., Padó, S., and Lenci, A. (2014). Logical metonymy resolution in a words-as-cues framework: evidence from self-paced reading and probe recognition. *Cognitive Science*, 38(5):973–996.

Zarcone, A. and Rüd, S. (2012). Logical metonymies and qualia structures: an annotated database of logical metonymies for German. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 1799–1804, Istanbul, Turkey.

Zarcone, A., Utt, J., and Lenci, A. (2012c). Logical metonymy from type clash to thematic fit. Poster presented at the 18th Conference on Architectures and Mechanisms for Language Processing, Riva del Garda, Italy.

Zarcone, A., Utt, J., and Padó, S. (2012d). Modeling covert event retrieval in logical metonymy: probabilistic and distributional accounts. In *Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics*, pages 70–79, Montréal, Canada.

Zubizarreta, M. L. (1987). *Levels of Representation in the Lexicon and in the Syntax*. Foris Publications, Dodrecht, Holland.

Lebenslauf

	Persönliche Daten
Name:	Alessandra Zarcone
Geburtsdatum und -ort:	29.07.1985, Palermo, Italien
Staatsangehörigkeit:	italienisch
E-mail:	a.zarcone@gmail.com
	Schulausbildung & Studium
06/2003	Liceo Classico Statale "Vittorio Emanuele II", Palermo, Italien Allgemeine Hochschulreife (100/100)
10/2003 – 06/2009	Stipendiatin der Scuola Normale Superiore, Pisa, Italien *Italienische Literaturwissenschaft und Linguistik*
11/2009	Diplom an der Scuola Normale (70/70 mit Auszeichnung)
10/2003 – 10/2006	Studium der *Digital Humanities* an der Università di Pisa, Italien
10/2006	Bachelor of Science (B. Sc.): 110/110 (mit Auszeichnung) Thema: "La classificazione azionale del verbo italiano primi esperimenti computazionali" (Betreuer: Alessandro Lenci)
10/2006-03/2009	Studium der *Linguistik* an der Università di Pisa, Italien
03/2009	Master of Arts (M.A.): 110/110 (mit Auszeichnung) Thema: "Empirical correlates of event types - a priming study" (Betreuer: Alessandro Lenci)
09/2007 – 03/2008 und 03/2009 – 08/2009	Austauschstudium der *Computerlinguistik* an der Universität des Saarlandes
Seit 08/2009	Promotionstudium in *Linguistik* am Institut für Maschinelle Sprachverarbeitung der Universität Stuttgart (Betreuer: Sebastian Padó)
	Berufliche Tätigkeiten
12/2007 – 02/2008	Fremdsprachenassistentin für *Italienisch* an der Universität des Saarlandes
08/2009 – 07/2014	Wissenschaftliche Mitarbeiterin am Institut für Maschinelle Sprachverarbeitung der Universität Stuttgart Project D6 / SFB 732
Seit 08/2014	Wissenschaftliche Mitarbeiterin an der Universität des Saarlandes Project A2 und A3 / SFB 1102