Scuola Universitaria Superiore IUSS Pavia

COMPUTATIONAL TOOLS AS PARTICIPANTS AND MODELS IN METAPHOR RESEARCH: FROM DIACHRONIC CHANGE TO BRAIN RESPONSES

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosophy in

THEORETICAL AND EXPERIMENTAL LINGUISTICS

by

Veronica Mangiaterra

Supervisor: Prof. Valentina Bambini

Co-Supervisor: Dr. Chiara Barattieri di San Pietro

January, 2026

She smiled innocuously – at variance with her words.
At this point he could not discern her degree of seriousness.
A topic of world-shaking importance, yet dealt with facetiously;
an android trait, possibly, he thought.
No emotional awareness, no feeling-sense of the actual meaning of what she said.
Only the hollow, formal, intellectual definitions of the separate terms.

Philip Dick "Do Androids Dream of Electric Sheep?" (1968)

ABSTRACT

Computational tools, such as word embeddings or Large Language Models (LLMs), have been shown to align with humans in certain linguistic behaviors, suggesting that they can offer a novel perspective in tackling open issues in linguistics and cognitive science. In this thesis, I present five studies that employed computational tools to investigate different aspects of metaphor processing, namely the change over time of processing demands and the operations underlying electrophysiological components. Beyond the theoretical questions, the role that computational tools can take in metaphor research was explored by employing them as participants and models.

Study 1 tested the validity and reliability of LLMs in providing ratings for metaphorical expressions along a series of dimensions (familiarity, comprehensibility, and imageability) for Italian and English. Our results showed that machine-generated ratings can approximate human ones and validly substitute them in statistical analysis predicting behavioral and neural responses. Moreover, LLMs' ratings seem to be highly reliable, as they were stable across independent prompting sessions. Some weaknesses emerged when generating ratings of metaphors requiring the integration of sensorimotor knowledge to be interpreted and when rating metaphors for their imageability. These results suggested that LLMs can capture at least some aspects of human metaphorical competence, making them suitable for augmenting human data as artificial participants.

Study 2 employed temporal word embeddings to simulate how the costs associated with metaphor processing changed over time. Specifically, semantic similarity between topics and vehicles of a set of 515 Italian literary metaphors was taken as a proxy of its processing demands and was computed in 19th-century and 21st-century vector space models, extending previous distributional approaches to the semantic shift of single words. Within each epoch, we also separately considered literary and nonliterary texts to test if the change in processing demands was also influenced by the textual genre. We found that literary metaphors followed the general pattern of the Italian language. While

metaphors were associated with equal processing demand in literary and nonliterary texts of the 19th century (when the two genres did not differ for stylistic features), they become more difficult in 21st-century literary texts (where the plain style of modern literature makes the metaphor more striking) and less difficult in 21st-century nonliterary texts (where the creative use of language makes less demanding to connect metaphor's conceptual domains). In Study 3, this method was extended to English literary metaphors, demonstrating its cross-linguistic applicability and revealing language-specific patterns. Here, metaphors were shown to differ in processing demands only based on genre, with no diachronic effect, as the English language underwent very little change in the past two centuries compared to Italian.

In Study 4, the interest shifted to using different computational models to disentangle the different theoretical accounts used to support the functional interpretation of electrophysiological components associated with metaphor processing. In particular, three computational approaches were employed, namely semantic similarity from word embeddings, surprisal from LLMs, and a novel Bayesian pragmatic measure inspired by the Rational Speech Acts (RSA) framework, taken to model three different theoretical views, linked respectively to semantic, context-based prediction, and inferential mechanisms. Results revealed that the N400 is mainly affected by surprisal, supporting the predictive nature of this component, while the P600 is influenced by both surprisal and the Bayesian pragmatic measure, suggesting that this component, while tied to contextual prediction, also incorporates the signature of pragmatic inferences.

Finally, Study 5 reported the preliminary results of a systematic review and meta-analysis investigating, with an innovative approach based on figure digitalization, the robustness of electrophysiological components reported for figurative language across metaphor, irony, and idioms. This study showed that the N400 is deeply linked to pragmatic processing, emphasizing the role of context in deriving all types of figurative meaning, and that the pragmatic P600 emerged

as a function of type of pragmatic phenomenon and its conventionality, being highly robust for irony (92%) but more transient for metaphor (2% overall, but 88% for novel metaphors).

Overall, these studies highlight the fertile integration of computational tools in metaphor research and characterize metaphor processing as a dynamic interplay of prediction and inference, deeply linked to the broad contextual and stylistic features of the language.

TABLE OF CONTENTS

LIST OF TABLES

STUDY 1:

STUDY 2:

STUDY 3:

STUDY 4:

STUDY 5:

LIST OF FIGURES

STUDY 1:

STUDY 2:

STUDY 3:

STUDY 4:

STUDY 5:

ACKNOWLEDGMENTS

I would like to begin this thesis by thanking the people who have contributed to this work and to my development as a student and researcher. My first thanks go to my PhD supervisor, Prof. Valentina Bambini, who made me fall in love with pragmatics since my first undergrad course with her, guided me in the never-ending discovery of how to do and communicate research, and encouraged me when I wanted to try exploring new directions.

I would also like to thank my PhD co-supervisor, Dr. Chiara Barattieri di San Pietro, for being one of the pillars of my doctoral years with her steady support and human and scientific guidance.

Even if not with a formal role, my PhD has deeply enriched by the presence of Dr. Paolo Canal; his scientific competence and his irony have provided the best support for all the (not only) statistical challenges along this path.

Also, my journey could not have been the same without my fellow colleagues (and friends) at the Neurolinguistics and Experimental Pragmatics Lab. I start by thanking Dr. Federico Frau and Fabrizio Luciani, who directly collaborated on one of the studies in this thesis, and Dr. Luca Bischetti, Dr. Chiara Pompei, and Maddalena Bressler, with whom I worked on side projects that I very much enjoyed on language acquisition and development of language resources. I learned a lot from all of you. I am also thankful to all the NEP members for the time shared together that made these years precious - and definitely not lonely.

During the PhD, I had the great opportunity to be hosted at two other Labs, well complementing my training. I profoundly thank Dr. Hamad Al-Azary, who hosted me at Lawrence Technological University in the US. Our scientific conversations have been full of insights and will be among my dearest memories of this PhD. I also thank Prof. Walter Schaeken, who hosted me at KU in Leuven, for his scientific support and for the stimulating lab meetings, where thanks also to his students, I discovered approaches to language and cognition completely novel to me.

I also thank Prof. Paolo Mazzarello, Stefano Maretti, Edoardo Razzetti, Paolo Guaschi, and all the staff at the Kosmos Museum, who hosted me for my internship and gave me the opportunity to present our work to the public.

Outside of university, I was lucky to have the invaluable support of family and friends. Thanks for believing in me and always pushing me to do my best.

As a final note, I have to thank Prof. Andrea Moro, who unknowingly led me where I am today. On summer of 2016, as a high school student, I listened to his lecture on neurolinguistics at the Scuola di Orientamento Universitario in Pisa. It was as if all the pieces of the puzzle of my interests had formed a complete picture that I had not known before, and led me to try to pursue this journey.

This work was conducted with financial support from Unione europea - Next Generation EU, Missione 4 Componente 1 CUP I13C22000330001.

INTRODUCTION

1. Metaphors

Language is rich in implicit meanings. Resolving what is implicated in texts and conversational exchanges goes far beyond the compositional combination of the meanings of the single words in the utterance. On the contrary, we are required to combine the semantics of the sentence with a series of additional knowledge we have about context, about the world, about the speaker, and what we think she may want to communicate.

One fascinating case of implicit meaning is represented by metaphors. Metaphors are a distinctive feature of all human languages and are used in conversations, newspapers, as well as in the literature (Steen et al., 2010). Yet they are far more than a mere linguistic ornament. As suggested by Lakoff & Johnson (1980) in the field of Cognitive Linguistics, metaphors represent a fundamental cognitive mechanism, which allows us to conceptualize complex abstract concepts in terms of more accessible concrete ones (for example, LIFE in terms of JOURNEY).

From this first *cognitive* shift, which brought metaphors from the domain of literary and rhetorical studies to that of cognitive science and linguistics, many theoretical accounts have proposed their own definition of metaphors. In the theoretical framework provided by the post-Gricean account of *Relevance Theory* (Sperber & Wilson, 1986), the gap between the literal and the intended meaning is said to be resolved through the inferential process, a cognitive operation that allows the derivation of the *implicature*, namely the part of meaning which is implicated (Allott, 2010). In this framework, metaphors such as "Jane is a computer", are one example of nonliteral and loose use of language (together with hyperboles and approximations, for instance) where the term used metaphorically (in this example, "computer", the so-called "vehicle" of the metaphor) undergoes a process of lexical adjustment, by narrowing and/or broadening its literal sense (Sperber & Wilson, 2012; Wilson & Carston, 2007). Thus, from the lexically encoded meaning of "computer",

an *ad-hoc* concept is derived, in which some features of the concept are promoted (for example, "it is precise", "lacks emotions"), while some others are discarded (for example, "it has a keyboard"). The context in which the sentence is embedded can facilitate and guide the interpretation. For instance, if we are talking about Jane's ability to complete a complex task, the interpretation "precise" will be preferred, while if we are talking about Jane's behavior with her friends, the interpretation "lacks emotion" will be more salient.

Metaphors can vary across a number of dimensions that deeply impact the way they are processed, such as imageability (how easily they can evoke a mental image), meaningfulness (how interpretable the sense of the expression is), and familiarity (how familiar the association between terms sounds). Along this last dimension, some metaphors, due to the frequency of use – their "career" (Bowdle & Gentner, 2005) - became conventionalized and are considered easier to process. On the other side of the familiarity spectrum, unique and creative metaphors lie, such as *literary metaphors,* which will be examined in two studies of this thesis. Written by novelists and poets within the context of literary texts, these metaphors do not point toward a unique and clear interpretation, but they are reported to elicit a wide array of weak implicatures. For instance, a metaphor such as "fog of melancholy" can evoke many interpretations in the mind of the reader, and the cost of exploring them and keeping them in memory is what elicits the *poetic effect* (Pilkington, 2000).

1.1. Metaphor processing

How the metaphorical meaning is derived, the effort compared to the derivation of literal sentences, and the role of the literal meaning in metaphor interpretation have been extensively studied with a variety of empirical approaches. These encompass reaction times (Bambini et al., 2013), eye-tracking (Columbus et al., 2015; Werkmann Horvat et al., 2023), magnetic resonance imaging (Bambini et al., 2011; Rapp et al., 2004), and neurophysiological methods, such as Event-Related Potential (ERP) approaches (Bambini et al., 2016; Lai & Curran, 2013). These diverse

methodologies have yielded converging evidence about the time course and neural substrates of metaphor comprehension, though debates remain about whether metaphor processing requires additional cognitive effort compared to literal language processing (Giora, 2003) and whether the literal meaning must be accessed before the metaphorical interpretation can be derived (Gernsbacher et al., 2001).

The electroencephalography (EEG) has been particularly popular in psycho/neurolinguistic approaches to metaphor, as it offers a high-temporal-resolution window into the distinct stages of metaphor comprehension (Coulson, 2008). From ERP studies, one primary component has been consistently reported and emerged as crucial in metaphor processing (the N400), a negative-going deflection peaking around 400ms post-stimulus onset that is thought to index semantic integration difficulty or violations of semantic expectancy (Kutas & Federmeier, 2011). The following time window has been variously characterized by either a Late Positivity (P600/LPC) or a sustained negativity (Baiocco et al., 2026; Canal & Bambini, 2023), both less precisely characterized in functional terms and reported less consistently.

1.2. Metaphors and machines

Since the first descriptions of the linguistic behavior of machines, researchers have noticed that metaphorical uses of language (e.g., "The car drank gasoline") represented a violation of *selectional preference*, difficult to incorporate into the semantic representation of machines, which were based on strict rules to assess the well-formedness of sentences (Fass & Wilks, 1983). Consequently, much work has been devoted to developing computer programs able to deal with figurative aspects of language, for instance by correctly identifying and interpreting metaphors (Fass, 1991; Martin, 1992). In these first phases, the idea was to install a metaphorical competence in machines by adding more rules that could include these apparently ambiguous uses of language. For instance, allowing an inanimate entity (e.g., "car") to be the agent of a verb like "drink", typically associated

with animate agents. After the manual definition of the new metaphorical rules, subsequent approaches have relied on Distributional Semantics (Reid & Katz, 2018). Based on the Distributional Hypothesis (Harris, 1954), namely that similar words tend to occur in similar contexts, computational models such as Latent Semantic Analysis (Landauer & Dumais, 1997) or word2vec (Mikolov et al., 2013), represented words as high-dimensional vectors (word embeddings) in the semantic space (Lenci, 2018). In this framework, metaphors have often been modeled as the semantic distance between the topic and the vehicle composing them. This strand of research has benefited from theoretical insights and from what was known about metaphors in human minds and brains (Tong et al., 2021), moving mostly from theory to machine implementation. For instance, methods of automatic metaphor identification have been inspired by Conceptual Metaphor Theory (Ge et al., 2022) or by the Metaphor Identification Procedure elaborated by the Pragglejaz Group, 2007 (Mao et al., 2019), or relied on semantic distance thresholds.

Nowadays, with the development of Large Language Models (LLMs), namely massive models of language trained on billions of tokens to learn word co-occurrences, metaphorical competence seems to emerge from the statistical semantic representation of these models without explicit instructions. For instance, LLMs reported great accuracy both in metaphor identification (Fuoli et al., 2025) and interpretation (Barattieri di San Pietro et al., 2023; Ichien et al., 2024), without any phase of specific metaphorical fine-tuning.

The impressive performance of LLMs on language tasks, despite some weaknesses, for instance related to the sensorimotor aspects of language (Borghi et al., 2023; Chemero, 2023), have yielded a deep debate on how fruitful the employment of computational tools can be for the advancement of scientific practice in general (Binz et al., 2025; Birhane et al., 2023) and for what we know about human linguistic and cognitive skills (Abdurahman et al., 2024; Wulff & Mata, 2025). On one side, some scientists claimed that LLMs can be of little help in answering questions on the nature of

language and its processing and acquisition, because they show only superficial similarities with human language but with completely different underlying mechanisms (Birhane & McGann, 2024; Bolhuis et al., 2024; Cuskley et al., 2024). On the other side, it has been claimed that LLMs solve long-standing questions about language (Contreras Kallens et al., 2023; Piantadosi, 2024). On an intermediate position, some researchers, while acknowledging their limitations, highlighted their potential implications for several aspects of language science. For instance, LLMs can inform the debate on the learnability of language, or quantify aspects, such as *surprisal*, namely the predictability of a word given its context (Slaats & Martin, 2025; Smith & Levy, 2013), which predict behavioral and neural responses, and provide data annotation that can be used to answer theoretical questions (M. C. Frank & Goodman, 2025; Levy et al., 2025).

The potential of computational tools and LLMs to explain human metaphorical abilities has been only partially explored. A pivotal approach was the *Predication Algorithm* proposed by Kintsch (2000), which employed LSA (Landauer & Dumais, 1997) to simulate how the human mind selects context-relevant features and inhibits irrelevant ones, and test whether his view on activation of word meaning (Kintsch, 1988) is supported by the LSA mathematical process. An extension of this approach was carried out by Utsumi (2011), which contrasted three theories on metaphor processing (Bowdle & Gentner, 2005; Glucksberg & Haught, 2006; Utsumi, 2007), simulating their interpretation via LSA. Only recently, *surprisal* has entered the field, acting as a relevant metric also in metaphor research (Momen et al., 2026).

In this work, while recognizing that LLMs do not provide a comprehensive model of language as it is in human brains and minds and that alignment does not imply cognitive plausibility, I tried to explore this other direction of the relationship between metaphors and machines, namely, not how we can make machines better metaphorical agents using theoretical and psycholinguistic evidence but which insights we can derive for human processing from the (emergent and incomplete) representation that machines have of metaphorical expressions.

2. Aims and thesis

This thesis develops along two main paths.

The first path is a methodological one and aims at integrating computational tools in metaphor research by considering them as "models" and "participants". By *models*, I mean quantity estimation tools (Hu, 2023) that can allow us to quantitatively define certain features of metaphors to test a certain theoretical hypothesis. For example, from distributional semantics, we can derive the semantic similarity between the two terms of a metaphor, and from LLMs, we can derive the predictability of the vehicle of a metaphor given the preceding metaphorical context. Manipulating these measures, we can see which role the similarity of conceptual domains or the predictability of association can have in metaphor processing. By *participants*, I mean the possibility of approximating human behavior with the aim of either probing processing in participants that we can no longer access, such as readers of past epochs, or augmenting datasets of human responses to accelerate the creation and the norming of experimental stimuli.

The second path is a theoretical one and aims at examining metaphor processing from different perspectives. One perspective explores how the overlap of cultural and linguistic references between the person who produces the metaphor and the person who experiences it can influence the processing demands of metaphors. Another perspective is how metaphors are processed in the brain: which are the operations behind the neural signatures that emerged from electrophysiological studies, and how stable these signatures are across studies and figurative language phenomena.

The thesis is articulated in five studies, exploring the role of computational tools in disentangling the theoretical issues of metaphor processing.

- **Study 1** investigated the validity and reliability of LLM-generated norms for metaphorical stimuli. Different GPT models were prompted to produce ratings of

familiarity, comprehensibility, and imageability for Italian and English metaphors, with distinct prompting settings. We tested whether the LLM-generated ratings approximated human ones and whether they yielded the same results when used in statistical analysis predicting EEG responses and reaction times.

- **Study 2** investigated how the processing demands of Italian literary metaphors changed from the time in which they were created to the present day, when they are experienced by contemporary readers. We operationalized the processing demands as the semantic similarity between the topic and the vehicle of a metaphor, and we tested the diachronic evolution by training temporal word embeddings on corpora of different epochs.

- **Study 3** expanded the previous study by testing the cross-linguistic applicability of the methodology to examine the evolution of processing demands of literary metaphors. Here, we investigated the temporal trajectories of English literary metaphors to see whether the language-specific patterns could emerge when taking another language into consideration.

- **Study 4** employed different computational models to explore the operations underlying two EEG components typically associated with metaphor processing, namely the N400 and the P600. The three models operationalized distinct theoretical approaches, focusing on semantic, inferential, and context-based predictive mechanisms, respectively. Our results pointed toward a functional distinction of the two components, with N400 being modulated by the predictive measure and the P600 associated with both predictive and inferential measures.

- **Study 5** reported preliminary results of a meta-analysis assessing the robustness of ERP components related to figurative language, from metaphor to irony. Although many technical issues related to the nature of EEG studies have so far prevented a systematic meta-analysis of the literature, we developed a novel methodology based on figure

digitalization that enabled us to overcome these issues and re-analyze the N400 and P600 components across multiple studies, pragmatic phenomena, and experimental settings.

STUDY ONE

CAN GPT REPLACE HUMAN RATERS? VALIDITY AND RELIABILITY OF MACHINE-GENERATED NORMS FOR METAPHORS[1]

Abstract

As Large Language Models (LLMs) are increasingly being used in scientific research, the issue of their trustworthiness becomes crucial. In psycholinguistics, LLMs have been recently employed in automatically augmenting human-rated datasets, with promising results obtained by generating ratings for single words. Yet, performance for ratings of complex items, i.e., metaphors, is still unexplored.

Here, we present the first assessment of the validity and reliability of ratings of metaphors on familiarity, comprehensibility, and imageability, generated by three GPT models for a total of 687 items gathered from the Italian *Figurative Archive* and three English studies. We performed a thorough validation in terms of both alignment with human data and ability to predict behavioral and electrophysiological responses.

We found that machine-generated ratings positively correlated with human-generated ones. Familiarity ratings reached moderate-to-strong correlations for both English and Italian metaphors, although correlations weakened for metaphors with high sensorimotor load. Imageability showed moderate correlations in English and moderate-to-strong in Italian. Comprehensibility for English metaphors exhibited the strongest correlations. Overall, larger models outperformed smaller ones and greater human-model misalignment emerged with familiarity and imageability. Machine-generated ratings significantly predicted response times and

[1] This chapter is a manuscript currently submitted and under review as Mangiaterra, V., Al-Azary, H., Barattieri di San Pietro, C., Canal, P. & Bambini, V., *Can GPT replace human raters? Validity and reliability of machine-generated norms for metaphors*" to Humanities and Social Science Communications.

the EEG amplitude, with a strength comparable to human ratings. Moreover, GPT ratings obtained across independent sessions were highly stable.

We conclude that GPT, especially larger models, can validly and reliably replace – or augment - human subjects in rating metaphor properties. Yet, LLMs align worse with humans when dealing with conventionality and multimodal aspects of metaphorical meaning, calling for careful consideration of the nature of stimuli.

1.1. Introduction

Large Language Models (LLMs) have gained popularity in the last few years, especially after the release of ChatGPT. Their outstanding abilities in producing language in a human-like way led to a number of new research questions within the psycholinguistic community, concerning the impact that LLMs could have on the study of the nature of language. While the more theoretical debate on LLMs' validity as models of human cognition is still ongoing (Bolhuis et al., 2024; Cuskley et al., 2024; M. C. Frank & Goodman, 2025), a more practical aspect worth exploring is LLMs' role as research tools in the experimental pipeline. A recent survey reported that up to 80% of researchers across diverse fields of study have used LLM-based tools in their research (Liao et al., 2024), calling for a thorough evaluation of the benefits and risks associated with each phase of this integration (Binz et al., 2025; Charness et al., 2025; Messeri & Crockett, 2024).

In language sciences and psycholinguistics, integrating LLMs has often resulted in their use to generate ratings for experimental stimuli (Conde, Grandury, et al., 2025; Guenther & Cassani, 2025). Collecting ratings from human participants is time and resource-consuming (with large-scale datasets requiring from 800 to 2000 participants and from 6 to 15 weeks of data collection, see Kuperman et al. 2012; Warriner et al. 2013; Brysbaert et al. 2014), yet nonetheless essential. Indeed, as demonstrated by various psycholinguistic experiments, the speed and accuracy of word recognition is affected by several linguistic properties. A frequent word like *cat* is processed faster than a less frequent word like *opal* (Brysbaert et al., 2018), and a concrete word like *bookcase* is processed faster than an abstract word like *knowledge* (Barber et al., 2013). Consequently, many normed datasets of words along a variety of dimensions were created (Stadthagen-Gonzalez and Davis 2006; Brysbaert et al. 2014; Scott et al. 2019; Lynott et al. 2020), allowing the scientific community to rely on pre-collected norms, with benefits from the perspective of reproducibility and reuse of existing resources. With the advent of LLMs, researchers explored the possibility of automatically generating ratings by directly prompting the models using natural language (Trott

2024a; Martínez et al. 2024b, a; Brysbaert et al. 2024; Xu et al. 2025; Kewenig et al. 2025). These studies prompted LLMs to provide ratings for single words or multi-word expressions for dimensions such as concreteness, arousal, and familiarity, and compared them to commonly used human-rated datasets. LLMs, despite exhibiting some biases, overall reported good validity.

The effect of psycholinguistic features on language processing becomes even more complex when moving from the study of single words to multi-word expressions. In this case, we see the effect not only of the characteristics of individual words, but also of the entire expression (Arnon and Snider 2010), with relevant features such as syntactic complexity and cloze probability. For figurative expressions, such as metaphors, other additional features come into play, such as familiarity or imageability. Focusing on metaphors (e.g., expressions such as *Lawyers are sharks*), they are a paradigmatic case of non-literal use of language, where interpretation needs to incorporate contextual aspects to go beyond the literally encoded meaning to derive the intended figurative meaning (Wilson and Carston 2007). The role of psycholinguistic features in metaphor processing has been widely explored across a variety of empirical approaches. One of the most consequential variables is metaphor familiarity, which has been reported to influence comprehension speed, with more familiar metaphors leading to shorter response times (Blasko and Briihl 1997). Also, familiarity was shown to affect brain activity, as shown by studies using both Event-Related Potentials, or ERPs (Canal & Bambini, 2023; Coulson, 2008) and functional neuroimaging (Schmidt and Seger 2009; Bambini et al. 2011). For instance, familiar metaphors show a reduced negative deflection of ERPs around 400 *ms* after stimulus presentation - the so-called N400 time window, which reflects early semantic processing (Lai et al. 2009). Another relevant dimension is the involvement of sensorimotor properties, which are known to be activated when processing novel metaphorical meanings (Al-Azary and Katz 2021), with more concrete metaphors eliciting a more negative peak in the N400 window (Canal et al. 2022).

To allow researchers to investigate the patterns of these effects in metaphor processing, several datasets in many languages have been developed (Katz et al. 1988; Bambini et al. 2014; Campbell and Raney 2016; Cardillo et al. 2017; Citron et al. 2020; Huang et al. 2024; Milenković et al. 2024; Bressler et al. 2026), providing up to 1,000 metaphorical expressions, as in the recent Italian *Figurative Archive* (Bressler et al., 2026). However, besides being time consuming (with a number of participants ranging from 60 to 600 subjects, see Katz et al. 1988; Cardillo et al. 2010; Bambini et al. 2014), these datasets remain limited in terms of the number of items compared to the wide range of possible metaphorical sentences, the high variety of syntactic structures in which metaphors can appear (nominal predicative, e.g., *The lawyer is a shark*, predicate, e.g., *He runs away from his problems*, adjectival, e.g., *silky sunsets*), and the many contexts in which they can be embedded. Therefore, it is particularly relevant to explore automatic approaches to generate ratings, which would allow researchers to efficiently and rapidly extract norms for new stimuli suitable for their specific research questions.

Due to their contextualized representations, LLMs have shown good capacity for capturing the context-dependent meaning of figurative expressions, as indicated by their high accuracy (e.g., 78%, reported by Barattieri di San Pietro et al., 2023) when prompted to interpret metaphors (Barattieri di San Pietro et al. 2023; Ichien et al. 2024). This evidence suggests that LLMs could demonstrate acceptable metaphor ratings. Notably, previous studies employed LLMs to generate creativity scores for metaphors and reported strong performance, yet only through fine-tuned models rather than prompting alone (DiStefano et al. 2024). It remains unexplored whether these models can generate ratings via simple prompts in natural language, which is the most accessible means for researchers. Furthermore, a previous study demonstrated that LLMs do not rely on sensorimotor features when producing metaphors (Mangiaterra et al. 2025), highlighting a weakness of models in representing embodied aspects of language, which was also found in metaphor interpretation (Barattieri di San Pietro et al. 2023). Hence, in addition to the question on the accuracy of model performance via simple prompting, it is unclear how LLMs can deal with

certain properties of metaphors, particularly with sensorimotor aspects, given that the integration of multimodal sources of information still represents a challenge for AI models (Barattieri di San Pietro et al. 2023; Chakrabarty et al. 2024).

In this work, we aimed to extend the study of LLMs as rating generators from the domain of single words to metaphors, testing the accuracy of the models as a tool to complement or substitute human ratings in generating norms for metaphorical stimuli. In particular, we tested validity, intended as the capacity to approximate human performance, and test-retest reliability, defined as the stability of the results over time, of psycholinguistic ratings generated by three recent GPT models (GPT3.5-turbo, GPT4o-mini, and GPT4o - Achiam et al., 2024) via prompting. We also tested the relation of machine-generated ratings with human behavioral and electrophysiological responses, which, to the best of our knowledge, is an unexplored domain within the field of LLMs' ratings, both for single words and more complex expressions. Furthermore, we conducted a separate analysis on the familiarity ratings obtained for subsets of metaphors displaying different embodiment features.

Given LLMs' promising performance for single-word ratings, we expect good validity of machine-generated ratings for metaphors as well. However, we expect shortcomings to emerge for metaphors with high sensorimotor load and for the imageability dimension, as these aspects of language are still challenging for LLMs, both for single words and metaphors (Barattieri di San Pietro et al. 2023; Mangiaterra et al. 2025; Xu et al. 2025). Also, we expected poor test-retest reliability, as GPT models have shown patterns of inconsistency in their responses (Khademi 2023, but see Hackl et al., 2023) for more positive results), with high sensitivity to small changes in prompts (Zhuo et al. 2024).

1.2. Methods

1.2.1. Material (human ratings)

As a benchmark for machine-generated ratings, we extracted human ratings for three dimensions (familiarity, imageability, and comprehensibility) for 687 metaphors, of which 469 are in Italian from the *Figurative Archive* (Bressler et al., 2026), originally employed in five studies (Bambini et al. 2013, 2014, 2024; Canal et al. 2022; Bressler et al. 2026), and 218 are in English from three original studies (Campbell and Raney 2016; Al-Azary and Buchanan 2017; Cardillo et al. 2017). We also retrieved ratings for 48 anomalous and 168 literal statements in English and for 48 anomalous and 48 literal statements in Italian. All rating studies were acquired from samples of university students, native speakers of the language under consideration. Familiarity ratings were available for seven studies (639 metaphors), imageability ratings for two studies (178 metaphors), and comprehensibility ratings for one study (48 metaphors). Most of the studies collected ratings on a 7-point Likert scale, while Bambini et al. (2013; 2014) employed a 5-point Likert scale and Al-Azary & Buchanan (2017) a 6-point Likert scale. When needed to compute the overall correlations, ratings were standardized to a 7-point Likert scale. A summary of the material is reported in Table 1.1, with a specification of which datasets (items and ratings) were available online before GPT models' knowledge cutoff (October 2023). More information on the participants in each study can be found in Appendix A, Supplementary Table 1.1.

Table 1.1. Summary of the datasets used in the study.

Measure	Definition	Language	Studies	N° item	Form
Familiarity	Frequency of experience of the expression	English	Campbell & Raney (2016)* Cardillo et al. (2017)*	170 metaphors 120 literals	(adj) X is (adj) Y *The incriminating files were a poison arrow.*
		Italian	*Figurative Archive*, originally Bambini et al. (2013*, 2014*, 2024); Canal et al. (2022); Bressler et al. (2026)	469 metaphors 46 anomalous 46 literals	X is Y(Bambini et al. 2013; Canal et al. 2022; Bressler et al. 2026) *Actors are masks* X – Y(Bambini et al. 2024) *Language - bridge* X of Y(Bambini et al. 2014) *Fog of melancholy*
Imageability	Ease with which each expression evoked a visual mental image	English	Campbell & Raney (2016)*	50 metaphors	X is Y *Wound is a fjord*
		Italian	*Figurative Archive*, originally Bambini et al. (2024)	128 metaphors	X – Y *Language - bridge*
Comprehensibility	How suitable or natural the expression is	English	Al-Azary & Buchanan (2017)	48 metaphors 48 anomalous 48 literals	X is Y *Sarcasm is a knife*

Note: Items and ratings from studies marked with * were available online before GPT models' knowledge cutoff.

Among metaphors rated for familiarity, 308 could be split into different subsets according to their sensorimotor properties and were used to compare the performance of LLMs for metaphors with varying embodiment features. Specifically, these included 62 metaphors classified as mental, i.e., where the relation between the concepts is based on psychological characteristics, e.g., *I genitori sono scudi* (Eng. Tr.: "Parents are shields") and 62 metaphors classified as physical, i.e., where the relation is based on physical characteristics, e.g., *Certi cantanti sono usignoli* (Eng. Tr.: "Some singers are nightingales"), from Canal et al. 2022); 60 metaphors based on motion words, e.g., *His thoughts*

were a pendulum, and 60 metaphors based on auditory words, e.g., *Her chores were a sad tune* from Cardillo et al. (2017); 32 metaphors with topics referring to body parts, e.g., *Quei bicipiti sono sassi* (Eng. Tr.: "Those biceps are stones") and 32 metaphors describing objects, e.g., *Quella casa è un gioiello* (Eng. Tr.: "That house is a jewel") from the IUSS NEPLab MetaBody study (Bressler et al., 2026).

Moreover, we retrieved behavioral responses (response times) for 64 metaphors from the IUSS NEPLab MetaBody study and electrophysiological responses (EEG) for 252 metaphors from Canal et al. (2022) and Bambini et al. (2024), to investigate the validity of machine-generated ratings with respect to human cognitive and neural processing measures. Canal et al. (2022) and Bambini et al. (2024) explored metaphor processing through Event Related Potentials and included mean EEG amplitudes for each metaphor in the N400 time window for frontal and centro-parietal electrodes. The IUSS NEPLab MetaBody study (Bressler et al., 2026) investigated metaphor processing through a sensicality task (participants were asked to say whether the expression made sense in a time-constrained setting) and included response times (RTs) for each metaphor.

1.2.2. Models

We prompted three GPT models (GPT3.5-turbo; GPT4o-mini; GPT4o - OpenAI et al. 2024) through the API and one GPT model (GPT4o-mini, as it was the only one freely accessible at the time of data collection) through the ChatGPT interface. Models were prompted in June 2025.

While acknowledging the limitations of using closed-source LLMs (Manchanda et al., 2025; Ravelli & Bolognesi, 2024), our choice was motivated by their extensive use in studies concerning both pragmatic abilities in LLMs (Hu et al. 2022; Barattieri di San Pietro et al. 2023) and the integration of LLMs in scientific pipelines as raters (Gilardi et al. 2023; Trott 2024a; Martínez et al. 2024b;

Brysbaert et al. 2024), as well as for their superior performance, compared to other LLMs, on rating tasks and metaphor applications (Fuoli et al., 2025; Q. Xu et al., 2025).

1.2.3. Prompting Procedure

To generate ratings, we prompted the models with the same instructions given to human participants in each original study, with minimal adaptation. Slight modifications with respect to the original instructions were introduced to remove any text referring to practical aspects of the experiments (such as the key to press to continue) and to add the specification to limit the answer to the rating value, given the tendency of the models to provide verbose answers. Building upon previous studies (Gilardi et al. 2023; Trott 2024a), we intentionally avoided using any prompt-engineering techniques tailored specifically to GPT models to ensure both the comparability of responses between GPT models and human participants and the reproducibility of findings for psycholinguistic research. An example of a prompt is provided in Appendix A, Supplementary Table 1.2. The prompting procedure was performed through both the API and the ChatGPT web interface, in two independent sessions for each model.

1.2.3.1. API parameters

Parameters were set to optimize performance, following previous work. Temperature, the parameter modulating the degree of determinism in models' behavior, was set at 0 to reduce randomness in the output and ensure consistent responses (Binz and Schulz 2023; Kosinski 2024; Xu et al. 2025). In addition to the instructions in the prompt to answer with only the rating value, to further limit verbosity, the maximum token number was set at 1. The number of top-k most likely tokens to return was set at 3. This was done with the aim of computing an overall rating, resulting from the combination of the rating token and the token probability. The overall rating, in addition to providing a more precise estimate (Martínez et al. 2024b), could mirror the

continuous nature of human ratings, which were obtained by averaging across participants. So, exploiting the API's possibility to extract the associated log probabilities for the most likely tokens (Hill and Abadkat 2023), we derived the overall rating by weighing each of the three most likely ratings for their log probabilities. For example, for the metaphor "Evolution is a lottery", GPT3.5-turbo provided the top three most likely outputs "2", "3", and "1" with log probabilities of 0.768, 0.195, and 0.037, respectively, resulting in an overall rating of 2.158.

1.2.4. Statistical Analysis

To assess the validity of machine-generated ratings, namely the possibility of approximating the human gold standard, we computed a correlation analysis and a substitution analysis.

First, we computed Spearman correlations between human-generated ratings and machine-generated ratings for all items (687 metaphors, 214 literal statements and 94 anomalous statements) separately for each dimension (familiarity, imageability, and comprehensibility) and for the two languages. Then, we computed separate Spearman correlations for each of the subsets of metaphors characterized by different sensorimotor load: mental and physical in Canal et al. (2022), auditory and motion in Cardillo et al. (2017), and body-related and object-related in the IUSS NEPLab MetaBody study (Bressler et al., 2026).

Second, we tested whether machine-generated metaphor ratings hold the same explanatory power as human-generated ratings in predicting human behavioral and ERP responses as recorded in the three studies with RT and ERP measures. To do so, in line with Trott (2024a), we replicated the statistical analysis of the three original studies (Canal et al. 2022; Bambini et al. 2024; Bressler et al. 2026) and substituted human-generated ratings with machine-generated ratings. Finally, we compared the models' goodness-of-fit in terms of Akaike Information Criterion (Bozdogan 1987) and R squared.

Specifically, to replicate the analysis predicting response times in the IUSS NEPLab MetaBody study (Bressler et al., 2026), we fitted separate Linear Mixed-Effects Models using *lme4* and *lmerTest* packages (Bates et al. 2015) for each of the GPT models and for human familiarity ratings, considering human response times from the original study as the dependent variable and GPT or human-generated ratings of familiarity as continuous predictors. Then, we also considered familiarity ratings for the two subsets of metaphors (body-related and object-related) separately.

To replicate the analysis predicting the N400 amplitude in Bambini et al. (2024), we first fitted separate Linear Mixed-Effects Models using *lme4* and *lmerTest* packages (Bates et al. 2015) for each of the GPT models and human familiarity and imageability ratings, considering the ERP response (both in frontal and centro-parietal electrodes) in the N400 window as the dependent variable and human and GPT-generated metaphor ratings of familiarity and imageability as continuous predictors. Then, we substituted GPT-generated familiarity and imageability separately in a more complex Linear Mixed-Effects Model, used in the original study, which also included a series of other human ratings (i.e., metaphoricity, semantic distance, number and strength of metaphorical interpretation).

To replicate the analysis predicting the N400 amplitude in Canal et al. (2022), we first assessed the effect of metaphor familiarity on the ERP response, by fitting separate Linear Mixed-Effects Models using *lme4* and *lmerTest* packages (Bates et al. 2015) for each of the GPT models and human familiarity ratings, with EEG amplitude as dependent variable and ratings of familiarity as continuous predictor. Then, we substituted GPT-generated familiarity in the Linear Mixed-Effects Models from the original study, which included a series of other ratings and subjects' task scores (i.e., word frequency and subjects' score at two Theory of Mind tasks).

To assess the reliability of machine-generated metaphor ratings, we computed Spearman correlations between machine-generated ratings obtained in two independent sessions (each following the same prompting procedure described in Section 2.3.).

Finally, in line with Trott (2024), we conducted an exploratory analysis to examine where models differ most from human ratings, operationalized as absolute error between human and GPT ratings. We fitted a Linear Mixed-Effects Model using *lme4* and *lmerTest* packages (Bates et al. 2015), with absolute error as the dependent variable and the original human ratings as a predictor in interaction with dimension (familiarity, imageability, and comprehensibility) and GPT model (GPT3.5-turbo, GPT4o-mini, GPT4o), to investigate if the GPT models report larger errors for high (or low) human ratings on a certain dimension, and the potential impact of the model used to elicit the ratings.

All analyses were performed in R (R Core Team 2025).

1.3. Results

1.3.1. Correlation analysis

For metaphor familiarity ratings, we found positive correlations, with coefficients ranging from 0.50 to 0.64 for English and from 0.20 to 0.65 for Italian. The larger model (GPT4o) showed the best performance compared to the other two, exhibiting a strong correlation with humans both for Italian and English metaphors. Results obtained with smaller models (GPT3.5-turbo and GPT4o-mini) were comparable to GPT4o for English, with moderate-to-strong correlations, while falling behind the larger model for Italian, with weak-to-moderate correlations.

In the imageability dimension, moderate correlations were obtained for English metaphors, with coefficients ranging from 0.37 to 0.56, while moderate-to-strong correlations were obtained for Italian metaphors, with coefficients ranging from 0.38 to 0.65.

For comprehensibility ratings, which were available for English metaphors only, we found strong positive correlations between human-generated and machine-generated ones, with coefficients ranging from 0.69 to 0.79.

Table 1.2 displays all correlations between human and machine-generated ratings for the three dimensions for metaphors (for more details on the results for the single studies, see Appendix A, Supplementary Table 1.3).

To better contextualize these results and provide an upper bound on achievable model performance, we report the reliability of the human ratings from Cardillo et al. (2017), which were the only dataset reporting this measure. They reported high inter-rater reliability across dimensions, with intraclass correlation coefficients (ICCs) ranging from .857 to .975. These values indicate that the measures are highly consistent across participants and therefore impose only a limited constraint on the maximum attainable model–data correspondence.

Table 1.2. Correlations between human and machine-generated metaphor ratings for the three dimensions and the two languages.

Measure	**Language**	**GPT3.5-turbo**	**GPT4o-mini (API)**	**GPT4o-mini (ChatGPT)**	**GPT4o**
Comprehensibility	English	0.69***	0.74***	0.78***	0.79***
Imageability	English	0.39**	0.42**	0.56***	0.37**
	Italian	0.38***	0.46***	0.45***	0.65***
Familiarity	English	0.64***	0.56***	0.50*	0.61***
	Italian	0.20***	0.42***	0.20**	0.65***

All models strongly aligned with humans when rating English literal and anomalous statements for comprehensibility (all *r*s > 0.96) and Italian literal and anomalous statements for familiarity (*r*s ranging from 0.82 to 0.92), and they moderately aligned when rating English literal statements for familiarity (*r*s ranging from 0.53 to 0.63). A complete report of correlations for literal and anomalous statements can be found in Appendix A, Supplementary Table 1.4.

The distribution of machine-generated and human-generated ratings across studies is displayed in Figure 1.1 (see Appendix A, Supplementary Figure 1.1 for density plots of single studies).

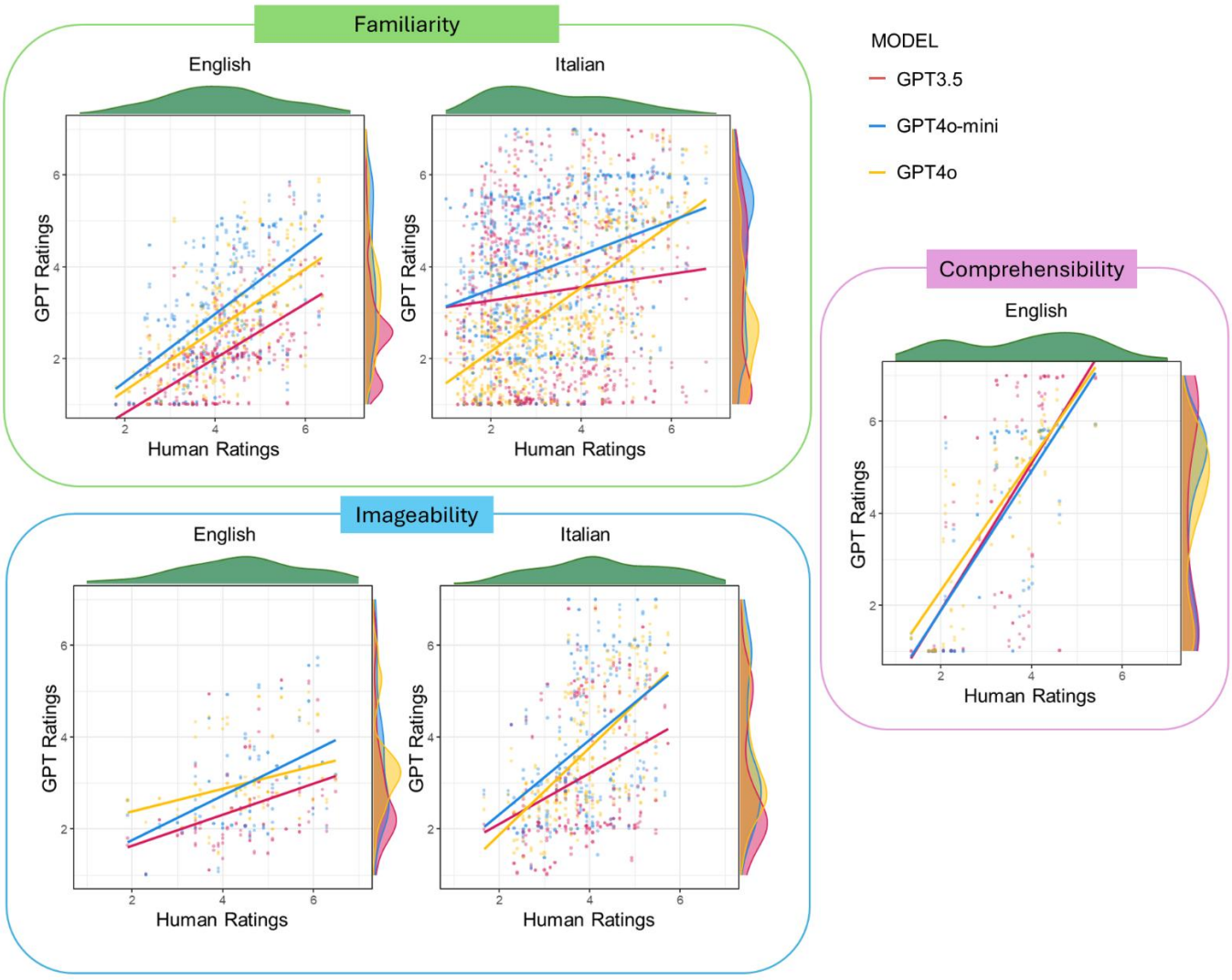


Figure 1.1. Distribution of metaphor ratings. The figure shows the relationship between GPT ratings (y-axis) and human ratings (x-axis) across three dimensions: Imageability (blue panel), Comprehensibility (purple panel), and Familiarity (green panel), for English and Italian metaphors, with marginal density plots illustrating rating distributions. Regression lines are color-coded by model: GPT3.5-turbo (red), GPT-4o-mini (teal), and GPT-4o (orange).

3.3.1.4. Correlations between familiarity ratings for metaphors characterized for sensorimotor properties

Correlation analysis (Figure 1.2) comparing human and machine-generated ratings of familiarity for subsets of metaphors characterized by different sensorimotor properties showed that, for English, ratings for motion metaphors reported strong correlations (GPT3.5: $r = 0.67$; GPT4o-mini: $r = 0.72$; GPT4o: $r = 0.71$), while ratings of auditory metaphors reported moderate to strong

correlations (GPT3.5: $r = 0.61$; GPT4o-mini: $r = 0.55$; GPT4o: $r = 0.56$). For Italian, GPT4o-mini and GPT4o showed moderate-to-strong correlations when generating familiarity ratings for object-related metaphors (GPT3.5: $r = 0.09$ (*ns*); GPT4o-mini: $r = 0.51$; GPT4o: $r = 0.71$) and weak-to-strong correlations when rating body-related metaphors (GPT3.5: $r = -0.16$; GPT4o-mini: $r = 0.37$; GPT4o: $r = 0.65$). Familiarity ratings for mental metaphors spanned from weak to strong correlations (GPT3.5: $r = 0.24$; GPT4o-mini: $r = 0.60$; GPT4o: $r = 0.77$), while physical metaphors reported weak-to-moderate performance (GPT3.5: $r = 0.29$; GPT4o-mini: $r = 0.54$; GPT4o: $r = 0.60$). Overall, numerically higher correlations were obtained by metaphors based on motion words, object-related metaphors, and mental metaphors, rather than metaphors based on auditory words, body-related metaphors, and physical metaphors.

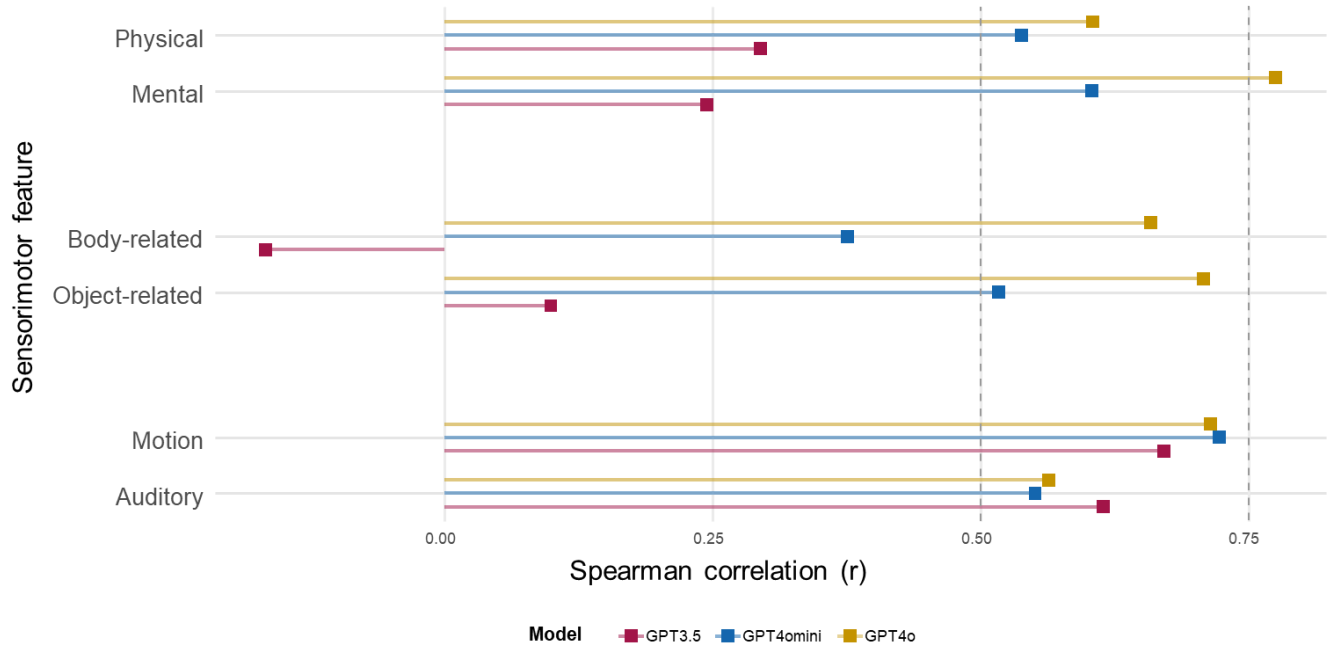


Figure 1.2. Correlation between human and GPT ratings of familiarity for metaphors characterized by sensorimotor load. Results of the correlation analysis between human-generated and machine-generated familiarity ratings for subsets of metaphors with different types of sensorimotor load (mental and physical from Canal et al. (2022); motion and auditory from Cardillo et al. (2017); object-related and body-related from IUSS NEPLab MetaBody study) for the three GPT models (GPT3.5-turbo, GPT4o-mini, GPT4o).

1.3.2. Substitution analysis

1.3.2.1. Response Times

Linear Mixed-Effects Models testing the effect of familiarity on response times (data from the IUSS NEPLab MetaBody study) showed a significant effect of machine-generated ratings for both GPT4o-mini prompted through the API (β = -0.044, t = -5.46, p < .001) and GPT4o (β = -0.038, t = -6.92, p < .001), with higher values of familiarity associated with shorter reaction times. This pattern mirrored what we observed in the statistical model with human-generated ratings (β = -0.046, t = -7.09, p < .001). Explained variance was comparable across models (R^2 = 0.30), yet the model with human familiarity had the best goodness of fit (Human: AIC = 1329, GPT4o: AIC = 1331, GPT4o-mini: AIC = 1347). No effect was found for GPT3.5-turbo and GPT4o-mini prompted through the ChatGPT interface (Figure 1.3a).

Looking at the two subsets of metaphors (body-related and object-related) available in the original study (Bressler et al., 2026), we found that human familiarity and GPT4o familiarity predicted RTs both for body-related (Human: β = -0.051, t = -4.58, p < .001; GPT4o: β = -0.044, t = -4.50, p < .001) and object-related metaphors (Human: β = -0.041, t = -3.34, p = .002; GPT4o: β = -0.056, t = -4.03, p < .001), while GPT4o-mini familiarity only predicted RTs for object-related metaphors (β = -0.040, t = -3.79, p < .001). Again, no effect was found for GPT3.5-turbo and GPT4o-mini prompted through the ChatGPT interface.

3.3.2.2. EEG response

The Linear Mixed-Effects Models examining the effect of familiarity on the N400 amplitude in centro-parietal electrodes (data from Bambini et al., 2024) showed a significant effect of machine-generated ratings (Figure 1.3b) for GPT3.5-turbo (β = 0.55, t = 2.84, p = .005), GPT4o-mini prompted through the API (β = 0.40, t = 2.06, p = .041) and GPT4o (β = 0.57, t = 3.03, p = .003).

No effect was found for GPT4o-mini prompted through the ChatGPT interface (p = .098). We found that more familiar metaphors were associated with reduced negativity, as observed for human-generated familiarity (β = 0.95, t = 3.72, p < .001) in the original study (Bambini et al. 2024). All models explained a comparable portion of variance (R^2 = 0.13), and AIC comparison indicated that the model with human familiarity provided the best fit (AIC = 9698.41), although differences in AIC across models were relatively modest (GPT4o: AIC = 9702.48).

Human familiarity significantly predicted EEG amplitude in frontal electrodes as well; however, we did not find an effect of machine-generated familiarity for those scalp locations.

The Linear Mixed-Effects Models examining the effect of imageability on EEG amplitude showed no significant effect of machine-generated ratings in the N400 window for either centro-parietal and frontal electrodes, contrasting with a significant effect of human familiarity in both areas (frontal: β = 0.66, t = 2.06, p =.04; centro-parietal: β = 0.74, t = 2.66, p < .01) in the original study (Bambini et al. 2024).

When considering the more complex Linear Mixed-Effects Models used in the original study (Bambini et al. 2024), we found an effect of familiarity for GPT3.5-turbo (β = 0.42, t = 2.53, p = .013) and GPT4o (β = 0.38, t = 1.99, p = .049) in the same direction as the human-generated familiarity (β = 0.83, t = 2.28, p = .024), again for centro-parietal electrodes only. All models explained a comparable portion of variance (R^2 = 0.14), and the model with familiarity by GPT3.5-turbo had the best goodness of fit (AIC = 19714), yet close to the other models (Human: AIC = 19715; GPT4o: AIC = 19716). Familiarity generated with GPT4o-mini did not significantly predict EEG response in this more complex statistical model. Following Trott (2024a), we checked the direction of the effect for this model, and, even if not significant, machine-generated familiarity showed the same direction as human-generated familiarity. Neither human nor machine-generated imageability reached significance in the more complex statistical model.

For the metaphors from Canal et al. (2022), human familiarity did not significantly predict the EEG amplitude in the N400 window in the original study, and, in line with that, we did not find an effect of machine-generated familiarity as well.

To sum up, machine-generated familiarity ratings, especially from GPT4o, can predict N400 amplitude in centro-parietal electrodes, aligning with human ratings, while no machine-generated imageability ratings reported a significant effect, despite human imageability ratings being predictive.

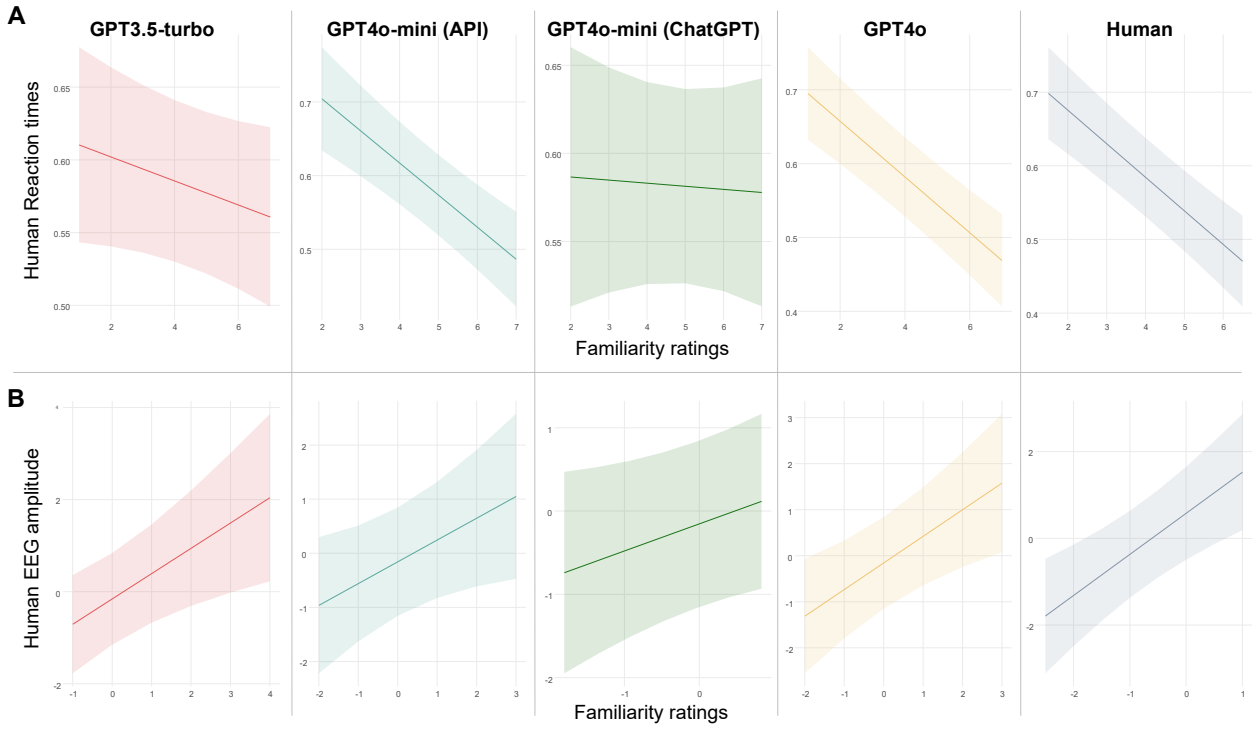


Figure 1.3. Results of the substitution analyses. Panel A shows the effect of metaphor familiarity ratings generated by GPT3.5-turbo, GPT4o-mini, prompted through the API and ChatGPT, GPT4o, and human participants on Response Times (RTs). Panel B shows the effect of human and machine-generated (GPT3.5-turbo, GPT4o-mini through the API and ChatGPT interface, and GPT4o) metaphor familiarity ratings on the amplitude of the EEG response in the N400 window for centro-parietal electrodes.

3.3.3. Reliability

The correlations between machine-generated metaphor ratings from two independent sessions showed very high reliability for the three models prompted through the API in all studies, with all correlation coefficients above 0.90 (Table 3). Ratings collected through the ChatGPT interface reported moderate to high reliability, with correlation coefficients ranging from 0.66 (for imageability ratings of English metaphors) to 0.91 (for comprehensibility ratings of English).

Table 1.3. Correlations between GPT-generated metaphor ratings obtained in two independent sessions for the three dimensions.

Measure	**Language**	**GPT3.5-turbo**	**GPT4o-mini (API)**	**GPT4o-mini (Interface)**	**GPT4o**
Familiarity	English	0.99***	0.99***	0.68***	0.98***
	Italian	0.98***	0.99***	0.83***	0.98***
Imageability	English	0.99***	0.98***	0.66***	0.97***
	Italian	0.99***	0.99***	0.67***	0.98***
Comprehensibility	English	0.99***	0.99***	0.91***	0.99***

A visual comparison between the reliability of the two prompting methods is shown in Figure 1.4. For more details on the results for the single studies, see Appendix A, Supplementary Table 1.5.

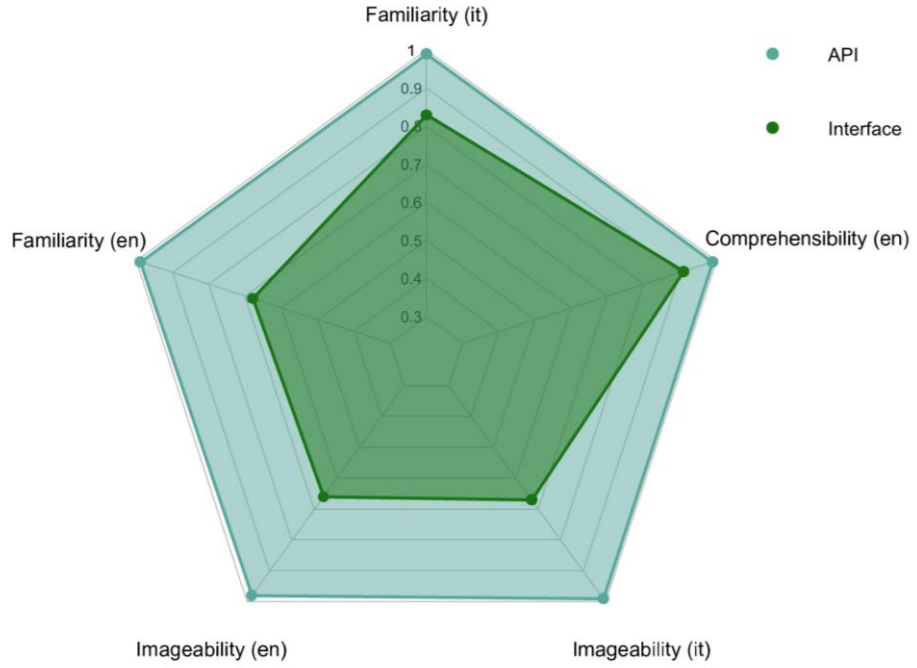


Figure 1.4. Reliability comparison between ChatGPT interface and API. Visual comparison of the reliability of machine-generated metaphor ratings obtained from GPT4o-mini prompted through the API or the ChatGPT interface.

1.3.4. Exploratory source of error analysis

The exploratory analysis of systematic sources of error aimed at testing where human and machine-generated ratings differ. It revealed a main effect of the original human ratings for the metaphors on the absolute error between human and GPT ratings (β = 0.37, t = 13.11, p < .001), with metaphors rated higher by humans associated with higher error. Both the GPT models used to generate ratings and the psycholinguistic dimensions moderated the effect of the original ratings. Specifically, we observed significant interactions between original human ratings and GPT model (GPT4o: β = –0.14, t = -4.28, p < .001; GPT4o-mini: β = –0.31, t = -9.46, p < .001), suggesting that while GPT3.5-turbo aligned less with humans when generating ratings for more familiar, imageable and comprehensible metaphors, more advanced GPT models are less impacted by the original ratings of the metaphor. Trend analyses using the *emmeans* package further supported this

finding: the slope of the relationship between human ratings and error was steepest for GPT3.5-turbo ($\beta = 0.37$) and significantly flatter for GPT4o ($\beta = 0.23$) and GPT4o-mini ($\beta = 0.06$).

We also found that the relationship between human ratings and error varied depending on the psycholinguistic dimension, as emerged from the significant interactions between original human ratings and both familiarity ($\beta = 0.15$, $t = 2.72$, $p < .01$) and imageability ($\beta = 0.22$, $t = 3.52$, $p < .001$). Estimated slopes showed that the error increased more steeply with human ratings in the imageability dimension ($\beta = 0.33$), followed by familiarity ($\beta = 0.25$), and was the weakest for comprehensibility ($\beta = 0.10$). Pairwise comparisons confirmed that the slope for comprehensibility was significantly smaller than both familiarity ($\Delta = –0.15$, $p = .018$) and imageability ($\Delta = –0.22$, $p = .001$), whereas familiarity and imageability did not significantly differ ($\Delta = –0.08$, $p = .24$). The GPT models (and especially GPT3.5-turbo) showed lower alignment with humans when generating ratings for high imageable and high familiar metaphors (Figure 1.5).

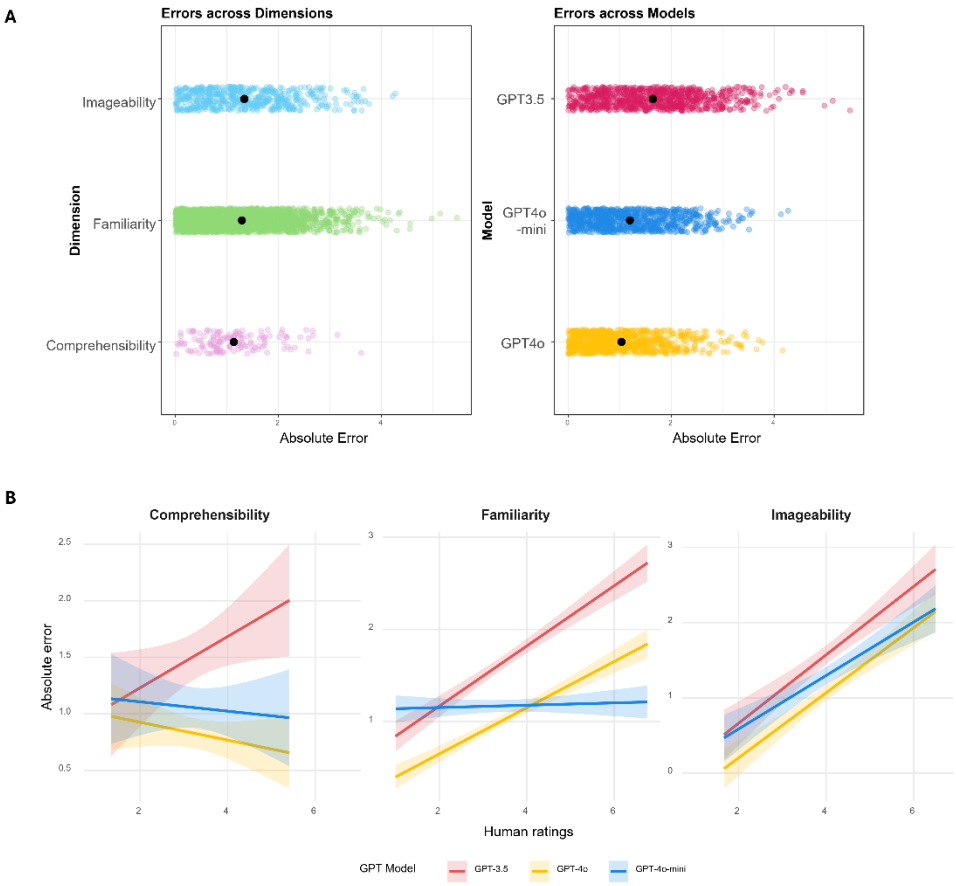


Figure 1.5. Absolute error between GPT and human metaphor ratings. Panel A shows the distribution of errors for each linguistic feature (Familiarity, Imageability, and Comprehensibility) and each model (GPT-3.5, GPT-4o-mini, GPT-4o). Black points indicate the mean error. Panel B shows absolute error between GPT and human metaphor ratings as a function of original human ratings. Each subpanel shows one dimension (Comprehensibility, Familiarity, Imageability), and each line represents a different GPT model (GPT3.5-turbo, GPT4o, GPT4o-mini).

1.4 Discussion

The increasing employment of LLMs as annotators or research assistants in the experimental pipeline calls for a systematic assessment of their trustworthiness to build guidelines for researchers that highlight the benefits as well as the potential risks associated with their integration in research. In this study, we extended the line of research investigating the viability of LLMs as raters for norming stimuli in psycholinguistic studies, examining the case of complex figurative expressions such as metaphors. Specifically, we evaluated the validity and reliability of machine-generated

ratings across three psycholinguistic dimensions, namely familiarity, imageability, and comprehensibility for metaphorical expressions in English and Italian. To generate ratings, we prompted three GPT models (GPT3.5-turbo, GPT4o-mini, and GPT4o) with the original instructions given to human participants. Models were prompted through both the API and the ChatGPT web interface, allowing for a comparison of prompting settings. To assess performance, we checked whether machine-generated ratings correlate with human ratings and whether they hold the same explanatory power as human ratings in predicting human behavioral and electrophysiological responses. Given the non-converging evidence on the consistency of LLMs' output (Khademi 2023; Hackl et al. 2023), we also examined the stability of ratings across separate sessions.

Our results showed that machine-generated ratings can largely approximate human ratings of metaphors, with high positive correlations emerging for all three dimensions. The larger and most recent model GPT4o reported the highest validity, both for Italian and English, obtaining a strong alignment with human ratings in both languages. Smaller models (GPT3.5-turbo and GPT4o-mini) showed good performance in rating English metaphors, but a sparser pattern of results emerged for Italian metaphors, with two datasets not reporting any association between machine and human ratings. The lower performance of GPT3.5-turbo on Italian is in line with similar findings observed for languages other than English for less recent models, whose training sets were heavily based on English (Rathje et al. 2024).

One issue to be considered when reporting these results is data contamination (Conde, Grandury, et al., 2025; Ravelli & Bolognesi, 2024), namely, whether the good performance obtained by GPT models is linked to the presence of the datasets under consideration in their training set. Given that GPT models' training data has not been released, we could not conclude that some of the datasets in our study could be part of it. However, four out of eight datasets (364 metaphors out of 687) were not available online before the models' knowledge cutoff and GPT-generated ratings

from these datasets consistently reported comparable, when not superior, alignment with human participants then datasets publicly available before the models' knowledge cutoff (see Supplementary Table 1.3). This supports the soundness of our results despite the possible data contamination issue for a portion of our materials.

The validity of GPT-generated metaphor ratings is further supported by the substitution analyses. Indeed, when put in relation to human behavioral and electrophysiological responses, machine-generated familiarity (except ratings obtained through the ChatGPT web interface) predicted response times and EEG amplitude comparable to human-generated familiarity ratings, demonstrating the possibility of substituting human ratings also in complex statistical analyses to model processing patterns. In the literature, many studies reported significant associations of the measure of lexical probability (or surprisal) from LLMs with neural and behavioral measures of processing costs (de Varda et al. 2023; Michaelov et al. 2024). Even if measures generated via prompting are known to be less reliable than measures derived by accessing internal states of LLMs, such as lexical probability (Hu and Levy 2023), our results support the validity of GPT-generated ratings in terms of ability to capture processing mechanisms, opening the possibility to model behavioral and brain data based on a series of other relevant dimensions – figurative familiarity and comprehensibility - rather than only widely used lexical surprisal.

Contrary to our predictions, machine-generated metaphor ratings reported excellent reliability for models prompted through the API. However, the model prompted through the ChatGPT interface showed a sparser pattern of reliability, highlighting the importance of controlling parameters, such as temperature, to ensure a more stable and deterministic behavior, a step possible only when prompting the models through the API.

While these results confirm the promising use of LLMs as raters reported for single words (Trott 2024a; Brysbaert et al. 2024) and fixed multi-word expressions (Martínez et al. 2024b), a number of weaknesses emerged that deserve special consideration. First, as highlighted by the error

analysis, higher misalignment between humans and machines emerged for more familiar metaphors. While humans spanned across all the values in the scale and assigned high ratings to conventional metaphors by treating them as fully acceptable expressions, models assigned lower values to conventional figurative expressions, maintaining a clearer boundary between metaphorical and literal language. This might suggest that models adopt a more rigid distinction between literal and figurative, being less sensitive than humans to the shades of this continuum as a function of conventionality (Bowdle and Gentner 2005; Mashal and Faust 2009; Sperber and Wilson 2012) and more to regularization (Ilievski et al. 2025). The excellent performance obtained by the models for literal and anomalous statements is in harmony with this view, indicating the models' ability in recognizing the two extremes of the scale of sense. Future research could further test this speculation by assessing, for instance, the validity of LLM-generated ratings for highly conventional yet figurative expressions such as idioms (for initial evidence, see O'Reilly et al., 2025).

Perhaps the most relevant limitation of LLMs as metaphor raters regards their ability to capture the embodied aspects of meaning. This emerged both in the analysis on the subset of metaphors with different sensorimotor features and in the error analysis. In the former case, models showed strong performance in rating familiarity for mental metaphors but only moderate performance for physical metaphors, suggesting a lack of perceptual experience hampers a closely human-like representation of meaning. Within concrete metaphors, the models showed strong performance for the motion and the object-related items and moderate-to-strong performance for the auditory and body-related items, showing that when the perceptual features are more represented in the lexicon (see Winter et al., 2018 for the greater prevalence of vision words in the lexicon compared to auditory), models can align better to human representations. In the error analysis, models exhibit low alignment with humans when providing ratings for imageability, where greater error between humans and machines was reported for highly imageable metaphors. Overall, these results are in line with evidence from multiple studies, which found that LLMs have impoverished

representations of sensorimotor aspects of language (Conde et al. 2025a; Xu et al. 2025), as they rely on linguistic more than sensorimotor features of words (Mangiaterra et al. 2025; Lee et al. 2025) and provide more accurate interpretations of metaphorical expressions that do not require embodied simulations (Barattieri di San Pietro et al. 2023, but see Wicke, 2023 for diverging evidence). This experimental evidence aligns with theoretical claims that identify the lack of grounding as one of the major points of distance between humans and LLMs (Borghi et al. 2023; Chemero 2023, but see Pavlick, 2023 for an alternative perspective).

Beyond these aspects, some other limitations should be considered when using LLMs to generate ratings. First, even if we can simulate the continuous nature of the ratings with an *ad-hoc* setting of hyperparameters, LLMs at this point can only approximate an average human participant, or the *wisdom of the crowd* (Trott 2024b). This does not allow for focus on individual variability, for which the recruitment of human participants is still essential (Qiu et al. 2025). Second, the average human participant that LLMs mimic is representative of the perspective of only a certain demographic. Casola et al. (2024) found that LLMs align with the perspective of young participants, while Martínez et al. (2024a) reported that LLMs align less with children's and extra-European participants' ratings. Given that this is a limitation of existing human-normed datasets as well, as most participants of rating studies were university students (Bressler et al., 2026), our results indicate that on the one hand, LLMs can approximate existing datasets by aligning with their predominant demographic, but on the other hand, continue to overrepresent certain groups to the disadvantage of less investigated samples of participants (Wang et al. 2025). Following promising results showing that LLMs could mirror different demographics when prompted to impersonate different types of participants (Puccetti et al., 2025), future directions could include the evaluation of machine-generated ratings mimicking children or older adults' ratings.

Given the limitations above, and capitalizing on the experience gained in this study, in Table 1.4 we provide a summary of recommendations to guide psycholinguists towards a careful and evidence-based integration of LLMs into their experimental pipelines.

Table 1.4. Summary of recommendations for the use of LLMs to generate metaphor ratings.

Models and parameters	- Prefer larger LLMs, such as GPT4o. - Access LLMs through APIs, carefully setting hyperparameters to ensure reliability (e.g., temperature set at 0). - Compute overall scores by incorporating log probabilities to obtain continuous ratings. - Check the performance of LLMs in the language of interest (e.g., check that training data contained text in that language).
Type of stimuli	- Prefer non-conventional items. - Prefer items with low sensorimotor load. - If available, generate ratings for anomalous and literal statements as benchmarks.
Type of linguistic features	- Prefer dimensions based on occurrence, such as familiarity and comprehensibility, rather than embodiment, such as imageability.
Prompt	- Use a prompt as close as possible to human instructions. - Limit models' verbosity by requiring only the rating as output.

1.5. Conclusion

The integration of large language models (LLMs) into psycholinguistic research, and cognitive science more broadly, has generated considerable debate (Dillion et al. 2023; Abdurahman et al. 2024; Bisbee et al. 2024; Harding et al. 2024) regarding the extent to which human data can be

replaced or augmented by artificial intelligence. Following the question posed by Dillion et al. (2023), "Can AI models replace human participants?", we argue that while humans must remain the primary subjects of investigation when studying how metaphorical expression are processed, LLMs can serve as valuable tools to augment human data, particularly in those stages of the experimental pipeline that precede analyses of human processing, such as collecting ratings for stimuli. This could enable the creation of larger and more diverse materials, including in languages beyond English, allowed by larger and multilingual models, and supporting further research on metaphor and human language processing.

Appendix A.

These supplementary materials contain:

i) Information about participants of each rating study

ii) Examples of prompt provided to GPT models

iii) Validity of GPT-generated metaphor ratings presented separately for each study

iv) Validity of GPT-generated ratings for anomalous and literal statements

v) Reliability of GPT-generated metaphor ratings presented separately for each study

Supplementary Table 1.1. The table shows demographic information on the sample of raters for each study.

Study	Participants
Al-Azary & Buchanan (2017)	52 fluent speakers of English (age > 18)
Bambini et al. (2013)	85 native speakers of Italian (42F; age: M = 26.85, SD = 3.80; education in years: M = 18.02, SD = 2.04)
Bambini et al. (2014)	105 native speakers of Italian (83F; age: M = 23.00, SD = 4.31)
Bambini et al. (2024)	122 native speakers of Italian (68F, age: M = 24.34, SD = 1.97)
Campbell & Raney (2016)	90 fluent speakers of English
Canal et al. (2022)	53 native speakers of Italian (40F; age: M = 23.91, range: 21–32; education in years: M = 15.83, range: 13–18)
Cardillo et al. (2017)	40 native speakers of English (21F; age: M = 21.3, education in years: M = 15.25)
IUSS NEPLab MetaBody study	49 native speakers of Italian (27F; age: M = 27.35, SD = 3.55; education in years: M = 15.82, SD = 2.76)

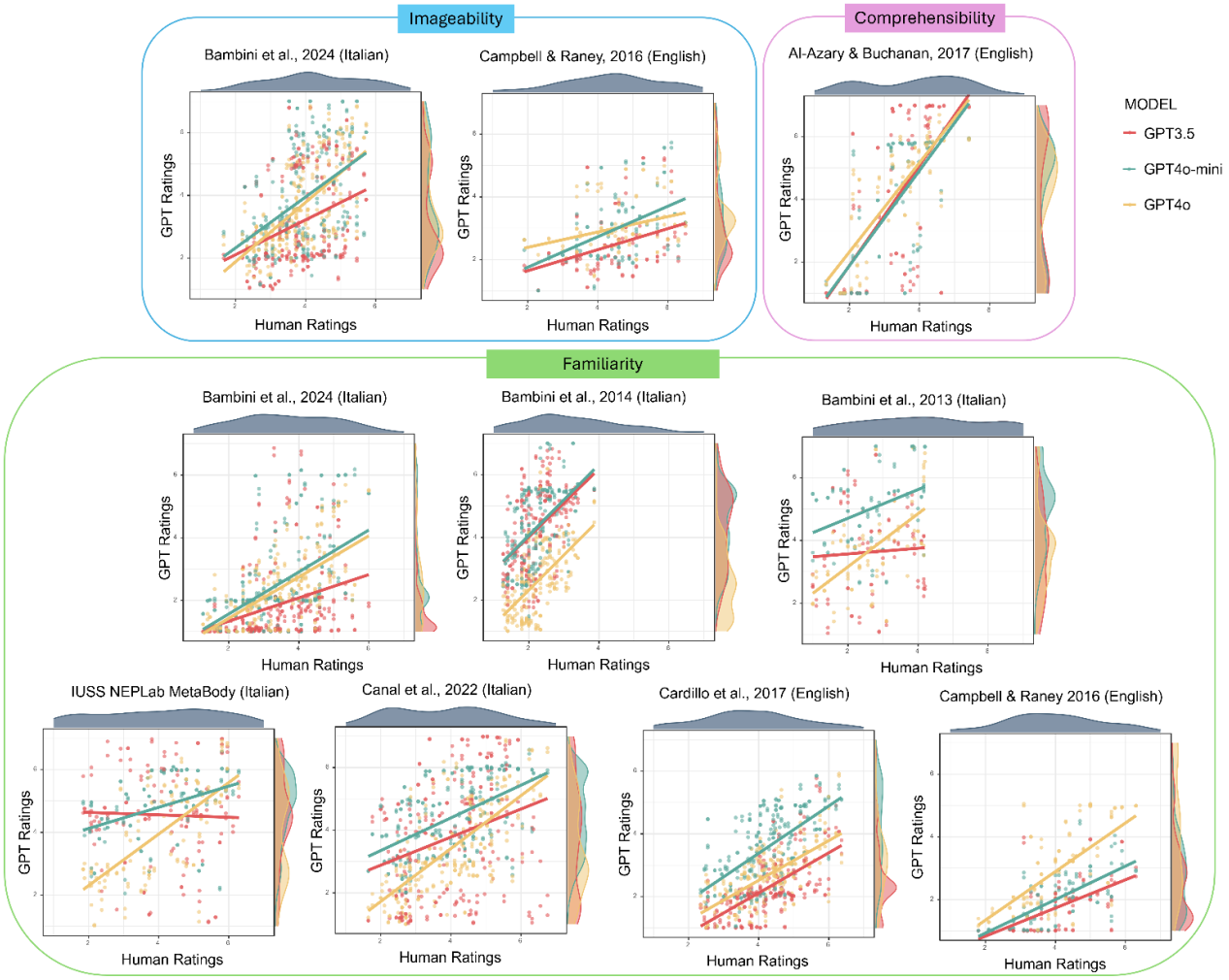
Supplementary Table 1.2. Example of prompt provided to models compared to human instructions.

Instructions to human participants	Prompt to Large Language Models
Your task will be to rate how suitable or natural a series of statements are. The statements will either be nonsensical, literal, or figurative. For example, "A Sheep is a Hill" is a nonsensical statement; "A Circle is a Shape" is a literal statement; "Love is a Journey" is a figurative statement. **Use the number pad to indicate your rating** from 1 to 6 (1 being very unsuitable/unnatural and 6 being very natural/suitable). **Please read the statements carefully before making a response. Press the space bar to begin. There will be three practice trials followed by the rest of the experiment. When you are finished, a thank-you message will appear. Afterward, you may exit the room.**	Your task will be to rate how suitable or natural a series of statements are. The statements will either be nonsensical, literal, or figurative. For example, "A Sheep is a Hill" is a nonsensical statement; "A Circle is a Shape" is a literal statement; "Love is a Journey" is a figurative statement. **Answer with your rating** from 1 to 6 (1 being very unsuitable/unnatural and 6 being very natural/suitable). **Answer with only the number from 1 to 6, do not add more.**

Note: Changes are highlighted in bold.

Supplementary Table 1.3. Validity of GPT-generated metaphor ratings for each study. Correlations between human-generated ratings and ratings generated by GPT3.5-turbo, GPT4o-mini (prompted through API and ChatGPT interface), and GPT4o for the three dimensions (comprehensibility, imageability, and familiarity).

Measure	Language	Study	GPT3.5-turbo	GPT4o-mini	GPT4o	ChatGPT
Comprehensibility	English	Al-Azary & Buchanan (2017)	0.69***	0.74***	0.79***	0.78***
Imageability	English	Campbell & Raney (2016)	0.39**	0.42**	0.36**	0.56***
	Italian	Bambini et al. (2024)	0.38***	0.46***	0.65***	0.45***
Familiarity	Italian	Bambini et al. (2013)	0.02	0.37**	0.56***	0.31*
	Italian	Bambini et al. (2014)	0.50***	0.55***	0.61***	0.29**
	Italian	Bambini et al. (2024)	0.49***	0.57***	0.64***	0.45***
	English	Campbell & Raney (2016)	0.65***	0.67***	0.62***	0.73***
	Italian	Canal et al. (2022)	0.32***	0.54***	0.70***	-0.09
	English	Cardillo et al. (2017)	0.65***	0.62***	0.63***	0.39***
	Italian	IUSS NEPLab MetaBody study	-0.02	0.47***	0.68***	-0.11



Supplementary Figure 1.1. Scatter plots showing the relationship between GPT ratings (y-axis) and human ratings (x-axis) across three dimensions: Imageability (blue panel), Comprehensibility (purple panel), and Familiarity (green panel), in Italian or English, with marginal density plots illustrating rating distributions. Regression lines are color-coded by model: GPT-3.5 (red), GPT-4o-mini (teal), and GPT-4o (orange)

Supplementary Table 1.4. Validity of GPT ratings for literal and anomalous statements. Correlations between human-generated ratings and ratings generated by GPT3.5-turbo, GPT4o-mini (prompted through API and ChatGPT interface), and GPT4o

Measure	Study	Items	Language	GPT3.5-turbo	GPT4o-mini (API)	GPT4o-mini (Interface)	GPT4o
Familiarity	Cardillo et al. (2017)	Literal	English	0.53***	0.58***	0.58***	0.61***
	Bambini et al. (2013)	Anomalous and literal	Italian	0.83***	0.82***	0.90***	0.93***
Comprehensibility	Al-Azary & Buchanan (2017)	Anomalous and literal	English	0.97***	0.98***	0.98***	0.96***

Supplementary Table 1.5. Reliability for each study. Correlations between ratings generated by GPT3.5-turbo, GPT4o-mini (prompted through API and ChatGPT interface), and GPT4o obtained in two independent sessions for the three dimensions (comprehensibility, imageability, and familiarity).

Measure	Language	Study	GPT3.5-turbo	GPT4o-mini	GPT4o	ChatGPT
Comprehensibility	English	Al-Azary & Buchanan (2017)	0.99***	0.99***	0.99***	0.91***
Imageability	English	Campbell & Raney (2016)	0.99***	0.98***	0.97***	0.66***
	Italian	Bambini et al. (2024)	0.99***	0.99***	0.98***	0.67***
Familiarity	Italian	Bambini et al. (2013)	0.98***	0.99***	0.96***	0.67***
	Italian	Bambini et al. (2014)	0.97***	0.99***	0.96***	0.88***
	Italian	Bambini et al. (2024)	0.98***	0.99***	0.99***	0.87***
	English	Campbell & Raney (2016)	0.98***	0.99***	0.99***	0.86***
	Italian	Canal et al. (2022)	0.97***	0.99***	0.99***	0.74***
	English	Cardillo et al. (2017)	0.98***	0.99***	0.98***	0.63***
	Italian	IUSS NEPLab MetaBody study	0.93***	0.99***	0.98***	0.47***

STUDY TWO

METAPHORS' JOURNEYS ACROSS TIME AND GENRE: TRACKING THE EVOLUTION OF LITERARY METAPHORS WITH TEMPORAL EMBEDDINGS[2]

Abstract

Metaphors are a distinctive feature of literary language, yet they remain less studied experimentally than everyday metaphors. Moreover, previous psycholinguistic and computational approaches overlooked the temporal dimension, although many literary metaphors were coined centuries apart from contemporary readers. This study innovatively applies tools from diachronic distributional semantics to assess whether the processing costs of literary metaphors varied over time and genre. Specifically, we trained word embeddings on literary and nonliterary Italian corpora from the 19th and 21st centuries, for a total of 124 million tokens, and modelled changes in the semantic similarity between topics and vehicles of 515 19th-century literary metaphors, taking this measure as a proxy of metaphor processing demands. Overall, semantic similarity, and hence metaphor processing demands, remained stable over time. However, genre played a key role: metaphors appeared more difficult (i.e., lower topic-vehicle similarity) in modern literary contexts than in 19th-century literature, but easier (i.e., higher topic-vehicle similarity) in today's nonliterary language (e.g., the Web) than in 19th-century nonliterary texts. This pattern was further shaped by semantic features of metaphors' individual terms, such as vector coherence and semantic neighborhood density. Collectively, these findings align with broader linguistic changes in Italian, such as the stylistic simplification of modern literature, which may have increased metaphor processing demands, and the high creativity of the Web's language, which seems to render metaphor more accessible.

[2] This chapter is a manuscript in preparation for submission to a peer-review journal as "Mangiaterra, V., Barattieri di San Pietro, C., Canal, P., & Bambini, V. "Metaphors' Journeys Across Time and Genre: Tracking the Evolution of Literary Metaphors with Temporal Embeddings".

2.1. Introduction

Research on metaphors, both from the perspective of cognitive linguistics (Lakoff & Johnson, 1980) and pragmatics (Sperber & Wilson, 2012), has largely contributed to depicting metaphors as a pervasive phenomenon in language and cognition, rather than just a tool for poets. Similarly, empirical research as pursued within Experimental Pragmatics, psycholinguistics, and neurolinguistics has focused on everyday manifestations of metaphors. Nonetheless, metaphor remains a stylistic hallmark of poetic and narrative language (Steen et al., 2010). Understanding the mechanisms through which we comprehend metaphors in literature captures, therefore, a relevant aspect of metaphor processing. To date, however, experimental research on literary metaphors remains a niche.

Theoretically, literary metaphors hold a special status compared to everyday metaphors. According to Relevance Theory, while everyday metaphors point toward a straightforward interpretation (e.g., 'That lawyer is a shark', meaning 'The lawyer is aggressive'), literary metaphors are held to generate a wide range of weak implicatures (e.g., the metaphor 'His ink is pale' by Gustave Flaubert, meaning that 'His writing lacks contrast, may fade, will not last'[3]). Processing this wide range of assumptions is supposed to require an additional cognitive effort that generates the so-called *poetic effect* of literary metaphors (Pilkington, 2000). On the quantitative side, corpus-linguistics studies found that metaphors are well represented in fiction, where they constitute 11% of the words (Steen et al., 2010), and that metaphors in literary texts are, more frequently than in other genres, non-lexicalized, i.e., with a novel association between topic and vehicle (Goatly, 1997).

From the experimental point of view, the few available studies on literary metaphors led to heterogeneous results. Katz et al. (1988) collected ratings from human participants for literary and everyday metaphors, finding no difference along ten psycholinguistic dimensions and suggesting that they do not differ in nature. Conversely, other studies reported that literary metaphors are less familiar, more open-ended, and more difficult to understand than journalistic metaphors (Semino & Steen, 2008; Steen, 1994). In

[3] Example from Pilkington (2000).

addition, Bambini et al. (2014) found that ratings are modulated by context, and specifically that when literary metaphors are presented together with their context, people tend to consider them more meaningful and less concrete, less difficult, and less familiar, compared to when presented without context. Consistently, electrophysiological evidence showed that literary metaphors trigger a sustained brain response, possibly linked to the manipulation of multiple meanings (Bambini et al., 2019), supporting the theoretical claims that literary metaphors evoke multiple weak implicatures and require a greater cognitive effort compared to non-literary metaphors.

A less considered factor in the study of literary metaphors is the temporal dimension: literary metaphors used in empirical studies were created by authors who lived well before the participants who took part in the experiments investigating the processing of these expressions. Whether this temporal gap between the time of production of literary metaphors and the time of their processing contributes to the cognitive effort involved in interpreting metaphors, and whether readers today need an additional effort compared to readers contemporary to the time of metaphor production, has never been investigated. A preliminary answer may come from the replication of the Katz et al. (1988) study by Campbell & Raney (2016). They recollected ratings for the same metaphors used in Katz et al. twenty-five years later, and found that judgments were consistent over time. However, twenty-five years is a limited time span, while the time in which metaphors were created and contemporary readers might be centuries apart.

The issue of the temporal gap becomes particularly relevant when we consider that both language and society have deeply changed over the last two centuries, potentially modifying the conceptual and contextual factors underlying metaphor interpretation. Focusing on Italian, studies on the history of language noted that in the 19th century, written language was still an élite product. Despite some innovative thrusts, Italian was characterized by a style close to the courtly and illustrious tradition both in literature (Beccaria, 1993; Serianni, 1993) and in journalistic and nonfiction language (Marazzini, 2002; Masini, 1977). In the 21st century, instead, language has evolved, particularly in literary contexts, toward a strong shift to oral and informal styles (Coletti, 2022). Moreover, in the 21st century, much of the exposure to written language occurs on the Web, as highlighted, for instance, by recent reports on news

consumption that indicate that web-based sources are used by 68% of the Italian sample, compared to 12% for print media (Newman et al., 2025). The language of the web, or *netspeak* (Crystal, 2006), has its own specificities: it is based on informal speech, combining terms from different registers and neologisms (Cerruti & Onesti, 2013), and allows users to experiment with new forms of language creatively (Fiorentino, 2018; Goddard, 2015).

2.1.1. A distributional approach to metaphor evolution

To understand how literary metaphors were perceived at the time of their creation, distributional semantics techniques be of into help, as they allow us to construct representations of meaning mimicking the semantic networks of readers of different epochs and to compare the status of the metaphors in the two representations. Two key features of these techniques make them suitable for answering our research question: i) their ability to approximate human representation of meaning and ii) the possibility to build models starting from historical corpora and therefore represent meaning in specific time points. Distributional semantics is based on the theoretical intuition, the so-called *distributional hypothesis* (Harris, 1954), that states that words occurring in similar contexts have similar meanings. Implementing this hypothesis, Vector Space Models (VSMs) can be trained on large text corpora to learn co-occurrences of words. In VSMs, words are represented as vectors, whose coordinates are derived from their co-occurrences in the corpora. Words whose vectors are closer in the VSM tend to be closer in meaning. In these models, the semantic similarity between two words can be operationalized as the cosine of the angle between the vector of word$_1$ (v_{w1}) and the vector of word$_2$ (v_{w2}), as defined by the following formula:

$$Cosine\ Similarity = \cos(v_{w1}, v_{w2}) = \frac{v_{w1}.v_{w2}}{\|v_{w1}\| \,.\, \|v_{w2}\|}$$

2.1.1.1. Applications of distributional semantics to psycholinguistic and metaphor research

Since their inception, distributional semantic approaches have been closely tied to psychological and psycholinguistic research, given that VSMs can model various dimensions of human semantic knowledge, capturing how word meaning is acquired, processed, and stored in the brain (Bhatia et al., 2019; Günther et al., 2016; M. N. Jones et al., 2015). Distributional semantics tools can be a valid application to quantify also aspects of figurative language processing (Reid & Katz, 2018). In these applications, a widely studied measure is the semantic similarity between the two terms of a metaphor. Well before the diffusion of computational tools, the relation between topic, namely the subject of the metaphor (e.g., *lawyer* in 'That lawyer is a shark'), and vehicle, namely the term used metaphorically (e.g., *shark*), was recognized as crucial to metaphor comprehension and appreciation, with rating-based semantic similarity being linked to a number of dimensions, such as metaphoricity, ease of interpretation, and goodness of nonliterary and literary metaphors (Katz et al., 1985; Marschark et al., 1983). The computational operationalization of semantic similarity via word embeddings has been largely applied to metaphor research (Bolognesi & Aina, 2019; Brglez & Vintar, 2025; Utsumi, 2011). Applications ranged from automatic metaphor identification (based on defining a cosine similarity threshold below which expressions are considered metaphorical, see Mao et al., 2018; Shutova, 2015; Su et al., 2017) to modeling metaphor processing costs (where cosine similarity is considered an approximation of human judgments). For instance, McGregor et al. (2019) found that contextual embeddings, i.e., dynamic representation of words based on surrounding contexts, can efficiently model human ratings of *metaphoricity*, *meaningfulness*, and *familiarity,* and Winter & Strik-Lievers (2023) showed that semantic similarity between the topic and the vehicle of synesthetic metaphors mirrors their degree of *metaphoricity* and *creativity*. Converging evidence comes from the *Figurative Archive* (Bressler et al., 2026), a recent resource collecting approximately 1,000 metaphors together with human ratings, where significant correlations between semantic distance and both familiarity and metaphoricity were reported. In particular metaphors with less semantically similar topics and vehicles are considered less familiar and more difficult to process.

From word embeddings, other measures relevant for language and metaphor processing can be derived. Among those, *Semantic Neighborhood Density* (SND), i.e., the number of words that are similar to the target one in terms of meaning, was shown to play a role in language processing, being able to predict performances on psycholinguistic tasks, such as lexical decision and word naming (Buchanan et al., 2001). In metaphor research, metaphors composed of low-SND words were considered more comprehensible than metaphors with high-SND words (Al-Azary & Buchanan, 2017). Moreover, the SND of the vehicle allows to discriminate between different types of metaphors: literary metaphors have vehicles with higher SND compared to nonliterary metaphors (Reid et al., 2023).

2.1.1.2. Applications of distributional semantics to semantic change

Semantic similarity has found applications in historical linguistics and semantic change research. The *distributional hypothesis* can indeed be adapted to the diachronic perspective: changes in a word's co-occurrences reflect changes in its meaning (Hilpert, 2008a). Hence, by looking at how word co-occurrences change through time, it is possible to track the evolution of meaning in time. Operationally, to create time-characterized word embeddings, i.e., embeddings representing the word's meaning at each specific time point, it is necessary to train VSMs on corpora of different epochs. However, given the stochastic nature of VSMs, every time a training process is initialized, different spatial coordinates are used, making it impossible to compare word vectors across spaces. For the VSMs to be comparable, they need to have the same spatial coordinates, i.e., to be aligned. Different procedures (Gulordava & Baroni, 2011; Kulkarni et al., 2015) have been proposed to train aligned VSMs, among which the *Temporally aligned Word Embeddings with a Compass* (TWEC) model (Di Carlo et al., 2019).

By employing aligned time-locked semantic representations of a word, it is possible to track its semantic shift over time by computing the cosine similarity between the embedding at time$_1$ (v_{wt1}) and at time$_2$ (v_{wt2}), a measure called *vector coherence* (Cassani et al., 2021; Hamilton et al., 2016). As a result, a word whose meaning has shifted over time would exhibit a lower semantic similarity between v_{wt1} and v_{wt2}

(i.e., lower vector coherence) compared to a word whose meaning has remained consistent, which would exhibit a higher semantic similarity between v_{wt1} and v_{wt2} (higher vector coherence). To date, most studies in diachronic distributional semantics have focused on identifying semantic changes of single words (Cain & Ryskin, 2025; Charlesworth et al., 2022; Garg et al., 2018; Hamilton et al., 2016; Xu & Kemp, 2015), by examining variations in vector coherence across epochs. Only recently these approaches have been applied to investigate how the relation between different words changes over time (Jenkins et al., 2025), a perspective that is crucial for tracing the evolution of complex expressions, such as metaphors.

2.1.2. The present study

In this study, we aimed to explore whether Italian literary metaphors are associated with different processing demands for today's readers compared to 19th-century readers, contemporary to the time of metaphors' original creation. To do so, we innovatively extended diachronic VSMs from word-level to the case of multi-word expressions, namely metaphors. Specifically, we trained VSMs on historical corpora from the 19th century and on contemporary corpora, mimicking the linguistic input available to present and past readers. Given that (i) the evolution of Italian is closely intertwined with textual genres, and (ii) different interpretative attitudes and processing modes may be activated depending on the literariness of a text (Steen, 1989), we incorporated the genre dimension by training separate models for literary and nonliterary corpora within each epoch. Then, we compared the semantic similarity between topics and vehicles of a set of literary metaphors in each diachronic VSM, taking this measure as a proxy of the evolution of metaphors' processing demands. Moreover, we examined whether metaphors' evolution is further shaped by lexical-semantic features of individual topics and vehicles, such as their stability over time (vector coherence), their semantic neighborhood density, and their frequency.

We expected that the temporal dimension, when assessed over a sufficiently long span of time, would significantly influence the processing costs of metaphors. In particular, we hypothesized that metaphors would yield higher elaboration demands in present-day readers than in readers of the past, who shared the conceptual and contextual environment in which the metaphors were originally created. Also, given that in the past both literary and nonliterary texts were intended for educated readerships and exhibited similar stylistic features (Aprile, 2014), we expected no difference across genres in the past. In the 21st century, instead, literary language and web-based language constitute distinct varieties of Italian, which led us to expect differences between the two genres in the processing demands of metaphors.

2.2. Methods

2.2.1. Metaphor dataset

A set of 19th-century Italian literary texts was retrieved from Project Gutenberg (https://www.gutenberg.org/), an initiative started in the 1970s, intending to collect digital versions of books that have never been copyrighted or whose copyright has lapsed. From the selected texts, "A di B" (Eng. "A of B") strings were extracted. Using the spaCy package (Honnibal & Montani, 2017), we performed a Part-Of-Speech (POS) tagging, and all the "NOUN of NOUN" strings were selected. Even though metaphors come in many shapes and forms, we chose to focus on the "NOUN of NOUN" structure to avoid any possible confounding factors due to differences in the number and type of their constituting elements. Following (Hanks, 2006), a set of keywords from semantic classes that are considered productive sources of metaphors, such as natural events and locations (river, storm, rain) and emotions (anger, joy), was then employed to filter the resulting list, yielding a set of N = 400 metaphors. The dataset was further enriched by adding an existing collection of "A of B"[4] metaphors (Bambini et al., 2014). The final dataset included a total of N = 515 metaphors in the form of "A of B" (e.g., "Capelli

[4] Thirty-seven of these metaphors displayed a definite artile in the propositional phrase, i.e., "A of *the* B" (e.g., "Abbraccio del sonno", Eng. "Embrace of the sleep").

di fiamma", Eng."Hair of flame") and is fully available in the *Literary Metaphors* module of the *Figurative Archive* . While metaphors can come in different syntactic constructions, we chose the genitive one because it clearly displayed the two terms (topic and vehicle) of the metaphors. The order of the terms can vary, with some metaphors displaying a topic-vehicle (TV) order, with the first term being the topic and second one being the vehicle, e.g., "Capelli di fiamma" (Eng. "Hair of flame"), and others displaying a vehicle-topic (VT) order, with the first term being the vehicle and the second one being the topic, e.g., "Grumo di nuvole" (Eng. "Clump of clouds"). Table 2.1 reports four examples of the metaphors included in the dataset, together with metadata regarding author, source, year of publication, topic, vehicle and their order.

Table 2.1. Examples of literary metaphors included in the final dataset of the study.

Metaphor	Author	Source/Book	Year	Topic	Vehicle	Order
Cielo di perla (Eng. sky of pearl)	Giovanni Pascoli	Myricae	1891	Cielo (Eng. Sky)	Perla (Eng. Pearl)	TV
Grumo di nuvole (Eng. clump of clouds)	Federico De Roberto	L'Illusione	1891	Nuvole (Eng. Clouds)	Grumo (Eng. Clump)	VT
Capelli di fiamma (Eng. hair of flame)	Sibilla Aleramo	Il Passaggio	1919	Capelli (Eng. Hair)	Fiamma (Eng. Flame)	TV
Nebbia di malinconia (Eng. Fog of melancholy)	Federico De Roberto	Documenti Umani	1888	Malinconia (Eng. Melancholy)	Nebbia (Eng. Fog)	VT

Note. T = Topic, V = Vehicle.

2.2.2. Training sets

To examine the diachronic evolution of metaphors, considering also the possible effect of textual genre, we collected a set of Italian corpora, for a total of 124 millions tokens, composed of literary and

nonliterary texts written in the 19th and the 21st century. Table 2.2 reports a summary of the characteristics of the collected corpora.

Table 2.2. **The composition of the literary and nonliterary sections of 19th-century and 21st-century corpora.**

Corpus	Size	Description	Genre
21st Century			
Itwac	42 M	Web-crawled corpus (Baroni et al., 2009).	Nonliterary
Contemporary Literature	20 M	Literary texts published between 2000 and 2020, used in accordance with the "fair use" principle of copyright law.	Literary
19th century			
Gutenberg Nonliterary (GNL)	10 M	Nonliterary texts downloaded from Project Gutenberg on topics such as botany, agriculture, and science.	Nonliterary
Lessico dell'Italiano Scritto (LIS)	6 M	Diachronic corpus of written Italian texts from 1850 to 1940 (Accademia della Crusca, 2013).	Nonliterary
ChronicItaly (CI)	16 M	Corpus of Italian immigrant newspapers published in the United States between 1898 and 1920 (Viola, 2021).	Nonliterary
Gutenberg Literary	30 M	Literary texts downloaded from Project Gutenberg, both prose and poetry.	Literary

Note. Size is in millions of tokens.

From the corpora, four training sets were built: 19th-Century Literary training set, 19th-Century Nonliterary training set, 21st-Century Literary training set, 21st-Century Nonliterary training set. The training sets included from 20 to 42 million tokens (19th century literary: 30 M; 19th century nonliterary:

32 M; 21st century literary 20 M; 21st century nonliterary: 42 M), in line with the mean size of large corpora employed by Di Carlo et al. (2019). The internal diversity of these datasets was designed to provide a proxy for the linguistic input accessible to the readers in each respective epoch. In the 19th-century training set, we included prose and poetry texts for the literary section and texts from technical manuals, newspapers, and diaries for the non-literary section. In the 21st-century training set, we included prose and poetry for the literary section and a collection of texts taken from the web, such as newspaper sites, blogs, and educational sites, for the non-literary section.

2.2.3. Training Aligned Spaces

Employing the *Temporally aligned Word Embeddings with a Compass* (TWEC) model (Di Carlo et al., 2019), we trained four sets of word embeddings on the four training sets previously outlined. The word embeddings provide a semantic representation of words in the Italian language that differs by epoch (19th and 21st centuries) and genre (Literary and Nonliterary).

The TWEC model exploits the dual representation of words derived from a word2vec model (Mikolov et al., 2013) based on a *Continuous Bag of Words* (CBOW) architecture, a feed-forward neural network trained to predict a target word given its context, relying on the theoretical assumption that most words do not change over time, and that words with a shifted meaning will appear in the context of words that did not change. This theoretical assumption is reflected in the creation of an atemporal VSM, called *compass*, based on which the spatial coordinates of all the other temporal VSMs are then subsequenttly defined. In other words, a *compass* model is first trained on the entire corpus, providing the semantic representation of words independently of time. After that, the *compass*'s context matrix is used to initialize the training of a time- and genre-specific target matrix on each corpus, which allows us to extract the temporal word embeddings.

Operationally, we trained the model on the whole training set (resulting from the combination of all training sets), and we extracted the resulting two atemporal matrixes, an input-weight *Context* matrix (C)

and an output-weight *Target* matrix (U). The target matrix U was used as "a compass", i.e., it served as a reference to initialize all the other VSMs along the same coordinate system. We hence trained a set of context matrices C_i on each slice of the training set (namely, 19th-century Literary training set, 19th-Century Nonliterary training set, 21st-Century Literary training set, 21st-Century Nonliterary training set). As a result, the context embeddings can differ according to the co-occurrence frequencies that are specific to a given temporal period. Each of the four resulting sets of word embeddings provides a representation of word meaning in the 19th century, in the 21st century, and in the literary and nonliterary sections of the training sets.

2.2.4. Measures of Diachronic Change

To semantically characterize the metaphors across time, we computed four measures of interest using the obtained sets of word embeddings, one at the metaphor level (Cosine Similarity between topic and vehicle - CS) and three at the word level, for each term of the metaphor (Semantic Neighborhood Density – SND, Vector Coherence – VC, and Frequency – Freq, see Table 2.3). Each measure was computed, for each metaphor, in all time and genre slices, to obtain a semantic characterization of the metaphor and its terms in each epoch and each textual genre. A graphical representation of the approach, including corpora collection, VSM training, and features computation, is depicted in Figure 2.1.

Table 2.3. Measures of metaphor change, with corresponding formula and interpretation.

Level	Measures	Formula	Description	Interpretation
Metaphor	**Cosine Similarity between topic and vehicle (CS)**	$CS_{(tv)} = \cos(v_t^{tn}, v_v^{tn})$	Cosine Similarity between the vector of the metaphor's Topic (v_t) and the vector of the metaphor's Vehicle (v_v) in all time and genre slices (t_n).	CS can be considered as a proxy of the processing costs of the metaphor; from a diachronic perspective, it can be used to describe the temporal trajectories of the metaphors' demands.
Word	**Semantic Neighborhood Density (SND)**	$SND_t = \frac{\sum_{i=1}^{n=500} \cos(v_t^{tn}, v_i^{tn})}{n}$ $SND_v = \frac{\sum_{i=1}^{n=500} \cos(v_v^{tn}, v_i^{tn})}{n}$	Semantic Neighborhood Density of the Topic (v_t) and the Vehicle (v_v) in all the slices (t_n) was computed as the average of the cosine similarities between the word and its *n* closest neighbors.	SND refers to the average proximity of a word vector to its closest neighbors as computed using a language model. It provides a measure of the word's position in VSMs relative to its nearest neighbors. A word with many close neighbors is considered semantically denser than a word with fewer close neighbors. Metaphors with denser vehicles are considered less comprehensible (Al-Azary & Buchanan, 2017).
	Vector Coherence (VC)	$VC_t = \cos(v_t^{t1}, v_t^{t2})$ $VC_v = \cos(v_v^{t1}, v_v^{t2})$	Vector Coherence of the Topic and the Vehicle was obtained by computing the cosine similarity between a word vector in the 19th century (v_w^{t1}) and its vector in the 21st century (v_w^{t2}).	VC is a measure of the stability of word meaning over time. A word with high VC has maintained a stable meaning because the word vector at t_1 is very close to the word vector at t_2. A word with low VC has changed meaning because the word vector at t_1 is quite far from the word vector at t_2.
	Frequency (Freq)	$Freq_t = \log\left(\frac{number\ of\ occurrences\ topic}{total\ number\ of\ token}\right)$ $Freq_v = \log\left(\frac{number\ of\ occurrences\ vehicle}{total\ number\ of\ token}\right)$	Logarithmic Relative Frequency of the Topic ($Freq_t$) and the Vehicle ($Freq_v$) in all slices.	Word Frequency is a key measure both in psycholinguistics and in diachronic semantic shift research (Baumann et al., 2023). Regarding word processing, higher-frequency words are elaborated faster than lower-frequency ones (Brysbaert et al., 2018). In diachrony, frequent words show the tendency to change more slowly (Hamilton et al., 2016).

Note. The value of *n* was set to 500, as in previous computational approaches to metaphors (Kintsch, 2000).

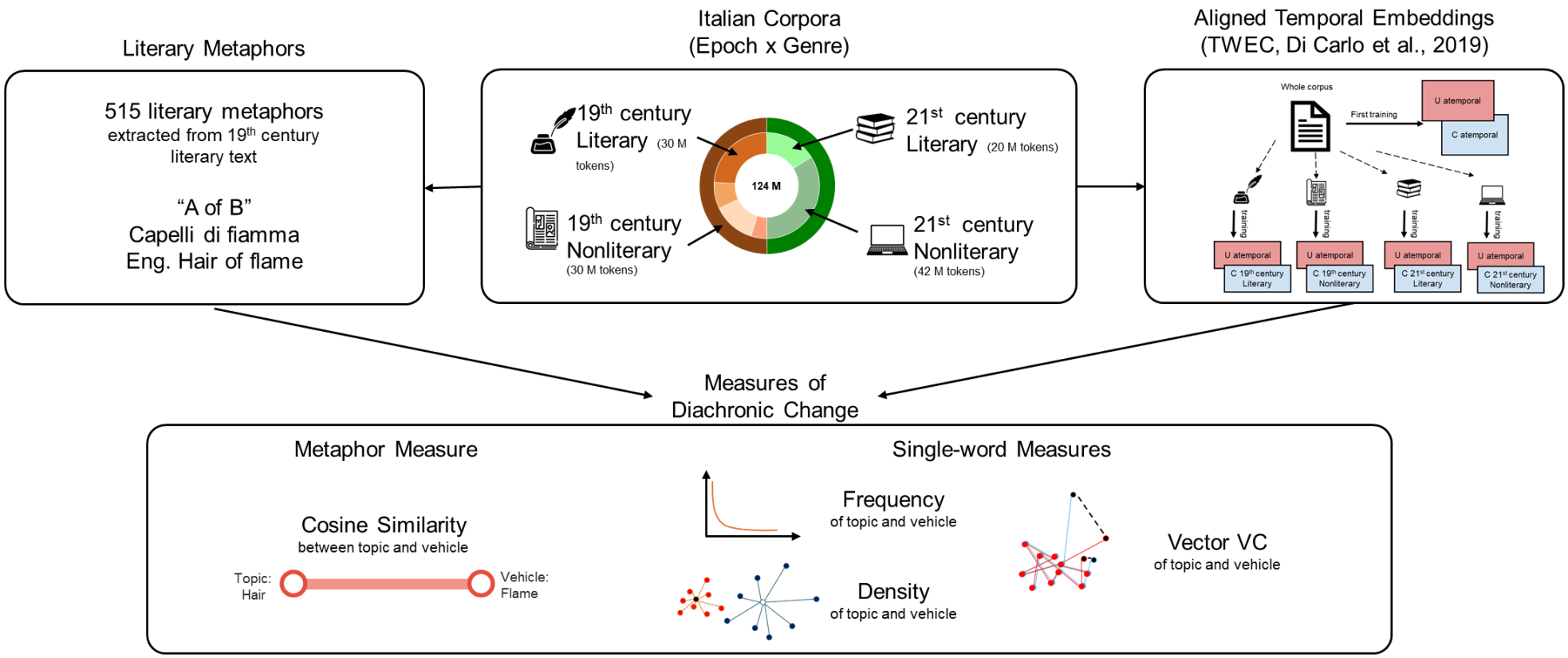


Figure 2.1. Workflow for the Diachronic Analysis of Italian Literary Metaphors. The process begins with the collection of Italian corpora stratified by epoch (19th vs. 21st century) and genre (literary vs. non-literary). From the 19th-century literary corpus, 515 metaphors (e.g., Hair of flame) were extracted. Then, the corpora were used to train aligned temporal embeddings (using the TWEC methodology). Diachronic change is then quantified using a metaphor-level measures (Cosine Similarity) and single-word measures, including Frequency, Semantic Neighborhood Density (SND), and Vector Coherence (VC).

2.2.5. Statistical analyses

To understand the relationship between word- and metaphor-level measures, we computed a set of Pearson correlations between CS, SND, VC, and Freq, corrected for multiple comparisons.

To test the hypothesis that CS varied through time and across different genres, we fitted a Linear Mixed-Effects Model (LMM – Pinheiro & Bates, 2000) using the *lme4* package (Bates et al., 2015), considering CS as a continuous dependent variable, and genre and epoch as interacting categorical predictors. Then, to explore whether the evolution of metaphors is further shaped by different word-level variables, we fitted a series of LMMs using again the *lme4* package, considering CS as a continuous dependent variable, genre and epoch as interacting categorical predictors, and word-level variables as continuous predictors. Specifically, to define the final statistical model, we incrementally added the word-level variables (i.e., first topic and vehicle SND, Freq and VC separately, then pairs of topic and vehicle features and finally all the variables), and we tested if the addition of variables contributed to explaining the data variance by comparing the models' goodness-of-fit in terms of Akaike Information Criterion (AIC), Bayesian Information Criteria (BIC), and Log-likelihood (Bozdogan, 1987; Neath & Cavanaugh, 2012). All models included a random intercept to account for the variability of individual metaphors. All analyses were performed in R (R Core Team, 2025).

2.3. Results

2.3.1. Descriptive statistics

Table 2.4 reports descriptive statistics of the measures of interest as computed in the four corpora and sets of embeddings. The distribution of each variable is displayed in Figure 2.2. While measures like CS, SND and Freq maintain relatively normal distributions with consistent peaks across both centuries, VC exhibits a significant departure from normality. Specifically, in the literary genre, both topics and vehicles

demonstrate high VC (approaching 1.0), suggesting that words in literary texts did not report drastic changes in meaning. Conversely, the words in the nonliterary genres show a great semantic shift.

Table 2.4. Descriptive statistics of word-level and metaphor-level measures reported as Mean (SD).

	19th Lit	**21st Lit**	**19th NonLit**	**21st NonLit**
CS	0.31 (0.21)	0.27 (0.21)	0.30 (0.22)	0.33 (0.20)
SND Topic	0.69 (0.05)	0.69 (0.05)	0.68 (0.05)	0.71 (0.05)
SND Vehicle	0.71 (0.05)	0.69 (0.05)	0.70 (0.05)	0.72 (0.05)
VC Topic	0.85 (0.1)		-0.08 (0.17)	
VC Vehicle	0.84 (0.12)		-0.12 (0.18)	
Freq Topic	-9.62 (1.45)	-9.90 (1.67)	-10.16 (1.47)	-10.69 (1.58)
Freq Vehicle	-10.53 (1.45)	-10.68 (1.49)	-11.09 (1.56)	-11.57 (1.51)

Note: CS = Cosine Similarity, computed between the Topic and the Vehicle of the metaphor; SND = Semantic Neighborhood Density, computed as average cosine similarity between the word and its 500 closest neighbors; VC = Vector Coherence, computed as cosine similarity between the word at t_1 and the word at t_2; Freq = Frequency, computed as the logarithm of the relative frequency of the word.

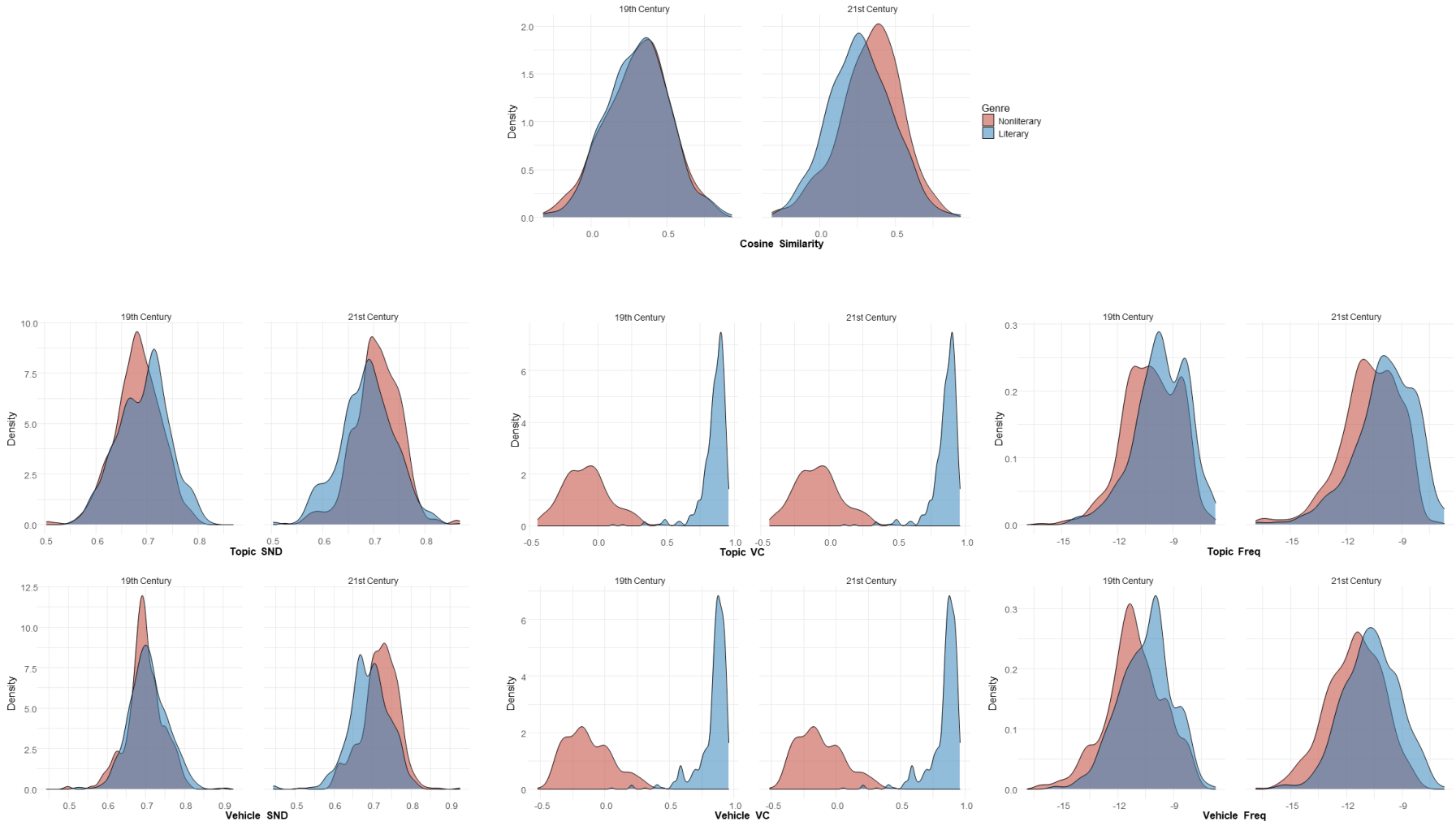


Figure 2.2. Density plots of metaphor and single-word variables. The plots show the density distribution of the metaphor feature (cosine similarity) and single-word features (semantic neighborhood density, vector coherence, and frequency) for both topic and vehicle across time and genre. Note: SND = Semantic Neighborhood Density, VC = Vector Coherence, Freq = Frequency.

2.3.2. Correlation analysis

Results of the correlation analyses between single-word and metaphor-level measures are reported in Figure 2.3.

CS, SND, and Freq showed overall strong positive correlations across genre and epoch, as highlighted by the black triangles on the diagonal. VC, however, reported low correlations confirming the different patterns of semantic change between literary and nonliterary genres.

Regarding word-level variables, SND and Freq of both topics and vehicles were negatively correlated in all slices (see the blue squares), indicating that less frequent words tend to be more semantically dense across time and genre. A positive correlation emerged between Freq and VC (see the light blue rectangles), especially in literary corpora, indicating that more frequent words have a more stable meaning. Moreover, SND positively correlated with VC (see the purple rectangle), suggesting that denser words have a more stable meaning.

Regarding the relations between word-level variables and CS (see green rectangles), we found that the latter was positively correlated with topic VC in literary corpora, suggesting that when the meaning of the topics tends to change over time, metaphors are characterized by more semantically distant terms. Moreover, in nonliterary corpora, we found a positive correlation between both topic and vehicle SND and CS, indicating that when metaphors are constituted by high-density terms, they tend to have a greater topic-vehicle similarity.

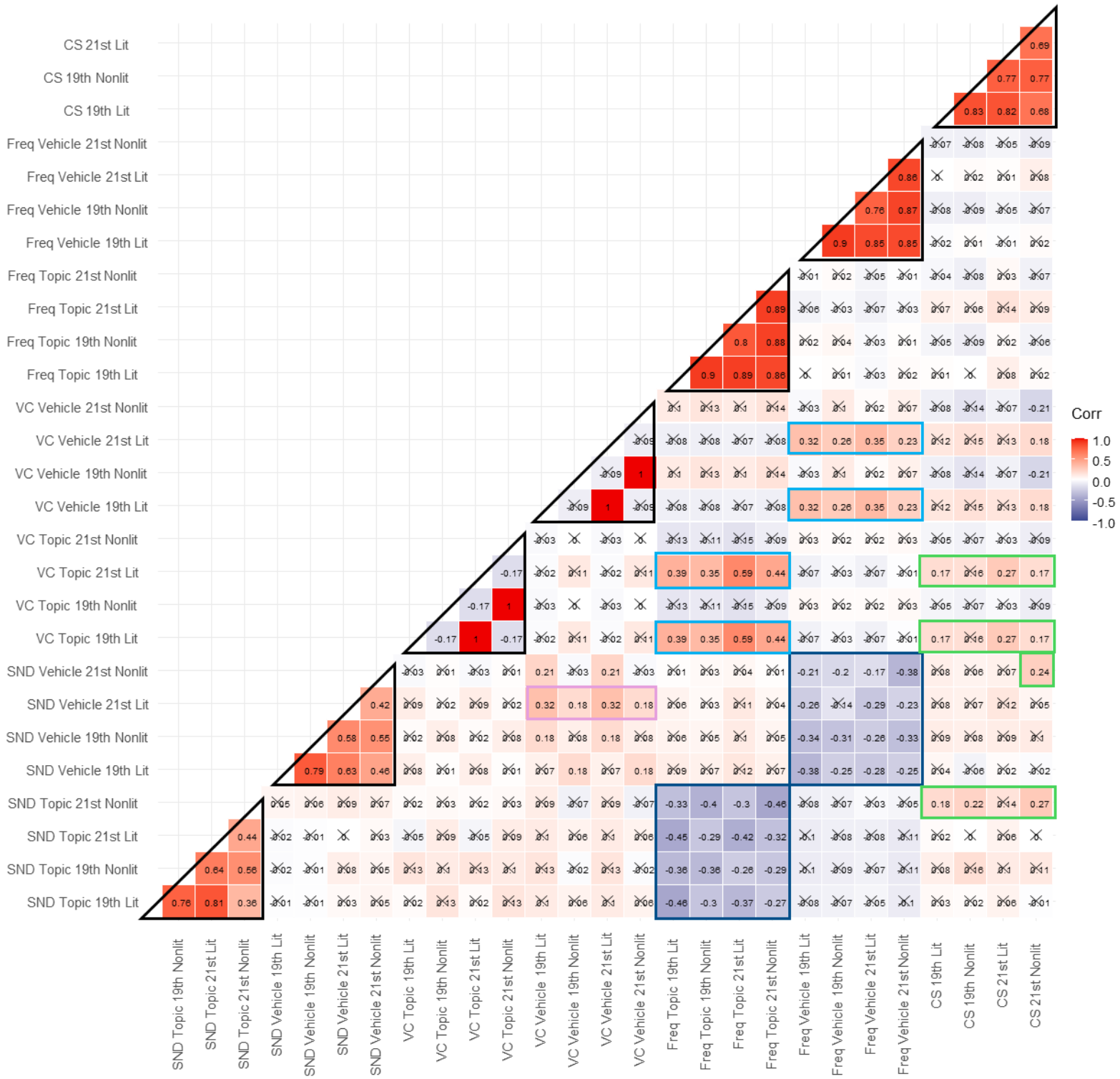


Figure 2.3. Correlations between word-level measures and metaphor-level measures. Positive correlations are displayed in red and negative correlations in blue. The color intensity is proportional to the correlation coefficients. Non-significant correlations are marked with a cross. All corrections are corrected for multiple comparisons.

2.3.3. Linear Mixed-Effects Models

The simple LMM including epoch and genre as interacting predictors revealed an effect of Genre (β = -0.03, t = -6.85, p < 0.001), with metaphors reporting a lower semantic similarity in Literary corpora compared to Nonliterary ones, further qualified by its interaction with Epoch (β = -0.07, t = -7.82, p < 0.001). No main effect of Epoch was reported (p = 0.35). A diverging trend emerged: metaphors' terms become increasingly distant in literary texts going from the 19th century to the 21st century, while in nonliterary texts, metaphors' terms showed an increasing semantic similarity (Table 2.6, Figure 2.4).

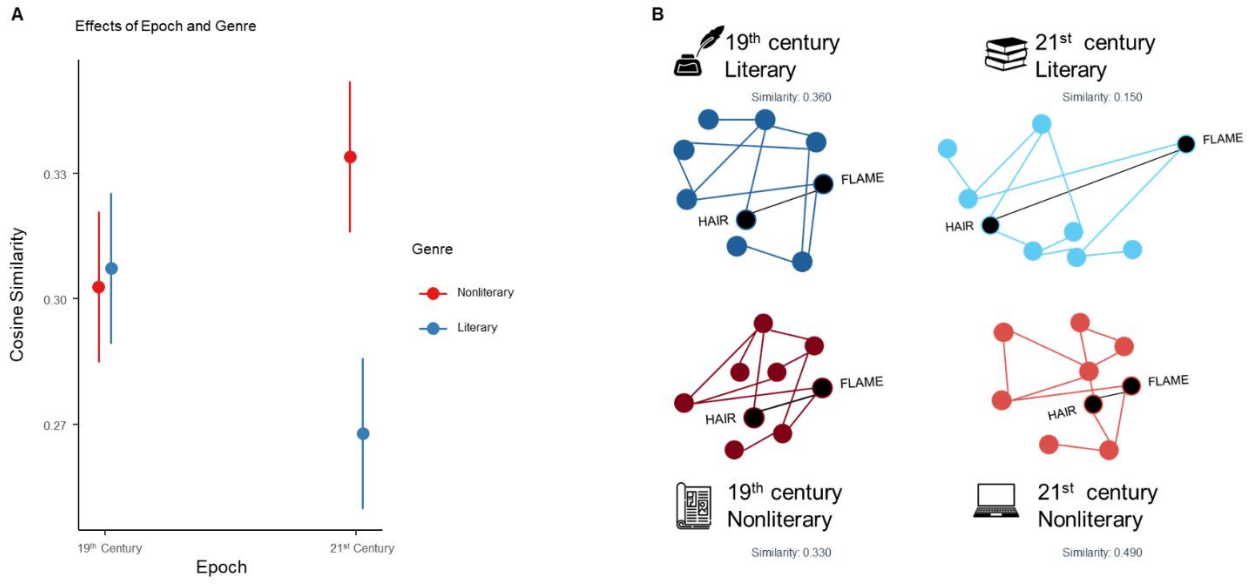


Figure 2.4. Effects of epoch and genre on cosine similarity between the topics and vehicles of metaphors. Panel A shows the effects of epoch and genre as emerged from the LMM. Panel B shows a graphical representation of how the relationship between topic and vehicle of a representative metaphor ("Capelli di fiamma", Eng. "Hair of flame") changed.

Moving to the investigation of how the diachronic evolution of metaphors is further shaped by lexical-semantic features of topic and vehicle, the comparison of AIC, BIC, and Log-likelihood showed that the best fit to the data is obtained by adding both topic and vehicle SND and to the simple model (see Table 2.5).

Table 2.5. Summary statistics of LMMs presenting AIC, BIC, and Log-Likelihood.

Model	*df*	*AIC*	*BIC*	*LogLik*
Epoch * Genre	6	-2,185.8	-2,152.1	1,098.9
Epoch * Genre * (SND topic + SND vehicle)	14	-2,321.5	-2,242.8	1,174.8
Epoch * Genre * (Freq topic + Freq vehicle)	14	-2,212.3	-2,133.5	1,120.1
Epoch * Genre * (VC topic + VC vehicle)	14	-2,231.3	-2,152.6	1,129.7
Epoch * Genre * (VC topic + VC vehicle + SND topic + SND vehicle)	**22**	**-2,357.1**	**-2,233.4**	**1,200.5**
Epoch * Genre * (VC topic + VC vehicle + Freq topic + Freq vehicle)	22	-2,245.3	-2,121.6	1,144.7
Epoch * Genre * (SND topic + SND vehicle + Freq topic + Freq vehicle)	22	-2,328.0	-2,204.3	1,186.0
Epoch * Genre * (VC topic + VC vehicle + SND topic + SND vehicle + Freq topic + Freq vehicle)	30	-2,349.8	-2,181.2	1,204.9

Note: AIC = Akaike's information criterion; BIC = Bayesian Information Criterion; Freq = word Frequency; LogLik = Log-Likelihood; SND = Semantic Neighborhood Density; VC = Vector Coherence.

The LMM showed a significant three-way interaction between topic VC, Genre, and Epoch ($\beta = 0.225$, $t = 3.15$, $p = 0.002$), suggesting that the coherence of the topic influenced metaphor CS in contemporary literary texts but not in the other slices (Table 2.6, Figure 2.5a). Moreover, the model showed a significant interaction between vehicle SND, Genre, and Epoch ($\beta = -0.602$, $t = -3.25$, $p = 0.001$), in addition to the main effect of vehicle SND ($\beta = 0.58$, $t = 7.90$, $p < 0.001$). While the main effect indicated that higher SND generally predicts higher CS, this relationship became stronger specifically in 21st century nonliterary texts (Table 2.6, Figure 2.5b).

Table 2.6. Outputs of the base LMM model and the best LMM model with Single-Word Features on CS.

	Base Model			***Model with Single-Word Predictors***		
Predictors	*Estimates*	*Statistic*	*p*	*Estimates*	*Statistic*	*p*
Epoch	-0.00	-0.93	0.353	-0.28	-2.99	**0.003**
Genre	-0.03	-6.85	**<0.001**	0.15	1.55	0.121
Epoch*Genre	-0.07	-7.82	**<0.001**	0.11	0.59	0.552
SND Topic				0.57	7.63	**<0.001**
SND Vehicle				0.58	7.90	**<0.001**
VC Topic				0.04	1.92	0.056
VC Vehicle				-0.04	-2.12	**0.034**
Epoch*SND Topic				0.02	0.24	0.811
Epoch* SND Vehicle				0.23	2.48	**0.013**
Epoch * VC Topic				0.12	3.30	**0.001**
Epoch * VC Vehicle				-0.04	-1.44	0.150
Genre * SND Topic				-0.25	-2.63	**0.008**
Genre * SND Vehicle				-0.16	-1.72	0.086
Genre * VC Topic				0.23	4.35	**<0.001**
Genre * VC Vehicle				0.08	1.76	0.079
Epoch * Genre * SND Topic				0.18	0.99	0.321
Epoch * Genre * SND Vehicle				-0.60	-3.25	**0.001**
Epoch * Genre * VC Topic				0.23	3.15	**0.002**
Epoch * Genre * VC Vehicle				0.02	0.33	0.740

Note: SND = Semantic Neighborhood Density; VC = Vector Coherence.

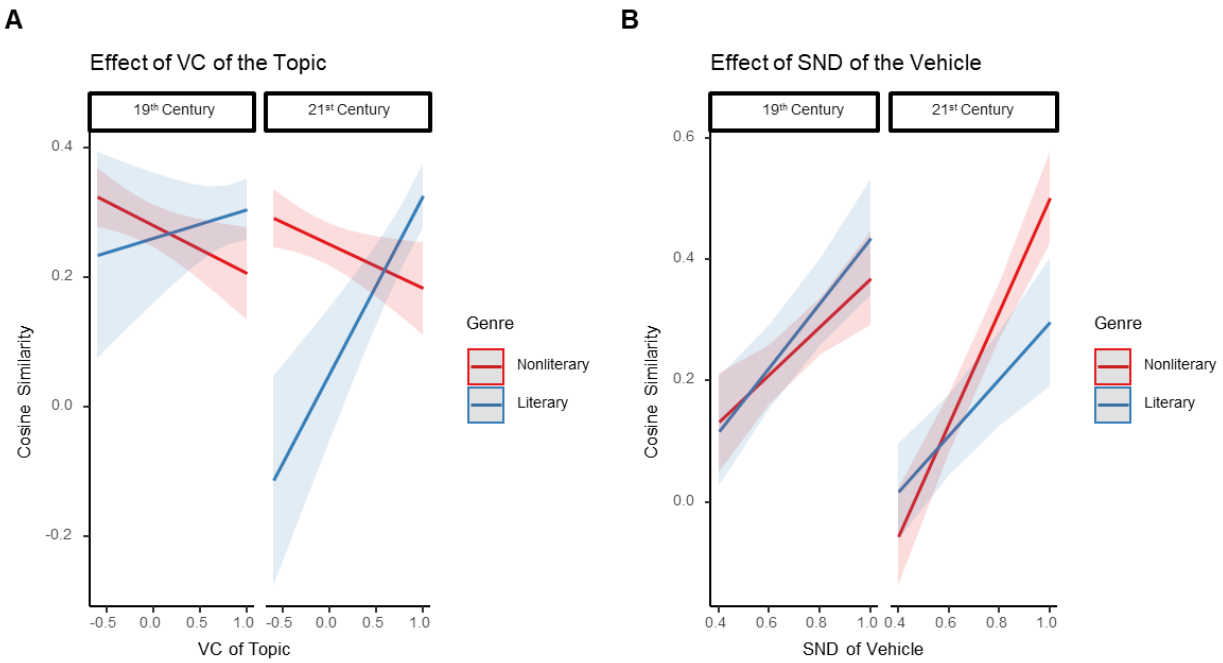


Figure 2.5. Significant effects of lexical-semantic features of topic and vehicle. Panel A illustrates the effect of Vector Coherence (VC) of the topic on Cosine Similarity (CS) across Epoch and Genre. Panel B shows the effect of Semantic Neighborhood Density (SND) of the vehicle on CS across Epoch and Genre.

2.4. Discussion

In the present study, we investigated whether the processing demands of Italian literary metaphors changed over time and as a function of textual genres and lexical-semantic features of the topic and vehicle. To do so, we collected 19th-century and 21st-century corpora (composed of literary and nonliterary texts), taken to reflect the linguistic input of today's and past readers. Then, we used the corpora to train different temporal vector space models, obtaining representations of word meanings in the two time points. Then, we examined how the semantic similarity between topics and vehicles of a set of 515 Italian metaphors, extracted from literary texts of the 19th century, varied. Crucially, we took semantic similarity as a proxy of metaphor processing difficulty, thereby associating metaphors with semantically closer terms to lower processing demands. Importantly, we did not limit the analysis to semantic similarity across epochs and genres, but we

also included three single-word variables relevant in historical and metaphor research. Specifically, we considered i) a measure of the stability of word meaning over time (*vector coherence*; Cassani et al., 2021; Hamilton et al., 2016b; Rodda et al., 2017); ii) *word frequency*, as it affects both semantic shift (Englhardt et al., 2020; Feltgen et al., 2017; Sagi et al., 2009) and metaphor processing (Littlemore et al., 2018); iii) Semantic Neighborhood Density (SND), which has been shown to impact metaphor processing, with high density hindering metaphor comprehensibility (Al-Azary & Buchanan, 2017; Reid et al., 2023). Given the significant stylistic change and simplification of Italian over the past two centuries (Coletti, 2022), we expected that today's readers may experience classical literary metaphor differently from the readers of the 19th century. In particular, we assumed that today's readers do not share the cultural background of the authors of classical literary metaphors, and, as such, we expected that metaphors could entail greater processing costs for contemporary readers.

Our analysis partially confirmed our predictions. Overall, the semantic relationship between topics and vehicles of metaphors did not change from the 19th century to today, suggesting that the processing of metaphors has not become more costly as we moved away from the time these metaphors were originally written. However, a different representation of the metaphors across time emerges depending on the genre (literary or non-literary). Indeed, the significant interaction between Epoch and Genre indicates that the semantic relationship between topic and vehicle followed a genre-depending diverging trajectory over time. Starting from an equivalent semantic representation in 19th-century literary and nonliterary VSMs, metaphors' terms in the 21st-century became increasingly distant (less similar) in literary texts, while they showed an increasing semantic similarity in nonliterary texts. This interaction suggests that the effort required to process these metaphors has actually decreased in modern nonliterary contexts; conversely, in the literary domain, these metaphors have become more semantically distant over time.

This result can be linked to the nature of the corpora considered as literary and nonliterary texts. The modern nonliterary corpus consisted of web-crawled texts (Baroni et al., 2009), a linguistic variety that differs substantially from literature (Pistolesi, 2014). Scholars have noted the language of the Web is characterized by greater variability, brevity, fragmentation, fluidity, and loose use of language with an open, hybrid, and changeable nature (Santini, 2007), while modern Italian literature is marked by a loss of literariness and a shift toward orality, highlighted by lexicon and discursive signals specific to a standardized spoken language (Dardano, 2014). In the modern nonliterary corpus, the vectors of topics and vehicles tend to be closer, indicating more entrenched associations and easier connections between distant concepts. In the literary modern corpus, by contrast, the vectors tend to be more distant, suggesting that novel associations are less frequent and therefore more striking. We can conclude that in contemporary nonliterary language, novel metaphorical associations required reduced processing demand due to increased frequency of novel associations between words, while in literary language, they remain marked and, hence, cognitively demanding.

It is noteworthy that, while in the 21st century the status of metaphors is quite distinct across genres, in line with the marked stylistic differences between literary and nonliterary texts reported for modern Italian (Aprile, 2014), this distinction was not present in the 19th century texts. Indeed, the interaction effect suggests that in the 19th century, metaphors were processed similarly regardless of the textual genre. This can be linked to the shared stylistic features of 19th-century literary and nonliterary texts. Texts from both genres were cultural products designed for the educated portion of the population and therefore characterized by high-register language, with echoes of the classical tradition (Aprile, 2014). Although elements of stylistic innovation can be found starting from the 19th century, language essays, manuals, and newspaper texts, which compose our nonliterary training sets, continue the expressive heritage of earlier prose and poetry (Masini, 1994).

Interestingly, in a proof-of-concept study applying the same methodology to English literary metaphors (Mangiaterra et al., 2024), we found a different pattern of results. English metaphors were associated with higher processing costs in nonliterary text compared to literary ones, irrespective of the epoch, confirming the stable nature of the English language – and its stylistic remarks, such as metaphors – in the last two centuries. These cross-linguistic differences suggest that the patterns emerging from this kind of analysis are language-specific and that metaphor evolution follows the broader stylistic trajectories of the language in which they are embedded.

2.4.1. Single-word features and their impact on metaphor evolution

The claim that metaphor evolution is embedded in the broader temporal changes of the language at stake is further strengthened when we consider the role of single-word variables. First, it is important to note that single-word variables showed known patterns of relationships, confirming the validity of our vector space models. In particular, vector coherence was associated with word frequency, consistent with the findings that frequent words change less over time (Hamilton et al., 2016b), and with semantic neighborhood density, consistent with the findings that words change more in the sparse portions of the semantic network (Ryskina et al., 2020). We also found that word frequency was negatively correlated with SND (consistent with Buchanan et al., 2001), suggesting that highly frequent words tend to occur in a variety of contexts and develop more loosely related relationships with their neighbors (less dense meaning), while low-frequency words tend to develop a higher level of specificity and tigher relationship with their neighbors (more dense meaning, see Rambelli & Bolognesi, 2024).

Moving to the analysis of the contribution of lexical-semantic features of topics and vehicles in shaping the diachronic evolution of metaphors, both the correlation analysis and the linear mixed models highlighted the centrality of the vehicle semantic neighborhood density and

the topic vector coherence. Semantic neighborhood density of the vehicle had an overall effect, which appeared stronger in 21st century nonliterary text, as denser vehicles were associated with metaphors with semantically closer terms. As suggested by Al-Azary & Buchanan (2017), when a word has many close semantic neighbors, it lacks the flexibility to acquire new metaphorical associations. Instead, a vehicle with low SND, with its looser semantic association, may evoke the wide array of weak implicatures characteristic of literary metaphors, which resulted in greater appreciation (Reid et al., 2023), but also higher processing costs, as suggested by our results. In the 21st century nonliterary text, high SND pulls the terms together into the entrenched, easily processed associations reflected in our high similarity results, such as in the highly dense vehicle "flour" in the metaphor "sky of flour", as compared to sparser vehicle "inebriation" in the metaphor "inebriation of light".

Moreover, we found that in 21st-century literary texts, where metaphors are associated with higher processing costs, these demands seem to be driven by the vector coherence of the topic. In deriving the meaning of a metaphor, the function of the topic is to help promote the salient features of the vehicle necessary to reach the intended interpretation. Topics whose meanings have changed greatly over time have probably acquired a sparser set of semantic relationships and are considered less concrete (Azarbonyad et al., 2017) and acquired later (Cassani et al., 2021). All these features contribute to making the shifted meaning of these topics less accessible and less prone to providing the straightforward contextual constraints necessary to guide the reader through the process of meaning derivation of complex literary metaphors. An example of this process can be provided by the metaphor "vento di lode" (Eng. "wind of praise"). The topic "lode" (Eng. "praise") shifted greatly between literary texts of the 19th-century (where the meaning was in the semantic domain of "kindness", "gratitude", "appreciation") and 21st-century (where the meaning is in the semantic domain of "report card", "degree", "graduated", given the prominence of the use in the expression "con lode" – Eng. "cum laude), and therefore represented

a less clear guide for the selection of the appropriate meaning of the metaphor, increasing its processing demands.

Overall, it emerged that the topic should have a stable meaning to guide the derivation of metaphor interpretation over time and reduce processing costs, and, at the same time, the vehicle should have a flexible meaning to allow the metaphor to emerge.

2.5. Conclusions

This study examined the evolution of processing costs of literary metaphors in terms of semantic similarity between topics and vehicles, expanding the diachronic application of word embeddings to the analysis of complex expressions such as metaphors. Our results showed that, while overall the processing demands of metaphors did not change from the past to today, they vary in relation to different textual genres, following the broader patterns of stylistic evolution in the Italian language. We can therefore argue that since literary and nonliterary language were very similar in the 19th century, readers did not process metaphors with different levels of effort depending on genre. Today's readers, however, have to make a greater effort to process metaphors in current literary texts, which have a plainer and simplified language, while they can more easily activate the connection between distant concepts in the creative nonliterary language of the Web. Possibly, this distinction may also hint at the effect of different linguistic exposure and its impact of processing in contemporary speakers with different backgrounds.

Methodologically, this work confirms the possibility of applying temporal embeddings to examine the evolution of multi-word expressions (Jenkins et al., 2025), extending their scope to figurative expressions. Regarding metaphor processing, our results empirically highlight the need to account for the textual context in which metaphors are embedded, which could determine different processes of elaboration (Steen, 1989). The point of optimal distance between topics and vehicles,

which "should be sufficiently distant to emphasize differences but sufficiently close to maintain similarities" (Katz, 1989), seems to be the result of a complex balance, modulated by the nature of the words composing the metaphors but also sensitive to the broader stylistic features of the language. Ultimately, this study demonstrates that what makes a metaphor difficult to process is not an inherent property, but a dynamic process influenced by diachronic genre-based shifts and by the structure of single-word semantic networks.

Data availability statement

Temporal vector space models and metaphor datasets used in the study are available at 10.5281/zenodo.18523747.

STUDY THREE

TEMPORAL WORD EMBEDDINGS IN THE STUDY OF METAPHOR CHANGE OVER TIME AND ACROSS GENRES: A PROOF-OF-CONCEPT STUDY ON ENGLISH[5]

Abstract

Temporal word embeddings have been successfully employed in semantic change research to identify and trace shifts in the meaning of words. In a previous work, we developed an approach to study the diachrony of complex expressions, namely literary metaphors. Capitalizing on the evidence that measures of semantic similarity between the two terms of a metaphor approximate human judgments of the difficulty of the expression, we used time-locked measures of similarity to reconstruct the evolution of processing costs of literary metaphors over the past two centuries. In this work, we extend this approach previously used on Italian literary metaphors and we present a proof-of-concept study testing its crosslinguistic applicability on a set of 19th-century English literary metaphors. Our results show that the processing costs of metaphors changed as a function of textual genre but not of epoch: cosine similarity between the two terms of literary metaphors is higher in literary compared to nonliterary texts, and this difference is stable across epochs. Furthermore, we show that, depending on the metaphor structure, the difference between genres is affected by word-level variables, such as the frequency of the metaphor's vehicle and the stability of the meaning of both topic and vehicle. In a broader perspective, general considerations can be drawn about the history of literary and nonliterary English language and the semantic change of words.

[5] This chapter has been published as Mangiaterra, V., Barattieri di San Pietro, C. & Bambini, V., *Temporal word embeddings in the study of metaphor change over time and across genres: a proof-of-concept study on English* in Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024), pages 548–555, Pisa, Italy. CEUR Workshop Proceedings.

3.1. Introduction

Does the metaphor "The wind is a wrestler" convey the same feeling today, as it did in the 1888 when Gerard Manley Hopkins used it in the poem "That nature is a Heraclitean Fire and of the comfort of the Resurrection" (Gardner & MacKenzie, 1967)? The answer to this question is not trivial: human languages evolve constantly, alongside with the society in which they are used, so much so that the concepts associated with each word, as well as their semantic associations with other words, have changed to different degrees (Hamilton et al., 2016).

Studies on lexical semantic change have a long tradition (Fortson, 2017; Traugott & Dasher, 2001) but, with the increasing availability of historical language data and the development of new digital tools, they radically opened up to new approaches coming from computational linguistics and distributional semantics (Kutuzov et al., 2018; Tahmasebi et al., 2019; Tang, 2018). In the diachronic declination of the Distributional Hypothesis (Harris, 1954), it is said that changes in the contexts in which a word occurs over time may reveal a change in meaning (Hilpert, 2008b). Operatively, this means that by training vector space models on historical text corpora from different epochs, it is possible to create time-locked representations of words: if the meaning of a word changed over time, its vectorial representation at $t1$ will be different from its vectorial representation at time $t2$; conversely, if the two vectors of the same word at $t1$ and $t2$ are in close proximity, the meaning of the word has remained stable. Comparing words vectors diachronically, however, is not effortless and requires the temporal vector space models to be aligned. Alignment is a crucial step in diachronic distributional semantics and it has been tackled by different approaches (Di Carlo et al., 2019; Gulordava & Baroni, 2011; Kulkarni et al., 2015). Previous studies employing temporal embeddings have found that more frequent words change slower than less frequent words, and that polysemous words change faster than monosemous words (Hamilton et al., 2016), while synonyms tend to change meaning comparably (Y. Xu & Kemp, 2015). However, temporal word embeddings have been mostly applied to the study of the semantic change of single words and only marginally to complex linguistic expressions leaving the field with a knowledge gap on the evolution of meaning of a widespread linguistic and textual phenomenon such as, for instance, metaphors.

Within the theoretical framework of Relevance Theory (Wilson & Carston, 2007), metaphors are non-literal uses of language involving a conceptual adjustment described as context-driven broadening of lexically denoted meaning of words. In terms of linguistic structure, metaphors normally involve two terms, the topic and the vehicle: for example, in the metaphor 'Sally is a chameleon', the topic Sally is described by the broadened vehicle chameleon, to indicate a person who changes attitude/behavior to fit their surroundings. While metaphors are broadly used in everyday communication, they are certainly a distinctive feature of literary texts, as long evidenced in stylistics (Fludernik et al., 1999). Past studies on literary metaphors, however, report mixed results. The rating study by Katz et al. (1988) found no difference between literary and everyday metaphors, while other studies showed that the former type is less familiar and more open-ended than the latter (Semino & Steen, 2008), but literary metaphors are rated as less difficult and more familiar when presented together with their original context (Bambini et al., 2014). Moreover, the processing of literary metaphors seems to be particularly effortful, given the multitude of possible meanings they evoke (Bambini et al., 2019). Therefore, open questions remain regarding how literary metaphors are processed. It must be also underlined that the literary metaphors used in previous studies were written tens or hundreds of years ago. Yet, the effect of this diachronic dimension on their processing costs, as well as its interplay with textual genre in which metaphors are embedded, remains an open question.

In addition to its diachronic application, the use of vector space models can help characterize metaphors thanks to the ability of these models to approximate human performance in psycholinguistic tasks. Measures derived from vector space models were shown to be able to approximate how humans process word meaning (Bhatia et al., 2019; Günther et al., 2016; M. N. Jones et al., 2015) and, more specifically to correlate with how humans perceive metaphorical expressions in terms of metaphoricity, difficulty, and other psycholinguistic dimensions (McGregor et al., 2019; Reid et al., 2023; Winter & Strik-Lievers, 2023). In particular, semantic similarity, operationalized in vector space models as cosine similarity (CS) between topic and vehicle, has long been considered relevant for metaphor studies (Katz et al., 1985) and, more recently, for automatic metaphor identification (Shutova, 2015).

In a previous study on Italian (Mangiaterra et al., *in preparation*, see §Study 1), we developed a novel method, employing the Temporal Word Embeddings with a Compass (TWEC) model (Di Carlo et al., 2019) as training procedure, to capture the temporal dynamics of literary metaphors. This method combines the computational models' abilities to approximate human judgments and their diachronic applications, allowing to track the diachronic evolution of how literary metaphors are perceived by readers over the course of 200 years. In the present proof-of-concept study, we apply this approach to English, to test its crosslinguistic applicability and whether it can provide language-specific insights into the evolution of metaphors. We take the similarity between the topic and vehicle of a metaphor as a proxy for its difficulty, and we analyze how it varies across time and textual genres. We also consider the role of word frequency (WF) and vector coherence (VC), two widely used measures in the study of semantic change (Englhardt et al., 2020; Feltgen et al., 2017), as well as semantic neighborhood density (SND) in shaping the difficulty of the expression. WF and VC were considered to assess the effect of the semantic change of the single word on the evolution of whole metaphor understanding, while SND was considered to analyze the impact of a measure known to synchronically impacts metaphor understanding (Al-Azary & Katz, 2023; Reid et al., 2023) on its diachronic unfolding.

3.2. Methods

3.2.1. Dataset of metaphors

The study focuses on "classic" literary metaphors (i.e., metaphors found in 19th-century literary texts). In terms of metaphor structure, we focused on metaphors in the form of 'A is B' (e.g. "Stars are dancers") and 'A of B' (e.g., "Clouds of melancholy"), as they clearly display the two metaphorical elements (topic and vehicle) and allow to avoid possible confounding factors (length of expression, intervening words, etc.). Twenty- four (24) 'A is B' metaphors were taken from the dataset in Katz et al. (1988) and 115 metaphors in the form 'A of B' were retrieved from a collection of literary texts of the 19th century. These latter were identified by PoS-tagging a corpus of literary texts from the 19th century (see below) with

spaCy (Honnibal & Montani, 2017), and then extracting only the 'NOUN of NOUN' constructions. The resulting list was then further reduced by manually searching for words belonging to known sources of metaphors, such as atmospheric events (e.g., 'rain') or physical locations (e.g., 'river') (Hanks, 2006), following the methodology in Bambini et al. (2014).

3.2.2. Corpora and training

To test whether the processing costs of metaphors changed as a function of epoch, we collected corpora from the 19th century and from the 21st century. We also included different textual genres (literary vs. nonliterary) of the corpora, to examine whether the difficulty of the figurative expression is modulated by the stylistic features of different types of language. Following previous work (Bambini & Trevisan, 2012), the corpora were built so as to be representative of the language to which speakers of the two epochs were exposed, and specifically by combining literary, nonfiction, and journalistic language for the 19th century, and literary and web language (which includes sections of newspapers, blogs, and other text types that can be found on the Internet) for the 21st century. Specifically, we trained four diachronic vector space models on four corpora:

• 19th-century literary corpus (32M tokens), consisting of a collection of literary texts (both narratives and poetry) retrieved from the Gutenberg project (gutenberg.org);

• a 19th-century nonliterary corpus (25M tokens), consisting of nonliterary texts, such as magazines or scientific essays, from the same online resource (gutenberg.org)

• a 21st-century literary corpus (16M tokens), collected from literary texts available on the web, employed without violating the "fair use" principle of copyright law;

• a 21st-century nonliterary corpus (46M tokens), collected from portions of the UMBC web- Base corpus (Han et al., 2013).

To train aligned temporal vector space models, we followed the procedure by (Di Carlo et al., 2019). The TWEC model is implemented on top of a Continuous Bag of Words (CBOW) architecture (Mikolov et al., 2013). The TWEC model exploits the double representation learned by the CBOW model: the target matrix and the context matrix. First, a model, the so-called "compass", is trained on the whole corpus, creating time-independent word embeddings. The context matrix of the compass is then maintained fixed to train on each corpus a time- and genre-specific target matrix from which we derive the temporal word embeddings. The four sets of embeddings obtained for the four corpora will represent the meaning of words in each time slice for the two genres. To validate our models, following previous studies (Hamilton et al., 2016), we computed the synchronic (within time period) accuracy of each vector space model against the MEN dataset (Bruni et al., 2012), which contains 3,000 pairs of words together with a semantic similarity score provided by humans. Finally, we tested whether our measure of metaphor difficulty (cosine similarity between topic and vehicle) correlated with the measure of difficulty in Katz et al., 1988a dataset.

3.2.3. Measures of interest and analyses

For each metaphor, we collected four measures of interest, at the metaphor- and word-level.

• Cosine similarity (CS): the similarity between the two terms of the metaphor (topic and vehicle). It is computed as the cosine of the angle between the vectorial representations of the two words. CS is here considered as a proxy value of difficulty of the metaphors.

• Semantic neighborhood density (SND): a measure of the density of the semantic space around a word. Words with many closely related words have a higher semantic density while words whose neighbors are more distant and are sparsely distributed have a lower density. It is computed as the mean cosine similarity between the target word and its 500 closest neighbors (standard size from previous work, see Kintsch, 2000).

• Vector coherence (VC): a measure of the stability of a word's meaning, computed as the cosine similarity between the target word at $t1$ the target word at $t2$. Words with a high vector coherence are considered

to have stable meaning through time, while a low vector coherence means that the word's meaning has changed.

• Word frequency (WF): computed as the logarithm of the frequency of the target word in the reference corpus.

Each measure was collected for all the temporal slices, extracted from the temporal vector space models (CS, SND, and VC) or corpora (WF). To analyze how the understanding of metaphors changed over time and if it was affected by genre and word-level variables, we fitted a set of Linear Mixed Models (LMMs) using the R package lme4 (Bates et al., 2015). The two metaphorical structures were treated separately, fitting distinct models for 'A is B' and 'A of B' metaphors. The linear mixed model considers CS as dependent variable and the interaction between epoch and genre and word-level variables as predictors. In all models Items (metaphors) were added as random variables. The resulting formula was:

lmer(cosine ~ epoch * genre * (VC-topic + VC-vehicle + SND-topic + SND-vehicle + WF-topic + WF-vehicle) + (1 | Item).

Alpha level was set at .05.

3.3. Results

First, to test the validity of the meaning representation in the vector space models, we correlated the human scores of relatedness and the semantic similarity derived from our word embedding for each pair of words in the MEN dataset (Bruni et al., 2012) (Table 3.1). These results show strong correlations, comparable to the results obtained by Hamilton et al. (2016), indicating that the models accurately mimic humans' representation of maning (i.e., they have a good synchronic accuracy).

Table 3.1. Results of correlation between models' semantic similarity scores and MEN dataset's semantic similarity scores.

19th Literary	19th Nonliterary	21st Literary	21st Nonliterary
.55	.58	.61	.59

Note: All the correlations have a $p < .001$.

Secondly, we tested whether cosine similarity between the two terms of a metaphor correlated with the measure of difficulty from the dataset by Katz et al. (1988). Results showed a moderate correlation ($r(26) = .49$, $p < .05$): metaphors with higher semantic similarity between topic and vehicle were rated with lower values of difficulty by participants, coherently with previous studies.

Thirdly, we explored whether the change in the semantic similarity between the topics and the vehicles of literary metaphors is driven by the interaction between the Epoch, Genre and single-word variables. The results of our predictors of interest are reported below. Concerning the 'A of B' metaphors' mixed model, results showed a main effect of genre ($\beta = 0.81$, $t = 2.44$, $p = .01$) and a significant three-way interaction between epoch, genre and vector coherence, both of the topic ($\beta = 0.34$, $t = 2.018$, $p = .04$) and of the vehicle ($\beta = -1.715$, $t = -4.954$, $p < .001$). These results indicate that the cosine similarity of literary metaphors' terms did not change over time, but it changed as a function of textual genres, resulting in greater difficulty (lower cosine similarity) in nonliterary texts than in literary (Figure 3.1).

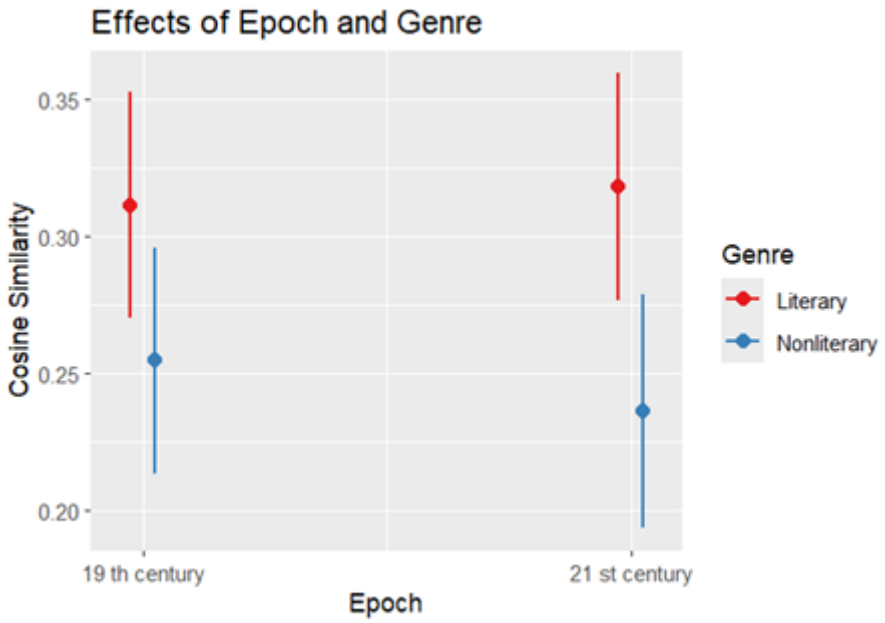


Figure 3.1. Effects of epoch and genre in defining the cosine similarity between the topic and vehicle of 'A of B' metaphors

As shown by the three-way interaction between Epoch and Genre and the single-word variables in Figure 3.2, the effect of VC acted differently in the two time points and in the two genres. VC of the vehicle did not affect CS in literary and non- literary texts in the past; conversely, more stable vehicles significantly lowered CS in present literary texts and increased CS in present nonliterary texts. A similar trend can be observed for VC of the topic, where its stability did not affect CS in the past, regardless of the literary genres. Conversely, stability of the topic contributed to significantly increase CS in present literary texts, but less so in nonliterary texts.

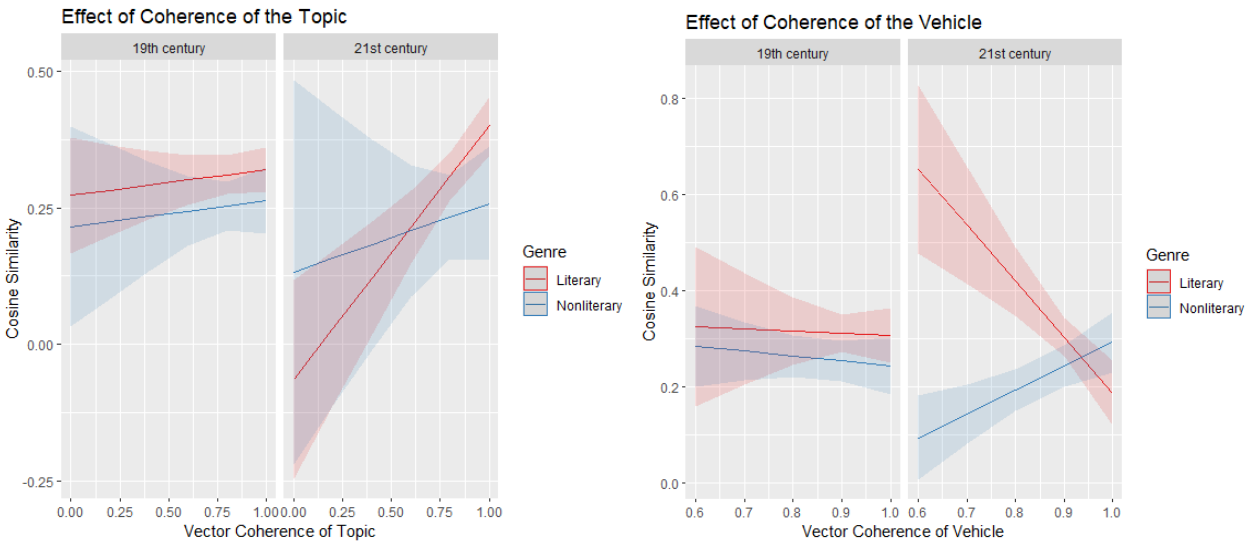


Figure 3.2. Effects of topic and vehicle VC in defining the cosine similarity between the topic and vehicle of 'A of B' metaphors

For 'A is B', the model revealed a significant three-way interaction between epoch, genre, and the frequency of the vehicle (β = 0.06, t = 2.077, p = .04), but no main effects. The effect of WF of the vehicle showed different patterns in the two time points and in the two genres (Figure 3.3): while WF of the vehicle did not affect CS in literary texts, both in the past and in the present, more frequent vehicles significantly increased CS in past nonliterary texts and lowered CS in present nonliterary texts.

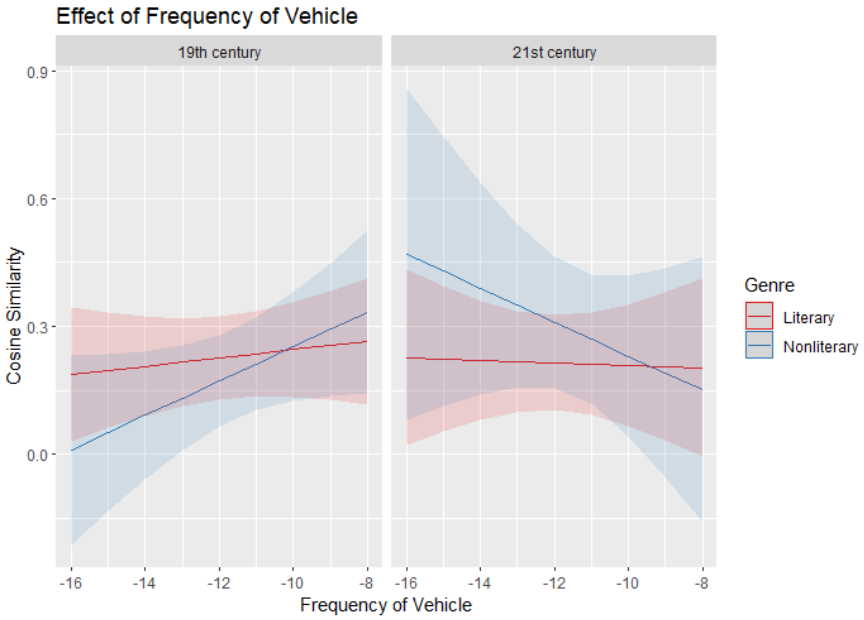


Figure 3.3. Effects of vehicle WF in defining the cosine similarity between the topic and vehicle of 'A is B' metaphors

3.4. Discussion

In this proof-of-concept study, we characterized the temporal dynamics of a set of English literary metaphors to understand whether their processing costs changed over time. We also explored whether this change was affected by the genre of the texts, as well as by the semantic properties of the constituting elements of the metaphors (topic and vehicle). By leveraging on the diachronic applications of distributional semantics and extending a method already applied to the study of Italian literary metaphors (Mangiaterra et al., *in preparation*, see §Study 1), we created a series of time-locked semantic representations of 139 English metaphors, from which we derived a measure of the cosine similarity between their terms (CS), taken as a proxy of their difficulty, together with semantic neighborhood density (SND), stability over time (VC), and, from four diachronic corpora, frequency (WF) of their topics and vehicles.

Results showed no effect of epoch for either 'A is B' or 'A of B' literary metaphors. Thus, no noticeable change in CS over time was revealed, suggesting that these metaphors come with similar processing costs for contemporary readers and for readers of the epoch in which the metaphors were created. The absence of an effect of epoch can be better understood by considering the historical evolution of the English

language, and specifically its early standardization. As stated by Wyld (1936), literary writing as early as the 18th century was considered 'English of our own age in all its essentials'. In line with this consideration, our results point to the stability of the main stylistic features of the English language in the last two centuries, including those related to metaphors.

While literary metaphors are not processed differently based on the epoch, the influence of the textual genre is noticeable. This factor emerged both as a main effect and in different interaction patterns with single-word variables, varying according to the type of metaphor.

For 'A of B' metaphors, results revealed that the difficulty of these metaphors changed as a function of the genre. In particular, they are perceived as less difficult when found in literary contexts, compared to when encountered in nonliterary texts. Hence, the difficulty of these metaphors is sensitive to the style of the text in which metaphors are found: when read in a text that has a literary style and aesthetic intent, the metaphor is less striking than the same metaphor in a nonliterary text.

Moreover, we found a strong effect of the stability of the meaning of the vehicle in interaction with epoch and genre. This suggests that 'A of B' metaphors with more unstable vehicles are perceived as less difficult than 'A of B' metaphors with vehicles whose meanings remained stable over time. We interpreted this result in light of Traugott (2017)'s theory of metaphorization, according to which the metaphorical use of a word can become one of its stable meanings. In the context of the present study, words that changed the most could have done so by incorporating meanings derived from their metaphorical uses. As a result, when these unstable and broadened vehicles are used, metaphors appear less difficult. The reader does not need to broaden the concept expressed by the vehicle to interpret the metaphor, because the metaphorical nuances have entered the standard meaning of the word. From a qualitative observation of the data, we can notice, for instance, that a metaphor such as "Wave of horror", where the vehicle wave incorporated the meaning of 'sudden increase in a particular phenomenon', is perceived as less metaphorical than "Clouds of doubt", whose vehicle "clouds" has maintained its original meaning.

For ‘A is B’ metaphors, instead, the statistical model highlighted an effect of the frequency of the vehicle in interaction with epoch and genre. In nonliterary texts, the perceived difficulty of ‘A is B’ metaphors differed as a function of the WF of their vehicle, to the point that metaphors showed opposite patterns in the past and in the present: in the past, the less frequent the vehicle, the more metaphorical the whole metaphorical expression; in the present, the less frequent the vehicle, the less metaphorical the metaphor. The pattern found in the 19th-century space model is in line with previous studies (Littlemore et al., 2018) that found that metaphors with less frequent vehicles are regarded as more metaphorical than those with highly frequent vehicles, indicating that the most metaphorical metaphors are those in which the vehicle communicates something new about the topic. Going back to Hopkins’ metaphor "The wind is a wrestler", the vehicle wrestler, as a particularly low frequency word in the 19th century, was indeed communicating something new about the topic “wind”. As such, the metaphors might have been perceived as more difficult and “more metaphorical”, leading to the creation of a new concept. The very same metaphor is nowadays perceived differently, because the frequency of the vehicle has changed: wrestler has become more frequent, and the whole expression has lost some of its metaphoricity for the 21st-century readers.

Overall, our results suggest that for the English language, metaphor processing costs are not affected by the temporal distance between the creation of metaphors, which occurred in the 19th century, and their processing by today’s readers. Instead, the key factor modulating metaphor processing costs seems to be the textual genre in which they appear. This modulation, however, occurs to a different extent depending on the syntactic structure of the metaphors and in interaction with single-word measures. Indeed, we observe that in defining what drives the difficulty of metaphors, different patterns emerged for the ‘A of B’ and ‘A is B’ structures. While for the former, in addition to the main effect of genre, we found the effect of vector coherence in interaction with epoch and genre, for the latter, the diachronic evolution of metaphor processing costs is related to the interaction of word frequency with epoch and genre.

While these differences might reflect genuine effects of the syntactic structure and how it impacts metaphorical predication (Bambini et al., 2013; Carston & Yan, 2023; Tonini et al., 2023), we must

acknowledge that the numerosity of the two sets of items varies, and this might obscure some of the effects in the less represented type (A is B). Future studies are needed to further explore the whole range of diachronic changes in processing related to structural differences.

In conclusion, this proof-of-concept study proposed an adaptation from Italian to English of a method employing temporal word embeddings to study the evolution of metaphors. Thanks to this approach, we could elucidate that the processing costs of English literary metaphors are stable over time (differently from Italian) but are dynamically affected by stylistic features of texts and by single-word measures. The proposed method seems to be sensitive to the specificities of the language under investigation, supporting its crosslinguistic applicability.

STUDY FOUR

BEYOND SURPRISAL: CAPTURING N400 AND P600 EFFECTS FOR METAPHOR VIA SEMANTIC, PRAGMATIC, AND PREDICTIVE COMPUTATIONAL MODELS[6]

Abstract

Accounts of metaphor processing have proposed different mechanisms underlying metaphor comprehension, variously emphasizing semantic integration, pragmatic inference, or, within broader theories of language processing, context-based prediction. These theoretical positions have guided the still-debated functional interpretation of the electrophysiological response to metaphor, typically characterized by an N400 often followed by later effects.

Here, we leveraged computational modeling to determine whether quantitative measures can clarify the operations underlying the different metaphor-triggered ERP components and account for individual differences. For a set of metaphor and literal sentences, we computed *semantic similarity* from word embeddings, *surprisal* from Large Language Models (LLMs), and a *Bayesian pragmatic measure* (BPM) inspired by the Rational Speech Acts framework, indexing semantic, predictive, and inferential processes, respectively. We then compared the modeling power of these measures on the N400 and P600 components, based on data from 55 participants.

We found a biphasic pattern of EEG response, with metaphors reporting an N400 followed by a P600. Among computational measures, surprisal had a strong overall effect on EEG amplitude, both N400 and P600. BPM, instead, emerged in the later window, with utterances with a greater pragmatic load eliciting more positive responses.

[6] This chapter is a manuscript in preparation for submission to a peer-review journal as "Mangiaterra, V., Canal, P., Barattieri di San Pietro, C., Vanooteghem, M., di Paola, S., Ricci, I., & Bambini, V. Beyond surprisal: Capturing N400 and P600 effects for metaphor via semantic, pragmatic, and predictive computational models"

These results suggest that, while predictive mechanisms have a general role throughout the time course of metaphor understanding, there is something more than context-based predictions. Specifically, pragmatic inference seems to emerge at a later stage, capturing the effort to derive the intended meaning.

4.1. Introduction

Metaphor processing has been explained through a range of theoretical accounts, differing in their assumptions about how the metaphorical meaning is represented and derived. Three main families of approaches can be distinguished. One family is represented by *semantic approaches*, including, for instance, structure-mapping theory (Bowdle & Gentner, 2005) and conceptual blending theory (Fauconnier & Turner, 1998). These accounts proposed that metaphor comprehension relies on aligning relational structure or integration of conceptual spaces, mapping aspects of the source onto the target to generate meaning. Instead, *pragmatic approaches* place emphasis on context-driven inferential processes. Early work by Grice (1975) placed metaphors among cases of flouting the Maxim of Quality ("Do not say what you believe to be false"), requiring the listener to derive an implicature to understand the intended meaning. This view has been further developed within the post-Gricean Relevance Theory framework, which posits that metaphor comprehension involves the construction of an *ad hoc* concept, derived inferentially through context-based modulation of the linguistically encoded concept via processes of narrowing and/or broadening (Sperber & Wilson, 2012; Wilson & Carston, 2007). Concurrently, *predictive mechanisms* are increasingly recognized in psycho-neurolinguistics. In this framework, language processing is seen as continuous prediction, where upcoming input is probabilistically anticipated based on context, and processing difficulty arises when predictions are violated (Kuperberg & Jaeger, 2016; Nour Eddine et al., 2024). Within the domain of figurative language, predictive mechanisms are reported to play a role in modulating the processing of idioms and metaphors. Behavioral research indicates that when idiomatic expressions are predictable, the literal interpretation can be bypassed in favor of faster figurative access (Cacciari & Tabossi, 1988). This is further supported by electrophysiological evidence showing a shift from general probabilistic anticipation to the recognition of stored idiomatic templates (Vespignani et al., 2010). Similarly, Lago et al. (2024) found that linguistic expectations are at work in metaphor processing as well, supported by a network of core language regions, including the temporoparietal junctions, which help coordinate the brain's readiness to process non-literal input.

Insights from these different accounts have been variously used to explain the electrophysiological pattern observed for metaphor processing, an area that remains debated. Indeed, the functional interpretation of the Event-Related Potential (ERP) components associated with metaphor processing is far from being resolved, with scholars linking them to the inferential mechanisms (Bambini et al., 2016) and others interpreting them in terms of conceptual mapping and semantic blending (Coulson & Van Petten, 2002; Lai & Curran, 2013; Yang et al., 2013). Typically, metaphors are associated with a negative deflection reaching a maximum at 400 *ms* (the so-called N400 component), followed by either a positive component after 600 *ms* (P600 or Late Positive Component, LPC) or a sustained negative response. Traditionally, the N400 has been linked to the retrieval of conceptual knowledge from semantic memory and its integration with the current mental model of the sentence (Kutas & Federmeier, 2011), with competing accounts either emphasizing the effort of semantic integration of a word in the preceding context or context-driven predictive mechanisms (Lau et al., 2008). A similar divide is observed in metaphor research. Some authors emphasize the semantic nature of the N400 and its role in retrieving the stored conceptual knowledge (Goldstein et al., 2012), with a concurrent effects of semantic relatedness, familiarity and meaningfulness (Arzouan et al., 2007), while others highlight that context can facilitate or hinder the retrieval of semantic information, thereby modulating N400 amplitude to words that are more or less likely to continue a metaphorical expression (Canal & Bambini, 2023; Lago et al., 2024). The P600, originally reported for syntactic anomalies (Hagoort et al., 1993), was later found also for semantic anomalies (Kuperberg et al., 2003) and for well-formed but pragmatically complex expressions, such as irony and humor (Canal et al., 2019; Spotorno et al., 2013). Within metaphor research, some scholars linked the P600 to pragmatic inference and context-guided (re)interpretation of the sentence (Bambini et al., 2016; De Grauwe et al., 2010), while others highlighted the role of active semantic retrieval (Coulson & Van Petten, 2002). Studies that reported a sustained negative response in this later window, especially linked to novel metaphors, such as scientific or literary figurative expressions (Bambini et al., 2019; Rutter et al., 2012; Tang et al., 2017), associated it with the prolonged effort to integrate meaning and manipulate the multiple possible interpretations.

In this study, we propose a computational modeling approach to provide insights for the functional interpretation of metaphor processing dynamics, in line with a rich literature aiming to model accounts of language processing, and specifically, the cognitive processing behind ERP components (Brouwer et al., 2017; Frank et al., 2015; Rabovsky et al., 2018; Rabovsky & McRae, 2014). To do so, we will employ three distinct computational measures, derived from three complementary modeling traditions, as indices of the accounts presented above, namely, semantic, inferential, and predictive.

As concerns semantic accounts, we capitalized on distributional semantics, an influential computational approach to meaning based on the distributional hypothesis (Harris, 1954), which claims that words occurring in similar contexts tend to have similar meanings. Operationalizations of this hypothesis through word embeddings (Landauer & Dumais, 1997; Mikolov et al., 2013) have been able to capture many aspects of human language processing, such as word associations and priming effects (Cassani et al., 2023; Ettinger & Linzen, 2016; Mandera et al., 2017). While the literature is sparser compared to behavioral modeling, attempts to relate ERP components to semantic similarity have yielded interesting insights for language processing, revealing N400 behaviors not entirely explained by context-based predictability (Dudschig et al., 2025; Ettinger et al., 2016). Insights from this approach have also been applied to metaphor research, from the seminal work by Kintsch (2000), who introduced the predication algorithm to derive metaphor interpretations, to more recent approaches relating semantic similarity from word embeddings to metaphor features (Mangiaterra et al., 2024; McGregor et al., 2019).

When it comes to modeling pragmatic operations, one prominent account is the Bayesian pragmatics or Rational Speech Acts (RSA) framework. The RSA framework (Degen, 2023; M. C. Frank & Goodman, 2012) describes communication as a process of recursive social reasoning between interlocutors (Scontras et al., 2021). Specifically, its basic implementation assumes three main levels of inference, represented as a literal listener, a pragmatic speaker, and a pragmatic listener. The literal listener interprets sentences based only on the sentence's semantics and updates her prior expectations, considering the utterance as true. The pragmatic speaker, reasoning about the literal listener, chooses his utterances based on their utility, maximizing the probability that the literal listener understands the conveyed meaning. Finally, the

pragmatic listener interprets the utterance by reasoning about the choices of the pragmatic speaker and updates her prior beliefs based on the likelihood that the pragmatic speaker produced that utterance to convey a certain meaning. Starting from these layers of inference, many pragmatic phenomena have been modeled, including hyperbole (Kao et al., 2014), vague quantifiers (van Tiel et al., 2021), referring expressions (Degen et al., 2020), and metaphors (Carenini et al., 2023; Kao, Bergen, et al., 2014; Mayn & Demberg, 2022). Most RSA approaches to metaphors focused on their interpretation. For instance, given a metaphor such as *Workers are ants*, based on human typicality judgements of how typical certain features are (e.g., *organized*) for the two terms of a metaphor (the topic *workers* and the vehicle *ants*), Carenini et al. derived RSA-based metaphor interpretations, which showed a high degree of overlap with human ones. Recent developments have extended the framework to quantify the pragmatic load, expanding the application of RSA to EEG research (Werning et al., 2019; Werning & Cosentino, 2017). Werning et al. proposed a Bayesian pragmatic model, by integrating semantic similarity (as a prior) with context-driven relevance (as a likelihood), to model the N400 amplitude in response to agentive verbs in standard vs new contexts. They found that the Bayesian Pragmatic model better explained the EEG response, compared to the semantic similarity and relevance considered alone.

Finally, language models, and recently transformer-based LLMs, have played a crucial role in computational modeling, as their predictive nature allows testing the role of listeners' context-driven predictions in online language processing. Language models can estimate the probability of a word given its context and compute the information-theoretical measure of surprisal (Slaats & Martin, 2025). Brought into the spotlight by the so-called surprisal theory (Hale, 2001; Levy, 2008), this measure, computed as the negative log-transformation of word probability, has been linked to the processing effort, with less predictable words (i.e., words with higher surprisal) requiring a greater effort to be comprehended. Experimentally, surprisal has been shown to explain both behavioral responses and brain activity (Aurnhammer & Frank, 2019; Caucheteux et al., 2021; Smith & Levy, 2013; Wilcox et al., 2023) and within EEG research, it has been linked to N400 modulations, with more surprising sentences eliciting greater negative components (de Varda et al., 2023; Frank & Aumeistere, 2024; J. Michaelov & Bergen,

2020; H. Xu et al., 2024). The relation between surprisal and metaphor has only started to be explored, with recent results pointing toward a convergence between surprisal and metaphor novelty (Momen et al., 2026).

In order to test the contribution of semantic, inferential, and predictive mechanisms in metaphor processing, here we tested three measures in a single study. Following approaches that compared computational models indexing competing account of language processing, for instance contrasting contextual similarity with predictability (J. A. Michaelov et al., 2024), lexical with semantic prediction error (Lopopolo & Rabovsky, 2024) and context-based with world knowledge expectation (Venhuizen et al., 2019), we employed Large Language Models, distributional semantics, and RSA, to gain a deeper understanding of the processes occurring at the two main time windows associated with metaphor processing (around 400ms and around 600ms).

Specifically, we collected electrophysiological responses to metaphorical and literal statements and compared three computational measures (*surprisal* from LLMs, *semantic similarity* from word vectors, and an RSA-inspired *Bayesian pragmatic measure*) in modeling the resulting ERP components. A greater effect of *semantic similarity* would support the semantic account, where the processing effort indexed by the ERP components reflects the process of semantic retrieval and integration, while a greater effect of *surprisal* would go in the direction of predictive coding accounts, supporting a hindered lexical access for metaphors due to their unpredictability. Finally, a prominent role of the *Bayesian pragmatic measure* would favor the involvement of inferential mechanisms, as in the pragmatic accounts of metaphor comprehension. Based on the literature, we derived the following hypotheses. First, we expect to find a biphasic electrophysiological pattern, with metaphors associated with a greater negativity in the N400 window and a greater positivity in the P600 window. Second, we predict a combined effect of surprisal and Bayesian pragmatic measures in modulating ERPs amplitude, drawing on recent evidence that, while prediction does play a role, it does not provide a complete explanation of the operations at work for metaphor processing (Lago et al., 2024).

4.2. Methods

4.2.1 Participants

Fifty-five participants took part in the experiment (mean age = 25,01; SD = 3,96; age range: 19-35; mean education (in years) = 14.44; SD = 4.14; 27 F). They were right-handed (scores > 85 according to the Edinburgh Handedness Inventory; Oldfield, 1971) native speakers of Italian, with normal or corrected-to-normal vision, no reading difficulties, and no history of neurological/ psychiatric disorders.

The study was approved by the local Research Ethics Committee (Comitato Etico Area Vasta Nord Ovest, Azienda Ospedaliero-Universitaria Pisana). Written informed consent was obtained from every participant before the beginning of the experiment. Participants received monetary compensation for their participation.

4.2.2 Stimuli

In the ERP paradigm, the experimental material was composed of a set of 80 pairs of not-lexicalized metaphorical sentences and literal counterparts (160 sentences in total) and a set of 160 filler sentences in Italian. All metaphors were given in the X(s) is/are Y(s) nominal predicative form, embedded in a one-sentence context (e.g., "In hard times hopes are stars that light the soul up"; original Italian: "Nei momenti difficili le speranze sono stelle che illuminano l'anima"). The corresponding literal expressions were created by maintaining the vehicle and manipulating the topic of the metaphor and the context such that a literal interpretation was obtained (e.g., "Those lights in the night sky are stars of distant galaxies"; original Italian: "Quelle luci nel cielo notturno sono stelle di galassie lontane"). 42 metaphor-literal pairs were adapted from (Bambini et al., 2013), and 38 pairs were created ex novo. Filler sentences were literal sentences with the same syntactic structure as the experimental items.

Materials from Bambini et al. (2013) were already rated for meaningfulness, difficulty, and familiarity. For the de novo created metaphors, we conducted a new rating study assessing the same psycholinguistic

variables. The questionnaire was administered online to 49 participants (mean age = 21.69; SD = 1.38), who were asked to rate on a 1-to-5 Likert scale the meaningfulness (i.e., how meaningful the sentence was), difficulty (i.e., how difficult it was to judge the meaningfulness of the sentence), and familiarity (i.e., how familiar the expression was to them/ frequency of experience) of the sentences. The newly rated items had the same characteristics of the previous ones and overall metaphors were less familiar (metaphor mean = 2.81, literal mean = 3.41, $p < 0.001$), less meaningful (metaphor mean = 3.59, literal mean = 4.00, $p < 0.001$), and more difficult (metaphor mean = 1.89, literal mean = 1.69, $p = 0.002$) than their literal counterparts. Finally, the mean (SD) length of the overall set of materials was 11.17(1.53) number of words.

Sentences and relative rating can be retrieved from the Figurative Archive (Bressler et al., 2026) either by consulting the Zenodo repository at the following link: https://doi.org/10.5281/zenodo.14924804, selecting the "Dataset_MetaEducation.xlsx" file, or by accessing the dedicated shinyapp (link: https://neplab.shinyapps.io/FigurativeArchive/), by selecting the "Everyday metaphors" section and flagging "Yes" in the "IUSS NEPLab MetaEducation Study" column.

4.2.3 EEG Procedure and Assessment

The EEG recording session lasted about 45/50 minutes, during which participants sat in a comfortable chair in a dimly lit room at 100 cm from a 19in computer screen. To ensure participants' attention, on 1/3 of the stimuli, at the end of the trial, participants were asked to perform a word matching task in which they indicated by button press which word out of two best matched the previous sentence. The two alternatives consisted of one related word, depicting a feature of the target word relevant for the interpretation of the statement, and one unrelated alternative (e.g., glare vs. switch for the metaphorical item "In hard times hopes are stars that light the soul up").

The EEG session started with the presentation of written instructions followed by a short training phase that familiarized participants with the experimental procedure and the task. Each trial started with a

750ms fixation cross displayed at the center of the screen, followed by the experimental sentence presented word-by-word with a 600ms ISI and a 200ms blank screen. ERPs were time-locked to target words, corresponding to metaphors' vehicles and to their literal equivalents (e.g., "stars" in "In hard times hopes are stars that light the soul up" and "Those lights in the night sky are stars of distant galaxies"). The presentation of trials was randomized. Then, the two lexical alternatives of the word matching task were displayed on the left and right of the screen for a maximum time window of 3000ms. If participants' responses exceeded this time window, a new trial started. The position in the screen (i.e., right/left) of the related and unrelated word was randomized across trials. Both response times and accuracy were measured.

Participants were randomly assigned to one of two lists such that the same subject reads only one version of the 80 sentence pairs (i.e., either metaphoric or literal). Each list was composed of 40 metaphorical, 40 literal, and 80 non-metaphorical filler items. Participants were given two scheduled breaks and were asked to refrain from blinking and moving their eyes during the word-by-word presentation of the sentences.

4.2.4. EEG Recording and Data Pre-processing

The EEG was recorded from the scalp using 60 electrodes mounted in an appropriately sized cap (EasyCap Brain Products) according to the 10-20 International System: Fpz, Fp1/Fp2, Af3/Af4, Af7/Af8, Fz, F1/F2, F3/F4, F5/F6, F7/F8, Fc1/Fc2, Fc3/Fc4, Fc5/Fc6, Ft7/Ft8, Cz, C1/C2, C3/C4, C5/C6, T7/T8, Cpz, Cp1/Cp2, Cp3/Cp4, Cp5/Cp6, Tp7/Tp8, Tp9/Tp10, Pz, P1/P2, P3/P4, P5/P6, P7/P8, Poz, Po3/Po4, Po7/Po8, Oz, O1.

Horizontal eye movements were monitored using one electrode placed at the outer canthus of each eye. Vertical eye movements were monitored using two electrodes placed respectively over and beneath the right eye. Scalp electrode impedances were kept below 5 kΩ. Data from all scalp electrodes were referenced online to an electrode placed close to the vertex, and later offline referenced to the

mathematical average of the left and right mastoids. The raw EEG signal was collected in AC current with a low cut-off filter (time constant 10s) at a sampling rate of 500 Hz and was amplified using Brain Amp® passive amplifiers (Brain Products GmbH, Gilching, DE).

EEG data processing was performed using the EEGLAB and FieldTrip toolboxes for MATLAB. The continuous EEG signal was bandpass filtered between 0.1 and 40 Hz using a fourth-order Butterworth filter. Non-biological artifacts (e.g., pauses between trials) were removed, and noisy channels were automatically identified and rejected using the clean_rawdata plugin. Independent Component Analysis (ICA) was applied to the continuous data to correct for ocular artifacts. Components related to eye blinks and horizontal eye movements were identified using the ICLabel algorithm and subtracted from the data. Following ICA, the rejected channels were interpolated, and the signal was re-referenced to the average of the mastoids. The data were then segmented into epochs time-locked to the onset of the target word. Artifact rejection was performed using a semi-automatic procedure: epochs with amplitude fluctuations exceeding ±150 μV were flagged for rejection, and the remaining trials were visually inspected to remove any residual artifacts. The average trial rejection rate was 6%. Five participants were excluded from the final analysis due to a rejection rate exceeding 40%, and eight additional participants were excluded due to technical failures. Finally, ERPs were computed for a 2000 ms window relative to a 200 ms pre-stimulus baseline.

The EEG data were pre-processed using BrainVision Analyzer 2.0.3 (Brain Products GmbH). For each participant, epochs (from -600 to 1500ms) time-locked to the onset of the target word in the metaphorical and literal conditions were extracted. An independent component analysis (e.g., Groppe et al., 2009; Mennes et al., 2010) was carried out on the obtained epochs to identify and remove artifacts due to blinks and horizontal eye movements. The independent component analysis was conducted in a semi-automatic modality: the maximum amplitude range allowed in each epoch was fixed at 150 μV; the remaining epochs were visually inspected and corrected when contaminated by artifacts. After ICA correction, artifact inspection was conducted again to identify any residual artifacts, and the affected epochs were rejected. The average rejection rate was 6%. Participants with more than 40% rejection rate were excluded from

final data analyses. Based on this criterion, 5 participants were excluded from statistical analyses. Eight additional participants were excluded due to technical failure. ERPs were computed for 1450ms after the onset of the target word relative to a -200ms stimulus baseline.

4.2.5. Computational measures

For each metaphorical and literal expression, we computed three measures: semantic similarity from word embeddings, a Bayesian pragmatic measure, and surprisal from a set of Large Language Models. x

4.2.5.1. Semantic similarity

Semantic similarity was computed as the cosine of the angle between the word embedding of the target word (namely, the vehicle of the metaphor or the corresponding word in the literal statement, e.g., "stars") and the mean vector obtained by averaging across the word embeddings of its preceding context (e.g., "In hard times hopes are" or "Those lights in the night sky are"). Word embeddings were extracted from the Italian pre-trained semantic space by FastText (Grave et al., 2018), trained with a CBOW architecture on Common Crawl and Wikipedia. A higher semantic similarity corresponds to a closer semantic relationship between the target word and the preceding context.

4.2.5.2. Bayesian pragmatic measure

The Bayesian pragmatic measure was adapted from a previous study by Werning et al. (2019). Adjusted to our stimuli, the measure corresponds to the pragmatic load of updating the prior belief of the pragmatic listeners, based on the semantics of the metaphorical terms, with the likelihood of using a certain metaphor to communicate a certain characteristic of the topic. The prior belief was operationalized as a monotonic function of the semantic similarity between the vehicle of the metaphor and its preceding context, representing the lexical representation of the expression. The likelihood, corresponding to the informativeness of the utterance, was operationalized using familiarity ratings collected for the study (available within the *Figurative Archive,* Bressler et al., 2026). Following previous Bayesian and rational

modeling approaches (e.g., Scontras et al., 2021), the Bayesian pragmatic measure (BPM) was defined as a posterior score, computed as the product of exponentiated prior and likelihood terms:

BPM = f(Semantic Similarity (v, c) * g(familarity) $\propto$ exp (semantic similarity) * exp (familiarity)

To ensure numerical stability and to prevent extreme values from disproportionately influencing model estimates, both semantic similarity and familiarity were rescaled to the [0, 1] interval before exponentiation. A higher BPM was associated with lower pragmatic inferential load needed, while a lower BPM was associated with a greater pragmatic load.

4.2.5.3. Surprisal

Surprisal, defined as the log probability of the target word (namely, the vehicle of the metaphor or the corresponding word in the literal statement, e.g. "stars") given its preceding context (e.g., "In hard times hopes are" or "Those lights in the night sky are"), was estimated from four large language models: mGPT (Shliazhko et al., 2024), GPT2 adapted for the Italian language (de Vries & Nissim, 2021), Minerva 1B (Orlando et al., 2024) and Llama 3.2 (Meta AI, 2024). All models were available in Hugging Face and were accessed through the package *transformers* (Wolf et al., 2020). Criteria for models' choice included i) they are open access, ii) can be run on a standard portable computer, and iii) were trained, or adapted, on corpora that include Italian text. Multi-token words, which represent a large portion of Italian words, were handled by summing the log probabilities of each sub-word token, as stated by the chain rule. Code to estimate surprisal was adapted from (de Varda et al., 2023). A higher surprisal corresponds to a less predictable target word given the preceding context.

4.2.6. Statistical Analysis

ERPs were analyzed with Linear Mixed-effects Models (LMM; Pinheiro & Bates, 2000) in R (R Core Team, 2025) using *lme4* and *lmerTest* (Bates et al., 2015) packages. For the N400 window (300-500 *ms*), we considered a subset of centro-parietal electrodes (CPz, CP1, CP2, CP4, P1, Pz, P2, P4, PO3, PO4, POz),

while for the later time window (600-800 *ms*), we considered a cluster of frontal electrodes (AF7, AF8, AF3, AF4, Fz, F1, F3, F5, F7, F4, F6, F8). The time windows were chosen based on the most common ones employed in the literature (see §Study 5), while the choice of the clusters of electrodes was based on visual inspection.

We carried out two sets of analyses for each time window. The first analysis aimed to assess the effect of the condition (metaphors vs. literal statements) on the EEG amplitude. The categorical variable of condition was effect coded, and the random structure included random intercepts for items, subjects, and channel, and random slopes for condition by both subjects and items. The resulting formula was *lmer(EEG amplitude ~ condition + (1+ condition |subj) + (1+ condition |item) + (1|ch)*.

The second analysis focused on the contribution of the three computational measures (surprisal, semantic similarity, and the Bayesian pragmatic measure). The computational measures were centered on the mean, and the random structure included random intercepts for items, subjects, and channels, and random slopes for the computational measures by subjects. The resulting formula was *lmer(EEG amplitude ~ surprisal + BPM+ sim + (1+ surprisal + BPM + sim |subj) + (1 |trial) + (1|ch)*. To assess whether the effect of the three computational measures was present within each condition, we also fitted the same Linear Mixed-effects Model independently for the two subsets of metaphors and literal statements.

A schematic representation of the approach is reported in Figure 4.1.

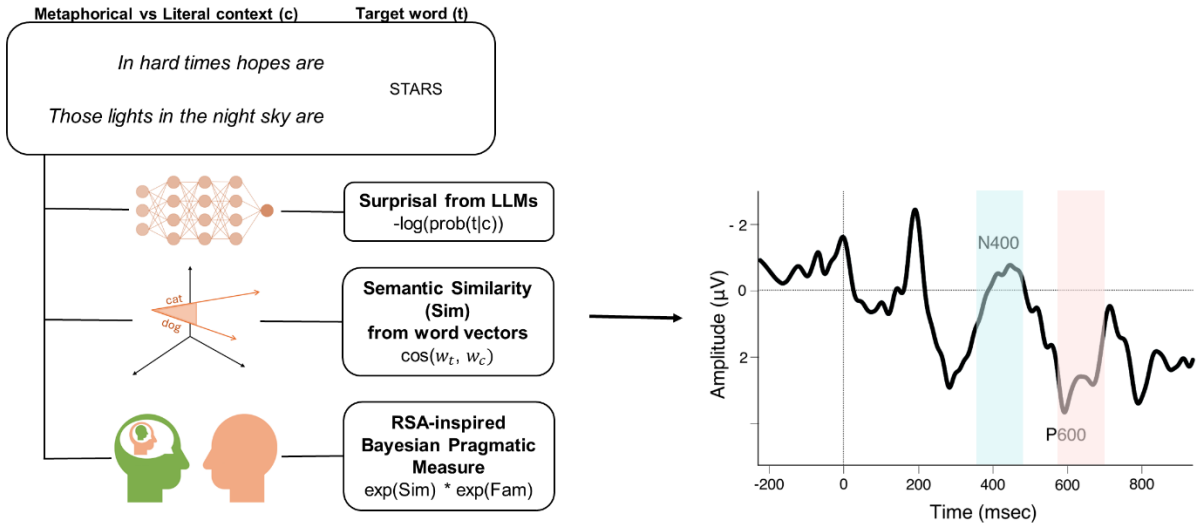


Figure 4.1. Rationale of the study. The study aims at modeling the EEG amplitude in the N400 and P600 windows, via three computational measures: surprisal, semantic similarity, and a Bayesian pragmatic measure.

4.3. Results

4.3.1. Computational measures

In Table 4.1, we report the descriptives of the computational measures across conditions. Overall, the two-tailed *t*-tests revealed significant differences across all measures (all *p*s < .001), with metaphors being associated with higher surprisal, lower semantic similarity, and lower BPM than literal statements (see also Figure 4.2B).

Moreover, correlation analysis reported expected patterns of relationship within measures. Surprisal values from distinct LLMs were positively correlated, while they negatively correlated with both semantic similarity and Bayesian pragmatic measures, with more surprising sentences being associated with lower similarity between the target word and its context, and lower pragmatic load (Figure 4.2A).

Measure	**Literal**	**Metaphors**	***t*-test**
Surprisal GPT2	17.63 (6.13)	20.79 (5.56)	-3.42 ***
Surprisal Llama 3.2	15.96 (3.34)	19.85 (3.93)	-6.75 ***
Surprisal Minerva	36.01 (3.56)	40.36 (4.12)	-7.16 ***
Surprisal mGPT	15.43 (3.38)	20.75 (4.10)	-8.96 ***
Semantic similarity	0.46 (0.09)	0.28 (0.09)	12.97 ***
BPM	3.50 (0.80)	2.22 (0.55)	11.88 ***

Note: For each measure, we report the mean (SD) and *t*-test between the literal and metaphorical conditions. *p<.05, **p<.01, ***p<.001

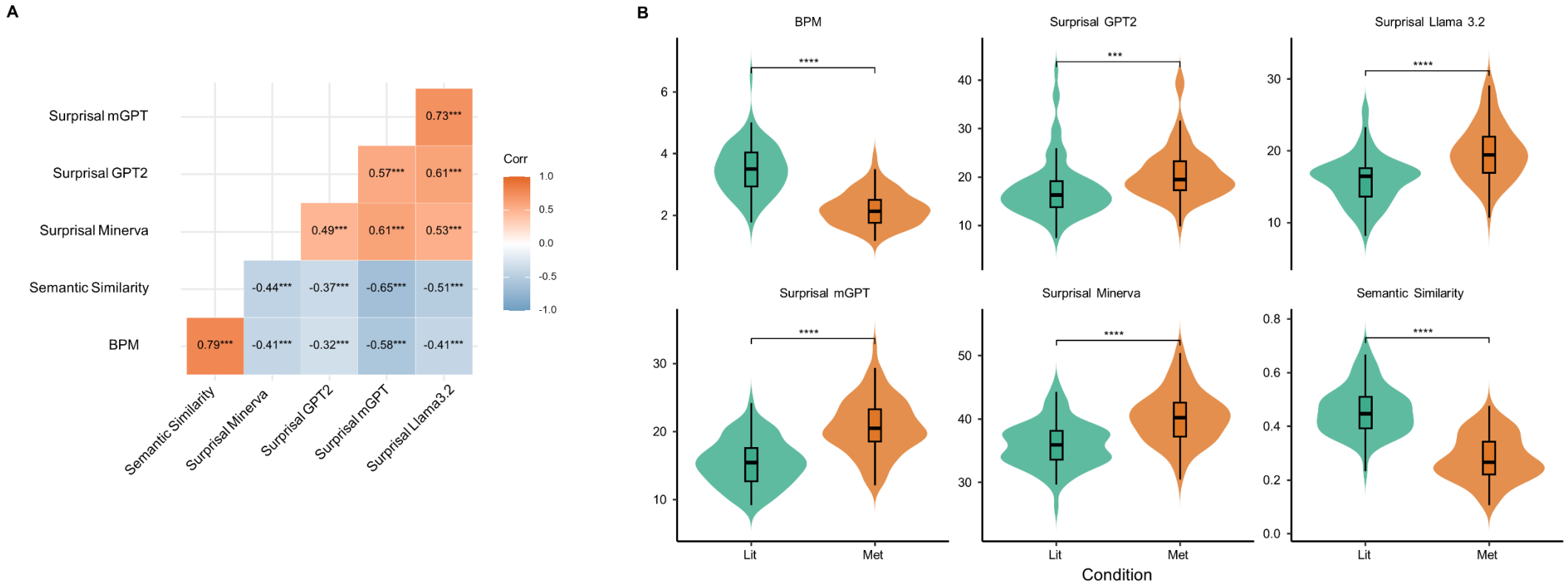


Figure 4.2. Computational measures for literal and metaphorical expressions. Panel A shows the correlation matrix between Bayesian Pragmatic Measure (BPM), surprisal values from various Large Language Models (GPT2, Llama 3.2, mGPT, Minerva), and semantic similarity. All correlations are corrected for multiple comparisons. Panel B shows the violin plots illustrating the distribution of these measures across literal (in green) and metaphorical (in orange) conditions. Asterisks indicate significant differences (*p<.05, **p<.01, ***p<.001).

4.3.2. Behavioural Results

The mean accuracy of correctly selecting the related-word option during the behavioural task was 93.15% and 93.31% in the metaphoric and literal conditions, respectively. Mean response times were 1.42 (SD = 0.47) and 1.41(SD = 0.68) in the metaphoric and literal conditions, respectively, with no significant difference across conditions (t = 0.25, p = 0.80). Overall, these results suggest that all participants paid attention to the presented stimuli while their EEG was recorded, obtaining a good task accuracy.

4.3.3. ERP results

Waveforms from 15 representative electrodes are shown in Figure 4.3. Visual inspection revealed differences in the brainwaves across conditions. Metaphors showed a negative peak around 400 *ms* and a positive trend around 600 *ms*, compared to their literal counterparts. The N400 peak emerged specifically in centro-parietal electrodes, while the P600 effect had a more frontal distribution. A complete summary of the results is reported in Figure 4.4, while the summaries of the models for Analysis 1 and 2 are reported in Table 4.2 for the N400 window and in Table 4.3 for the P600 window.

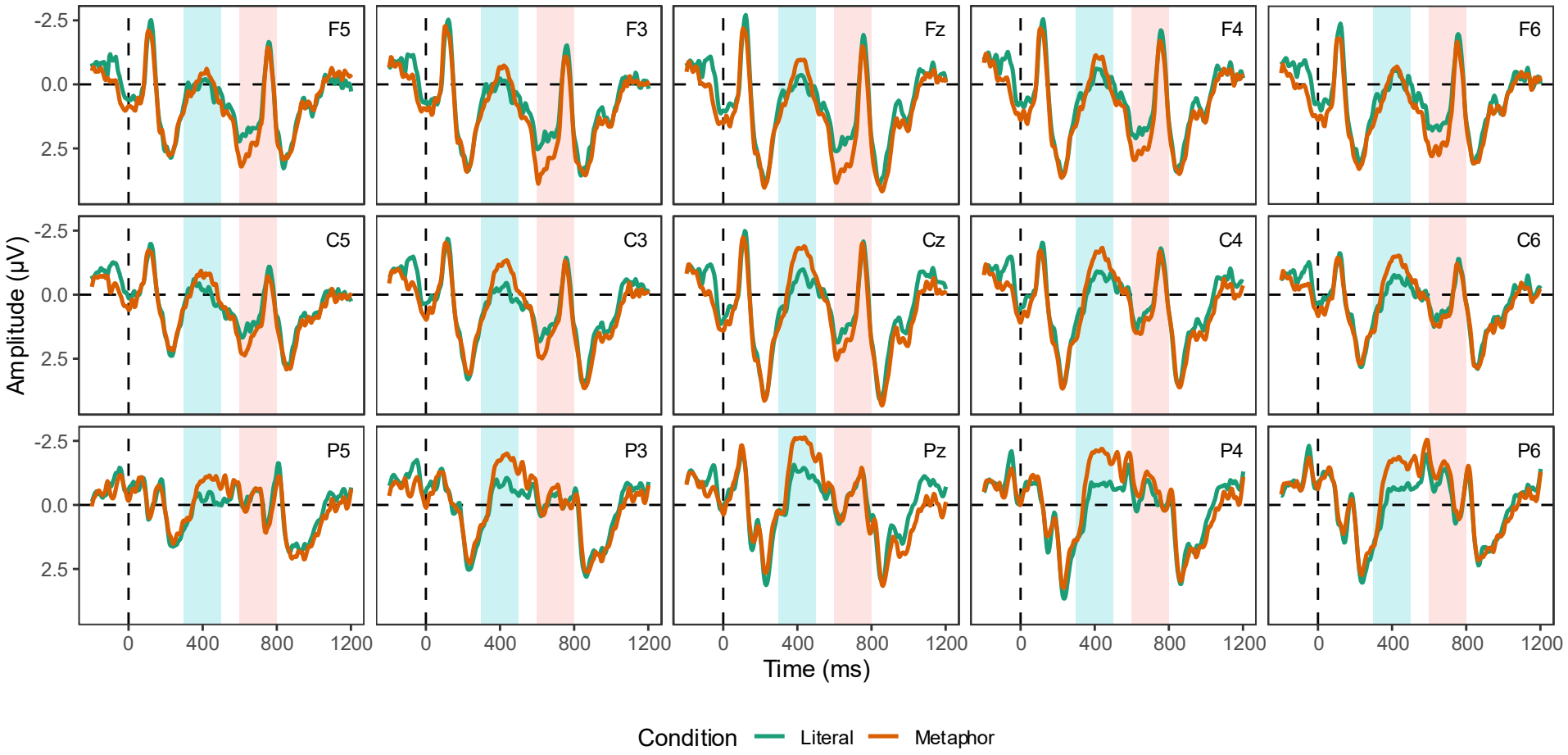


Figure 4.3. ERP Grand Averages. The figure depicts the ERPs from a set of 15 representative electrodes. Voltage amplitudes are displayed from −200 to 1200 ms from the onset of the target word. Orange lines represent metaphors, while green lines represent literal statements. Negative is plotted upward.

4.3.3.1 ERP Analysis 1: Condition

In the first time window (ranging from 300 to 500 *ms*), we examined ERPs across the subset of centro-parietal electrodes. The Linear Mixed-Effects Model revealed a main effect of Condition (-0.84 μV, t = -3.39, p = 0.001), indicating that metaphors elicited significantly greater N400 compared to literal statements. In the later time window (ranging from 600 to 800 *ms*), we considered ERPs across the subset of frontal electrodes. We found a main effect of Condition (0.68 μV, t = 2.70, p = .007), with metaphors associated with greater P600 compared to literal statements.

4.3.3.2. Analysis 2: Computational measures

In the N400 time window, we observed a main effect of surprisal[7] (-0.38 μV, t = -2.81, p = 0.005), indicating that items with higher surprisal, namely, target words that were less predictable given the preceding context, are associated with more negative deflection. On the contrary, no significant effect emerged for semantic similarity (p = 0.97) and BPM (p = 0.16). Looking at the two subsets based on condition, we found that none of the three computational variables has an effect on the difference in EEG amplitude, either within metaphors or within literal statements.

In the later window, we observed a reverse pattern of results with respect to surprisal: this measure significantly predicted the ERPs amplitude (0.38 μV, t = 3.13, p = .002), with more surprising items (less predictable given the context) being associated with more positive responses. Additionally, we found a significant effect of BPM in this later window (β = -0.34 μV, t = 2.28, p = 0.023), suggesting that stimuli with greater pragmatic load (namely, a lower BPM) elicited more positive ERPs. No effect of semantic similarity was reported (p = 0.11). Looking at the two subsets based on condition, we found that, also in this later window, none of the three computational variables influences the difference in EEG amplitude, either within metaphors or within literal statements.

[7] While surprisal from all four LLMs showed the same pattern of results, in the Results section we included only surprisal from mGPT, which showed the best goodness of fit (see Supplementary Table 4.1).

Table 4.2. Outputs of the Linear Mixed-effects Models in the N400 window (centro-parietal electrodes)

	Analysis 1			**Analysis 2**			**Analysis 2: Metaphors**			**Analysis 2: Literals**		
Predictors	*Estimates*	*Statistic*	*p*	*Estimates*	*Statistic*	*p*	*Estimates*	*Statistic*	*p*	*Estimates*	*Statistic*	*p*
Intercept	-1.01	-3.74	**<0.001**	-1.01	-3.77	**<0.001**	-1.30	-3.28	**0.001**	-0.96	-2.94	**0.003**
Condition	-0.84	-3.39	**0.001**									
Surprisal mGPT				-0.38	-2.81	**0.005**	-0.19	-0.86	0.388	-0.51	-1.72	0.085
BPM				0.26	1.40	0.162	0.38	0.97	0.332	-0.09	-0.28	0.776
Semantic Similarity				0.01	0.03	0.973	-0.30	-1.04	0.300	0.21	0.50	0.618

Table 4.3. Outputs of the Linear Mixed-effects Models in the P600 window (frontal electrodes)

	Analysis 1			**Analysis 2**			**Analysis 2: Metaphors**			**Analysis 2: Literals**		
Predictors	*Estimates*	*Statistic*	*p*	*Estimates*	*Statistic*	*p*	*Estimates*	*Statistic*	*p*	*Estimates*	*Statistic*	*p*
Intercept	1.15	5.44	**<0.001**	1.15	5.57	**<0.001**	1.50	4.32	**<0.001**	0.90	2.93	**0.003**
Condition	0.68	2.70	**0.007**									
Surprisal mGPT				0.38	3.13	**0.002**	0.36	1.57	0.115	0.35	1.23	0.218
BPM				-0.34	-2.28	**0.023**	0.13	0.34	0.731	-0.32	-1.16	0.246
Semantic Similarity				0.26	1.58	0.114	0.21	0.65	0.515	0.45	1.21	0.227

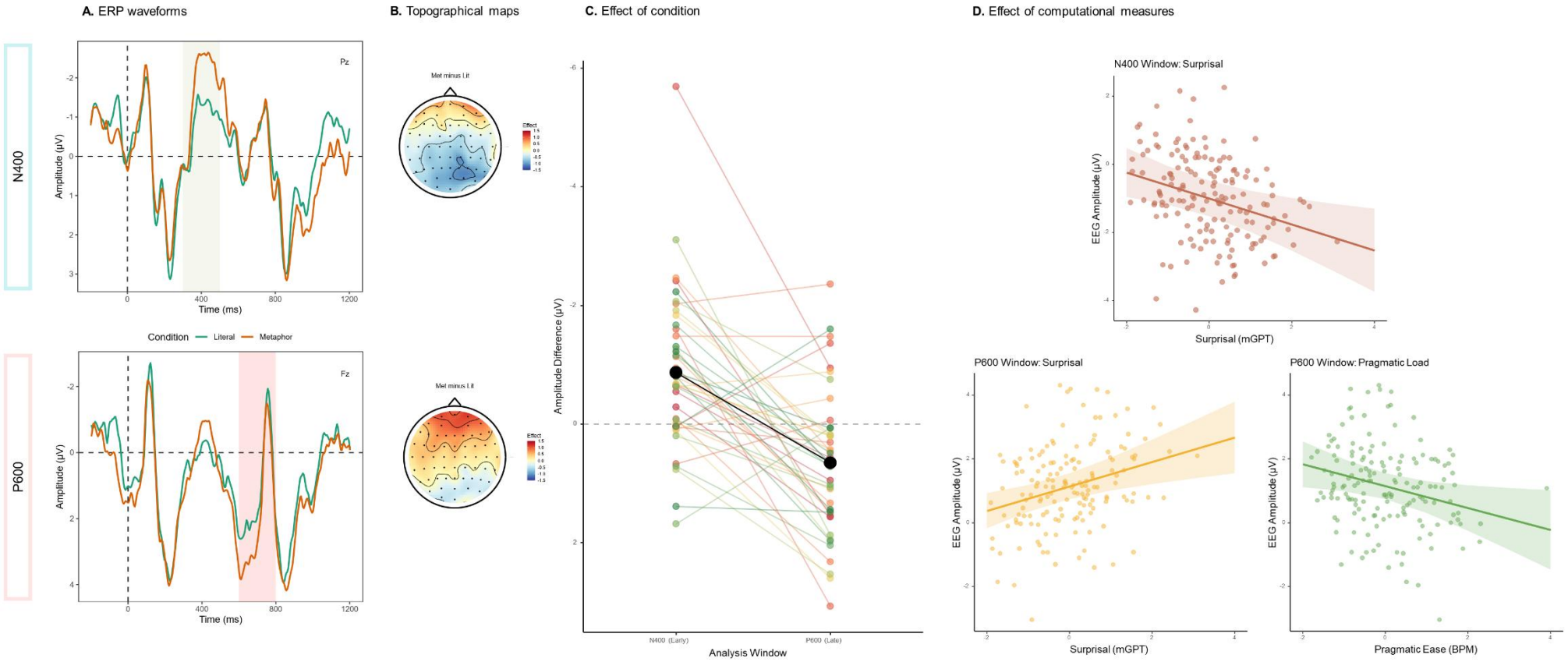


Figure 4.4. Electrophysiological patterns and role of computational measures in metaphor processing. Panel A shows the grand-average ERP waveforms at electrode Fz for literal (green) and metaphorical (orange) conditions, in the N400 (300–500 *ms*) and P600 (600–800 *ms*) windows. Panel B shows the topographical maps reporting the distribution of the difference (Metaphor minus Literal) for each window. Panel B shows the mean amplitude differences (Metaphors minus Literals) across analysis windows. Each point represents a participant, and the black dots represent the means of the sample. Scatter plots in Panel D illustrate the predictive power of computational measures on EEG amplitude: surprisal in the N400 window (top), and surprisal (bottom left) and Bayesian Pragmatic Measure (BPM, bottom right) in the P600 window.

4.4. Discussion

The aim of this work was to model electrophysiological responses to metaphorical expressions by comparing three complementary theoretical accounts that have guided the functional interpretation of ERP components in metaphor and language processing more broadly. These accounts have variously highlighted, as main processes underlying electrophysiological components, semantic operations of conceptual alignment and mapping, pragmatic inference, and predictive mechanisms. Operationally, we implemented these different accounts via three computational measures: semantic similarity from word embeddings, as an index of the pragmatic account, a Bayesian pragmatic measure based on the RSA framework, as an index of pragmatic inference, and surprisal from LLMs, as an index of the predictive approach.

Visual and statistical inspection of EEG data showed that metaphors were associated with a greater N400 component, distributed across a cluster of centro-parietal electrodes, and greater positivity in the subsequent time window, across a cluster of frontal electrodes. These results are in line with a biphasic pattern of metaphor processing, previously reported in the literature (Baiocco et al., 2025; Bambini et al., 2016; Coulson & Van Petten, 2002; De Grauwe et al., 2010).

Considering the modeling power of computational measures, we found that surprisal strongly accounted for the difference in the waveforms between metaphors and literal statements in both time windows. Metaphors are typically associated with higher surprisal values than literal statements, and this difference predicted with great accuracy both the greater negativity for metaphors in the N400 window and their greater positivity in the P600 window. As reported for the comparisons between literal and anomalous sentences or jokes (J. A. Michaelov et al., 2024; H. Xu et al., 2024), the predictability of an expression indexed by surprisal seems to be a main driver in the electrophysiological response. Thus, these results support the role of a general predictive mechanism distinguishing metaphor processing from the comprehension of literal statements and the predictive accounts of the N400, according to which early processing reflects the detection of violations of contextual expectations (Kuperberg & Jaeger, 2016; Van Petten & Luka, 2012).

In the P600 window, we found that, in addition to surprisal, the effect of the Bayesian pragmatic measure emerged. Specifically, lower BPM values, which indicated increased inferential pragmatic demands, were associated with greater positivity. Metaphors reported higher pragmatic inferential load than literal statements, which led to a greater re-analysis effort indexed by the greater late positivity. So, unlike the N400, which responds to violation of linguistic expectations, the P600 also appears to index the cognitive effort of inferring communicative intent beyond literal semantic content. This suggests that pragmatic inference plays a role in a later stage of metaphor processing, supporting the view of P600 as an index of pragmatic operations showed both in metaphor (Bambini et al., 2016) and irony processing (Regel et al., 2014; Spotorno et al., 2013).

Notably, while our measures distinguished metaphors from literal statements, we did not observe a graded effect within the metaphorical or literal sets. This is a limit previously highlighted for surprisal, which showed difficulty in capturing subtle graded effects (Krieger et al., 2025), but also possibly due to the intra-condition balance of our stimuli.

From a theoretical standpoint, what emerged clearly is a crucial role of context and expectations, declined in the N400 as context-based expectations, with a focus on predictions created by the purely linguistic context, and in the P600 as expectations about what the speaker wants to communicate and how, presuming that their semantic choices are guided by their utility to convey a certain meaning.

So, while metaphorical and literal statements are significantly different across all measures, their difference in terms of electrophysiological response is explained mostly by a different degree of predictability in the N400 window and by a combination of predictability and pragmatic load in the P600 window. These results point towards two main conclusions. First, we confirmed that predictive mechanisms strongly guide language processing, shaping the neural signature associated with metaphors yet representing only one facet of the process (Lago et al., 2024). After recognizing metaphors as unpredictable utterances, inferential operations take place, leading the subjects to reason not only on context-based expectations but also on their expectation about the communicative intent of the speaker, thereby inferring the intended meaning. Second, these results help us understand the debated nature of

the P600, variously linked to semantic retrieval or more pragmatic operations, providing evidence in favor of the latter.

While computational modeling could help us shed light on the mechanisms underlying ERP components, some limitations of our approach should be highlighted. On a more technical side, as surprisal from different LLMs showed different degrees of fit to the data, further research should include semantic similarities from semantic spaces other than FastText to allow a similar comparison. Moreover, it must be noted that not all computational models are comparable in terms of training data. For instance, while FastText and mGPT, which entered the final analysis, were both trained on Wikipedia and Common Crawl data and the amount of Italian token was comparable (in the order of tens of billions), mGPT is also trained on much more data in other 22 languages and that could drive the higher predictive power on human data (but see Oh & Linzen, 2025 for evidence that larger models do not necessarily align better with human brain data). On a more theoretical side, it is important to acknowledge that good predictive performance alone does not ensure cognitive or neural plausibility. Computational systems such as LLMs are trained to learn statistical regularities in large corpora, and their internal representations and learning mechanisms may differ substantially from those implemented in the human brain. Consequently, even when model-derived measures such as surprisal align well with electrophysiological data, this correspondence should be interpreted as descriptive evidence regarding underlying cognitive processes (Shah & Varma, 2025).

4.5. Conclusions

Taken together, the present findings support a temporally differentiated account of metaphor comprehension in which predictive and pragmatic mechanisms contribute at distinct stages of processing. Across both time windows, surprisal emerged as the most robust predictor distinguishing metaphors from literal statements, indicating that metaphorical expressions are primarily processed as globally less predictable inputs. This effect dominated the N400 window, consistent with predictive accounts, while in the later P600 window, the role of BPM became relevant as well. This suggested that, after recording a prediction error, inferential mechanisms are engaged to derive the intended meaning.

Together, these results reconcile competing interpretations of ERP components in metaphor research by showing that early effects primarily reflect predictive mechanisms, whereas later effects incorporate also more specific pragmatic operation. Rather than alternatives, distinct theoretical accounts, such as the general predictive model and post-Gricean views, seem to anchor different stages in metaphor processing.

Appendix A

Supplementary Table 4.1. AIC comparison between Linear Mixed-effects Models with surprisal from different LLMs.

Predictor	**df**	**AIC**
N400		
Surprisal mGPT	6	**219045.3**
Surprisal Minerva	6	219077.8
Surprisal GPT2	6	219109.4
Surprisal Llama 3.2	6	219153.8
P600		
Surprisal mGPT	6	**238701.3**
Surprisal Minerva	6	238728.4
Surprisal GPT2	6	238744.7
Surprisal Llama 3.2	6	238760.0

STUDY FIVE

ELECTROPHYSIOLOGICAL SIGNATURES OF FIGURATIVE LANGUAGE: PRELIMINARY FINDINGS FROM A SYSTEMATIC REVIEW AND DIGITALIZATION-BASED META-ANALYSIS[8]

Abstract

Figurative language processing has been associated, in Event Related Potential (ERP) studies, with two main components, namely the N400 and the P600. Particularly, the latter has not been consistently reported across studies, leaving open the question of whether we can talk of a pragmatic P600. However, many issues arising for meta-analytical approaches to ERP data, such as the multiple time windows employed or the lack of reference to effect sizes, have hindered the possibility of providing an answer. In this work, by employing a novel methodology based on figure digitalization, we present the first quantitative meta-analysis of ERP data for figurative language. Our results suggest that the N400 strongly emerges as a key response to figurative language, tightly linking pragmatics to the manipulation of context. The P600 is confirmed as a more nuanced component, deeply influenced by the specific phenomenon and its novelty.

[8] This chapter is a manuscript in preparation for submission to a peer-review journal as "Canal, P., Vespignani, F., Mangiaterra, V., Luciani, F., Frau, F., Bischetti, L., & Bambini, V. Electrophysiological signatures of figurative language: a systematic review and digitalization-based meta-analysis"

5.1. Introduction

Saying "What a beautiful day!" during a storm, using the word "shark" to describe a lawyer, or asking someone to "spill the beans" means communicating something that goes beyond the compositional meaning of the words in the utterances. All these sentences are different instances of non-literal uses of language, showing a gap between the literal meaning conveyed by the single words and the intended meaning, which can be derived through inferential processes. Among non-literal expressions, we find *irony*, a figure of speech aiming at communicating the opposite meaning of what is said, *metaphors*, describing one thing in terms of another to highlight a shared quality, and *idioms*, namely non-compositional expressions with a highly conventional meaning and recurrent structure.

Various models have been proposed to explain how listeners derive the interpretation of figurative language and the time course of this process (R. Gibbs & Colston, 2012). Some argued in favor of indirect access to the figurative meaning, after that the literal meaning has been processed and rejected (Grice, 1975), while others claimed that the figurative meaning can be directly accessed without a first literal stage (R. Gibbs, 1994). One methodology that allowed to shed light on neuro-chronometry of language elaboration in general, and also on the processing of nonliteral expressions, is electroencephalography (EEG). Electrophysiological studies have highlighted the role of two Event-Related Potentials (ERP) components for figurative language processing: the N400 and the P600.

The N400 is a negative deflection between 200 and 600 ms peaking at 400 *ms* after stimulus presentation, associated with the retrieval of conceptual knowledge and typically reported for semantically anomalous or unpredictable words in sentences (Kutas & Federmeier, 2011). In figurative language processing, it is usually reported for metaphors and metonymy, with amplitudes correlating to the novelty of the expression (Bambini et al., 2016; Lai & Curran, 2013; Weiland et al., 2014). The P600 has a more undefined nature. Originally, it was characterized as the Syntactic Positive Shift (SPS) due to its association with a number of syntactic violations (Hagoort et al., 1993), such as tense (Osterhout & Nicol, 1999) and agreement violations (Coulson et al., 1998). However, later it was associated with other types of semantic violations that did not elicit an N400, such as theme assignment violations, suggesting that P600's nature is not purely syntactic. Focusing on figurative expressions, the P600 has been variously attested for all the previous types of nonliteral language, even if not consistently (De Grauwe et al., 2010; Schumacher, 2011; Spotorno et al., 2013; Vespignani et al., 2010).

ERP researchers have attempted to relate their findings to the theoretical accounts of metaphor processing, yet many issues emerged. First, as noted by Bambini & Resta (2012), the same pattern (for instance, the biphasic pattern N400-P600) has been interpreted by Pynte et al. (1996) as evidence for the direct-access hypothesis, while De Grauwe et al. (2010) claimed that it supported the indirect-access view. Moreover, the electrophysiological response associated with pragmatic phenomena is not homogeneous.

The case of metaphor is paradigmatic of the inconsistency of results; as documented by Baiocco et al. (2026) some studies reported a biphasic pattern, others a sustained negativity after the N400, and in certain cases, the later window is not considered in the statistical analysis.

The variability of ERP findings in the literature mainly concerns the P600 effects. Thus, the current study aims at i) assessing the size and consistency of the N400 effects, and ii) evaluating the existence of a *pragmatic P600*, looking at ERP effects across studies on figurative language processing, namely those focusing on metaphor, metonymy, irony, idioms, and proverbs. To do this, we conducted a systematic review of the literature on figurative language processing with ERPs and a meta-analysis.

Quantitative meta-analyses of ERP studies are associated with a series of issues, which do not make it feasible to apply the standard meta-analytic approach. First of all, ERP data analysis is typically carried out on time windows selected by the authors, which may greatly change from study to study. Secondly, our main component of interest (i.e., the P600) may not always be the focus of ERP studies on figurative language processing, and effects at later windows may not be tested at all in the original paper. Third, researchers may have considered the topographic factors in different ways, for example, focusing on different subsets of electrodes, which could hinder the comparison of results across studies. Lastly, measures of effect size are rarely reported, and, when available, they are also dependent on the choice of the time window previously mentioned. These limitations have historically prevented ERP researchers from conducting the kind of quantitative meta-analyses that have become standard practice in fMRI research

In the present study, we applied a novel methodology developed by Vespignani (2020), which allowed us to overcome the issues previously outlined. This methodology, similar to the Great Grand Average approach proposed by Sambrook & Goslin (2015) and advocated by Moran et al. (2017), consists of digitizing the ERP waveforms depicted in the figures of eligible papers. From the digitized figure, it is possible to extract the coordinates of each waveform and coherently re-analyse the ERP data of all eligible studies with a uniform approach.

Based on the literature, we expected that figurative language would be strongly associated with the N400 response and that a more nuanced pattern would emerge for the P600.

5.2. Methods

5.2.1. Paper search and selection

Paper search was first performed in PubMed, Scopus, and Google Scholar databases in June 2024. An updated search was then carried out in September 2025 on PubMed and Scopus only.

The string was composed of 'ERP keywords' AND 'pragmatic phenomena keywords', resulting in ("ERP" OR "ERPs" OR "event-related potential*") AND ((("figurative" OR "nonliteral") AND ("language" OR "expression*")) OR "idiom*" OR "proverb*" OR "metaphor*" OR "metonym*" OR "irony" OR "ironic" OR "sarcasm" OR "sarcastic"). A slightly simplified version was adopted for Google Scholar, which does not allow combinations of keywords. For searches on PubMed and Scopus, search terms were applied to title, abstract, and keywords, without constraints to year of publication. Google Scholar, from which we retrieved the articles using the *Publish or Perish* software (Hazing, 2007) restraining year of publication between 2020 and 2024, does not permit the restriction of the fields to which the search terms are applied, resulting in a larger, but often less relevant, number of articles being retrieved.

Inclusion criteria for eligibility were:

- Study must be on language comprehension, in written (word-by-word) or auditory modality, with analyses focused on a specific target word. Paradigms that are far from reading or listening tasks (e.g., priming, hemifield presentation) should be excluded.
- Study must be carried out on a sample of healthy adults (between 16 and 60), all L1 subjects.
- The paper must have at least one ERP plot of a single channel along the midline (Oz, Pz, CPz, Cz, FCz, Fz...), or a cluster of channels including one midline electrode (excluding Occipital and Frontal).
- Plot(s) must show a pre-stimulus interval (baseline) ≥ 100ms.
- Plot(s) must show a post-stimulus activity ≥ 800ms.
- A literal control condition is needed.
- Data must be re-referenced to mastoids, ear lobes, or nearby sites, or Average. We will exclude unusual references (e.g., tip of the nose).

Exclusion criteria were:

- The paper was written in languages other than English or other languages known by the authors (Italian).
- The paper was not an original peer-reviewed research article (thus leaving out reviews, conference papers, book chapters, not-peer-reviewed preprints, etc.).

5.2.2. Coding procedures

Each paper was annotated following a coding scheme developed by two authors (PC, VM) as a JSON schema. The coding scheme aimed at collecting the main features of the papers across four domains: EEG recordings (online and offline reference, filters, sampling rate, epoch rejection, epoch segmentation, electrodes, hardware and laboratory environment), subjects characteristics (age, language spoken, handedness, n° of females), procedure (language, item characteristics – part-of-speech, context in which are embedded, type of manipulation, design, word duration, potential linguistic features collected) and statistical analysis (N400 and P600 window, subset of electrodes, statistic, degrees of freedom and reported amplitude). The annotation was carried out by three annotators (PC, FL, VM), and for the purposes of inter-annotator agreement, 10 papers were annotated independently by all three annotators.

5.2.3. Digitalization

For each paper, electronic copies of figures complying with the inclusion criteria at Section ***2.2*** were retrieved and uploaded to Web Plot Digitized (Rohatgi, 2025). The first phase of digitalization consisted of axis calibration, by marking two points on the X axis and two points in Y axis with a mouse click to define the coordinates of the figures (Figure 1A). When more than one graph is shown in the same figure, this procedure is repeated for each pair of axes, and each calibration is named as the electrode or cluster of electrodes shown in the graph. The second phase consisted of extracting the coordinates of each waveform by laying points along each line (Figure 1B). When the resolution of the image and the color-coding of conditions allowed it, the extraction was automatically performed using the dedicated tool on Web Plot Digitized that lays a point every 5 or 10 pixels, followed by a phase of manual adjustment. Otherwise, when the resolution was too low, or all the conditions were shown in the same colors with different types of dashing, which made the automatic procedure impractical, extraction was entirely manual. Each waveform was referred to the axis calibration of the corresponding electrode and named as the name of the electrode or cluster of electrodes, followed by the name of the condition represented by that waveform. For each figure, a JSON file reporting the axis calibrations and the coordinates of each waveform was created.

From the digitized coordinates, we linearly interpolated voltage between the digitalization sampling points.

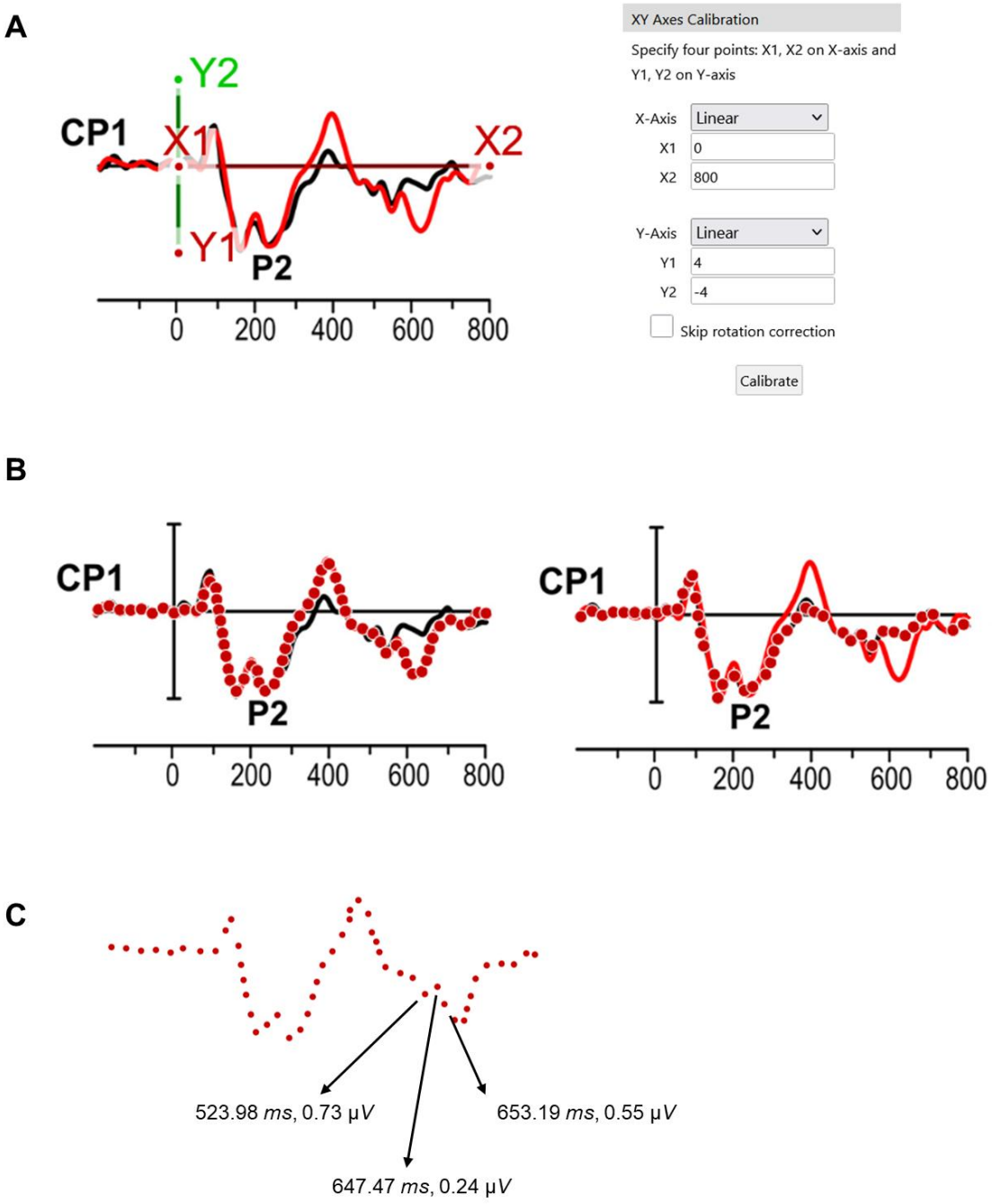


Figure 5.1. The process of ERP waveforms digitalization. Panel A shows the first phase of digitalization, namely axis calibration. Panel B shows the digitalization of the waveforms of each condition. Panel C shows the extraction of the waveforms' coordinates from the digitization.

5.2.4. Statistical analysis

To prepare the data for analysis, we first filtered the data into discrete spatial Regions of Interest (ROI). Each electrode location was moved to a common 10-05 (Oostenveld & Praamstra, 2001)bidimensional

space. A Frontal ROI was defined to occupy between 10 and 30% of y axis, and the posterior occupied between y < -0.1 and y >-.90), while restricting the lateral spread (|x|<0.75) to focus on mid-line activity.

Since ERP studies lacked a consistent effect size description, we accounted for the relative uncertainty of each study based on participant and item sample size, in particular we computed a composite weight to be used in modeling corresponding to the multiplication of the number of subjects by the square root of the number of items.

To test how phenomena influence EEG amplitude across the ROI regions in the N400 and P600 windows, we fitted Bayesian multilevel robust regression models using the *brms* package (Bürkner, 2017). Metonymy and proverbs were excluded from the analysis, given the limited number of eligible papers investigating those phenomena. For each Experiment in each Paper, we used the average voltage of all channels falling in the two ROIs to compute the voltage difference between figurative and literal conditions that served as the dependent variable (DV). The DV was weighed by the composite weight and tested against two factors: ROI and Phenomenon, allowing us to assess whether the spatial distribution of the ERP effect changes depending on the specific figurative language type. For metaphor studies, we fitted a Bayesian multilevel robust regression model distinguishing between the nature of the metaphorical stimuli (when reported in the original study), namely, conventional or novel expressions.

To account for the hierarchical structure of the meta-analytic data, we included nested random intercepts for individual experiments within their respective papers. We defined weakly informative, symmetric robust priors, which do not assume a specific polarity but constrain the effect to a plausible range (approximately ±6μV).

5.3. Results

5.3.1. Systematic review

The literature search returned 2667 records (PubMed: 250; Scopus: 291; Google Scholar: 1726), of which 316 were automatically excluded as duplicates, not original research articles, and not written in English or Italian. Screening for eligibility was then performed by VM (Figure 5.2) and yielded 66 papers. Some descriptives of the studies, as resulted from the annotation, are reported in Table 5.1 and in the Supplementary Materials.

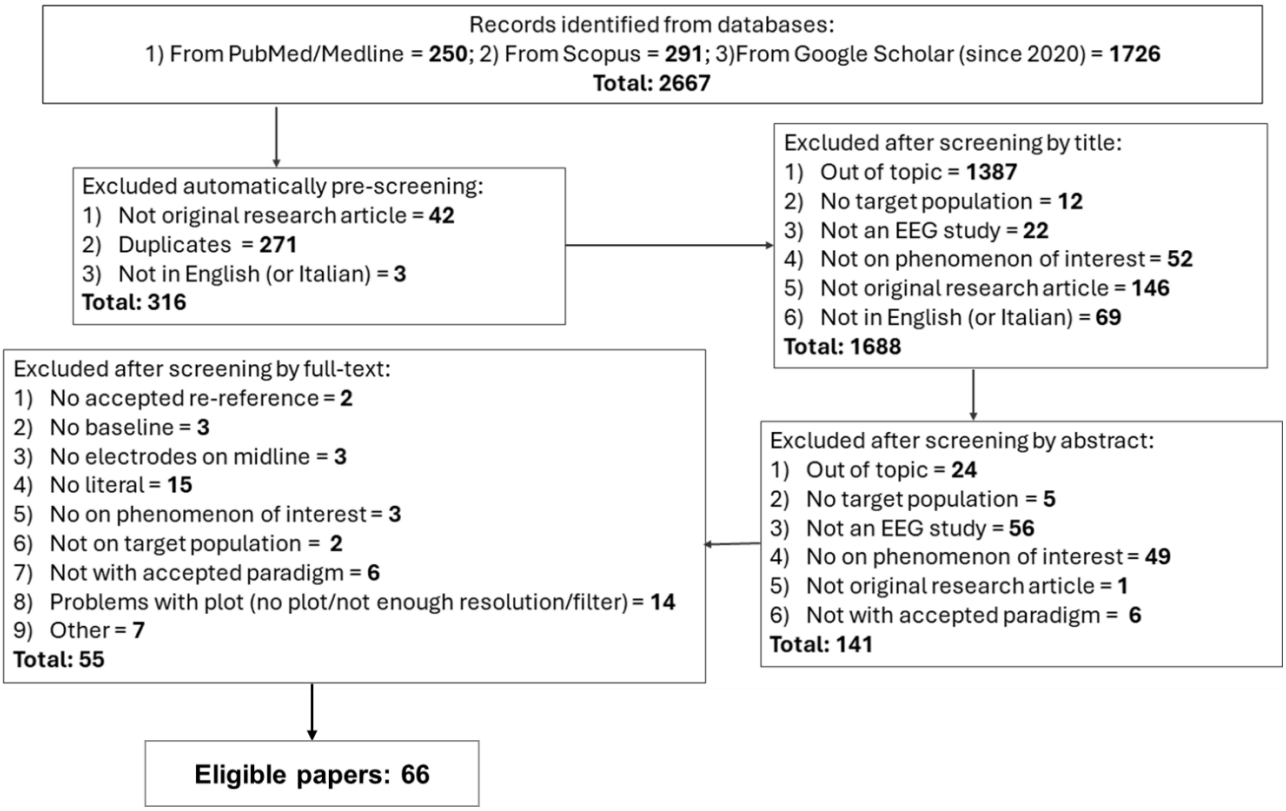


Figure 5.2. Steps of the systematic review. The figure reports the flowchart of paper selection for inclusion into the systematic review.

Table 5.1. Results of the systematic review. The table shows the main characteristics of eligible papers.

Papers	Phenomena	N° Subjects	Language	Modality	Context	Word Type
Abraham et al. (2021)	Metaphor	40	German	Written	Single Sentence - Minimal <8 Words	Verb
Aldunate et al. (2019)	Metaphor	18	Spanish	Written	Single Sentence - Minimal <8 Words	Noun
Arzouan et al. (2007a)	Metaphor	29	Hebrew	Written	Two-Words Paradigm	Mix
Arzouan et al. (2007b)	Metaphors	15	Hebrew	Written	Single Sentence - Minimal <8 Words	Mix
Baiocco et al. (2025)	Metaphor	39	English	Written	Single Sentence - Non Minimal >8 Words	Mix
Bambini et al. (2016)	Metaphor	13	Italian	Written	Single Sentence - Minimal <8 Words	Noun
Bambini et al. (2019)	Metaphor	22	Italian	Written	Naturalistic	Noun
Bambini et al. (2024)	Metaphor	32	Italian	Written	Two-Words Paradigm	Noun
Besson et al. (1997)	Proverbs	9	French	Written	Single Sentence - Minimal <8 Words	Mix
Bonnaud et al. (2002)	Metaphor	10	French	Written	Two-Words Paradigm	Noun
Caillies et al. (2019)	Irony	49	French	Spoken	Single Sentence - Minimal <8 Words	Adjective
Canal et al. (2017)	Idioms	42	Italian	Written	Single Sentence - Non Minimal >8 Words	Noun
Champagne-Lavau et al. (2023)	Irony	24	French	Written	Mini-Discourse	Adjective
Coopmans et al. (2022)	Idioms	40	Dutch	Spoken	Single Sentence - Non Minimal >8 Words	Verb

Cornejo et al. (2007)	Irony	20	Spanish	Written	Mini-Discourse	Noun
Coulson & Van Petten, 2002)	Metaphor	21	English	Written	Single Sentence - Non Minimal >8 Words	Noun
De Grauwe et al. (2010)	Metaphor	24	English	Written	Single Sentence - Minimal <8 Words	Noun
Ferretti et al. (2007)	Proverb	24	English	Written	Mini-Discourse	Mix
Filik et al. (2014)	Irony	32	English	Spoken	Dialogue	Adjective
Fondevila et al. (2016)	Metaphor	24	Spanish	Written	Single Sentence - Minimal <8 Words	Mix
Forgács et al. (2015)	Metaphor	42	English	Written	Two-Words Paradigm	Noun
Gold et al. (2010)	Metaphor	16	Hebrew	Written	Two-Words Paradigm	Mix
Goldstein et al. (2012)	Metaphor	28	Hebrew	Written	Two-Words Paradigm	Mix
Hubbard et al. (2023)	Idioms	24	English	Written	Single Sentence - Non Minimal >8 Words	Noun
Iakimova et al. (2005)	Metaphor	20	French	Written	Single Sentence - Minimal <8 Words	Noun
Jankowiak et al. (2021)	Metaphor	31	Polish	Written	Single Sentence - Minimal <8 Words	Noun
Ji et al. (2020)	Metaphor	25	Chinese	Written	Single Sentence - Minimal <8 Words	Verb
Jończyk et al. (2020)	Metaphor	43	English	Written	Single Sentence - Minimal <8 Words	Verb
Kazmerski et al. (2003)	Metaphor	48	English	Written	Single Sentence - Minimal <8 Words	Noun
Lai et al. (2009)	Metaphor	29	English	Written	Single Sentence - Minimal <8 Words	Mix
Lai & Curran (2013)	Metaphor	28	English	Written	Single Sentence - Minimal <8 Words	Mix

Lai et al. (2019)	Metaphor	28	English	Written	Dialogue	Mix
Laurent et al. (2006)		24	English	Written	Dialogue	Mix
Li et al. (2020)	Metaphor	28	English	Written	Single Sentence - Minimal <8 Words	Mix
Li et al. (2022)	Metaphor	34	English	Written	Single Sentence - Minimal <8 Words	Verb
Mauchand et al. (2021)	Idioms	30	French	Other	Single Sentence - Minimal <8 Words	Mix
Obert et al. (2018)	Irony	28	Finnish	Spoken	Mini-Discourse	Adjective
Pfeifer & Lai (2021)	Irony	44	English	Written	Mini-Discourse	Mix
Pfeifer et al. (2025)	Irony	44	English	Written	Two-Words Paradigm	Adjective
Proverbio et al. (2009)	Idiom	15	Italian	Written	Single Sentence - Minimal <8 Words	Noun
Pynte et al. (1996)	Metaphor	12	French	Written	Single Sentence - Minimal <8 Words	Noun
Regel et al. (2011)	Irony	44	English	Written	Mini-Discourse	Mix
Regel et al. (2010)	Irony	44	English	Written	Two-Words Paradigm	Adjective
Regel et al. (2014)	Idiom	15	Italian	Written	Single Sentence - Minimal <8 Words	Noun
Schmidt-Snoek et al. (2015)	Metaphor	12	French	Written	Single Sentence - Minimal <8 Words	Noun
Schneider et al. (2014)	Metaphor	26	German	Written	Dialogue	Mix
Schneider et al. (2015)	Metaphor	22	German	Written	Single Sentence - Minimal <8 Words	Noun
Schumacher, 2013)	Metonymy	24	German	Written	Dialogue	Mix
Shen et al. (2015)	Metaphor	26	Mandarin Chinese	Written	Single Sentence - Non Minimal >8 Words	Verb

Shen et al. (2022)	Metaphor	-23	Chinese	Written	Single Sentence - Minimal <8 Words	Noun
Shi & Li (2022)	Irony	33	Chinese	Written	Mini-Discourse	Adjective
Spotorno et al. (2013)	Irony	20	French	Written	Mini-Discourse	Mix
Sun et al. (2022)	Metaphor	48	Chinese	Written	Single Sentence - Minimal <8 Words	Noun
Tang, Qi, Wang, et al. (2017)	Metaphor	25	Chinese	Written	Single Sentence - Minimal <8 Words	Noun
Tang, Qi, Jia, et al. (2017)	Metaphor	20	Chinese	Written	Single Sentence - Minimal <8 Words	Noun
Tang et al. (2022)	Metaphor	20	Chinese	Written	Single Sentence - Minimal <8 Words	Noun
Tang et al. (2025)	Metaphor	31	Chinese	Written	Single Sentence - Minimal <8 Words	Noun
Tartter et al. (2002)	Metaphor	7	English	Written	Single Sentence - Non Minimal >8 Words	Noun
Vespignani et al. (2010)	Idiom	50	Italian	Written	Single Sentence - Minimal <8 Words	Verb
Wang et al. (2019)	Metaphor	40	Chinese	Written	Single Sentence - Minimal <8 Words	Noun
Wang et al. (2021)	Metaphor	30	Chinese	Written	Two-Words Paradigm	Noun
Wang et al. (2023)	Metaphor	60	Chinese	Other	Single Sentence - Minimal <8 Words	Noun
Weiland et al. (2014)	Metaphor	27	German	Spoken	Single Sentence - Non Minimal >8 Words	Noun
Weiland-Breckle & Schumacher (2017)	Metonymy	24	German	Written	Dialogue	Verb
Zhang et al. (2013)	Idioms	18	Chinese	Written	Single Sentence - Minimal <8 Words	Mix

Zhang et al. (2024)	Metaphor	25	Chinese	Written	Single Sentence - Minimal <8 Words	Verb

5.3.2 Bayesian regressions

The model in the N400 window suggested that metaphors have a 100% probability of eliciting a negative response both in the frontal (mean = -0.678, CI [-1.05, -0.31]) and in the posterior ROI (mean = -0.89, CI [-1.27, -0.53]). Irony showed a lower probability of displaying the N400 effects (frontal ROI 73%; posterior ROI 76%), yet a negativity emerged in both ROIs (frontal mean = -0.21, CI [-0.87, 0.43]; posterior mean = -0.24, CI [-0.90, 0.39]). Results for idioms are less straightforward, as the model suggested a 63% posterior probability of displaying a N400 response in the posterior ROI (mean = -0.16, CI [-1.11, 0.73]) and a 23% probability of displaying a positive response in the frontal ROI (mean = 0.35, CI [-0.58, 1.25]). The conditional means are shown in Figure 5.3, while the posterior probabilities are reported in Figure 5.4.

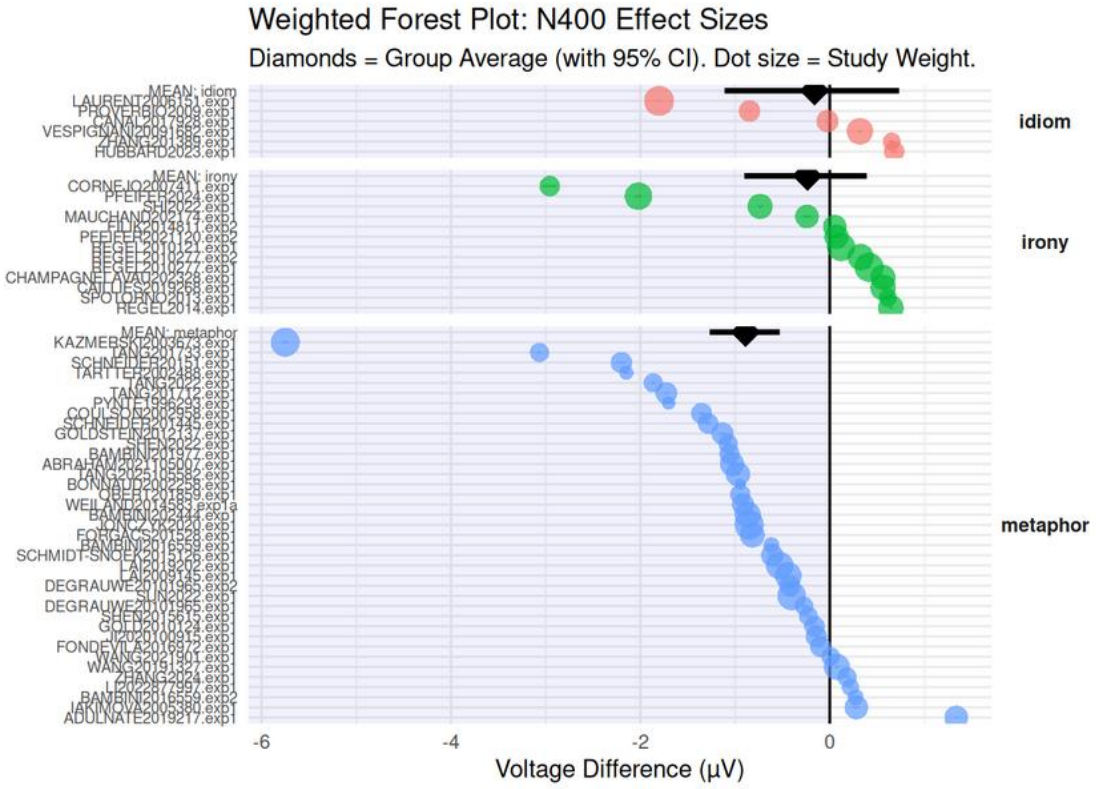


Figure 5.3. Forest plot in the N400 window.

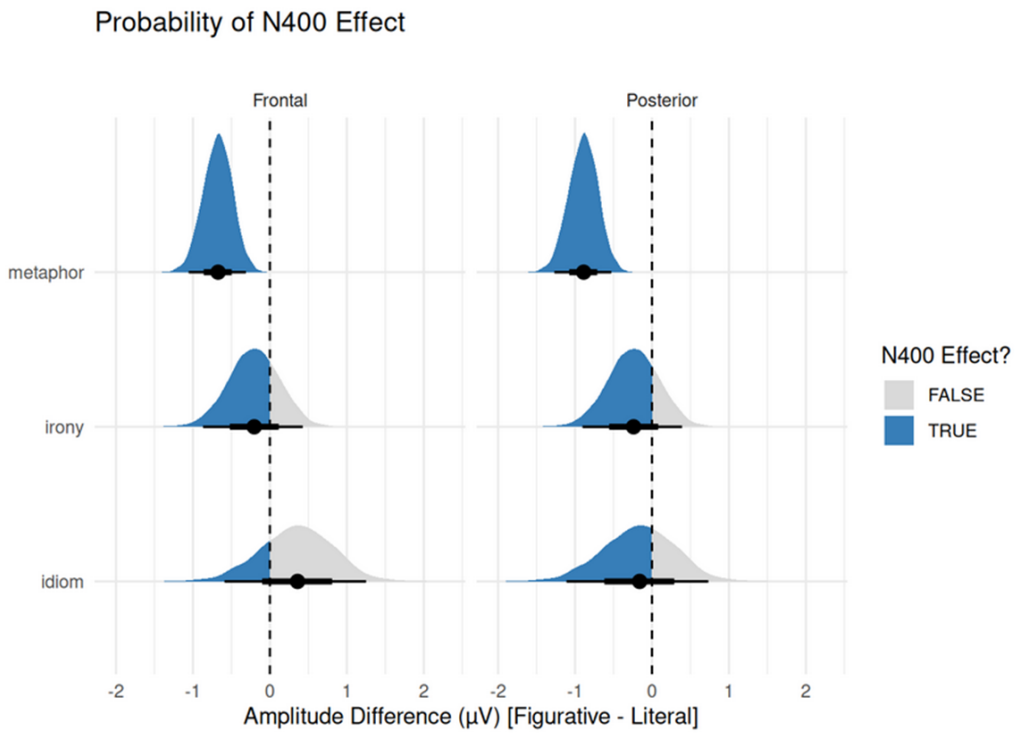


Figure 5.4. Posterior probabilities in the N400 window. The probability of reporting the effect is depicted in blue.

The model in the P600 window suggested that irony has a 99% probability of showing a positive peak in the posterior ROI (mean = 0.68, CI [0.14, 1.22]) and a 95% probability in the frontal ROI (mean = 0.45, CI [-0.08, 0.99]). Idioms showed a similar pattern with a 92% probability in the frontal ROI (mean = 0.52, CI [-0.22, 1.29]) and an 82% probability in the posterior ROI (mean = 0.34, CI [-0.40, 1.11]) of eliciting a P600. The pattern for metaphors was less conclusive, with a 2% probability of showing a negative response in posterior electrodes (mean -0.29, CI [-0.61, 0.01]) and a 31% probability of reporting a tiny negativity (mean = 0.07, CI [-0.38, 0.22]). The conditional means are shown in Figure 5.5, while the posterior probabilities are reported in Figure 5.6.

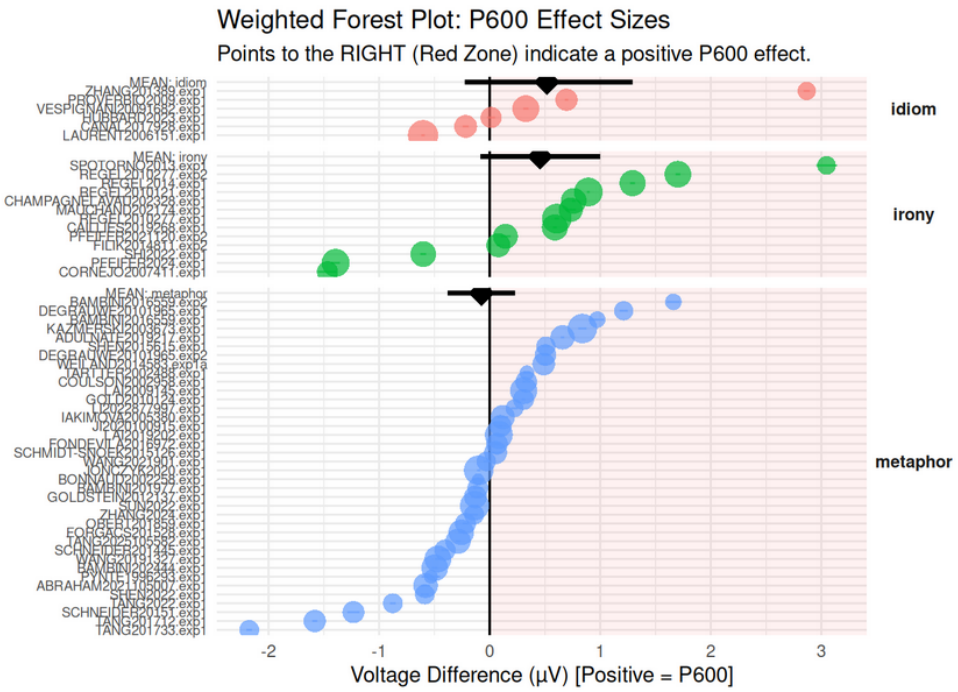


Figure 5.5. Forest plot in the P600 window.

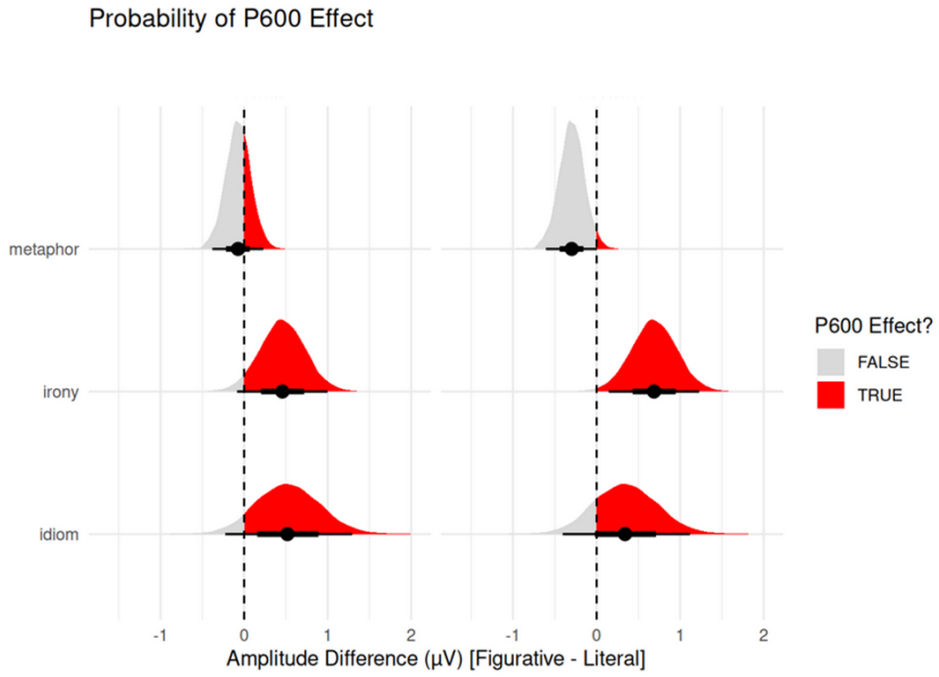


Figure 5.6. Posterior probabilities in the P600 window. The probability of reporting the effect is depicted in red.

The model investigating the role of the nature of the metaphorical stimulus (novel vs. conventional) suggested that novel metaphors have an 88% probability of reporting a P600 in the frontal ROI (mean = 0.15, CI [-0.10, 0.39]), while conventional metaphors have a 60% probability (mean = 0.03, CI [-0.21, 0.27]). Metaphors that were not labeled either as novel or conventional (labeled "Intermediate" in the

figures) had low probabilities (< 6%) of reporting an effect in the P600 time window. The conditional means are shown in Figure 5.7, while the posterior probabilities are reported in Figure 5.8.

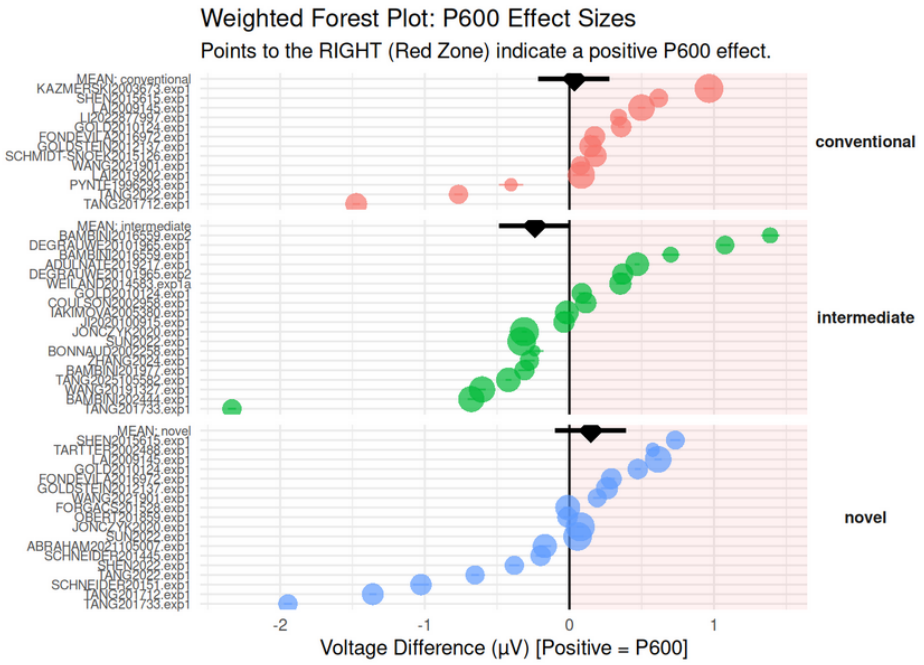


Figure 5.7. Forest plot of different types of metaphors in the P600 window.

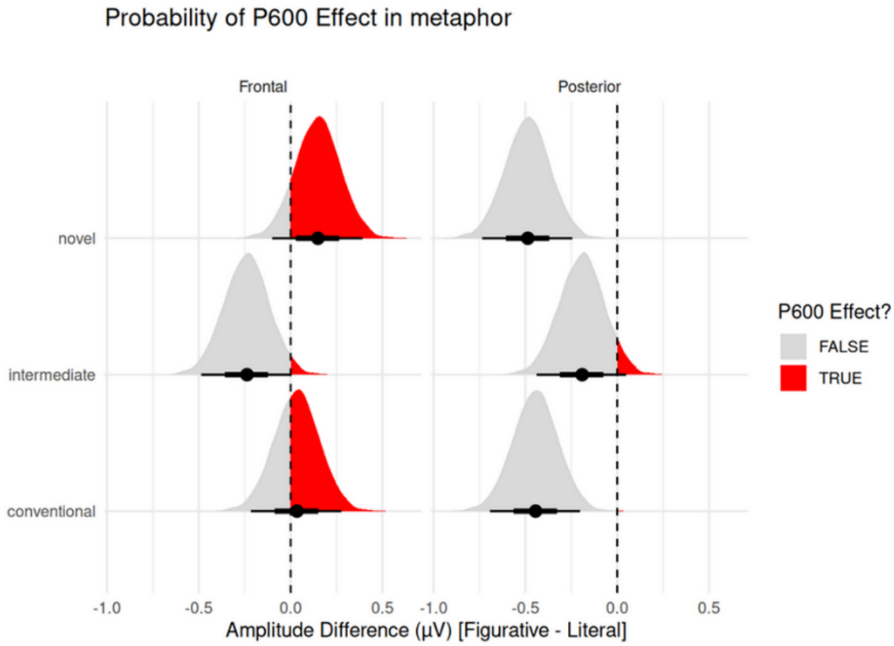


Figure 5.8. Posterior probability of different types of metaphors in the P600 window. The probability of reporting the effect is depicted in red.

5.4. Discussion

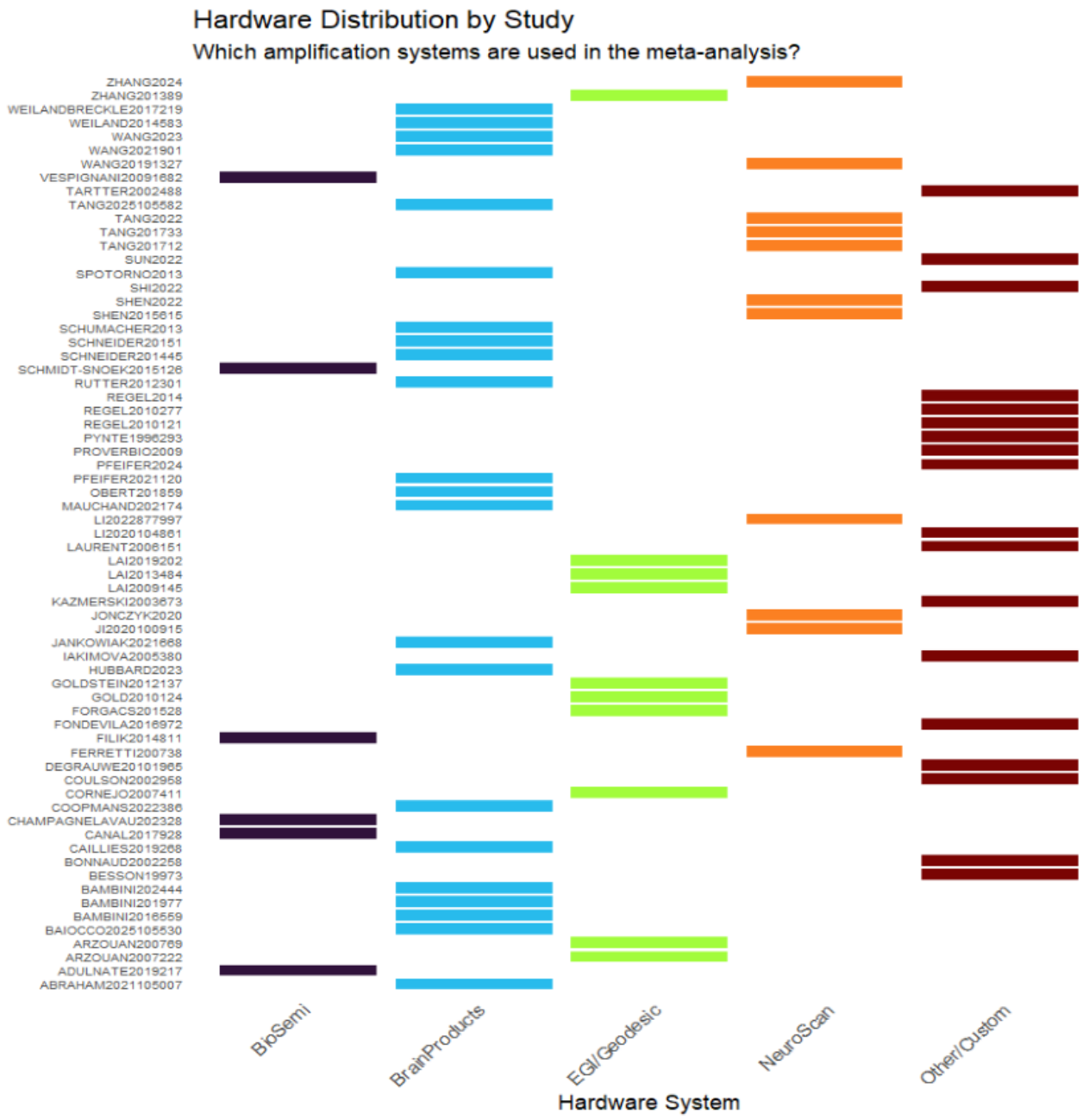
In this study, we presented the first study that, via a systematic review and a novel meta-analytic approach, provided a quantitative analysis of the time course of figurative language. In particular, we aimed to assess the robustness of ERP components typically associated with figurative expressions and whether we can talk of a pragmatic P600. The novel meta-analytic approach allowed us to overcome limitations associated with meta-analysis of ERP studies and to re-analyze EEG data across studies on figurative language processing. Specifically, we digitized the waveforms of figures in the papers retrieved during the systematic review to reconstruct ERP data via interpolation and analyze it with consistent time windows and regions of interest.

Preliminary results showed that metaphor consistently reported a strong N400 effect, both in frontal and posterior clusters of electrodes. The negativity in this time window was reported for irony and idioms as well, even if with a lower amplitude and probability, suggesting the semantic and contextual operations indexed by the N400 component are less strong for this type of figurative expression.

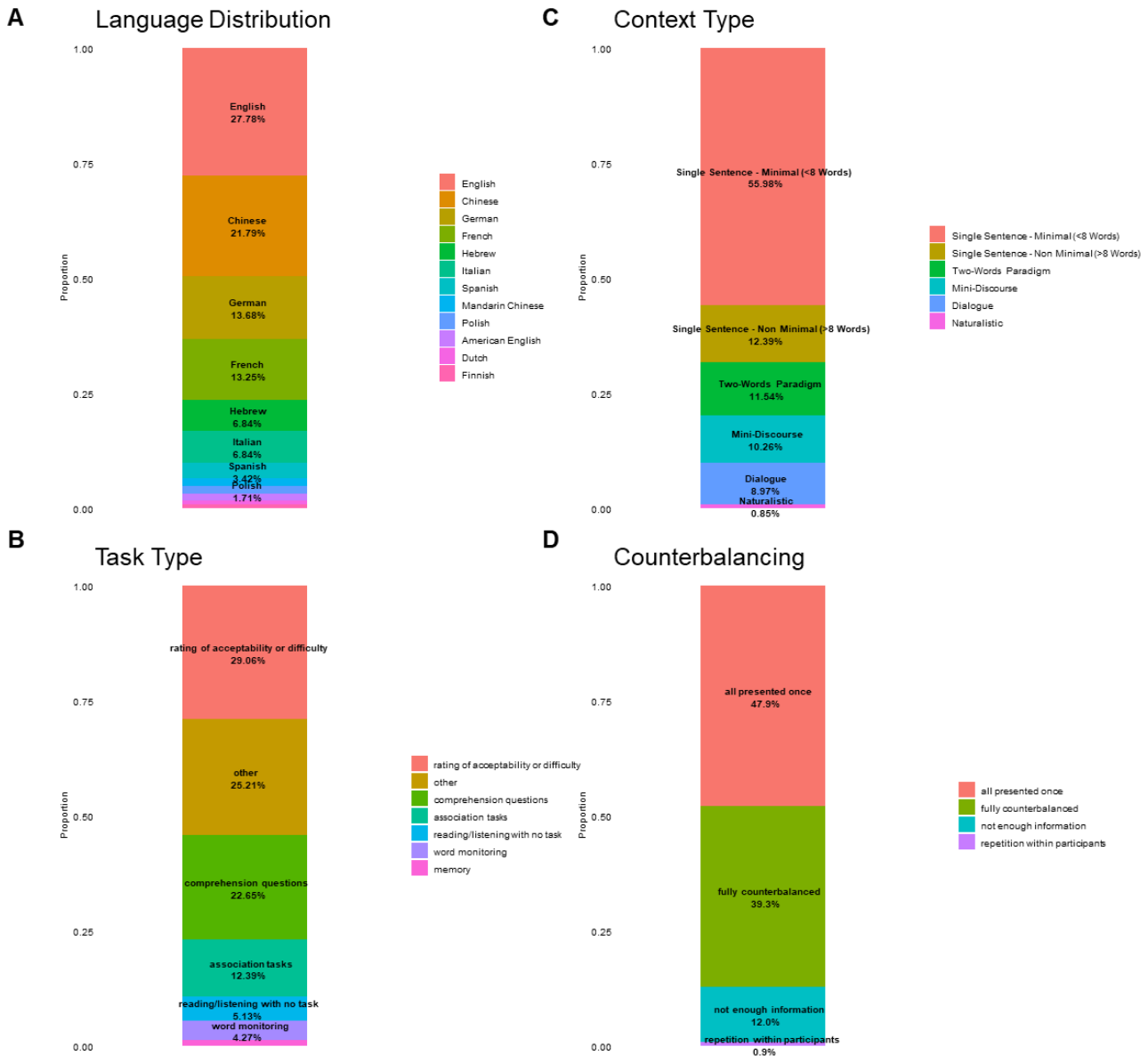
Regarding the P600, the effect did not emerge clearly across phenomena. While it is strongly elicited by ironic and idiomatic statements, it seemed not to be consistently reported for metaphors. The P600 effect for metaphors was, however, present in specific subsets, such as novel metaphors, and presented a more frontal distribution.

The present work allowed us to shed light on a number of issues in figurative language electrophysiology. First, N400 is the most robust electrophysiological response associated with pragmatic phenomena, while the P600 is confirmed as a more transient component. Assuming that N400 is mostly an index of context-based predictions (Lau et al., 2008), our results confirmed that pragmatics is strictly related to context and that semantic processes and context-guided adaptations underlie the processing of all pragmatic phenomena. However, many differences arose between the types of figurative language. For instance, irony is characterized by a strong P600 response and a less consistent N400, while the opposite is true for metaphors, emphasizing also the existence of phenomenon-specific mechanisms. Thus, the P600 emerged as a component linked to the re-analysis of the linguistic stimulus in a broader sense, extending far beyond the original scope as revision of the syntactic structure. This component seems to be linked to a high-level revision, which can take the form of an inference, of the stimulus, triggered not by explicit lexical markers, but by the necessity of integrating the socio-communicative intent. To answer our main research question—whether a pragmatic P600 exists—our results suggest that we can indeed speak of a pragmatic P600, but one that is deeply influenced by many factors, such as the specific phenomenon and its conventionality. Ultimately, the neural signature of pragmatics is distributed and multifaceted, representing a shift from purely linguistic processing to a broader, inference- and context-based reconstruction of intended meaning.

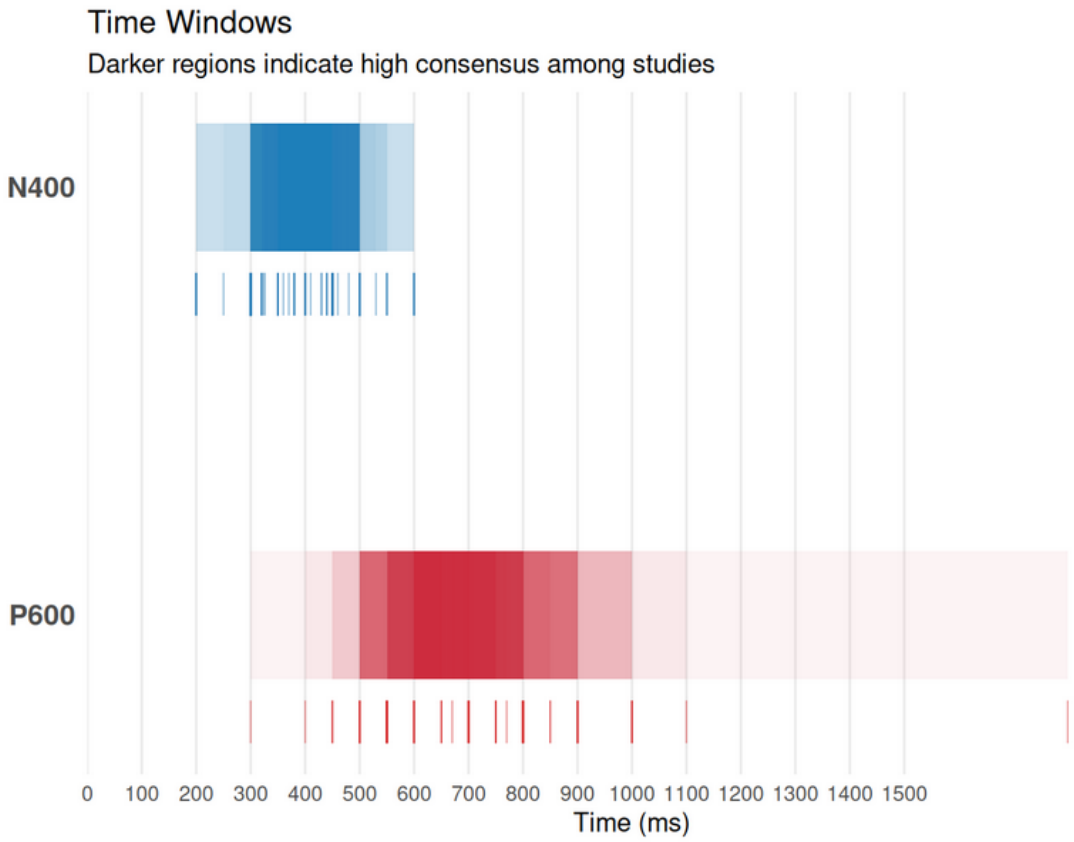
Appendix A



Supplementary Figure 5.1. Hardware distribution by study.



Supplementary Figure 5.2. Distribution of study characteristics across the dataset. Panel A shows the distribution of languages used in the studies. Panel B shows the distribution of task type. Panel C shows the distribution of the types of context in which metaphors are embedded. Panel D shows the distribution of the type of counterbalancing.



Supplementary Figure 5.3. Time windows employed across studies.

CONCLUSIONS

This thesis aimed to explore questions related to two domains of metaphor processing: i) how the processing demands of metaphorical expressions coined centuries ago changed as a function of the time elapsed between their creation and the time they are experienced, as well as of the textual genre within which they appear, ii) how the processing of metaphors unfolds in the brain, namely which operations are indexed by the Event-Related Potential (ERP) components that are typically reported in the literature, and how we can quantify metaphor features that influence the processing.

To answer these questions, the present work capitalized on the rising potential of computational tools such as word embeddings derived from Vector Space Models (Lenci, 2018) and the recent Large Language Models (LLMs, see Naveed et al., 2025), bringing both theoretical issues about metaphor processing as well as some methodological concerns about the validity and trustworthiness of these tools when applied to the study of human language. In particular, I explored two roles that computational tools can have in cognitive and language sciences: that of "participants", approximating human behavior on a certain task regardless of the mechanisms that lead to that outcome, and that of "models", providing a quantitative measure of a certain theoretical prediction.

In the following paragraphs, I will outline the main findings that emerged from the studies in this thesis, addressing first the theoretical inquiries into metaphor processing and then the methodological implications of using computational tools.

1. Metaphor processing in diachrony

The roles of the context and of the knowledge about the world in interpreting metaphors have been extensively investigated (Bambini et al., 2016; Briner et al., 2018; Jończyk et al., 2020; Shinjo & Myers, 1987). However, the combined effects of time and of semantic change in modifying the

shared coordinates of cultural context and world knowledge necessary to interpret metaphors written many centuries ago have never been tackled. **Study 2** and **Study 3** addressed this issue in two languages (Italian and English) to probe whether language-specific patterns influence the cost associated with metaphor processing and their changes in time. Our results suggested that metaphor processing is a dynamic operation, influenced by the broader diachronic changes happening in the language *tout court*. In Italian, the influence of time emerged based on the textual genre. Since literary language and everyday language were very similar in the past, metaphors were not processed with different efforts depending on genre (Steen, 1989). Today, however, readers have to make a greater effort to process metaphors in literary texts, which have a plainer and simpler language, while they are able to more easily activate the connection between distant concepts in the creative nonliterary language of the Web. English, on the other hand, achieved stability well before Italian, as did the distinction between genres, which remains the only factor driving a different processing demand for metaphors.

2. Metaphor processing in the brain

The study of the electrophysiology of metaphor processing started thirty years ago with the seminal work by Pynte et al. (1996). Since then, several studies have explored the time course of metaphor processing, with converging results and less consistent ones. (Bambini et al., 2019; Coulson & Van Petten, 2002; De Grauwe et al., 2010; Tang et al., 2017). **Study 4** and **Study 5** tried to shed some light on the existing literature via computational modeling and a novel meta-analystic approach. Specifically, in **Study 4** I modeled EEG data with three computational models (surprisal from LLMs, semantic similarity from word embeddings and a Bayesian pragmatic measure inspired by the Rational Speech Acts framework) implementing the most common views on the operations taking place during metaphor processing, namely predictive, semantic and inferential ones. In

Study 5, we conducted a systematic review of the literature and via digitalization we extracted EEG data from the original papers to provide a systematic re-analysis of accumulated EEG data.

Study 5 emphasized the stable role of the N400 component, which seems to be consistently elicited by metaphorical statements. **Study 4** linked it to context-based predictive operations, more than semantic or purely inferential ones, in line with predictive coding accounts (Nour Eddine et al., 2024) and confirming previous modeling studies relating surprisal to the N400 in other types of sentences (Frank et al., 2015; Michaelov & Bergen, 2020; Xu et al., 2024).

Regarding the P600, **Study 5** confirmed its unstable nature for metaphors (while it is more consistently elicited by other pragmatic phenomena, such as irony and idioms), possibly linked to the novelty of the expression, which enhances the probability to observe the late positivity. Its functional nature, however, seems to be associated not simply with context-based prediction, but also with pragmatic inferential processes, as the pragmatic measure emerged as a key predictor in this time window, alongside the stable predictive model (**Study 4**).

As a final remark, insights from **Study 1** suggest that collecting ratings to control for stimuli variability (a critical stage at the beginning of every EEG experiment) can be at least partially delegated to artificial agents, making it easier and faster to construct stimuli, thus giving further impetus to the study of ERP related to metaphor in new contexts and with new experimental manipulations.

3. Computational tools as participants

The possibility of using an artificial agent to augment or replace human participants is quite novel, and it emerged when it was possible to prompt chatbots directly in natural language, without the need to specifically train or fine-tune the models. The emergence of this possibility has generated much debate in the scientific community, with researchers emphasizing the alignment between

humans and models and others highlighting their lack of human-like errors and individual variability (Dillion et al., 2023; Harding et al., 2024; Wang et al., 2025). At the same time, the great accuracy obtained in single-word rating tasks (Brysbaert et al., 2024; Martínez, Conde, et al., 2024; Trott, 2024a) warrants the investigation of its validity for multifaceted expressions, such as metaphors. From the perspective of this thesis, the role of LLMs as participants is essentially utility-based. It does not aim to simulate how participants achieve a certain output, but simply to verify whether it is possible to replicate that output artificially. The results of **Study 1** supported the use of LLMs as artificial participants, given that machine-generated ratings closely approximate human ratings, also when included in statistical models predicting behavioral and electrophysiological responses. Limitations have emerged, as models align less with humans when they have to generate ratings on dimensions related to imaginability or for metaphors, particularly related to the sensorimotor domain, confirming that the lack of embodiment is one of the main domains of misalignment between humans and models (Borghi et al., 2023; Chemero, 2023).

Therefore, LLMs can indeed be used as participants, but as participants without body and sensorimotor experiences, as ratings related to the latter (such as imageability) do not seem to emerge as reliably as for other dimensions.

4. Computational tools as models

In a recent paper by van Rooij et al. (2024), the Authors cite an extract from Boden (2008): "Computers as such are in principle less crucial for cognitive science than computational concepts are". In this thesis, my approach to LLMs as models has tried to follow this kind of perspective. LLMs do not "make cognition" computationally; their emergent metaphorical abilities, *per se* tells us little about human metaphorical abilities. However, we can decompose our theoretical assumptions about metaphors in human minds in computationally feasible steps and probe this against the data. In **Study 4**, by operationalizing semantic, inferential, and predictive mechanisms,

we avoided the claim that the brain processed metaphors as an LLM or as a Vector Space Model and instead used these tools as quantity estimation devices to test the explanatory scope of different theories. What emerged from the experiment further confirms this stance: the predictability as computed with LLMs can capture a large part of the human electrophysiological patterns of metaphor processing, but leaves room to other processes, such as the inferential ones.

5. Cautionary remarks

While the results presented in the studies provide some positive evidence for the integration of LLMs in metaphor research, some caveats need to be highlighted. First, at the current stage of development, LLMs cannot be treated as a homogeneous or stable class of scientific objects. Consequently, the claims advanced in this thesis should be restricted to the specific models examined in each study, as generalization to LLMs as a whole is not warranted given the substantial variability in performance observed across models. Specifically, these findings should be interpreted as evidence for the conditional usefulness of LLMs as research participants, limited both to the specific tasks and the specific models under investigation , rather than as a general validation of their role as substitutes for human participants, which should continue to be investigated to provide data-driven recommendations and to avoid acritical, hype-drive adoption (Bender & Hanna, 2025).

Moreover, the limitations identified in relation to sensorimotor experience do not exhaust the potential sources of misalignment between human cognition and model behavior. Other forms of divergence may also be relevant to metaphor processing and warrant further investigation. For instance, abstraction and generalizability processes, which are related to the construction of metaphorical meanings (Bolognesi et al., 2020; Reijnierse et al., 2019), appear to be implemented differently in humans and artificial systems (Collacciani et al., 2024; Rambelli et al., 2024). This suggests again that a similar surface behavior (for instance, human-like accuracy in interpretation)

may be driven by different underlying mechanisms (Bender & Koller, 2020; Lake et al., 2017), thereby supporting descriptive rather than explanatory applications of these tools.

6. Final remarks

In this thesis, I attempted to provide some new insights into the complex interplay of factors underlying metaphor processing. In particular, metaphors emerged as a dynamic object, shaped by the broader cultural context, and consistent with a concurrent process of prediction and inference. Moreover, the results supported the implementation of (at least some) LLMs within metaphor research, not as substitutes for human participants, which remain the reference in the study of processing, but as potential "quantifiers" of human behavior and theoretical assumptions. However, the pitfalls reported in the studies should be taken into consideration when integrating insights from LLMs into psycholinguistics approaches. Interestingly, these divergences from human patterns, for instance in the management of sensorimotor aspects of language and inferential processes revealed maybe even more about human abilities than when they successfully approximate human behaviors.

BIBLIOGRAPHY

Abdurahman, S., Atari, M., Karimi-Malekabadi, F., Xue, M. J., Trager, J., Park, P. S., Golazizian, P., Omrani, A., & Dehghani, M. (2024). Perils and opportunities in using large language models in psychological research. *PNAS Nexus*, *3*(7), pgae245. https://doi.org/10.1093/pnasnexus/pgae245

Accademia della Crusca. (2013). *LIS - Lessico dell'Italiano Scritto*.

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., … Zoph, B. (2024). GPT-4 Technical Report. *ArXiv:2303.08774*.

Al-Azary, H., & Buchanan, L. (2017). Novel metaphor comprehension: Semantic neighbourhood density interacts with concreteness. *Memory & Cognition*, *45*(2), 296–307. https://doi.org/10.3758/s13421-016-0650-7

Al-Azary, H., & Katz, A. N. (2021). Do metaphorical sharks bite? Simulation and abstraction in metaphor processing. *Memory & Cognition*, *49*(3), 557–570. https://doi.org/10.3758/s13421-020-01109-2

Al-Azary, H., & Katz, A. N. (2023). On choosing the vehicles of metaphors 2.0: the interactive effects of semantic neighborhood density and body-object interaction on metaphor production. *Frontiers in Psychology*, *14*, 1216561. https://doi.org/10.3389/fpsyg.2023.1216561

Allott, Nicholas. (2010). *Key terms in pragmatics*. Continuum.

Aprile, M. (2014). Trattatistica. In G. Antonelli, M. Motolese, & L. Tomasin (Eds.), *Storia dell'italiano scritto: II* (pp. 73–118). Carocci.

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, *62*(1), 67–82. https://doi.org/10.1016/j.jml.2009.09.005

Arzouan, Y., Goldstein, A., & Faust, M. (2007). Brainwaves are stethoscopes: ERP correlates of novel metaphor comprehension. *Brain Research*, *1160*, 69–81. https://doi.org/10.1016/j.brainres.2007.05.034

Aurnhammer, C., & Frank, S. L. (2019). Evaluating information-theoretic measures of word prediction in naturalistic sentence reading. *Neuropsychologia*, *134*, 107198. https://doi.org/10.1016/j.neuropsychologia.2019.107198

Azarbonyad, H., Dehghani, M., Beelen, K., Arkut, A., Marx, M., & Kamps, J. (2017). Words are Malleable. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 1509–1518. https://doi.org/10.1145/3132847.3132878

Baiocco, L., Kielar, A., & Lai, V. T. (2026). Metaphor comprehension in neurotypical adults: A scoping review of event-related potential studies. *Cognitive, Affective, & Behavioral Neuroscience*. https://doi.org/10.3758/s13415-025-01376-z

Baiocco, L., Pfeifer, V. A., & Lai, V. T. (2025). Metaphor processing is influenced by stimulus emotionality and task demands: Evidence from ERPs. *Brain and Language*, *261*, 105530. https://doi.org/10.1016/j.bandl.2024.105530

Bambini, V., Bertini, C., Schaeken, W., Stella, A., & Di Russo, F. (2016). Disentangling Metaphor from Context: An ERP Study. *Frontiers in Psychology*, *7*, 559. https://doi.org/10.3389/fpsyg.2016.00559

Bambini, V., Canal, P., Resta, D., & Grimaldi, M. (2019). Time Course and Neurophysiological Underpinnings of Metaphor in Literary Context. *Discourse Processes*, *56*(1), 77–97. https://doi.org/10.1080/0163853X.2017.1401876

Bambini, V., Gentili, C., Ricciardi, E., Bertinetto, P. M., & Pietrini, P. (2011). Decomposing metaphor processing at the cognitive and neural level through functional magnetic resonance imaging. *Brain Research Bulletin*, *86*(3–4), 203–216. https://doi.org/10.1016/j.brainresbull.2011.07.015

Bambini, V., Ghio, M., Moro, A., & Schumacher, P. B. (2013). Differentiating among pragmatic uses of words through timed sensicality judgments. *Frontiers in Psychology*, *4*, 938. https://doi.org/10.3389/fpsyg.2013.00938

Bambini, V., Ranieri, G., Bischetti, L., Scalingi, B., Bertini, C., Ricci, I., Schaeken, W., & Canal, P. (2024). The costs of multimodal metaphors: comparing ERPs to figurative expressions in verbal and verbo-pictorial formats. *Discourse Processes*, *61*(1–2), 44–68. https://doi.org/10.1080/0163853X.2023.2282895

Bambini, V., & Resta, D. (2012). Metaphor and experimental pragmatics: when theory meets empirical investigation. *Humana.Mente Journal of Philosophical Studies*, *23*, 37–60.

Bambini, V., Resta, D., & Grimaldi, M. (2014). A Dataset of Metaphors from the Italian Literature: Exploring Psycholinguistic Variables and the Role of Context. *PLoS ONE*, *9*(9), e105634. https://doi.org/10.1371/journal.pone.0105634

Bambini, V., & Trevisan, M. (2012). Esploracolfis: Un'interfaccia web per le ricerche sul cor- pus e lessico di frequenza dell'italiano scritto (colfis). *Quaderni Del Laboratorio Di Linguistica* , *11*, 1–16.

Barattieri di San Pietro, C., Frau, F., Mangiaterra, V., & Bambini, V. (2023). The pragmatic profile of ChatGPT: Assessing the communicative skills of a conversational agent. *Sistemi Intelligenti*, (2), 379–400. https://doi.org/10.1422/108136

Barber, H. A., Otten, L. J., Kousta, S.-T., & Vigliocco, G. (2013). Concreteness in word processing: ERP and behavioral effects in a lexical decision task. *Brain and Language*, *125*(1), 47–53. https://doi.org/10.1016/j.bandl.2013.01.005

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, *43*(3), 209–226. https://doi.org/10.1007/s10579-009-9081-4

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using **lme4**. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Baumann, A., Hofmann, K., Marakasova, A., Neidhardt, J., & Wissik, T. (2023). Semantic micro-dynamics as a reflex of occurrence frequency: a semantic networks approach. *Cognitive Linguistics*, *34*(3–4), 533–568. https://doi.org/10.1515/cog-2022-0008

Beccaria, G. L. (1993). La letteratura in versi. Dal Settecento al Novecento. In L. Serianni & P. Trifone (Eds.), *Storia della lingua italiana* (p. 682). Einaudi.

Bender, E., & Hanna, A. (2025). *The AI con: How to fight big tech's hype and create the future we want*. Random House.

Bender, E., & Koller, A. (2020). Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185–5198). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.463

Bhatia, S., Richie, R., & Zou, W. (2019). Distributed semantic representations for modeling human judgment. *Current Opinion in Behavioral Sciences*, *29*, 31–36. https://doi.org/10.1016/j.cobeha.2019.01.020

Binz, M., Alaniz, S., Roskies, A., Aczel, B., Bergstrom, C. T., Allen, C., Schad, D., Wulff, D., West, J. D., Zhang, Q., Shiffrin, R. M., Gershman, S. J., Popov, V., Bender, E. M., Marelli, M., Botvinick, M. M., Akata, Z., & Schulz, E. (2025). How should the advancement of large language models affect the practice of science? *Proceedings of the National Academy of Sciences*, *122*(5), e2401227121. https://doi.org/10.1073/pnas.2401227121

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, *120*(6), e2218523120. https://doi.org/10.1073/pnas.2218523120

Birhane, A., Kasirzadeh, A., Leslie, D., & Wachter, S. (2023). Science in the age of large language models. *Nature Reviews Physics*, *5*(5), 277–280. https://doi.org/10.1038/s42254-023-00581-4

Birhane, A., & McGann, M. (2024). Large models of what? Mistaking engineering achievements for human linguistic agency. *Language Sciences*, *106*, 101672. https://doi.org/10.1016/j.langsci.2024.101672

Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B., & Larson, J. M. (2024). Synthetic Replacements for Human Survey Data? The Perils of Large Language Models. *Political Analysis*, *32*(4), 401–416. https://doi.org/10.1017/pan.2024.5

Blasko, D. G., & Briihl, D. S. (1997). Reading and Recall of Metaphorical Sentences: Effects of Familiarity and Context. *Metaphor and Symbol*, *12*(4), 261–285. https://doi.org/10.1207/s15327868ms1204_4

Boden, M. A. (2008). *Mind as machine: A history of cognitive science*. Oxford University Press.

Bolhuis, J. J., Crain, S., Fong, S., & Moro, A. (2024). Three reasons why AI doesn't model human language. *Nature*, *627*(8004), 489–489. https://doi.org/10.1038/d41586-024-00824-z

Bolognesi, M., & Aina, L. (2019). Similarity is closeness: Using distributional semantic spaces to model similarity in visual and linguistic metaphors. *Corpus Linguistics and Linguistic Theory*, *15*(1), 101–137. https://doi.org/10.1515/cllt-2016-0061

Bolognesi, M., Burgers, C., & Caselli, T. (2020). On abstraction: decoupling conceptual concreteness and categorical specificity. *Cognitive Processing*, *21*(3), 365–381. https://doi.org/10.1007/s10339-020-00965-9

Borghi, A., De Livio, C., Mannella, F., Tummolini, L., & Nolfi, S. (2023). Exploring the prospects and challenges of large language models for language learning and production. *Sistemi Intelligenti*, (2), 361–378. https://doi.org/10.1422/108135

Bowdle, B. F., & Gentner, D. (2005). The Career of Metaphor. *Psychological Review*, *112*(1), 193–216. https://doi.org/10.1037/0033-295X.112.1.193

Bozdogan, H. (1987). Model Selection and Akaike's Information Criterion (AIC): The General Theory and its Analytical Extensions. *Psychometrika*, *52*(3), 345–370. https://doi.org/10.1007/BF02294361

Bressler, M., Mangiaterra, V., Canal, P., Frau, F., Luciani, F., Scalingi, B., Barattieri di San Pietro, C., Battaglini, C., Pompei, C., Romeo, F., Bischetti, L., & Bambini, V. (2026). Figurative Archive: an open dataset and web-based application for the study of metaphor. *Scientific Data*. https://doi.org/10.1038/s41597-025-06459-7

Brglez, M., & Vintar, Š. (2025). In search of semantic distance: metaphoric and non-metaphoric constructions in static and contextual representations. *Journal of Language Modelling*, *13*(2). https://doi.org/10.15398/jlm.v13i2.437

Briner, S. W., Schutzenhofer, M. C., & Virtue, S. M. (2018). Hemispheric processing in conventional metaphor comprehension: The role of general knowledge. *Neuropsychologia*, *114*, 101–109. https://doi.org/10.1016/j.neuropsychologia.2018.03.040

Brouwer, H., Crocker, M. W., Venhuizen, N. J., & Hoeks, J. C. J. (2017). A Neurocomputational Model of the N400 and the P600 in Language Processing. *Cognitive Science*, *41*(S6), 1318–1352. https://doi.org/10.1111/cogs.12461

Bruni, E., Boleda, G., Baroni, M., & Tran, N.-K. (2012). Distributional Semantics in Technicolor. In H. Li, C.-Y. Lin, M. Osborne, G. G. Lee, & J. C. Park (Eds.), *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 136–145). Association for Computational Linguistics. http://www.vlfeat.org/

Brysbaert, M., Mandera, P., & Keuleers, E. (2018). The Word Frequency Effect in Word Processing: An Updated Review. *Current Directions in Psychological Science*, *27*(1), 45–50. https://doi.org/10.1177/0963721417727521

Brysbaert, M., Martínez, G., & Reviriego, P. (2024). Moving beyond word frequency based on tally counting: AI-generated familiarity estimates of words and phrases are an interesting additional index of language knowledge. *Behavior Research Methods*, *57*(1), 28. https://doi.org/10.3758/s13428-024-02561-7

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911. https://doi.org/10.3758/s13428-013-0403-5

Buchanan, L., Westbury, C., & Burgess, C. (2001). Characterizing semantic space: Neighborhood effects in word recognition. *Psychonomic Bulletin & Review*, *8*(3), 531–544. https://doi.org/10.3758/BF03196189

Bürkner, P.-C. (2017). **brms** : An *R* Package for Bayesian Multilevel Models Using *Stan*. *Journal of Statistical Software*, *80*(1). https://doi.org/10.18637/jss.v080.i01

Cacciari, C., & Tabossi, P. (1988). The comprehension of idioms. *Journal of Memory and Language*, *27*(6), 668–683. https://doi.org/10.1016/0749-596X(88)90014-9

Campbell, S. J., & Raney, G. E. (2016). A 25-year replication of Katz et al.'s (1988) metaphor norms. *Behavior Research Methods*, *48*(1), 330–340. https://doi.org/10.3758/s13428-015-0575-2

Canal, P., & Bambini, V. (2023). *Pragmatics Electrified* (M. Grimaldi, E. Brattico, & Y. Shtyrov, Eds.; pp. 583–612). https://doi.org/10.1007/978-1-0716-3263-5_18

Canal, P., Bischetti, L., Bertini, C., Ricci, I., Lecce, S., & Bambini, V. (2022). N400 differences between physical and mental metaphors: The role of Theories of Mind. *Brain and Cognition*, *161*, 105879. https://doi.org/10.1016/j.bandc.2022.105879

Canal, P., Bischetti, L., Di Paola, S., Bertini, C., Ricci, I., & Bambini, V. (2019). 'Honey, shall I change the baby? – Well done, choose another one': ERP and time-frequency correlates of humor processing. *Brain and Cognition*, *132*, 41–55. https://doi.org/10.1016/j.bandc.2019.02.001

Cardillo, E. R., Schmidt, G. L., Kranjec, A., & Chatterjee, A. (2010). Stimulus design is an obstacle course: 560 matched literal and metaphorical sentences for testing neural hypotheses about metaphor. *Behavior Research Methods*, *42*(3), 651–664. https://doi.org/10.3758/BRM.42.3.651

Cardillo, E. R., Watson, C., & Chatterjee, A. (2017). Stimulus needs are a moving target: 240 additional matched literal and metaphorical sentences for testing neural hypotheses about metaphor. *Behavior Research Methods*, *49*(2), 471–483. https://doi.org/10.3758/s13428-016-0717-1

Carenini, G., Bischetti, L., Schaeken, W., & Bambini, V. (2023). Towards a Better Rational Speech Act Framework for Context-Aware Modeling of Metaphor Understanding. *Roceedings of the First Workshop on Theory of Mind in Communicating Agents at ICML*.

Carston, R., & Yan, X. (2023). Metaphor processing: Referring and predicating. *Cognition*, *238*, 105534. https://doi.org/10.1016/j.cognition.2023.105534

Casola, S., Frenda, S., Lo, S. M., Sezerer, E., Uva, A., Basile, V., Bosco, C., Pedrani, A., Rubagotti, C., Patti, V., & Bernardi, D. (2024). MultiPICo: Multilingual Perspectivist Irony Corpus. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 16008–16021. https://doi.org/10.18653/v1/2024.acl-long.849

Cassani, G., Bianchi, F., Attanasio, G., Marelli, M., & Guenther, F. (2023). *Meaning Modulations and Stability in Large Language Models: An Analysis of BERT Embeddings for Psycholinguistic Research*. https://doi.org/10.31234/osf.io/b45ys

Cassani, G., Bianchi, F., & Marelli, M. (2021). Words with Consistent Diachronic Usage Patterns are Learned Earlier: A Computational Analysis Using Temporally Aligned Word Embeddings. *Cognitive Science*, *45*(4). https://doi.org/10.1111/cogs.12963

Caucheteux, C., Gramfort, A., & King, J.-R. (2021). Disentangling Syntax and Semantics in the Brain with Deep Networks. In M. Meila & T. Zhang (Eds.), *Proceedings of the 38th International Conference on Machine Learning* (pp. 1336–1348). PMLR.

Cerruti, M., & Onesti, C. (2013). Netspeak: a language variety? Some remarks from an Italian sociolinguistic perspective. In E. Miola (Ed.), *Languages go Web: Standard and non-standard languages on the Internet* (pp. 23–39). Edizioni dell'Orso,.

Chakrabarty, T., Laban, P., Agarwal, D., Muresan, S., & Wu, C.-S. (2024). Art or Artifice? Large Language Models and the False Promise of Creativity. In F. F. Mueller, P. Kyburz, J. R. Williamson, C. Sas, M. L. Wilson, P. Toups Dugas, & I. Shklovski (Eds.), *Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 1–34). ACM. https://doi.org/10.1145/3613904.3642731

Charness, G., Jabarian, B., & List, J. A. (2025). The next generation of experimental research with LLMs. *Nature Human Behaviour*, *9*(5), 833–835. https://doi.org/10.1038/s41562-025-02137-1

Chemero, A. (2023). LLMs differ from human cognition because they are not embodied. *Nature Human Behaviour*, *7*(11), 1828–1829. https://doi.org/10.1038/s41562-023-01723-5

Citron, F. M. M., Lee, M., & Michaelis, N. (2020). Affective and psycholinguistic norms for German conceptual metaphors (COMETA). *Behavior Research Methods*, *52*(3), 1056–1072. https://doi.org/10.3758/s13428-019-01300-7

Coletti, V. (2022). *Storia dell'italiano letterario Dalle origini al XXI secolo*. Einaudi.

Collacciani, C., Rambelli, G., & Bolognesi, M. (2024). Quantifying Generalizations: Exploring the Divide Between Human and LLMs' Sensitivity to Quantification. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 11811–11822. https://doi.org/10.18653/v1/2024.acl-long.636

Columbus, G., Sheikh, N. A., CÃ´tÃ©-Lecaldare, M., HÃ¤user, K., Baum, S. R., & Titone, D. (2015). Individual differences in executive control relate to metaphor processing: an eye movement study of sentence reading. *Frontiers in Human Neuroscience*, *8*, 1057. https://doi.org/10.3389/fnhum.2014.01057

Conde, J., González, M., Grandury, M., Martínez, G., Reviriego, P., & Brysbaert, M. (2025). Psycholinguistic Word Features: a New Approach for the Evaluation of LLMs Alignment with Humans. In O. Arviv, M. Clinciu, K. Dhole, R. Dror, & S. Gehrmann (Eds.), *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)* (pp. 8–17). Association for Computational Linguistics.

Conde, J., Grandury, M., Fu, T., Arriaga, C., Martínez, G., Clark, T., Trott, S., Green, C. G., Reviriego, P., & Brysbaert, M. (2025). Adding LLMs to the psycholinguistic norming toolbox: A practical guide to getting the most out of human ratings. *ArXiv:2509.14405*. 2509.14405

Contreras Kallens, P., Kristensen-McLachlan, R. D., & Christiansen, M. H. (2023). Large Language Models Demonstrate the Potential of Statistical Learning in Language. *Cognitive Science*, *47*(3). https://doi.org/10.1111/cogs.13256

Coulson, S. (2008). Metaphor Comprehension and the Brain. In R. W. Jr. Gibbs (Ed.), *The Cambridge Handbook of Metaphor and Thought* (pp. 177–194). Cambridge University Press.

Coulson, S., King, J. W., & Kutas, M. (1998). Expect the Unexpected: Event-related Brain Response to Morphosyntactic Violations. *Language and Cognitive Processes*, *13*(1), 21–58. https://doi.org/10.1080/016909698386582

Coulson, S., & Van Petten, C. (2002). Conceptual integration and metaphor: An event-related potential study. *Memory & Cognition*, *30*(6), 958–968. https://doi.org/10.3758/BF03195780

Crystal, D. (2006). *Language and the Internet. 2nd edition.* Cambridge University Press.

Cuskley, C., Woods, R., & Flaherty, M. (2024). The Limitations of Large Language Models for Understanding Human Language and Cognition. *Open Mind*, *8*, 1058–1083. https://doi.org/10.1162/opmi_a_00160

Dardano, M. (2014). Romanzo. In G. Antonelli, M. Motolese, & L. Tomasin (Eds.), *Storia dell'italiano scritto* (pp. 359–420).

De Grauwe, S., Swain, A., Holcomb, P. J., Ditman, T., & Kuperberg, G. R. (2010). Electrophysiological insights into the processing of nominal metaphors. *Neuropsychologia*, *48*(7), 1965–1984. https://doi.org/10.1016/j.neuropsychologia.2010.03.017

de Varda, A. G., Marelli, M., & Amenta, S. (2023). Cloze probability, predictability ratings, and computational estimates for 205 English sentences, aligned with existing EEG and reading time data. *Behavior Research Methods*, *56*(5), 5190–5213. https://doi.org/10.3758/s13428-023-02261-8

de Vries, W., & Nissim, M. (2021). As Good as New. How to Successfully Recycle English GPT-2 to Make Models for Other Languages. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 836–846. https://doi.org/10.18653/v1/2021.findings-acl.74

Degen, J. (2023). The Rational Speech Act Framework. *Annual Review of Linguistics*, *9*(1), 519–540. https://doi.org/10.1146/annurev-linguistics-031220-010811

Degen, J., Hawkins, R. D., Graf, C., Kreiss, E., & Goodman, N. D. (2020). When redundancy is useful: A Bayesian approach to "overinformative" referring expressions. *Psychological Review*, *127*(4), 591–621. https://doi.org/10.1037/rev0000186

Di Carlo, V., Bianchi, F., & Palmonari, M. (2019). *Training Temporal Word Embeddings with a Compass*. www.aaai.org

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences*, *27*(7), 597–600. https://doi.org/10.1016/j.tics.2023.04.008

DiStefano, P. V., Patterson, J. D., & Beaty, R. E. (2024). Automatic Scoring of Metaphor Creativity with Large Language Models. *Creativity Research Journal*, 1–15. https://doi.org/10.1080/10400419.2024.2326343

Dudschig, C., Günther, F., & Mackenzie, I. G. (2025). Cognitive plausibility of count-based versus prediction-based word embeddings: A large-scale N400 study. *Biological Psychology*, 109079. https://doi.org/10.1016/j.biopsycho.2025.109079

Englhardt, A., Willkomm, J., Schäler, M., & Böhm, K. (2020). Improving semantic change analysis by combining word embeddings and word frequencies. *International Journal on Digital Libraries*, *21*(3), 247–264. https://doi.org/10.1007/s00799-019-00271-6

Ettinger, A., Feldman, N. H., Resnik, P., & Phillips, C. (2016). Modeling N400 amplitude using vector space models of word representation. In A. Papafragou, D. J. Grodner, D. Mirman, & J. Trueswell (Eds.), *Proceedings of the Annual Meeting of the Cognitive Science Society, 38* (pp. 1445–1450). Cognitive Science Society.

Ettinger, A., & Linzen, T. (2016). Evaluating vector space models using human semantic priming results. *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, 72–77. https://doi.org/10.18653/v1/W16-2513

Fass, D. (1991). met*: A Method for Discriminating Metonymy and Metaphor by Computer. *Computational Linguistics*, *17*(1), 49–90.

Fass, D., & Wilks, Y. (1983). Preference semantics, ill-formedness, and metaphor. *Computational Linguistics*, *9*(3–4), 178–187.

Fauconnier, G., & Turner, M. (1998). Conceptual integration networks. *Cognitive Science*, *22*(2), 133–187. https://doi.org/10.1016/S0364-0213(99)80038-X

Feltgen, Q., Fagard, B., & Nadal, J.-P. (2017). Frequency patterns of semantic change: corpus-based evidence of a near-critical dynamics in language change. *Royal Society Open Science*, *4*(11), 170830. https://doi.org/10.1098/rsos.170830

Fiorentino, G. (2018). Sociolinguistica della scrittura: varietà del web nel repertorio linguistico italiano. *CLUB Working Papers in Linguistics*, *2*, 40–60.

Fludernik, M., Freeman, D. C., & Freeman, M. H. (1999). Metaphor and beyond: An introduction. *Poetics Today*, *20*, 383–396.

Fortson, B. W. (2017). An Approach to Semantic Change. In *The Encyclopedic Dictionary of Applied Linguistics: A Handbook for Language Teaching* (pp. 648–666). Blackwell Publishing Ltd. https://doi.org/10.1002/9781405166201.ch21

Frank, M. C., & Goodman, N. D. (2012). Predicting Pragmatic Reasoning in Language Games. *Science*, *336*(6084), 998–998. https://doi.org/10.1126/science.1218633

Frank, M. C., & Goodman, N. D. (2025). Cognitive Modeling Using Artificial Intelligence. *Annual Review of Psychology*. https://doi.org/10.1146/annurev-psych-030625-040748

Frank, S. L., & Aumeistere, A. (2024). An eye-tracking-with-EEG coregistration corpus of narrative sentences. *Language Resources and Evaluation*, *58*(2), 641–657. https://doi.org/10.1007/s10579-023-09684-x

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, *140*, 1–11. https://doi.org/10.1016/j.bandl.2014.10.006

Fuoli, M., Huang, W., Littlemore, J., Turner, S., & Wilding, E. (2025). *Metaphor identification using large language models: A comparison of RAG, prompt engineering, and fine-tuning*.

Gardner, W. H., & MacKenzie, N. H. (Eds.). (1967). *The Poems of Gerard Manley Hopkins*. Oxford University Press.

Ge, M., Mao, R., & Cambria, E. (2022). Explainable Metaphor Identification Inspired by Conceptual Metaphor Theory. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*(10), 10681–10689. https://doi.org/10.1609/aaai.v36i10.21313

Gernsbacher, M. A., Keysar, B., Robertson, R. R. W., & Werner, N. K. (2001). The Role of Suppression and Enhancement in Understanding Metaphors☆. *Journal of Memory and Language*, *45*(3), 433–450. https://doi.org/10.1006/jmla.2000.2782

Gibbs, R. (1994). *The Poetics of Mind: Figurative Thought, Language, and Understanding*. Cambridge University Press.

Gibbs, R., & Colston, H. (2012). *Interpreting Figurative Meaning*. Cambridge University Press.

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, *120*(30), e2305016120. https://doi.org/10.1073/pnas.2305016120

Giora, R. (2003). *On our mind*. http:/10.1093/acprof:oso/9780195136166.001.0001

Glucksberg, S., & Haught, C. (2006). On the Relation Between Metaphor and Simile: When Comparison Fails. *Mind <html_ent Glyph="@amp;" Ascii="&"/> Language*, *21*(3), 360–378. https://doi.org/10.1111/j.1468-0017.2006.00282.x

Goddard, A. (2015). Creativity and Internet communication . In R. H. Jones (Ed.), *The Routledge Handbook of Language and Creativity* (pp. 367–382). Routledge.

Goldstein, A., Arzouan, Y., & Faust, M. (2012). Killing a novel metaphor and reviving a dead one: ERP correlates of metaphor conventionalization. *Brain and Language*, *123*(2), 137–142. https://doi.org/10.1016/j.bandl.2012.09.008

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). *Learning Word Vectors for 157 Languages*.

Grice, P. (1975). Logic and Conversation. In *Speech Acts* (pp. 41–58). BRILL. https://doi.org/10.1163/9789004368811_003

Groppe, D. M., Makeig, S., & Kutas, M. (2009). Identifying reliable independent components via split-half comparisons. *NeuroImage*, *45*(4), 1199–1211. https://doi.org/10.1016/j.neuroimage.2008.12.038

Guenther, F., & Cassani, G. (2025). *Large Language Models in psycholinguistic studies*. https://doi.org/https://doi.org/10.31234/osf.io/cvnam_v1

Gulordava, K., & Baroni, M. (2011). A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. In S. Pado & Y. Peirsman (Eds.), *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics* (pp. 67–71). Association for Computational Linguistics.

Günther, F., Dudschig, C., & Kaup, B. (2016). Latent semantic analysis cosines as a cognitive similarity measure: Evidence from priming studies. *Quarterly Journal of Experimental Psychology*, *69*(4), 626–653. https://doi.org/10.1080/17470218.2015.1038280

Hackl, V., Müller, A. E., Granitzer, M., & Sailer, M. (2023). Is GPT-4 a reliable rater? Evaluating consistency in GPT-4's text ratings. *Frontiers in Education*, *8*, 1272229. https://doi.org/10.3389/feduc.2023.1272229

Hagoort, P., Brown, C., & Groothusen, J. (1993). The syntactic positive shift (sps) as an erp measure of syntactic processing. *Language and Cognitive Processes*, *8*(4), 439–483. https://doi.org/10.1080/01690969308407585

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. *Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies 2001 - NAACL '01*, 1–8. https://doi.org/10.3115/1073336.1073357

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). *Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change*.

Han, L., Kashyap, A., Finin, T., Mayfield, J., & Weese, J. (2013). UMBC EBIQUITY-CORE: Semantic Textual Similarity Systems. In M. Diab, T. Baldwin, & M. Baroni (Eds.), *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity* (Vol. 1, pp. 44–52). Association for Computational Linguistics.

Hanks, P. (2006). Metaphoricity is gradable. In Stefanowitsch A & Gries S (Eds.), *Corpora in Cognitive Linguistics - Vol. 1: Metaphor and Metonymy*. Mouton de Gruyter. .

Harding, J., D'Alessandro, W., Laskowski, N. G., & Long, R. (2024). AI language models cannot replace human research participants. *AI & SOCIETY*, *39*(5), 2603–2605. https://doi.org/10.1007/s00146-023-01725-x

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

Hazing, A. W. (2007). *Publish or perish*. Https://Harzing.Com/Resources/Publish-or-Perish.

Hill, J., & Abadkat, S. (2023). *Using logprobs*. OpenAI. https://cookbook.openai.com/examples/using_logprobs

Hilpert, M. (2008a). *Germanic Future Constructions: A usage-based approach to language change*. Benjamin.

Hilpert, Martin. (2008b). *Germanic future constructions : a usage-based approach to language change*. John Benjamins Pub. Co.

Honnibal, M., & Montani, I. (2017). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.*

Hu, J. (2023). *Neural language models and human linguistic knowledge*. Massachusetts Institute of Technology.

Hu, J., Floyd, S., Jouravlev, O., Fedorenko, E., & Gibson, E. (2022). A fine-grained comparison of pragmatic language understanding in humans and language models. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 4194–4213). Association for Computational Linguistics.

Hu, J., & Levy, R. (2023). Prompting is not a substitute for probability measurements in large language models. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 5040–5060). Association for Computational Linguistics.

Huang, J., Chen, L., Huang, Y., Chen, Y., & Zou, L. (2024). COGMED: a database for Chinese olfactory and gustatory metaphor. *Humanities and Social Sciences Communications*, *11*(1), 1080. https://doi.org/10.1057/s41599-024-03593-2

Ichien, N., Stamenković, D., & Holyoak, K. J. (2024). Large Language Model Displays Emergent Ability to Interpret Novel Literary Metaphors. *Metaphor and Symbol*, *39*(4), 296–309. https://doi.org/10.1080/10926488.2024.2380348

Ilievski, F., Hammer, B., van Harmelen, F., Paassen, B., Saralajew, S., Schmid, U., Biehl, M., Bolognesi, M., Dong, X. L., Gashteovski, K., Hitzler, P., Marra, G., Minervini, P., Mundt, M., Ngomo, A.-C. N., Oltramari, A., Pasi, G., Saribatur, Z. G., Serafini, L., … Villmann, T. (2025). Aligning generalization between humans and machines. *Nature Machine Intelligence*, *7*(9), 1378–1389. https://doi.org/10.1038/s42256-025-01109-4

Jenkins, C., Miletić, F., & Schulte Im Walde, S. (2025). Multi-word Measures: Modeling Semantic Change in Compound Nouns. In W. Che, J. Nabende, E. Shutova, & M. Taher Pilehvar (Eds.), *Findings of the Association for Computational Linguistics: ACL 2025* (pp. 10850–10864). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.findings-acl.566

Jończyk, R., Kremer, G. E., Siddique, Z., & van Hell, J. G. (2020). *Engineering* creativity: Prior experience modulates electrophysiological responses to novel metaphors. *Psychophysiology*, *57*(10). https://doi.org/10.1111/psyp.13630

Jones, M. N., Willits, J., & Dennis, S. (2015). Models of semantic memory. In J. Busemeyer, Z. Wang, J. Townsend, & A. Eidels (Eds.), *Oxford handbook of mathematical and computational psychology* (pp. 232–254). Oxford University Press.

Kao, J. T., Bergen, L., & Goodman, N. (2014). Formalizing the Pragmatics of Metaphor Understanding. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the Annual Meeting of the Cognitive Science Society 36* (Number 36).

Kao, J. T., Wu, J. Y., Bergen, L., & Goodman, N. D. (2014). Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*, *111*(33), 12002–12007. https://doi.org/10.1073/pnas.1407479111

Katz, A. N. (1989). On choosing the vehicles of metaphors: Referential concreteness, semantic distances, and individual differences. *Journal of Memory and Language*, *28*(4), 486–499. https://doi.org/10.1016/0749-596X(89)90023-5

Katz, A. N., Paivio, A., & Marschark, M. (1985). Poetic comparisons: Psychological dimensions of metaphoric processing. *Journal of Psycholinguistic Research*, *14*(4), 365–383. https://doi.org/10.1007/BF01067881

Katz, A. N., Paivio, A., Marschark, M., & Clark, J. M. (1988). Norms for 204 Literary and 260 Nonliterary Metaphors on 10 Psychological Dimensions. *Metaphor and Symbolic Activity*, *3*(4), 191–214. https://doi.org/10.1207/s15327868ms0304_1

Kewenig, V., Skipper, J. I., & Vigliocco, G. (2025). A multimodal transformer-based tool for automatic generation of concreteness ratings across languages. *Communications Psychology*, *3*(1), 100. https://doi.org/10.1038/s44271-025-00280-z

Khademi, A. (2023). Can ChatGPT and Bard generate aligned assessment items? A reliability analysis against human performance. *Journal of Applied Learning & Teaching*, *6*(1). https://doi.org/10.37074/jalt.2023.6.1.28

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, *95*(2), 163–182. https://doi.org/10.1037/0033-295X.95.2.163

Kintsch, W. (2000). Metaphor comprehension: A computational theory. *Psychonomic Bulletin & Review*, *7*(2), 257–266. https://doi.org/10.3758/BF03212981

Kosinski, M. (2024). Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, *121*(45), e2405460121. https://doi.org/10.1073/pnas.2405460121

Krieger, B., Brouwer, H., Aurnhammer, C., & Crocker, M. W. (2025). On the limits of LLM surprisal as a functional explanation of the N400 and P600. *Brain Research*, *1865*, 149841. https://doi.org/10.1016/j.brainres.2025.149841

Kulkarni, V., Al-Rfou, R., Perozzi, B., & Skiena, S. (2015). Statistically Significant Detection of Linguistic Change. *Proceedings of the 24th International Conference on World Wide Web*, 625–635. https://doi.org/10.1145/2736277.2741627

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, *31*(1), 32–59. https://doi.org/10.1080/23273798.2015.1102299

Kuperberg, G. R., Sitnikova, T., Caplan, D., & Holcomb, P. J. (2003). Electrophysiological distinctions in processing conceptual relationships within simple sentences. *Cognitive Brain Research*, *17*(1), 117–129. https://doi.org/10.1016/S0926-6410(03)00086-7

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978–990. https://doi.org/10.3758/s13428-012-0210-4

Kutas, M., & Federmeier, K. D. (2011). Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP). *Annual Review of Psychology*, *62*(1), 621–647. https://doi.org/10.1146/annurev.psych.093008.131123

Kutuzov, A., Øvrelid, L., Szymanski, T., & Velldal, E. (2018). *Diachronic word embeddings and semantic shifts: a survey*.

Lago, S., Zago, S., Bambini, V., & Arcara, G. (2024). Pre-Stimulus Activity of Left and Right TPJ in Linguistic Predictive Processing: A MEG Study. *Brain Sciences*, *14*(10), 1014. https://doi.org/10.3390/brainsci14101014

Lai, V. T., & Curran, T. (2013). ERP evidence for conceptual mappings and comparison processes during the comprehension of conventional and novel metaphors. *Brain and Language*, *127*(3), 484–496. https://doi.org/10.1016/j.bandl.2013.09.010

Lai, V. T., Curran, T., & Menn, L. (2009). Comprehending conventional and novel metaphors: An ERP study. *Brain Research*, *1284*, 145–155. https://doi.org/10.1016/j.brainres.2009.05.088

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, e253. https://doi.org/10.1017/S0140525X16001837

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. University of Chicago Press.

Landauer, T. K., & Dumais, S. T. (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. https://doi.org/10.1037/0033-295X.104.2.211

Lau, E. F., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics: (de)constructing the N400. *Nature Reviews Neuroscience*, *9*(12), 920–933. https://doi.org/10.1038/nrn2532

Lee, J., Park, D., Lee, J., Choi, H., & Lee, S.-E. (2025). *Exploring Multimodal Perception in Large Language Models Through Perceptual Strength Ratings*. https://doi.org/10.1109/ACCESS.2025.3618700

Lenci, A. (2018). Distributional Models of Word Meaning. *Annual Review of Linguistics*, *4*(1), 151–171. https://doi.org/10.1146/annurev-linguistics-030514-125254

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006

Levy, R., Kim, Y., & Fox, D. (2025). The Science of Language in the Era of Generative AI. *An MIT Exploration of Generative AI*. https://doi.org/10.21428/e4baedd9.f6a0052d

Liao, Z., Antoniak, M., Cheong, I., Cheng, E. Y.-Y., Lee, A.-H., Lo, K., Chang, J. C., & Zhang, A. X. (2024). LLMs as Research Tools: A Large Scale Survey of Researchers' Usage and Perceptions. *ArXiv:2411.05025*.

Littlemore, J., Sobrino, P. P., Houghton, D., Shi, J., & Winter, B. (2018). What makes a good metaphor? A cross-cultural study of computer-generated metaphor appreciation. *Metaphor and Symbol*, *33*(2), 101–122. https://doi.org/10.1080/10926488.2018.1434944

Lopopolo, A., & Rabovsky, M. (2024). Tracking Lexical and Semantic Prediction Error Underlying the N400 Using Artificial Neural Network Models of Sentence Processing. *Neurobiology of Language*, *5*(1), 136–166. https://doi.org/10.1162/nol_a_00134

Lynott, D., Connell, L., Brysbaert, M., Brand, J., & Carney, J. (2020). The Lancaster Sensorimotor Norms: multidimensional measures of perceptual and action strength for 40,000 English words. *Behavior Research Methods*, *52*(3), 1271–1291. https://doi.org/10.3758/s13428-019-01316-z

Manchanda, J., Boettcher, L., Westphalen, M., & Jasser, J. (2025). *The Open Source Advantage in Large Language Models (LLMs)*.

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, *92*, 57–78. https://doi.org/10.1016/j.jml.2016.04.001

Mangiaterra, V., Barattieri di San Pietro, C., & Bambini, V. (2024). Temporal word embeddings in the study of metaphor change over time and across genres: a proof-of-concept study on English. In F. Dell'Orletta, A. Lenci, S. Montemagni, & R. Sprugnoli (Eds.), *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)* (pp. 548–555).

Mangiaterra, V., Barattieri Di San Pietro, C., Frau, F., Bambini, V., & Al-Azary, H. (2025). On choosing the vehicles of metaphors without a body: evidence from Large Language Models. In G. Rambelli, F. Ilievski, M. Bolognesi, & P. Sommerauer (Eds.), *Proceedings of the 2nd Workshop on Analogical Abstraction in Cognition, Perception, and Language (Analogy-Angle II)* (pp. 37–44). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.analogyangle-1.4

Mao, R., Lin, C., & Guerin, F. (2018). Word embedding and wordnet based metaphor identification and interpretation. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1222–1231.

Mao, R., Lin, C., & Guerin, F. (2019). End-to-End Sequential Metaphor Identification Inspired by Linguistic Theories. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3888–3898. https://doi.org/10.18653/v1/P19-1378

Marazzini, C. (2002). *La lingua italiana: profilo storico*. Il Mulino.

Marschark, M., Katz, A. N., & Paivio, A. (1983). Dimensions of metaphor. *Journal of Psycholinguistic Research*, *12*(1), 17–40. https://doi.org/10.1007/BF01072712

Martin, J. H. (1992). Computer Understanding of Conventional Metaphoric Language. *Cognitive Science*, *16*(2), 233–270. https://doi.org/10.1207/s15516709cog1602_4

Martínez, G., Conde, J., Reviriego, P., & Brysbaert, M. (2024). AI-generated estimates of familiarity, concreteness, valence, and arousal for over 100,000 Spanish words. *Quarterly Journal of Experimental Psychology*. https://doi.org/10.1177/17470218241306694

Martínez, G., Molero, J. D., González, S., Conde, J., Brysbaert, M., & Reviriego, P. (2024). Using large language models to estimate features of multi-word expressions: Concreteness, valence, arousal. *Behavior Research Methods*, *57*(1), 5. https://doi.org/10.3758/s13428-024-02515-z

Mashal, N., & Faust, M. (2009). Conventionalisation of novel metaphors: A shift in hemispheric asymmetry. *Laterality: Asymmetries of Body, Brain and Cognition*, *14*(6), 573–589. https://doi.org/10.1080/13576500902734645

Masini, A. (1977). *La lingua di alcuni giornali milanesi dal 1859 al 1865*. La Nuova Italia.

Masini, A. (1994). La lingua dei giornali dell'Ottocento. In L. Serianni & P. Trifone (Eds.), *Storia della lingua italiana: 2: Scritto e Parlato* (pp. 635–665). Einaudi.

Mayn, A., & Demberg, V. (2022). Pragmatics of Metaphor Revisited: Modeling the Role of Degree and Salience in Metaphor Understanding. In J. , Culbertson, H. Rabagliati, V. Ramenzoni, & A. Perfors (Eds.), *Proceedings of the Annual Meeting of the Cognitive Science Society 44* (p. 3154). Cognitive Science Society.

McGregor, S., Agres, K., Rataj, K., Purver, M., & Wiggins, G. (2019). Re-Representing Metaphor: Modeling Metaphor Perception Using Dynamically Contextual Distributional Semantics. *Frontiers in Psychology*, *10*, 765. https://doi.org/10.3389/fpsyg.2019.00765

Mennes, M., Wouters, H., Vanrumste, B., Lagae, L., & Stiers, P. (2010). Validation of ICA as a tool to remove eye movement artifacts from EEG/ERP. *Psychophysiology*. https://doi.org/10.1111/j.1469-8986.2010.01015.x

Messeri, L., & Crockett, M. J. (2024). Artificial intelligence and illusions of understanding in scientific research. *Nature*, *627*(8002), 49–58. https://doi.org/10.1038/s41586-024-07146-0

Meta AI. (2024). *Llama 3.2: Revolutionizing edge AI and vision with open, customizable models.* . Meta AI Blog.

Michaelov, J. A., Bardolph, M. D., Van Petten, C. K., Bergen, B. K., & Coulson, S. (2024). Strong Prediction: Language Model Surprisal Explains Multiple N400 Effects. *Neurobiology of Language*, *5*(1), 107–135. https://doi.org/10.1162/nol_a_00105

Michaelov, J., & Bergen, B. (2020). How well does surprisal explain N400 amplitude under different experimental conditions? In R. Fernández & T. Linzen (Eds.), *Proceedings of the 24th*

Conference on Computational Natural Language Learning (pp. 652–663). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.conll-1.53

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*.

Milenković, K., Tasić, M., & Stamenković, D. (2024). Influence of translation on perceived metaphor features: quality, aptness, metaphoricity, and familiarity. *Linguistics Vanguard*, *10*(1), 285–296. https://doi.org/10.1515/lingvan-2023-0086

Momen, O., Sitter, E., Herrmann, B., & Zarrieß, S. (2026). *Surprisal and Metaphor Novelty: Moderate Correlations and Divergent Scaling Effects*.

Moran, T. P., Schroder, H. S., Kneip, C., & Moser, J. S. (2017). Meta-analysis and psychophysiology: A tutorial using depression and action-monitoring event-related potentials. *International Journal of Psychophysiology*, *111*, 17–32. https://doi.org/10.1016/j.ijpsycho.2016.07.001

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2025). A Comprehensive Overview of Large Language Models. *ACM Transactions on Intelligent Systems and Technology*, *16*(5), 1–72. https://doi.org/10.1145/3744746

Neath, A. A., & Cavanaugh, J. E. (2012). The Bayesian information criterion: background, derivation, and applications. *WIREs Computational Statistics*, *4*(2), 199–203. https://doi.org/10.1002/wics.199

Newman, N., Ross Arguedas, A., Robertson, C. T., Nielsen, R. K., & Fletcher, R. (2025). *Digital news report 2025*. Reuters Institute for the Study of Journalism .

Nour Eddine, S., Brothers, T., Wang, L., Spratling, M., & Kuperberg, G. R. (2024). A predictive coding model of the N400. *Cognition*, *246*, 105755. https://doi.org/10.1016/j.cognition.2024.105755

Oh, B.-D., & Linzen, T. (2025). *To model human linguistic prediction, make LLMs less superhuman*.

Oostenveld, R., & Praamstra, P. (2001). The five percent electrode system for high-resolution EEG and ERP measurements. *Clinical Neurophysiology*, *112*(4), 713–719. https://doi.org/10.1016/S1388-2457(00)00527-7

O'Reilly, D., Onysko, A., Rasse, C., Papitsch, L., Colston, H., & van der Horst, I. (2025). High ceilings and ingenuine allies: tapping into the idiom meaning knowledge of first and second language speakers of English. *Language and Cognition*, *17*, e72. https://doi.org/10.1017/langcog.2025.10029

Orlando, R., Moroni, L., Cabot, P.-L. H., Barba, E., Conia, S., Orlandini, S., Fiameni, G., & Navigli, R. (2024). Minerva LLMs: The First Family of Large Language Models Trained from Scratch on Italian Data. In F. Dell'Orletta, A. Lenci, S. Montemagni, & R. Sprugnoli (Eds.), *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)* (pp. 707–719). CEUR Workshop Proceedings. https://nlp.uniroma1.it/minerva

Osterhout, L., & Nicol, J. (1999). On the Distinctiveness, Independence, and Time Course of the Brain Responses to Syntactic and Semantic Anomalies. *Language and Cognitive Processes*, *14*(3), 283–317. https://doi.org/10.1080/016909699386310

Pavlick, E. (2023). Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *381*(2251), 20220041. https://doi.org/10.1098/rsta.2022.0041

Piantadosi, ST. (2024). Modern language models refute Chomsky's approach to language. In E. Gibson & M. Poliak (Eds.), *From Fieldwork to Linguistic Theory: A Tribute to Dan Everett* (pp. 353–414). Language Science Press.

Pilkington, A. (2000). *Poetic Effects: a Relevance Theory Perspective*. John Benjamins.

Pinheiro, J. C., & Bates, D. M. (2000). Linear Mixed-Effects Models: Basic Concepts and Examples. In *Mixed-Effects Models in S and S-PLUS* (pp. 3–56). Springer-Verlag. https://doi.org/10.1007/0-387-22747-4_1

Pistolesi, E. (2014). Scritture digitali. In G. Antonelli, M. Motolese, & L. Tomasin (Eds.), *Storia dell'italiano scritto* (pp. 349–375). Carocci.

Pragglejaz Group. (2007). MIP: A Method for Identifying Metaphorically Used Words in Discourse. *Metaphor and Symbol*, *22*(1), 1–39. https://doi.org/10.1080/10926480709336752

Puccetti, G., Esuli, A., & Bolognesi, M. (2025). Wordnet and Word Ladders: Climbing the abstraction taxonomy with LLMs. *Proceedings of the 13th Global Wordnet Conference*, 51–65. https://doi.org/10.18653/v1/2025.gwc-1.7

Pynte, J., Besson, M., Robichon, F.-H., & Poli, J. (1996). The Time-Course of Metaphor Comprehension: An Event-Related Potential Study. *Brain and Language*, *55*(3), 293–316. https://doi.org/10.1006/brln.1996.0107

Qiu, M., Brisebois, Z., & Sun, S. (2025). *Can LLMs Simulate Human Behavioral Variability? A Case Study in the Phonemic Fluency Task*. ArXiv:2505.16164

R Core Team. (2025). *R: A language and environment for statistical computing* (3.5.2). R Foundation for Statistical Computing.

Rabovsky, M., Hansen, S. S., & McClelland, J. L. (2018). Modelling the N400 brain potential as change in a probabilistic representation of meaning. *Nature Human Behaviour*, *2*(9), 693–705. https://doi.org/10.1038/s41562-018-0406-4

Rabovsky, M., & McRae, K. (2014). Simulating the N400 ERP component as semantic network error: Insights from a feature-based connectionist attractor model of word meaning. *Cognition*, *132*(1), 68–89. https://doi.org/10.1016/j.cognition.2014.03.010

Rambelli, G., & Bolognesi, M. (2024). The Contextual Variability of English Nouns: The Impact of Categorical Specificity beyond Conceptual Concreteness. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 15854–15860). ELRA and ICCL.

Rambelli, G., Chersoni, E., Collacciani, C., & Bolognesi, M. (2024). Can Large Language Models Interpret Noun-Noun Compounds? A Linguistically-Motivated Study on Lexicalized and Novel Compounds. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 11823–11835. https://doi.org/10.18653/v1/2024.acl-long.637

Rapp, A. M., Leube, D. T., Erb, M., Grodd, W., & Kircher, T. T. J. (2004). Neural correlates of metaphor processing. *Cognitive Brain Research*, *20*(3), 395–402. https://doi.org/10.1016/j.cogbrainres.2004.03.017

Rathje, S., Mirea, D.-M., Sucholutsky, I., Marjieh, R., Robertson, C. E., & Van Bavel, J. J. (2024). GPT is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences*, *121*(34), e2308950121. https://doi.org/10.1073/pnas.2308950121

Ravelli, A. A., & Bolognesi, M. M. (2024). Yet another approximation of human semantic judgments using LLMs... but with quantized local models on novel data. *Italian Journal of Computational Linguistics*, *10*(2), 57. https://doi.org/10.17454/IJCOL102.04

Regel, S., Meyer, L., & Gunter, T. C. (2014). Distinguishing Neurocognitive Processes Reflected by P600 Effects: Evidence from ERPs and Neural Oscillations. *PLoS ONE*, *9*(5), e96840. https://doi.org/10.1371/journal.pone.0096840

Reid, N. J., Al-Azary, H., & Katz, A. N. (2023). Cognitive Factors Related to Metaphor Goodness in Poetic and Non-literary Metaphor. *Metaphor and Symbol*, *38*(2), 130–148. https://doi.org/10.1080/10926488.2021.2011285

Reid, N. J., & Katz, A. N. (2018). Vector Space Applications in Metaphor Comprehension. *Metaphor and Symbol*, *33*(4), 280–294. https://doi.org/10.1080/10926488.2018.1549840

Reijnierse, W. G., Burgers, C., Bolognesi, M., & Krennmayr, T. (2019). How Polysemy Affects Concreteness Ratings: The Case of Metaphor. *Cognitive Science*, *43*(8). https://doi.org/10.1111/cogs.12779

Rodda, M. A., Senaldi, M. S. G., & Lenci, A. (2017). Panta rei: Tracking Semantic Change with Distributional Semantics in Ancient Greek. *Italian Journal of Computational Linguistics*, *3*(1), 11–24. https://doi.org/10.4000/ijcol.421

Rohatgi, A. (2025). *WebPlotDigitizer*. Https://Automeris.Io.

Rutter, B., Kröger, S., Hill, H., Windmann, S., Hermann, C., & Abraham, A. (2012). Can clouds dance? Part 2: An ERP investigation of passive conceptual expansion. *Brain and Cognition*, *80*(3), 301–310. https://doi.org/10.1016/j.bandc.2012.08.003

Ryskina, M., Rabinovich, E., Berg-Kirkpatrick, T., Mortensen, D. R., & Tsvetkov, Y. (2020). *Where New Words Are Born: Distributional Semantic Analysis of Neologisms and Their Semantic Neighborhoods*. https://doi.org/10.7275/1jra-8m83

Sagi, E., Kaufmann, S., & Clark, B. (2009). Semantic Density Analysis: Comparing word meaning across time and phonetic space. In R. Basili & M. Pennacchiotti (Eds.), *Proceedings of the*

Workshop on Geometrical Models of Natural Language Semantics (pp. 104–111). Association for Computational Linguistics.

Sambrook, T. D., & Goslin, J. (2015). A neural reward prediction error revealed by a meta-analysis of ERPs using great grand averages. *Psychological Bulletin*, *141*(1), 213–235. https://doi.org/10.1037/bul0000006

Santini, M. (2007). Automatic genre identification: towards a flexible classification scheme. *BCS IRSG Symposium: Future Directions in Information Access 2007*.

Schmidt, G. L., & Seger, C. A. (2009). Neural correlates of metaphor processing: The roles of figurativeness, familiarity and difficulty. *Brain and Cognition*, *71*(3), 375–386. https://doi.org/10.1016/j.bandc.2009.06.001

Schumacher, P. B. (2011). *The hepatitis called …* (pp. 199–219). https://doi.org/10.1075/la.175.10sch

Scontras, G., Tessler, M. H., & Franke, M. (2021). *A practical introduction to the Rational Speech Act modeling framework*.

Scott, G. G., Keitel, A., Becirspahic, M., Yao, B., & Sereno, S. C. (2019). The Glasgow Norms: Ratings of 5,500 words on nine scales. *Behavior Research Methods*, *51*(3), 1258–1270. https://doi.org/10.3758/s13428-018-1099-3

Semino, E., & Steen, G. (2008). Metaphor in literature. In R. W. Gibbs (Ed.), *The Cambridge Handbook of Metaphor and Thought* (Vol. 6, pp. 57–70). Cambridge University Press.

Serianni, L. (1993). La prosa. In L. Serianni & P. Trifone (Eds.), *Storia della lingua italiana* (pp. 451–577). Einaudi.

Shah, R. S., & Varma, S. (2025). *The potential -- and the pitfalls -- of using pre-trained language models as cognitive science theories*.

Shinjo, M., & Myers, J. L. (1987). The role of context in metaphor comprehension. *Journal of Memory and Language*, *26*(2), 226–241. https://doi.org/10.1016/0749-596X(87)90125-2

Shliazhko, O., Fenogenova, A., Tikhonova, M., Kozlova, A., Mikhailov, V., & Shavrina, T. (2024). mGPT: Few-Shot Learners Go Multilingual. *Transactions of the Association for Computational Linguistics*, *12*, 58–79. https://doi.org/10.1162/tacl_a_00633

Shutova, E. (2015). Design and Evaluation of Metaphor Processing Systems. *Computational Linguistics*, *41*(4), 579–623. https://doi.org/10.1162/COLI_a_00233

Slaats, S., & Martin, A. E. (2025). What's Surprising About Surprisal. *Computational Brain & Behavior*, *8*(2), 233–248. https://doi.org/10.1007/s42113-025-00237-9

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319. https://doi.org/10.1016/j.cognition.2013.02.013

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition*. Harvard University Press.

Sperber, D., & Wilson, D. (2012). A deflationary account of metaphors. In D. Wilson & D. Sperber (Eds.), *Meaning and relevance* (pp. 97–122). Cambridge University Press.

Spotorno, N., Cheylus, A., Van Der Henst, J.-B., & Noveck, I. A. (2013). What's behind a P600? Integration Operations during Irony Processing. *PLoS ONE*, *8*(6), e66839. https://doi.org/10.1371/journal.pone.0066839

Stadthagen-Gonzalez, H., & Davis, C. J. (2006). The Bristol norms for age of acquisition, imageability, and familiarity. *Behavior Research Methods*, *38*(4), 598–605. https://doi.org/10.3758/BF03193891

Steen, G. (1989). Metaphor and literary comprehension: Towards a discourse theory of metaphor in literature. *Poetics*, *18*(1–2), 113–141. https://doi.org/10.1016/0304-422X(89)90025-9

Steen, G. (1994). *Understanding Metaphor in Literature: An Empirical Approach*. Longman Publishing.

Steen, G., Dorst, A. G., Herrmann, J. B., Kaal, A. A., & Krennmayr, T. (2010). Metaphor in usage. *Cogl*, *21*(4), 765–796. https://doi.org/10.1515/cogl.2010.024

Su, C., Huang, S., & Chen, Y. (2017). Automatic detection and interpretation of nominal metaphor based on the theory of meaning. *Neurocomputing*, *219*, 300–311. https://doi.org/10.1016/j.neucom.2016.09.030

Tahmasebi, N., Borin, L., & Jatowt, A. (2019). *Survey of Computational Approaches to Lexical Semantic Change*.

Tang, X. (2018). A state-of-the-art of semantic change computation. *Natural Language Engineering*, *24*(5), 649–676. https://doi.org/10.1017/S1351324918000220

Tang, X., Qi, S., Wang, B., Jia, X., & Ren, W. (2017). The temporal dynamics underlying the comprehension of scientific metaphors and poetic metaphors. *Brain Research*, *1655*, 33–40. https://doi.org/10.1016/j.brainres.2016.11.005

Tong, X., Shutova, E., & Lewis, M. (2021). Recent advances in neural metaphor processing: A linguistic, cognitive and social perspective. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4673–4686. https://doi.org/10.18653/v1/2021.naacl-main.372

Tonini, E., Bischetti, L., Del Sette, P., Tosi, E., Lecce, S., & Bambini, V. (2023). The relationship between metaphor skills and Theory of Mind in middle childhood: Task and developmental effects. *Cognition*, *238*, 105504. https://doi.org/10.1016/j.cognition.2023.105504

Traugott, E. C. (2017). Semantic Change. In *Oxford Research Encyclopedia of Linguistics*. Oxford University Press. https://doi.org/10.1093/acrefore/9780199384655.013.323

Traugott, E. C., & Dasher, R. B. (2001). *Regularity in Semantic Change*. Cambridge University Press. https://doi.org/10.1017/CBO9780511486500

Trott, S. (2024a). Can large language models help augment English psycholinguistic datasets? *Behavior Research Methods*, *56*(6), 6082–6100. https://doi.org/10.3758/s13428-024-02337-z

Trott, S. (2024b). Large Language Models and the Wisdom of Small Crowds. *Open Mind*, *8*, 723–738. https://doi.org/10.1162/opmi_a_00144

Utsumi, A. (2007). Interpretive Diversity Explains Metaphor–Simile Distinction. *Metaphor and Symbol*, *22*(4), 291–312. https://doi.org/10.1080/10926480701528071

Utsumi, A. (2011). Computational Exploration of Metaphor Comprehension Processes Using a Semantic Space Model. *Cognitive Science*, *35*(2), 251–296. https://doi.org/10.1111/j.1551-6709.2010.01144.x

Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology*, *83*(2), 176–190. https://doi.org/10.1016/j.ijpsycho.2011.09.015

van Rooij, I., Guest, O., Adolfi, F., de Haan, R., Kolokolova, A., & Rich, P. (2024). Reclaiming AI as a Theoretical Tool for Cognitive Science. *Computational Brain & Behavior*, *7*(4), 616–636. https://doi.org/10.1007/s42113-024-00217-5

van Tiel, B., Franke, M., & Sauerland, U. (2021). Probabilistic pragmatics explains gradience and focality in natural language quantification. *Proceedings of the National Academy of Sciences*, *118*(9). https://doi.org/10.1073/pnas.2005453118

Venhuizen, N. J., Crocker, M. W., & Brouwer, H. (2019). Expectation-based Comprehension: Modeling the Interaction of World Knowledge and Linguistic Experience. *Discourse Processes*, *56*(3), 229–255. https://doi.org/10.1080/0163853X.2018.1448677

Vespignani, F. (2020). *Digerps*. Ttps://Github.Com/Francesco-Vespignani/DigERPs.Git.

Vespignani, F., Canal, P., Molinaro, N., Fonda, S., & Cacciari, C. (2010). Predictive Mechanisms in Idiom Comprehension. *Journal of Cognitive Neuroscience*, *22*(8), 1682–1700. https://doi.org/10.1162/jocn.2009.21293

Viola, L. (2021). ChroniclItaly and ChroniclItaly 2.0: Digital Heritage to Access Narratives of Migration. *International Journal of Humanities and Arts Computing*, *15*(1–2), 170–185. https://doi.org/10.3366/ijhac.2021.0268

Wang, A., Morgenstern, J., & Dickerson, J. P. (2025). Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, *7*(3), 400–411. https://doi.org/10.1038/s42256-025-00986-z

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191–1207. https://doi.org/10.3758/s13428-012-0314-x

Weiland, H., Bambini, V., & Schumacher, P. B. (2014). The role of literal meaning in figurative language comprehension: evidence from masked priming ERP. *Frontiers in Human Neuroscience*, *8*, 3389. https://doi.org/10.3389/fnhum.2014.00583

Werkmann Horvat, A., Bolognesi, M., & Althaus, N. (2023). Attention to the source domain of conventional metaphorical expressions: Evidence from an eye tracking study. *Journal of Pragmatics*, *215*, 131–144. https://doi.org/10.1016/j.pragma.2023.07.011

Werning, M., & Cosentino, E. (2017). The Interaction of Bayesian Pragmatics and Lexical Semantics in Linguistic Interpretation: Using Event-related Potentials to Investigate Hearers' Probabilistic Predictions. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the Annual Meeting of the Cognitive Science Society 39* (pp. 3504–3509). Cognitive Science Society.

Werning, M., Unterhuber, M., & Wiedemann, G. (2019). Bayesian Pragmatics Provides the Best Quantitative Model of Context Effects on Word Meaning in EEG and Cloze Data. In A. K. Goel, C. M. Seifert, & C. Freksa (Eds.), *Proceedings of Annual Meeting of the Cognitive Science Society 42* (Number 0, pp. 3085–3091). Cognitive Science Society.

Wicke, P. (2023). LMs stand their Ground: Investigating the Effect of Embodiment in Figurative Language Interpretation by Language Models. *Findings of the Association for Computational Linguistics: ACL 2023*, 4899–4913. https://doi.org/10.18653/v1/2023.findings-acl.302

Wilcox, E. G., Pimentel, T., Meister, C., Cotterell, R., & Levy, R. P. (2023). Testing the Predictions of Surprisal Theory in 11 Languages. *Transactions of the Association for Computational Linguistics*, *11*, 1451–1470. https://doi.org/10.1162/tacl_a_00612

Wilson, D., & Carston, R. (2007). A unitary approach to lexical pragmatics: relevance, inference and ad hoc concepts. In N. Burton-Roberts (Ed.), *Pragmatics* (pp. 230–259). Palgrave-Macmillan.

Winter, B., Perlman, M., & Majid, A. (2018). Vision dominates in perceptual language: English sensory vocabulary is optimized for usage. *Cognition*, *179*, 213–220. https://doi.org/10.1016/j.cognition.2018.05.008

Winter, B., & Strik-Lievers, F. (2023). Semantic distance predicts metaphoricity and creativity judgments in synesthetic metaphors. *Metaphor and the Social World*, *13*(1), 59–80. https://doi.org/10.1075/msw.00029.win

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. Le, Gugger, S., … Rush, A. M. (2020). *HuggingFace's Transformers: State-of-the-art Natural Language Processing*.

Wulff, D. U., & Mata, R. (2025). *Advancing Cognitive Science with LLMs*.

Wyld, H. C. (1936). *A history of modern colloquial English*. Basil Blackwell.

Xu, H., Nakanishi, M., & Coulson, S. (2024). Revisiting Joke Comprehension with Surprisal and Contextual Similarity: Implication from N400 and P600 Components. In L. Samuelson, S. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the Annual Meeting of the Cognitive Science Society 46* (pp. 1449–1456). Cognitive Science Society.

Xu, Q., Peng, Y., Nastase, S. A., Chodorow, M., Wu, M., & Li, P. (2025). Large language models without grounding recover non-sensorimotor but not sensorimotor features of human concepts. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-025-02203-8

Xu, Y., & Kemp, C. (2015). A computational evaluation of two laws of semantic change. *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, 2703–2708.

Yang, F.-P. G., Bradley, K., Huq, M., Wu, D.-L., & Krawczyk, D. C. (2013). Contextual effects on conceptual blending in metaphors: An event-related potential study. *Journal of Neurolinguistics*, *26*(2), 312–326. https://doi.org/10.1016/j.jneuroling.2012.10.004

Zhuo, J., Zhang, S., Fang, X., Duan, H., Lin, D., & Chen, K. (2024). ProSA: Assessing and Understanding the Prompt Sensitivity of LLMs. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 1950–1976). Association for Computational Linguistics.