A COMPUTATIONAL APPROACH TO LINGUISTIC RELATIVITY WITH EMPIRICAL EXPERIMENTATION

NUREDDİN CÜNEYD ÜNAL

BOĞAZİÇİ UNIVERSITY

2025

A COMPUTATIONAL APPROACH TO LINGUISTIC RELATIVITY WITH EMPIRICAL EXPERIMENTATION

Thesis submitted to the

Institute for Graduate Studies in Social Sciences

in partial fulfillment of the requirements for the degree of

Master of Arts

in

Cognitive Science

by

Nureddin Cüneyd Ünal

Boğaziçi University

2025

A Computational Approach to Linguistic Relativity

with Empirical Experimentation

The thesis of Nureddin Cüneyd Ünal

has been approved by:

Assist. Prof. Ümit Atlamaz _____
(Thesis Advisor)

Prof. Tunga Güngör _____
(Thesis Co-Advisor)

Assoc. Prof. Emre Uğur _____

Assist. Prof. Ömer Demirok _____

Assist. Prof. Fatma Başak Aydemir _____
(External Member)

June 2025

DECLARATION OF ORIGINALITY

I, Nureddin Cüneyd Ünal, certify that

- I am the sole author of this thesis and that I have fully acknowledged and documented in my thesis all sources of ideas and words, including digital resources, which have been produced or published by another person or institution;
- this thesis contains no material that has been submitted or accepted for a degree or diploma in any other educational institution;
- this is a true copy of the thesis approved by my advisor and thesis committee at Boğaziçi University, including final revisions required by them.

Signature…….

Date ……

ABSTRACT

A Computational Approach to Linguistic Relativity

with Empirical Experimentation

This thesis investigates linguistic relativity (LR), a hypothesis proposing that language influences thought, through a computational framework utilizing large language models (LLMs). To empirically test this proposition, the study prompt engineers LLMs as proxies simulating bilingual speakers, each representing a collective linguistic identity corresponding to one of 24 distinct native language groups, whose members are characterized by varying levels of English proficiency. Semantic embedding vectors are generated by eliciting adjective associations for nouns from various LLM configurations, thereby forming language specific conceptual representations. These embeddings are used to calculate semantic distances through measures of cosine similarity, whereby all of these language's conceptual representations are mapped to form a semantic field. Further experiments are conducted on these representational bilingual proxies to include both replications of existing studies in LR literature and novel cognitive tasks across a variety of domains, with results analyzed for systematic variations in task performance. The findings indicate that while native speakers of diverse languages employ distinct lexical choices to articulate conceptualizations of identical entities even in a common medium language (English), experimental data did not show statistically significant variations at the cognitive level of thought. Consequently, it is concluded that these cross-linguistic distinctions remain confined to the linguistic surface instead of extending to the deeper cognitive level of thought.

Overall, by bridging the fields of linguistics and artificial intelligence, this study contributes to the ongoing investigation of LR within cognitive science studies by presenting a innovative and scalable methodological approach that addresses traditional limitations.

ÖZET

Dilsel Göreliliğe Bilişimsel bir Yaklaşım: Deneysel bir İnceleme

Bu tez, dilin düşünceyi etkilediğini öne süren dilsel görelilik (DG) hipotezini, büyük dil modellerini (BDM) kullanan bilişimsel bir çerçeve dahilinde incelemektedir. Bu hipotezi deneysel olarak sınamak amacıyla çalışma, istem mühendisliği yöntemini kullanarak BDM'leri, farklı derecelerde İngilizce yeterliliğine sahip oldukları belirtilen 24 farklı anadilden konuşurların her birinin ortak kimliğini modelleyen ve iki dilli konuşucuları simüle eden vekil modeller olarak yapılandırmaktadır. Anlamsal gömme vektörleri, farklı BDM yapılandırmalarından isimlere yönelik sıfat çağrışımlarının elde edilmesiyle oluşturulmakta ve böylece dillere özgü kavramsal modeller meydana getirilmektedir. Bu gömme vektörleri kullanılarak kosinüs benzerliği ölçütüne dayalı anlamsal mesafeler hesaplanmakta ve tüm dillerin konumlandırıldığı ortak bir semantik uzay haritası oluşturulmaktadır. Hazırlanan iki dilli model vekilleri üzerinde halihazırda mevcut DG çalışmalarının tekrarlanmasının yanı sıra çeşitli alanlardan yeni bilişsel görevleri kapsayan deneyler yapılmaktadır ve bu modellerin görev performanslarındaki sistematik farklılıklar açısından analizi yapılmaktadır. Bulgular, farklı dillerin anadil konuşucularının, ortak bir iletişim dili olan İngilizce ile bile aynı kavramları ifade ederken farklı sözcük tercihleri yaptığını göstermiştir; ancak deneysel veriler bu sözcüksel farklılıkların düşüncenin bilişsel düzeyinde istatistiksel açıdan anlamlı değişimlere işaret etmediğini ortaya koymuştur. Dolayısıyla, bu farklılıklar, dilsel yüzeyde kalmakta ve bilişin daha derin düşünce seviyesine uzanmamakta olduğu sonucuna ulaşılmıştır.

Neticede, dilbilim ve yapay zeka disiplinlerini birleştiren bu çalışma, geleneksel kısıtlamaları aşan yenilikçi ve ölçeklenebilir bir yöntemsel yaklaşım sunarak, bilişsel bilimlerde DG'nin süregelen araştırmalarına katkı sağlamaktadır.

ACKNOWLEDGMENTS

As I approached the final stages of writing my thesis, I had told my thesis advisor Dr. Ümit Atlamaz that this journey had been such a humbling experience. I mention this here because I believe it truly reflects what an invaluable learning endeavor this journey has been for me. Undoubtedly, this learning experience was made possible by the spectacular guidance of both of my advisors, Dr. Ümit Atlamaz and Prof. Dr. Tunga Güngör. It is with no exaggeration when I say how fortunate I feel to have had such world class scholars as my advisors.

Dr. Ümit Atlamaz has a unique talent of incorporating his students into a rigorous academic mindset by teaching critical reasoning skills, which have been paramount in my learning process. He is incredibly academically unbiased, having been patiently supportive of my ideas, for which his contributions have refined and led to their practical implementations, resulting in the realization of this thesis. He has been an amazing mentor and beyond that, a sincere friend, for which I thank him.

Prof. Dr. Tunga Güngör has been greatly supportive for the implementation of my thesis, always ready to provide feedback. I've learned a great deal from him beyond just his contributions to my thesis. His kindness and professionalism, along with his incredible scope of knowledge have been a source of inspiration for me.

Both Dr. Ümit Atlamaz and Prof. Dr. Tunga Güngör are special in their merging of professionalism and sincerity; incredibly knowledgeable yet humble and down to earth academicians. I must admit that this short acknowledgment doesn't do justice to my advisor's characters or contributions, but for the sake of brevity, I shall suffice with my sincere expression of gratitude to both of my valuable advisors.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION 1

1.1 Language and thought 1

1.2 Questions and hypothesis 13

1.3 Thesis roadmap 20

CHAPTER 2: LITERATURE REVIEW 26

2.1 Linguistic relativity 26

2.2 Empirical investigations into linguistic relativity 33

CHAPTER 3: LINGUISTIC REPRESENTATIONS 52

3.1 Introduction to computational framework 52

3.2 Semantic field construction 53

3.3 Semantic embedding generation 64

3.4 Semantic distance analysis 67

3.5 Semantic space visualization and mapping 72

CHAPTER 4: LINGUISTIC RELATIVITY EXPERIMENTS 112

4.1 Cross-linguistic multiple choice experiments 113

4.2 Cross-linguistic spatial reasoning experiments 130

4.3 Cross-linguistic color reasoning tasks 144

4.4 Experimentation results and discussion 149

CHAPTER 5: CONCLUSION 152

5.1 Methodology discussion 153

5.2 Results discussion 166

5.3 Future directions 176

APPENDIX A: LANGUAGES AND LANGUAGE FAMILIES.....180

APPENDIX B: LANGUAGES BY GRAMMATICAL GENDER 182

APPENDIX C: LIST OF NOUN CATEGORIES AND LEXICAL ITEMS 184

APPENDIX D: LINGUISTIC PROPERTIES BY LANGUAGE 186

APPENDIX E: FIGURES AND TABLES REFERENCE 193

APPENDIX F: EXPERIMENT QUESTIONS.....196

REFERENCES 210

LIST OF TABLES

Table 1. Average Cross-Linguistic Distances for Noun Categories.....81

Table 2. Cosine Distance Cohen's d Across French-Persian 83

Table 3. Noun Category Overview Analysis—All Language Pairs.....83

Table 4. Average Noun Category Distances - Turkish vs. Average 84

Table 5. Distance of "Love" between Arabic and All Languages.....85

Table 6. "Love" in Arabic's Distance from All Languages 85

Table 7. Cross-Linguistic Random Sample Distances.....88

Table 8. Cross-Linguistic Distances 94

Table 9. Each Language's Distance to Collective Mean Distance.....95

Table 10. Grammatical Gender Category Distances.....100

Table 11. Sample Grammatical Gender Language Level Distances.....101

Table 12. Reproduction of Boroditsky and Mickan's Research 102

Table 13. ANOVA Results of Our Computational Framework.....104

Table 14. Cohen's d for Cross-Model Language Mean Distances.....106

Table 15. Cohen's d: Cross-Proficiency Language Mean Distances 108

Table 16. Cohen's d: Cross-Proficiency Language Mean Distances 119

LIST OF FIGURES

Fig. 1 LM temperature configuration (Sanderson, 2024) 60

Fig. 2 Embedding vector semantic similarity visualization (Sanderson, 2024) 65

Fig. 3 Embedding vectors semantic (cosine) distances (Sanderson, 2024) 69

Fig. 4 3D t-NSE interactive semantic space of noun categories.....78

Fig. 5 3D PCA interactive semantic space of each nouns position.....79

Fig. 6 2D t-NSE interactive space of each nouns position 79

Fig. 7 “Food Items” cosine distances between languages 81

Fig. 8 “Human Roles” cosine distances between languages.....82

Fig. 9 “Cultural Artifacts” distances between languages.....82

Fig. 10 Average noun category distances between all languages 84

Fig. 11 Raw Correlation of cosine and Jaccard Distances.....86

Fig. 12 Noun category distances between Portuguese - all languages 87

Fig. 13 Nouns plotted with color coded noun categories by language 87

Fig. 14 Nouns plotted with color coded noun categories by language 88

Fig. 15 3D PCA interactive semantic space of each nouns position.....89

Fig. 16 3D UMAP interactive semantic space of each nouns position.....89

Fig. 17 Cosine - Jaccard pairs for all nouns in Japanese 90

Fig. 18 Cosine - Jaccard pairs for all language pairs 91

Fig. 19 2D UMAP interactive semantic space of languages.....92

Fig. 20 2D t-NSE interactive semantic space of languages 92

Fig. 21 2D UMAP interactive language space for “Time Concepts” 93

Fig. 22 2D UMAP interactive language space for “Time Concepts” 93

Fig. 23 2D t-NSE interactive semantic space of languages - “Food Items” 94

Fig. 24 Language distance heat map 96

Fig. 25 Language distance dendrogram 96

Fig. 26 Language family distance heat map 97

Fig. 27 2D PCA Language family distance semantic map 98

Fig. 28 Grammatical gender system distances heat map 99

Fig. 29 Grammatical gender system distances dendrogram 99

Fig. 30 Cross-LLM provider Cohen’s d heat map 105

Fig. 31 Cross-temperature Cohen’s d 107

Fig. 32 Cross-proficiency Cohen’s d 107

Fig. 33 Relative variability across proficiency levels for BB and TN 110

Fig. 34 Multiple choice response distribution by language 121

Fig. 35 Neutral animals cross-linguistic variance by response 130

Fig. 36 ASCII SRT-1 132

Fig. 37 ASCII SRT-2 136

Fig. 38 ASCII MN-1-2 140

Fig. 39 Spatial reasoning most common response (route) per language 141

Fig. 40 Color reasoning response distribution by language 148

ABBREVIATIONS

AI	Artificial Intelligence
CL	Computational Linguistics
CS	Cognitive Science
LM	Language Model
LMM	Large Language Model
LR	Linguistic Relativity
NLP	Natural Language Processing
SPH	Sapir-Whorf Hypothesis

CHAPTER 1

INTRODUCTION

1.1 Language and thought

The intricate relationship between language and thought constitutes an essential topic of inquiry in cognitive science (CS), with profound implications for understanding human cognition. Scholars representing diverse theoretical traditions have long debated the extent to which language serves merely as a vehicle for expressing pre-existing mental constructs, or whether language actively shapes cognitive structures, influencing how individuals perceive, conceptualize, and reason about the world around them (Lucy, 1992). Within the extensive discourse examining these two contrasting positions, the hypothesis of linguistic relativity (LR) emerges as particularly significant, positing that the structural and semantic characteristics inherent to specific languages systematically influence thought processes.

Since its inception in early 20th century writings—most prominently associated with Edward Sapir and Benjamin Lee Whorf—LR has attracted both enthusiastic supporters and cautious skeptics. Sapir, drawing upon ethnographic linguistic observations in his seminal work “Language: An Introduction to the Study of Speech” (1921), suggested that linguistic categories, such as grammatical structures and lexical distinctions, predisposed speakers toward certain cognitive orientations by shaping how the external world was parsed conceptually. He argued that language influences thought by providing a framework for organizing experience, as seen in his analysis of diverse linguistic communities. Whorf, building on Sapir’s ideas, further developed this perspective through comparative studies of

Native American languages in “Language, Thought, and Reality” (1956). He highlighted that linguistic structures influenced key cognitive domains like temporal cognition, event construal, and causal reasoning, proposing that these linguistic differences led to distinct worldviews among speakers.

However, these pioneering efforts by Sapir and Whorf were not without criticism, largely due to methodological limitations inherent to their qualitative ethnographic approaches, which often relied on subjective observations and lacked standardized experimental controls (Lucy, 1992). Critics, such as Pinker (1994), argued that their studies suffered from ambiguity in defining and quantifying cognitive effects across diverse linguistic communities, making it challenging to establish causal links between linguistic structures and thought processes. Additionally, the absence of rigorous statistical analysis and replicable methodologies in their work led to skepticism about the generalizability of their findings (Pullum, 1991).

The formal emergence and subsequent development of CS provided essential frameworks to revisit LR from multiple disciplinary perspectives. Central to CS’s objectives is the systematic study of mental processes—including language, reasoning, categorization, memory, and problem-solving—through integrative research methodologies (Thagard, 2023). As language constitutes a primary pathway for externalizing internal cognitive states, it affords valuable empirical perspective for analyzing thought, effectively providing a “window into the human mind.” Essentially, intricate linguistic activities reflect underlying cognitive architectures and semantic structures that researchers strive to characterize empirically. Thus, careful linguistic analysis significantly enriches our understanding of thought,

offering profound insights into fundamental cognitive processes. As CS evolved in this way during the twentieth century, debates shifted towards more controlled and empirically rigorous paradigms, adopting experimental methodologies informed by cross-linguistic comparisons, as evidenced by studies in LR and cognitive diversity (Lucy, 1992; Levinson, 2003; Gentner & Goldin-Meadow, 2003).

Recent developments in artificial intelligence (AI) have facilitated a new approach in the the analysis of cross-linguistic studies in a computational environment. Most recently, natural language processing (NLP), a subfield of AI has undergone substantial advancements through the emergence of transformer-based large language models (LLMs), with the attention mechanism at its core (Vaswani et al., 2017; Devlin et al., 2019; Liu et al., 2019; Brown et al., 2020). Contemporary computational linguistics (CL) paradigms conceptualize language as multidimensional representational spaces, comprised of lexical elements whose semantic meanings derive substantially from statistical and contextual co-occurrences. Within such distributional semantic models, words—and even entire languages—can be quantitatively represented as vector embeddings, encoding semantic relations within high-dimensional mathematical spaces (Mikolov et al., 2013; Peters et al., 2018).

Such computational semantic approaches facilitate objective, quantifiable, and systematic analyses of linguistic semantic structures, making large-scale cross-linguistic comparisons empirically practical and scientifically robust. Researchers have recently begun taking advantage of the capabilities of these LLMs in cross-linguistic contexts, leveraging multilingual LLMs to explore how semantic and structural representations vary across languages. For instance, Dubossarsky et al.

(2020) present a large-scale study demonstrating that the spectral similarity of monolingual embedding spaces correlates strongly with performance across various cross-lingual tasks—such as bilingual lexicon induction, part-of- speech tagging, parsing, and machine translation, subfields of NLP—suggesting that quantifiable structural properties of languages influence task success. In a related but distinct line of inquiry, Gonen et al. (2022) demonstrated that multilingual language models encode gender distinctions with both shared and language-specific components, revealing how these internal representations enable cross-linguistic transfer while limiting the transferability of gender removal interventions. Together, these studies exemplify how CL driven methodologies can reveal both universal and language-specific patterns in semantic representation. This line of inquiry has proven especially promising in investigating how different linguistic systems encode meaning, revealing patterns of both convergence and divergence in multilingual semantic modeling.

This shift naturally opens a compelling avenue for advancing research on LR by equipping CL with tools capable of empirically testing conceptual variation across languages. Through the use of transformer-based generative models, computational linguists can construct controlled, reproducible frameworks that simulate cross-linguistic semantic behavior. These frameworks allow for the systematic quantification of conceptual distance between languages based on their modeled linguistic outputs. By doing so, they enable direct operationalization of LR claims within computational environments, transforming what were once qualitative, interpretive challenges into measurable, replicable experimental designs.

Following this new approach to assessing cross-linguistic studies in a CL framework, this thesis explicitly situates LR within bilingual semantic modeling frameworks as a core methodological tool to probe the relationship between linguistic representations and thought empirically and quantitatively. The principal research problem addressed herein aims at empirically assessing, measuring, and systematically analyzing hypothesized cognitive variability consequential to differences in linguistic properties across multiple languages. By adopting a CL paradigm, one that leverages sophisticated LLMs as operational simulations of human behavior as a proxy for semantic cognition, this thesis comprehensively investigates potential linguistic effects on thought processing through carefully engineered bilingual experimental paradigms.

Computational semantics, realized through large-scale linguistic modeling, enables the direct quantification and measurement of conceptual variability across languages. This methodological approach uses the aforementioned computationally derived semantic representations as proxies that model and capture human behavior in an attempt to reflect deeper thought processes. This enables controlled simulations of bilingual cognitive scenarios to systematically test whether cognitive behaviors respond in consistent manner to quantifiable structural differences across linguistic profiles. Given this methodological emphasis, the thesis specifically addresses the complex interaction between language and thought in computational terms, centering the discourse firmly within interdisciplinary research areas bridging linguistics and AI.

Thus, this thesis integrates foundational theoretical perspectives derived predominantly from anthropology and cognitive linguistics with applied

computational methodologies emerging from contemporary AI, striving toward an objective, empirically rigorous, and theoretically informed exploration of the language-thought relationship. In doing so, it navigates a scholarly debate characterized historically by theoretical nuance and methodological complexity, explicitly articulating precise operational definitions and methodological assumptions designed to enhance empirical clarity.

In subsequent sections of this introductory chapter, core notions of language and thought are systematically defined and elaborated within the context of CS and the research pertaining to this thesis. This is followed by the fundamental questions this research seeks to answer, as well as a clearly defined hypothesis regarding the implications of assessing LR through a computational proxy of linguistic representations. Ultimately, such an integrative exploration strives to contribute toward refining understandings of whether and to what extent cross-linguistic properties might systematically influence thought processes, engaging actively with contemporary debates while simultaneously providing empirically and computationally grounded insights that extend the interdisciplinary understanding of the relationship between language and thought.

1.1.1 Defining language

Language is a uniquely human cognitive capacity characterized as a structured symbolic system used for communication and representation, enabling individuals to encode, convey, and interpret complex ideas, emotions, intentions, and knowledge. This system, as Saussure (1916) emphasized, relies on arbitrary signs to form a structured code, while Chomsky (1965) highlighted its generative nature, rooted in

innate cognitive structures. Pinker (1994) further underscored language's role as a faculty for expressing and sharing complex mental representations. Linguistic systems consist of multiple interrelated levels of representation, including phonological (sound-based), morphological (word-formation), lexical-semantic (word meanings and conceptual categories), syntactic (sentence structure), and pragmatic (contextual and intended meanings) organization, all integrated under the umbrella of their grammars (Jackendoff, 2002; Halliday & Matthiessen, 2004; Van Valin & LaPolla, 1997). Jackendoff (2002) articulates a parallel architecture model that integrates phonology, syntax, and semantics, while Halliday and Matthiessen (2004) provide a systemic functional perspective that encompasses grammar, semantics, and discourse. Van Valin and LaPolla (1997) offer a role and reference grammar framework that connects syntax, semantics, and pragmatics within a unified system.

Within the framework of CS, and specifically for exploring LR, language functions as a critical window into cognitive architecture. By systematically analyzing linguistic properties and their structures, researchers attempt to access underlying cognitive representations and infer mental processes which are otherwise opaque and challenging to investigate directly. It may therefore be assumed that language takes on dual roles: on one hand it serves as a medium to externalize and communicate pre-existing cognitive states, and on the other as a potentially active agent in structuring and guiding cognitive processes themselves; the latter of which this thesis attempts to empirically assess the legitimacy and degree of.

Moreover, as this thesis specifically examines LR utilizing CL methodologies, the semantic dimension of language emerges as particularly salient,

with other aspects of language embedded with particular linguistic systems potentially influencing the semantic representations for their respective speakers. Semantics is the level of linguistic apprehension intimately connected with conceptual representation in the human cognitive architecture (Jackendoff, 1990). Fundamental to semantics are categorization processes, which determine how distinct languages partition and classify the continuous and multidimensional external sensory environment into discrete, linguistically encoded categories, an aspect of languages that we will delve deeper into further on in this research. Through such linguistic categories and cross-linguistic lexical distinctions, semantic analyses can be systematically mapped onto conceptual structures within cognitive space.

Comparative linguistic studies have demonstrated substantial cross-linguistic variation in domains such as grammatical structure, grammatical gender and natural classification systems, spatial, temporal and quantitative reasoning, color categorization, and auditory reasoning as is extensively detailed in (Section 2.2). These diverse categorization systems provide an empirical foundation for investigating linguistic influences on thought through behavioral assessments reflecting thought processes.

In the context of the central research question pursued in this thesis, which concerns the relationship between cross-linguistic properties and structures and conceptual cognition, language is operationalized explicitly as the quantifiable semantic and conceptual distinctions encoded in lexical and semantic representations that systematically vary cross-linguistically. Consequently, language within this thesis becomes a scientifically precise and empirically operationalized construct,

specifically focused on semantic categorization and conceptual-structural characteristics that may be reliably captured in CL representations in the form of embedding vectors (Mikolov et al., 2013; Dubossarsky et al., 2020, Gonen et al., 2022). Thus, language is not considered as a monolithic entity, but rather as a collection of systematic semantic distinctions, embedded within it linguistic properties, structures and cultural aspects which, when quantified computationally by means of a cosine distance metric within a shared cross-lingual vector embedding space, these systematically varying semantic distinctions offer an empirical basis for investigating whether—and to what extent—linguistic diversity influences experimental behavior and thought induced cognitive variability across speakers from distinct linguistic backgrounds. This carefully defined notion of language serves as a foundation for exploring how language-specific semantic structures and categorical systems might interact with, influence, or shape cognitive processes, thus providing a comprehensive empirical framework to rigorously examine LR through computational modeling and cross-linguistic cognitive experimentation.

1.1.2 Defining thought

In the context of investigating LR and bilingual semantic modeling, precisely defining and operationalizing thought processes is fundamental, as these cognitive phenomena, as functions of behavioral analyses, are the central targets of empirical examination and computational simulation in the present study. The construct of "thought" broadly encapsulates the ensemble of internal cognitive processes underlying the generation, manipulation, and application of mental representations that allow humans to perceive, interpret, and engage meaningfully with the external

and internal environment (Barsalou, 2008). Thought encompasses a wide range of cognitive functions, including conceptual categorization, reasoning, decision making, problem-solving, inference-making, perception, memory retrieval, and executive control mechanisms—each reflecting complex and intertwined mental activities that remain central to human intelligence and behavior (Anderson, 1983; Miyake et al., 2000).

Within CS, thought can be more specifically understood as structured cognitive operations on internal representations of knowledge and experience that permit individuals to reason, deliberate, and act effectively within their environments. These internal representations, which cognitive scientists frequently characterize as mental concepts or conceptual structures, are specifically grounded within multi-layered cognitive architectures, and serve as fundamental components facilitating semantic categorization, logical inference, analogical mapping, abstraction formation, and adaptive cognitive functioning (Vygotsky, 1978).

LR hypotheses center upon how the structural and semantic properties—discussed in length in future chapters—of particular languages may potentially shape, constrain, or guide these cognitive representations and their processing dynamics. Hence, cognitive investigations anchored in LR typically concentrate on elucidating precisely how language properties and structures may systematically influence underlying representational patterns and cognitive operations across linguistic communities. Lucy (1997) provides a reformulated framework for LR, emphasizing empirical studies that demonstrate how linguistic structures, such as grammatical and lexical categories, shape cognitive processes like categorization and reasoning across diverse linguistic groups.

Consequently, relevant thought processes for the present thesis are those empirically observable, quantifiable, and experimentally tractable cognitive tasks that explicitly reflect underlying conceptual and semantic structures—thus making them amenable to systematic variation and cross-linguistic comparative analyses. A broad typology of thought processes relevant to the specific aims of this thesis includes the following:

Classification and Categorization: These cognitive processes involve identifying, distinguishing, grouping, organizing and relating objects, events, and abstract concepts within mental hierarchies and semantic frameworks. Classification tasks assess how cognitive systems assign conceptual categories and distinguish conceptual boundaries. Linguistic influences on classification are claimed to be observable in semantic categorization tasks that reflect lexical and grammatical distinctions, such as color categorization, spatial categorization, numerical scales, or grammatical gender assignments (Lakoff, 1987; Lucy, 1992; Boroditsky, 2001).

Reasoning and Logical Inference: Reasoning refers to cognitive operations used for systematically processing evidence, deducing conclusions, inferring causal and functional relationships, and generating coherent justifications or solutions (Anderson, 1983). Subtypes of reasoning specifically relevant include spatial reasoning (processing spatial relations and orientational configurations), temporal reasoning (representing temporal sequences, durations, and causal chains), causal reasoning (inferring causal structures and relationships), and analogical reasoning (generalizing concepts and relational mappings from one domain to another). Linguistically encoded frameworks are claimed to actively influence reasoning by

highlighting certain semantic distinctions that orient and bias cognitive inference patterns (Boroditsky et al, 2011; Levinson, 2003).

Problem-solving involves recognition of an initial state, delineation of a goal state, and the application of logical processes to navigate from the former toward the latter (Newell & Simon, 1957). Decision-making encompasses evaluative choice phenomena based on assessing alternatives against cognitive criteria, outcomes, preferences, and constraints (Kahneman & Tversky, 1979). Linguistic differences in categorization, conceptual encoding, and grammatical structure may influence problem-solving style, heuristics, strategies, and decision preferences, particularly by making certain cognitive paths more accessible or cognitively salient in specific language contexts (Boroditsky, 2010).

In the context of empirical research probing LR, these categories of cognitive processes—classification, reasoning, and decision-making strategies—provide operationally precise targets for the examination of cognitive variability potentially influenced by linguistic structures across languages. Integrating recent computational advances from AI, specifically those of CL, provides systematic frameworks for simulating, modeling, recording, and measuring such cognitive phenomena empirically through the lens of behavior as a proxy for underlying thought processes. The aforementioned computational approaches simulating bilingual scenarios relate conceptually relevant cognitive tasks to linguistic representation spaces, explicitly quantifying semantic differences between languages and operationalizing thought processes through empirical computational proxies. Thus, thought becomes definable operationally and quantitatively measurable through linguistic responses and

semantic associations recorded in structured computational experiments explicitly designed to probe LR hypotheses.

In summary, “thought”, as defined and operationalized in this thesis, is fundamentally conceptualized as structured and empirically measurable cognitive processes reflecting semantic structures and representational content shaped potentially by linguistic influences. Computational modeling methods, particularly those involving large-scale semantic embeddings and systematically designed cognitive tasks, facilitate empirical testing of potential differences in thought processes produced by linguistic structure. Ultimately, a detailed definition and thorough conceptualization of thought as an empirically accessible, measurable cognitive process by means of behavioral experimentation provide a theoretical foundation and methodological framework necessary to systematically evaluate hypotheses positing linguistic influences upon thought as a cognitive process in cross-linguistic semantic contexts. The behavioral task experiments designed to probe the language—thought relation are discussed in further detail in Section (1.3.2) and available in their entirety in Chapter 4.

1.2 Questions and hypothesis

This section presents the main questions and hypotheses of this thesis, which seek to explore the relationship between language and thought within the framework of LR. The investigation begins with an articulation of core questions designed to critically frame how language properties and structures may influence cognitive patterns among speakers from diverse linguistic backgrounds in (Section 1.2.1). Subsequent to this discussion, the next (Section 1.2.2) provides a hypothesis, outlining an

empirically testable prediction regarding the quantifiable cognitive differences expected to emerge from varying linguistic foundations. In formulating this hypothesis, particular emphasis is placed upon adopting a CL approach, specifically harnessing advanced generative language modeling techniques that permit the semantic characterization of languages and the empirical measurement of linguistic-thought influences as a function of distinct behavioral manifestations. Collectively, these research questions and the associated hypothesis provide the foundational logic guiding both the theoretical and empirical methodologies of this research, facilitating a rigorous exploration of LR through innovative computational means.

1.2.1 Research questions

The research questions outlined in this subsection articulate the core inquiries guiding the exploration of LR. Central to this inquiry is whether—and to what measurable degree—the intrinsic properties of a given language influence cognitive functions and patterns of thought in its speakers. In particular, attention is directed toward determining whether linguistic diversity among distinct linguistic properties found in various languages lead to perceptible differences in how individuals interpret events, perceive experiences, and make decisions, as well as whether empirical methodologies can reliably detect and quantify these cross-linguistic cognitive disparities.

Building upon this fundamental premise, the research also explores specific linguistic dimensions—including lexical semantics, grammatical structure, grammatical gender and natural classification systems, spatial, temporal and quantitative reasoning, color categorization, and auditory reasoning—to identify

whether and to what extent any of these aspects cause cognitive variability across speakers of different languages. Furthermore, the research seeks to clarify whether effects associated with LR are more pronounced in particular semantic or cognitive domains—for instance, differentiating between abstract versus concrete, animate versus inanimate, and tangible versus intangible conceptualizations such as possible perception divergencies between languages for abstract concepts such as “freedom” versus inanimate objects such as “chair”. Another area of examination involves identifying whether there are recognizable patterns in cognitive influences within clusters of related languages compared to those from different language families. These aspects of our study are detailed extensively in Chapter 3.

Given that previous approaches to LR often faced limitations such as small sample sizes, simplified experimental designs, and overly generalized conclusions (Lucy, 1992), this project proposes methodological refinements and novel experimental paradigms aiming to clarify cognitive effects directly attributable to linguistic properties. As noted in the previous section, recent innovations in AI, particularly in the form of LLMs—serve as key methodological tools enabling more rigorous measurement and systematic analysis of LR. Unlike traditional methods, computational modeling approaches offer scalable quantitative frameworks with a systems level approach for capturing subtle variations in thought processes driven by linguistic features.

Thus, this thesis integrates insights from linguistics and AI under the umbrella of CS. It both replicates prior LR research on a transformed and enlarged computational scale, and introduces original experimental tasks specifically designed to probe aspects of language-thought interactions in new ways. These efforts aim to

critically assess the validity of previous LR research by providing novel contributions with empirical research while also contributing to an integrative, multidisciplinary understanding of how linguistic structures may influence human thought.

1.2.2 Hypothesis

This thesis hypothesizes that linguistic properties and structures influence thought processes to a meaningful degree, leading speakers of distinct languages to demonstrate quantifiable disparities in cognitive behavior—particularly in conceptual and semantic tasks. To empirically examine this hypothesis, LLMs are strategically prompt-engineered to emulate the performance of bilingual speakers, thereby serving as proxies for approximating human cognitive behavior. Variations observed in LLM performance across different languages will thus be interpreted as evidential indicators representing the impact exerted by native language structures on cognitive tasks performed through a shared second language medium. Moreover, this research further hypothesizes that computational representations of language, generated via innovative, large-scale semantic modeling techniques involving systematic adjustment of LLM parameters combined with meticulously designed prompts, may work to encapsulate the cognitive profiles typical of various linguistic backgrounds to a reliable degree of accuracy, as is apparent in the experimental results in Chapter 4. Specifically, employing deterministic parameters of model settings will effectively highlight intrinsic statistical co-occurrences unique to each language, thereby generating CL representations capable of capturing and revealing differential

behavioral responses to identical cognitive assessments as a proxy for language based cognitive variations. The core prediction thus stipulates:

When bilingual cognitive tasks are executed by LMs simulating speakers with distinct native languages proficient in English, there will be measurable variations in task performance and output choices, and these variations will exhibit a correlation with the semantic distances observed between corresponding language representations derived through generated embedding values, indicating an influence of language on the thought processes resulting in those behavioral discrepancies.

The hypothesis is grounded within the framework of LR, specifically the weaker form commonly recognized within contemporary CS (Lucy, 1997; Gentner & Goldin-Meadow, 2003), which holds that while thought is fundamentally universal, linguistic structures and practices nevertheless encourage or highlight distinct cognitive habits and biases across different languages. To provide an empirical basis to this hypothesis, this research has developed a novel computational approach leveraging the powerful representational capabilities of state of the art LLMs. The fundamental logic follows several steps: semantic modeling, quantitative assessment of linguistic distance, and systematic bilingual task evaluation. We gloss over them in the next subsection, and provide explanation on the logical reasoning behind these approaches and their detailed implementations in their respective sections.

1.2.3 Reconciling language and thought interactions

Empirical research on LR posits that language shapes cognitive processes, influencing perception, categorization, and reasoning, with effects observable at behavioral and neurological levels (Boroditsky, 2001; Lucy, 1992; Levinson, 2003),

as is elaborated on in (Section 2.2), noting that these effects emerge very early in the perceptual process as a claim for their effect on thought (Winawer, 2007).

For instance, Boroditsky (2001) found that Mandarin speakers' vertical metaphors for time (e.g., "past is above") enhance performance on vertical temporal tasks, with faster response times (~150–200 ms) and higher accuracy (95% vs. 85%) compared to English speakers. Similarly, Winawer et al. (2007) showed that Russian speakers' lexical distinctions for blue shades ("goluboy" vs. "siniy") improve color discrimination speed by ~100–150 ms, with Thierry et al. (2009) providing neurological evidence via event-related potentials (ERPs) showing pre-attentive categorical shifts in Greek speakers for light vs. dark blue at 150–250 ms, indicating language's influence on early visual processing. Roberson et al. (2000) further demonstrated that Himba speakers' five-term color lexicon enhances discrimination at language-specific boundaries, unlike English speakers' broader categories. Boroditsky et al. (2003) found that constraining language during reasoning tasks—by having participants repeat irrelevant utterances—impairs performance, suggesting that language facilitates cognitive processing. These findings claim that language influences thought early in perceptual processes, particularly in color and temporal domains, and supports task performance.

However, these claims are not without challenges from an opposing perspective as abundant research in this field act as an antithesis to these claims. The claimed performance impairment from irrelevant utterances may stem from cognitive resource competition due to increased task complexity, rather than the absence of linguistic input (MacLeod, 1991; Baddeley, 1992). For example, repeating irrelevant phrases could overload working memory, diverting resources from reasoning tasks,

reflecting a negative effect of cognitive load rather than language's necessity for cognition. Additionally, evidence for a dissociation between language and thought Ivanova (2024) contests stronger LR claims. It is argued that Broca's aphasia, where linguistic deficits do not impair non-linguistic tasks like numerical reasoning or visual categorization, suggests thought operates independently of language, and that it is apparent in toddlers that thought precedes language. Studies on pre-linguistic cognition further support this: deaf children with minimal linguistic input produce homesign systems, exhibiting complex categorization and problem-solving (Goldin-Meadow & Feldman, 1977; Senghas & Coppola, 2001). These findings imply that thought precedes and can function without language, which fundamentally disregards the concept of language influencing thought.

The weaker LR framework reconciles these perspectives, positing that language and thought interact dynamically, with linguistic structures guiding cognitive habits without imposing absolute constraints (Gentner & Goldin-Meadow, 2003). This interaction is claimed to be evident in numerous works regarding cross-linguistic behavioral differences, even if underlying cognitive mechanisms remain opaque. Analogous to the Turing Test, which infers intelligence from behavioral outputs without accessing internal processes (Turing, 1950), this approach sidesteps the metaphysical "ghost in the machine"—the notion of a separate cognitive essence critiqued by Ryle (1949)—by focusing on observable linguistic behaviors as proxies for cognitive differences. That is, just as Ryle argued that mental states are manifested in dispositions of actions, this thesis uses computational linguistic models to quantify semantic and behavioral variations in bilingual tasks, interpreting these as evidence of language-driven cognitive influences. For instance, LLM-generated

adjective associations for nouns across languages may reflect language-specific cognitive biases, with similar approaches for cross-linguistic tasks haven proven effective (Dubossarsky et al., 2020). Thus, while we may not be able to capture thought nakedly as an elusive intangible concept, we may utilize linguistic structures to measure the influence of language on cognitive performance, as evidenced by behavioral disparities in cross-linguistic contexts.

1.3 Thesis roadmap

Thus far in this thesis, the foundational groundwork has been laid down by defining language and thought independently within the research context, clarifying how they relate to one another from the perspective of CS, and highlighting key insights gained from two central fields within CS—namely, linguistics and AI—in order to frame the research question presented in Section 1.1, “Language and Thought”. Building upon this, in Section 1.2, “Questions and Hypothesis”, we have defined targeted research questions and expressed a hypothesis designed to empirically assess the validity of LR in the context of the research questions provided earlier. We proposed a novel computational framework employing LLMs configured to simulate bilingual cognitive profiles, which allows us not only to revisit previously conducted LR experiments at an unprecedented scale, but also to implement original tasks and systematically analyze cognitive divergences associated with distinct linguistic profiles.

In the current section, our aim is to outline clearly and concisely the structure of the remainder of this thesis, guiding the reader through the logic underlying subsequent chapters as a referential roadmap. We particularly focus on previewing

Chapters 3 and 4—titled "Linguistic representations" and "Linguistic relativity experiments," respectively—to contextualize the methodological implementation of our research approach. This overview aims to clarify the rationale and progression of empirical and computational investigations, setting a clear intellectual trajectory for the analyses and discussions that follow.

1.3.1 Linguistics representations

Initially, languages are modeled to generate computational semantic fields representing each language distinctly. This is accomplished through the implementation of multiple LLMs, which collectively represent diverse linguistic and cognitive sampling pools. This computational paradigm comprises an array of commercially available models, all of which have been systematically configured to exhibit distinct bilingual cognitive profiles. These profiles are precisely instantiated via carefully engineered system prompts which explicitly designate linguistic identities reflective of native speakers from diverse linguistic backgrounds acquiring English as a second language at varying proficiency levels. Through this carefully structured strategy, the models generate output effectively emulating bilingual language production, facilitating scalable and empirically informed analyses of how linguistic background shapes thought representation and linguistic performance.

Using an automated Python-based pipeline (available on GitHub[1]), language models are first assigned a specific native language and a defined level of English proficiency to represent bilingual individuals who have acquired English. These models are subsequently prompted to generate adjectives semantically associated

[1] https://github.com/nurcunal/linguistic-relativity-computational-framework

with a comprehensive set of nouns drawn from various categories. Rather than directly extracting noun embeddings from multilingual LLMs—an approach known to obscure subtle, language-specific features due to the alignment and merging of linguistic representations within a unified embedding space (Gonen et al., 2022)—our method employs a reconstructive procedure. Specifically, we reconstruct noun embeddings by aggregating the embeddings of their semantically associated adjectives using mean pooling, a technique effectively demonstrated in existing research (Mikolov et al., 2013; Teehan et al., 2024). Consequently, each noun is represented by a newly derived aggregate embedding. These embedding vectors are generated in embedder models distinct from the LLMs that the adjectives are acquired from as a means of added a level of abstraction away from biases that may be inherent in those LLMs. This methodological strategy explicitly seeks to simulate bilingual speakers to examine the extent to which structural elements from their native language persistently influence linguistic cognition—even within experimental tasks devoid of overt native-language cues—thus underscoring the enduring effects linguistic structures exert on cognitive processes, as further elaborated in Chapter 3. This framework is meticulously designed for quantitative representation, involving a large scale generation process. This methodological rigor has resulted in millions of adjective data points, which are subsequently embedded to high-dimensional vector representations. The vector dimensions succinctly encapsulate the nuanced semantic biases imparted by the native language when individuals employ English, replicating potential cognitive influences of each speaker's native language projected onto a common medium between all models.

These aggregated embeddings form a semantic representation for each noun as a conceptual fingerprint that projects the cognitive effect of LR onto noun conceptualization; each noun's aggregated embedding thus aims to represent a computational proxy for the representation of that noun elicited from an artificial bilingual speaker. Further aggregation of these noun embeddings provides a unified language-level representation. The semantic distance between these language representations is measured using cosine similarity, producing a reliable quantification of how far conceptually distinct languages are from each other. A semantic map is then constructed to depict the conceptual proximity or distance between languages at both the overall language level and the more granular noun level. Comparable assessments are also conducted for individual nouns at the categorical level between all LMs (e.g., abstract concepts vs. inanimate objects, etc.), providing nuanced insight into whether language differences exert a stronger cognitive influence over particular conceptual domains. Additional analysis is conducted on language families to evaluate potential clustering patterns within and across these families. These processes are detailed extensively in Chapter 3.

1.3.2 Linguistic relativity experiments

Having quantified and mapped semantic distances through comprehensive linguistic representations, the subsequent phase of research (detailed in Chapter 4) involves systematically employing these computational constructs in controlled cognitive experimentation. In particular, this thesis utilizes these bilingual linguistic representation models—configured to emulate speakers of different native languages proficient in English as mentioned previously—to execute a series of cognitive tasks

in English as a medium for cross-linguistic bilingual representation models. The primary methodological objective behind these experiments is to observe whether variations in cognitive performance and behavior are present and examine their correlation in a meaningful and consistent manner with the semantic distances derived from the previously established linguistic representations. The experimental designs implemented in this thesis are organized into two distinct categories:

i. Reimplementation of prior experiments with bilingual models:
   - Grammatical gender and natural classification tasks (influencing categorization and conceptualization)

ii. Novel cognitive tasks designed for bilingual models:
   - Multiple choice preference tasks
   - Spatial reasoning and representation tasks
   - Color perception and categorization

These cognitive experiments are designed to ensure methodological consistency and rigor, while preserving comparability of task parameters across language models. The outcomes from this phase will subsequently undergo quantitative analysis to investigate behavioral divergences in relation to defined semantic distances. Distinct responses from different linguistic representation models and a direct correlation with the generated semantic maps of the respective models would lend substantial empirical weight to the central hypothesis advanced earlier in this thesis, reinforcing the proposition that linguistic properties indeed exert measurable influence upon thought processes and the conceptual representations held by speakers of different linguistic backgrounds.

1.3.3 Significance and contribution of this research

The CL approach significantly contributes to and builds upon other previous research in this field. Employing computational language modeling as an avenue to rigorously test LR offers significant methodological advantages unavailable through traditional behavioral experimentation. The quantitative precision in the computational embedding approach provides an empirically objective, reproducible continuous scale of semantic comparison, reducing subjective interpretive biases inherent in manual qualitative linguistic analyses. The use of an automated API pipeline facilitates replication and momentous scaling of experiments beyond the limited subject pool constraints that have challenged previous LR research, thus directly countering sample size limitations (Lucy, 1992). Furthermore, commercial LLMs currently represent among the highest fidelity simulations of human linguistic processes available. Specifically instructing these models to replicate bilingual cognitive mechanisms by means of prompt engineering and customization of the model's parameters enables precise experimental control over proficiency levels, native speaker identity, and task execution conditions. This specific CL framework merges large language data with AI in a novel manner, producing methodological innovation and comprehensive empirical testing otherwise unavailable at such large scale.

CHAPTER 2

LITERATURE REVIEW

2.1 Linguistic relativity

LR encapsulates the theoretical proposition that the properties, structures and lexical distinctions inherent to a given language exert substantial influence over the cognitive processes of its speakers. Expanding upon anthropological and linguistic scholarship, particularly the seminal contributions of Edward Sapir and Benjamin Lee Whorf, this paradigm argues that language is not merely a passive instrument of thought but acts dynamically to shape perception, categorization, attention, memory, and reasoning. LR thus posits that linguistic diversity translates into meaningful cognitive variability, generating distinct conceptual frameworks and perceptual tendencies within different language communities (Sapir, 1921; Whorf, 1956; Lucy, 1997).

This thesis adopts LR—particular its weaker, nuanced formulation as detailed in Section 2.1.3—as the primary theoretical framework guiding hypotheses positing measurable, language-driven cognitive differences. Through computational semantic modeling and empirical cross-linguistic cognitive experiments (Chapters 3 and 4 respectively), the research investigates whether and to what extent linguistic structures contribute toward divergences in cognitive behavior. This section explicates the conceptual foundations established by Sapir and Whorf, delineating the distinction between stronger (deterministic) and weaker (relativistic) versions of their theory. Subsequently, in Section 2.2, it extensively evaluates the empirical evidence available from contemporary cognitive-psycholinguistic and

anthropological studies that have demonstrated linguistic influences within domains including lexical semantics, grammatical structure, grammatical gender and natural classification systems, spatial, temporal and quantitative reasoning, color categorization, and auditory reasoning.

2.1.1 The Sapir-Whorf hypothesis

The Sapir-Whorf hypothesis (SPH) encompasses the central claim of LR—namely, that language actively shapes perceptions and thought processes. Although earlier philosophers such as Humboldt (1836) and Wittgenstein (1922) recognized and discussed potential connections between language and thought—with many reflections on this subject predating even their contributions by millennia, exemplified by Plato's (1997) consideration of whether the Greek concept of "justice" carries nuances distinct from those found in other languages, and continuing across subsequent centuries through metaphors equating second language acquisition to the gain of a second soul, as articulated notably by Charlemagne (Pavlenko, 2014) and embedded in various idioms within diverse cultures—it was Sapir (1921, 1929) and Whorf (1956) who first systematically examined these ideas empirically to a degree, explicitly formulating them through comparative anthropological analyses of linguistic systems.

Scholarly interpretations of Sapir and Whorf's paradigm have diverged into two largely distinct forms: a strong deterministic interpretation, whereby language imposes strict cognitive limitations, and a weaker relativistic stance, under which language guides or biases thought without enforcing rigid constraints (Lucy, 1997).

2.1.2 Linguistic determinism: The strong hypothesis

The deterministic form of the SPH proposes that linguistic categories strictly limit the concepts available to speakers, such that cognition cannot extend beyond linguistic distinctions, thereby establishing the relationship between language and thought as language having an impeding effect on the spectrum of conceivable thought. This strong deterministic view was first articulated by Wilhelm von Humboldt, who argued that language shapes thought by imposing a unique worldview, stating, "Language is the formative organ of thought…the diversity of languages is not a diversity of signs and sounds but a diversity of views of the world…thought and language are therefore one and inseparable from each other" (Humboldt, 1836/1988), thus suggesting that speakers are confined to the cognitive boundaries set by their language's structure. Similarly, Wittgenstein's early philosophical reflection succinctly summarized this deterministic stance: "The limits of language… are the limits of what can be said, and therefore of what can be thought… the limits of my language mean the limits of my world" (Wittgenstein, 1922). However, it was Whorf that first began producing empirical research to substantiate this claim, arguing that languages like Hopi, lacking specific linguistic tense markers for linear temporal relations familiar to European languages, significantly constrain native speakers' conceptions of time (Whorf, 1956).

However, subsequent research has challenged strong determinism by highlighting how bilingual and multilingual speakers effortlessly conceive concepts across linguistic boundaries, demonstrating cognitive flexibility beyond strictly linguistic categories (Pavlenko, 1999). Likewise, evidence of cognitive universals—such as universal capacities for color discrimination—undermine claims that

linguistic structures entirely limit cognition (Berlin & Kay, 1969). Consequently, contemporary theorists (Boroditsky, 2001; Lee, 1996; Lucy 1997) and this thesis contest strong determinism in favor of the weaker, relativistic formulation.

2.1.3 Linguistic relativism: The weak hypothesis

The weak hypothesis of linguistic relativism, often termed the "thinking-for-speaking" perspective, posits that language subtly shapes cognitive processes by guiding attention, perception, and habitual thought patterns without imposing rigid constraints on cognition (Slobin, 1996; Gentner & Goldin-Meadow, 2003). Unlike the strong deterministic interpretation of the SPH, which claims language strictly limits conceivable thought, the weak hypothesis suggests that linguistic structures bias how speakers categorize, perceive, and reason about the world, while preserving universal cognitive capacities. This relativistic view, predominant among cognitive scientists, acknowledges that human cognition is fundamentally similar across linguistic communities, but language-specific structures systematically influence cognitive habits (Levinson, 2003; Lucy, 1997).

The intellectual foundations of the weak hypothesis trace back to anthropological and linguistic scholarship, particularly the work of Franz Boas, who argued that languages reflect culturally specific ways of classifying reality, thereby shaping habitual thought. Boas observed that Inuit languages' lexical distinctions for snow (e.g., terms for wet or falling snow) align with environmentally salient features, reinforcing their cognitive prominence for speakers (Boas, 1911). He asserted that "language is not simply a reproducing instrument for voicing ideas but rather is itself the shaper of ideas", suggesting that linguistic structures guide perception and

memory by emphasizing particular conceptual dimensions (Boas, 1938). The oft-cited “Eskimo snow” example, however, has been critiqued for exaggeration, with Pullum (1991) undermining the claim at its foundation, calling it a hoax and suggesting that Inuit languages have a modest number of distinctions comparable to English compound terms (e.g., “wet snow”), and Krupnik deeming it to be misconstrued as a justification for LR (Krupnik et al., 2010). Nevertheless, Boas’s broader claim that linguistic categories reflect culturally relevant distinctions remain influential in the shaping of the LR argument.

Edward Sapir, a student of Boas, extended this perspective by arguing that language serves as “a guide to social reality”, shaping how speakers conceptualize experience (Sapir, 1921, 1929). Sapir emphasized that grammatical categories and lexical semantics predispose speakers toward specific cognitive orientations. For instance, languages with tense-based versus aspect-based systems may encourage different ways of parsing temporal experience, influencing habitual conceptualizations (Sapir, 1921).

Benjamin Lee Whorf, Sapir’s student, further developed this view through comparative studies of languages like Navajo and Hopi. Whorf argued that Navajo’s verb-focused semantics, prioritizing processes over static entities, foster distinct conceptual predispositions compared to English’s noun-centric frameworks (Whorf, 1956). Whorf’s claim that Hopi’s lack of explicit tense markers leads to cyclical time conceptualization has sparked debate, with critics like Malotki (1983) challenging its extent. Similarly, Slobin (1996) proposed that languages differ in how they encode motion events, prompting distinct attentional patterns when describing movement.

Much research has been done in this field, primary instances of which are detailed in (Section 2.2).

2.1.4 Conflicting perspectives: LR criticism and responses

The claim of LR certainly can't be declared a consensus, but rather quite contentious and speculative. Indeed, LR has, and continues to face criticism from scholars who argue that linguistic influences on cognition are minimal or overstated compared to universal cognitive mechanisms. These critics contend that thought operates largely independently of language, with linguistic effects, if any, confined to superficial or task-specific contexts. They also emphasize cultural, environmental, or innate cognitive factors as primary drivers of cognitive differences, challenging the hypothesis's linguistic claims at its core. Key critics, including Steven Pinker, John McWhorter, Jerry Fodor, among others, have articulated various arguments against LR, supported by evidence from bilingualism, pre-linguistic cognition, and universal cognitive capacities.

Steven Pinker, a prominent advocate of cognitive universalism, argues that thought is fundamentally language-independent, rooted in a universal "mentalese" — a language of thought that operates prior to linguistic expression, with language acting solely as an output mechanism (Pinker, 1994). In "The Language Instinct" (1994) and "The Stuff of Thought" (2007), Pinker contends that LR exaggerates the role of language in shaping cognition, asserting that cognitive processes like categorization, reasoning, and perception are governed by innate, universal mechanisms. He suggests that any observed linguistic effects are superficial, limited to how thoughts are articulated rather than their underlying structure. For example,

Pinker cites cross-linguistic similarities in basic cognitive tasks, such as object recognition, to argue that language merely labels pre-existing concepts rather than shaping them (Pinker, 2007). He dismisses relativistic claims as exaggerations of linguistic diversity, arguing that humans share a common cognitive architecture regardless of their native language.

John McWhorter (2014) extends this critique, emphasizing that cultural and environmental factors, rather than language, drive most cognitive differences across populations. In "The Language Hoax", McWhorter argues that LR overstates language's influence by conflating it with cultural practices. He points to bilingual individuals who seamlessly navigate multiple linguistic frameworks without apparent cognitive disruption, suggesting that cognition remains consistent across languages. He contends that relativistic studies often fail to control for cultural variables, leading to misattributions of cognitive differences to language rather than context or experience.

Jerry Fodor, a staunch critic of LR, explicitly rejected the idea that language shapes thought, rather bluntly expressing his deep disdain for the concept (Fodor, 1985). In "The Language of Thought" (1975) and "The Modularity of Mind" (1983), Fodor argued that cognition operates through a universal, innate language of thought (LOT), a symbolic system independent of natural language. He asserted that mental representations are not contingent on linguistic structures, and thus, relativism—whether strong or weak—is fundamentally misguided. Fodor's modular view of the mind posits that cognitive processes like perception and reasoning are encapsulated, unaffected by linguistic variation (Fodor, 1985). He criticized relativistic claims as empirically weak and philosophically untenable, arguing that linguistic diversity does

not imply cognitive diversity. His rejection of LR presented it as undermining the universality of human cognition and unnecessarily complicating CS.

Empirical evidence supporting these critiques often draws on cases of pre-linguistic or non-linguistic cognition. Goldin-Meadow and Feldman (1977) studied deaf children who developed "home sign systems"—gestural communication without formal linguistic input—demonstrating complex categorization and problem-solving abilities. These findings suggest that cognitive capacities like conceptual organization precede and function independently of language, undermining claims of linguistic influence (Brentari et al., 2012; Senghas & Coppola, 2001). Similarly, studies of pre-linguistic infants show that they discriminate colors and spatial relations before acquiring language, indicating that core cognitive processes are innate and universal (Werker & Tees, 1984). Critics argue that such evidence weakens the weak hypothesis, as it demonstrates that thought can operate robustly without linguistic scaffolding.

Despite these critiques, the weak hypothesis provides compelling empirical evidence suggestive of the authenticity of its claims as is detailed in the next section. Furthermore, modern approaches compatible with interdisciplinary perspectives, including computational approaches using LLMs to model language-specific cognitive biases (Dubossarsky et al., 2020; Teehan et al., 2024), as implemented in this study, provide a new perspective into the realm of LR.

2.2 Empirical investigations into linguistic relativity

Empirical research from the fields of CS, psycholinguistics, and anthropology provide supporting evidence, indicative of the LR claim, providing a foundation to

test our hypothesis that linguistic properties and lexical distinctions influence on thought processes may be quantifiable in a computational environment. These studies collectively strengthen the claim for the weaker, relativistic version of the SPH, termed LR, demonstrating that language guides cognitive habits rather than imposing insurmountable constraints upon thought. To evaluate and contextualize these claims rigorously, the current section synthesizes and critically analyzes empirical findings across the following primary cognitive domains: lexical semantics, grammatical structure, grammatical gender and natural classification systems, spatial, temporal, and quantitative reasoning, color categorization, and auditory reasoning.

2.2.1 Cross-linguistic grammatical discrepancies

Empirical studies consistently demonstrate that grammatical systems—including tense markers, evidentiality encoding, verb aspect, grammatical voice, case distinctions, and verb encoding patterns—compel speakers to habitually direct their attention toward specific informational aspects of events, thereby influencing cognitive conceptualization, attention, and recall. These findings support our hypothesis that obligatory and language-specific grammatical structures condition cognitive preferences, influencing how speakers perceive and represent events.

Zhang et al.'s (2018) investigation into tense marking reveals how obligatory grammatical structures influence temporal cognition, providing evidence for LR. English, which requires explicit tense encoding in verbs with morphological markers (e.g., "walk" for present, "walked" for past), heightens speakers' sensitivity to temporal sequencing. In contrast, Mandarin, a tenseless language, relies on optional time adverbials (e.g., "zuó tiān" for yesterday) or context, leading to less focus on

strict chronology. In tasks testing temporal order violations in English sentences, English speakers showed robust neural responses (N400 modulation) to reference time misalignments, indicating strong temporal processing driven by tense. Mandarin-English bilinguals, however, exhibited weaker N400 responses, suggesting reduced sensitivity to tense-based temporal cues, even with high English proficiency. This discrepancy, observed despite the presence of temporal adverbials like “after”, underscores that English’s tense system enhances temporal cognition compared to Mandarin’s adverbial approach, supporting the notion that language shapes thought.

Boroditsky’s (2001) study on temporal cognition illustrates how language-specific spatial metaphors influence how speakers conceptualize time, providing empirical support for LR. English predominantly employs horizontal spatial metaphors for time, such as “the good times ahead of us” or “the hardships behind us”, which align with its use of front/back terms to describe temporal sequences. In contrast, Mandarin systematically uses both horizontal and vertical metaphors, with terms like “shàng” (up) for earlier events and “xià” (down) for later events. This linguistic difference is claimed to shape distinct cognitive biases: English speakers tend to think of time horizontally, while Mandarin speakers incorporate vertical orientations. To test this, Boroditsky employed spatial priming tasks where participants viewed images arranged horizontally or vertically before answering questions about temporal relationships, such as “Does March come before April?” The results revealed that English speakers responded faster and more accurately when primed with horizontal spatial cues, reflecting their language’s linear temporal framework. Mandarin speakers, however, showed a split pattern: horizontal primes facilitated responses to certain temporal questions, while vertical primes enhanced

others, mirroring their language's dual metaphorical system. Additionally, when English speakers were briefly trained to use vertical metaphors for time, their cognitive responses shifted to resemble those of Mandarin speakers, demonstrating that linguistic input can directly alter thought patterns. These findings suggest that the habitual use of spatial metaphors in language influences how speakers perceive and process abstract domains like time, thereby supporting LR's premise that language influences cognition. However, these claims are not without being challenged, as other researchers reports having failed to replicate Boroditsky's findings, expressing that their work offered no logical ground for Boroditsky's claim and were also unable to replicate her experiments in four different attempts. They concluded that Chinese speakers do not think about time in a different way than English speakers just because Chinese also uses the vertical spatial metaphors to express time (Chen, 2007).

Extending this perspective, Ünal and Papafragou (2018) demonstrate that evidential markers, such as the Turkish suffix "-miş" (indicating indirect or inferential information), shape attentional and memory-related processes by compelling speakers to routinely attend to the source of information. Similarly, Slobin's (1996) "Thinking for Speaking" hypothesis posits that language-specific grammatical requirements adapt thought processes to linguistic demands. For example, Turkish's obligatory evidential distinctions (e.g., "kaçı-yor-du" [was running, witnessed (obligatory detail)] vs. "düş-müş" [fell, non-witnessed (obligatory detail)]) force speakers to focus on the epistemic status of events, unlike English's optional evidential markers (e.g., "The child apparently (optional detail) fell"). This grammatical obligation leads to cross-linguistic differences in event

conceptualization, as Turkish speakers habitually encode and recall the source of evidence.

Grammatical voice also systematically influences perceptions of agency and causality. Fausey and Boroditsky (2010, 2011) found that English speakers, who typically employ active constructions (e.g., “I broke the vase”), are more inclined to identify agents in accidental events and attribute higher levels of blame. In contrast, Spanish speakers, who frequently use passive or agentless constructions (e.g., “Se rompió el jarrón” [The vase broke (itself)]), focus on the accidental nature of events and are less likely to assign personal responsibility. This linguistic difference results in systematic variations in blame attribution and actor recall, reflecting how grammatical structure modulates thought processes related to memory, judgment, and causal attribution. Filipović (2013) corroborates these findings, showing that Serbian speakers’ frequent use of passive voice (e.g., “Vaza se razbila” [The vase got broken]) obscures agents, affecting responsibility judgments in legal contexts. In experiments, Serbian speakers prioritized event outcomes (e.g., a broken window) over agents (e.g., “John”), assigning less blame in mock legal scenarios compared to English speakers’ agent-centric active-voice judgments (e.g., “John broke the window”). These linguistic tendencies have practical implications in multilingual legal settings, where passive-to-active translations may inflate perceived blame.

Further illustrating the cognitive impact of grammatical structures, Talmy’s (2000) motion event typology distinguishes verb-framed languages (e.g., Spanish, encoding path in verbs like “entrar corriendo” [enter running]) from satellite-framed languages (e.g., English, encoding manner in verbs like “run into”). These differences direct speakers’ attention to path or manner, respectively, influencing

cognitive construal and event recall. For instance, Spanish speakers prioritize the trajectory of motion (path), while English speakers focus on the style of motion (manner), leading to divergent mental representations of the same event. This grammatical variation shapes how speakers encode and retrieve motion events.

Cognitive linguistics, as advanced by scholars like Lakoff and Johnson (1980), complements these findings by highlighting how lexical semantics, through metaphors, interact with grammatical structures to shape thought. Their conceptual metaphor theory argues that metaphors, rooted in embodied experience, structure abstract conceptualization. For example, the metaphor "time is money" encourages English speakers to view time as a resource, influencing decision-making and prioritization, while "I'm feeling up" links spatial orientations to emotions (e.g., "happy is up"), fostering culturally distinct cognitive patterns. These metaphorical structures, embedded in language-specific lexical and grammatical systems, along with the other grammatical structures mentioned, reinforce the role of grammar in directing cognitive attention and shaping conceptual frameworks across diverse linguistic communities.

Furthermore, a magnetoencephalography (MEG) study by Coopmans and Martin (2024) reveals language-specific differences in syntactic processing during sentence comprehension. Analyzing Dutch listeners engaged with natural stories, the study compared three models of temporal syntactic parsing and found a highly predictive top-down approach, where the brain anticipates syntactic structures several words ahead of the speech signal, representing Dutch comprehension. This contrasts with English, where prior research typically supports a mildly predictive left-corner parsing strategy. These findings suggest that, even when syntactic

structures are similar, the timing of their neural projection differs across languages, with Dutch grammar fostering earlier structural anticipation. This aligns with the LR's exploration of how grammatical systems shape cognitive habits, as it indicates that Dutch speakers' predictive parsing biases attention and working memory deployment during real-time comprehension, consistent with LR's predictions of low-level processing differences.

2.2.2 Grammatical gender and conceptual representations

The presence of grammatical gender in languages such as Arabic, French or German imposes categorical frameworks that may influence speakers' conceptualizations and descriptions of objects and abstract concepts, which, if empirically verified, would have supporting implications for LR.

Konishi (1993) investigated this phenomenon in a cross-cultural study, testing whether grammatical gender carries connotative meanings of femininity and masculinity. In his experiment, 40 native German speakers and 40 native Spanish speakers rated 54 high-frequency translation-equivalent nouns on semantic differential scales, measuring evaluation, potency, and activity. These nouns were selected such that half were feminine in German but masculine in Spanish (Type I words, e.g., Luft [feminine, German] vs. aire [masculine, Spanish]), and half were masculine in German but feminine in Spanish (Type II words, e.g., Schlüssel [masculine, German] vs. llave [feminine, Spanish]). Konishi hypothesized that German speakers would rate Type II words higher in potency (associated with masculine traits like strength) and Type I words lower, while Spanish speakers would show the opposite pattern. His findings supported this, revealing a significant

language-by-word-type interaction for potency ($F(1, 76) = 8.51$, $p < .01$), with German speakers rating Type II words ($M = 4.32$) higher than Type I words ($M = 4.22$), and Spanish speakers rating Type I words ($M = 4.21$) higher than Type II words ($M=4.11$). This suggested that grammatical gender influences the perceived potency of inanimate nouns.

Boroditsky et al. (2003) further explored this effect, focusing on cognitive carryover in a non-gendered language. They tested Spanish and German speakers describing objects in English, which lacks grammatical gender. Their results indicated that native language gender assignments influenced descriptions. For example, the noun bridge (masculine el puente in Spanish, feminine die Brücke in German) elicited descriptions from Spanish speakers like “strong” and “sturdy”, reflecting masculine stereotypes, while German speakers used “beautiful” and “elegant”, aligning with feminine stereotypes. Similarly, key (masculine el llave in Spanish, feminine die Schlüssel in German) prompted Spanish speakers to use adjectives like “hard” and “heavy”, while German speakers favored “intricate” and “lovely”. These divergent descriptions in a common medium void of such categorization (i.e. English) underscore how grammatical gender may induce cognitive biases in object conceptualization.

This effect extends to abstract concepts. In French, freedom (feminine la liberté) is often personified as female, as seen in the Statue of Liberty, a French gift symbolizing liberty as a female figure. Boroditsky et al. (2003) noted that German speakers described death (masculine der Tod) with male-associated traits like “grim” and “harsh”, while Spanish speakers, for whom death is feminine (la muerte), used terms like “pale” and “sorrowful”. Similarly, French speakers described love

(masculine l'amour) as "passionate" and "intense", reflecting masculine stereotypes, while Italian speakers portrayed peace (feminine la pace) as "calm" and "gentle".

However, the claims by Konishi (1993) and Boroditsky et al. (2003) that grammatical gender shapes conceptual representations were not without criticism and faced significant scrutiny due to challenges in replicability and methodological robustness. Mickan et al. (2014), in their study "Key is a llave is a Schlüssel: A failure to replicate an experiment from Boroditsky et al. 2003," contested the findings of the above mentioned study. Their study involved 60 participants (30 German, 30 Spanish) rating 24 nouns on adjective scales associated with masculinity or femininity. Unlike Boroditsky et al., they found no evidence of grammatical gender influencing adjective choices, suggesting the original results may have been a statistical artifact or dependent on specific conditions.

Similarly, Elpers et al. (2022) conducted a replication of Boroditsky et al.'s study. Their experiment, involving 120 bilingual German-Spanish speakers tested in English to neutralize language-specific effects, examined similarity judgments for noun pairs and found no grammatical gender effects. Their findings highlighted issues like small sample sizes and potential confounds in stimuli selection in the original studies.

Additionally, Mahajan and Durgin (2025), in a forthcoming study, tested Boroditsky et al.'s claims across French and Italian speakers using implicit association tests and advanced statistical modeling with larger, diverse participant pools. Their results, controlling for cultural stereotypes, also failed to detect consistent gender-based conceptual effects, further questioning the generalizability of the original findings. These replication failures suggest that Konishi's (1993)

findings, while statistically significant, may also be limited by methodological factors, such as a limited sample size and lack of scrutiny in their methodology that may produce false positives.

However, there have been other works conducted that have independently verified the claims provided by Konishi and Boroditsky (among others). Similarly, Sato and Athanasopoulos (2018) presented evidence that grammatical gender can influence perceptual judgments in bilinguals. In their study, French-English bilinguals and native English monolinguals were consecutively presented with images of objects that varied in their conceptual and grammatical gender associations, followed by a target face on which they had to make a gender-related judgment. The results showed that when the association between the prime and the target was implicit, only the bilinguals' judgments were modulated by the grammatical gender of the objects, whereas monolinguals relied exclusively on conceptual gender. This suggests that grammatical gender provides a structural basis that can unconsciously bias perceptual processes, supporting the idea that linguistic categories shape cognitive representations in ways consistent with LR.

The back and forth debates regarding the implications of grammatical gender in the context of LR continue extensively, but the potential implications of grammatical gender—if verified— have profound effects. Languages with grammatical gender categorize their entire lexicon—encompassing tangible objects and intangible concepts—within a gendered framework, potentially influencing how speakers perceive and interact with their world. If validated, such effects would support LR's claim that linguistic structures shape conceptual representations. To address these inconsistencies in various researches and provide a novel approach to

the debate, our present study employs computational semantic embeddings and bilingual categorization tasks to probe whether gender-driven biases persist in a common second language among a generous variety of bilingual language pairs. By modeling semantic associations and designing tasks to test implicit gender biases in noun categorization, this study aims to provide an empirical foundation for assessing the impact of grammatical gender on thought. Grounded in LLM's statistical representation large sample size and quantifiable distances between a high magnitude of nouns and adjective pairs, these methods seek to overcome the methodological limitations of earlier studies and clarify whether linguistic gender systems influence conceptual thought.

2.2.3 Natural classification systems

Research on natural noun classification systems also offers valuable insight into LR. Lakoff's (1987) influential research into Australia's Dyirbal language, which classifies nouns into abstract conceptual clusters (notably grouping "women, fire and dangerous things" and "men, kangaroos and boomerangs" into two distinct categories), demonstrates linguistic-cognitive interface effects. Dyirbal speakers were shown to respond faster and more robustly to nouns sharing linguistic classifications reflecting conceptual links (e.g., danger, animacy), such as recalling "fire" and "scorpion" as related due to their shared "dangerous" category, unlike English speakers, who relied on taxonomic or functional groupings.

Comparable outcomes emerged from Young and Morgan's (1987) research on Navajo, which employs a classifier system based on shape and animacy to categorize nouns. Navajo verbs of handling, for instance, require classifiers that specify whether

an object is long and rigid (e.g., a stick, classified with the "-łéėch"), round and compact (e.g., a ball, "-łéėł"), or animate (e.g., a person, "-łéí"). Young and Morgan conducted memory and categorization experiments, finding that Navajo speakers organized objects according to these linguistic classifiers rather than functional or taxonomic criteria. For example, when asked to group items like a stick, a rope, and a basket, Navajo speakers consistently paired the stick and rope as "long and flexible" items, reflecting the classifier "-łéėch", while English speakers grouped the basket and rope based on their functional use (e.g., as containers or tools). This research demonstrates that Navajo's classifier system influences thought processes by directing attention to shape and animacy, structuring cognitive organization and categorization, and enhancing memory for linguistically salient features. These effects contrasted with English speakers' functional strategies, highlighting language-driven cognitive variability.

Research by Lakoff (1987) and Lucy (1992) further underscores how linguistic classifications—including gender and classifier systems—systematically influence cognition. Lucy's work with Yucatec Maya, which employs classifiers emphasizing material over shape (e.g., distinguishing "wooden" vs. "clay" objects), revealed that speakers prioritized substance-based organization. In sorting tasks, Yucatec speakers grouped items like a wooden spoon and a wooden chair together due to their shared material, while English speakers grouped the spoon with a metal fork based on function (utensils). These findings from Dyirbal, Navajo, and Yucatec Maya illustrate that natural noun classification systems influence cognition, guiding categorization, memory, and attention toward linguistically salient features while emphasizing variability in thought across languages.

2.2.4 Spatial reasoning

Spatial reasoning exemplifies LR's cognitive impact. Levinson (2003) and Majid et al. (2004) show that geocentric languages with absolute spatial frameworks (e.g., Guugu Yimithirr's cardinal directions) enhance orientation and navigation compared to egocentric systems with relevant terms (e.g., English's left-right). Geocentric language speakers outperform egocentric speakers in recall and navigation tasks, with effects attributed to linguistic encoding, not biology (Haviland, 1993; Boroditsky & Gaby, 2010), supporting the thesis's hypothesis of language-driven cognitive variability.

Levinson (2003) studied Guugu Yimithirr, which uses cardinal directions (e.g., "north") instead of egocentric terms. In a rotation task, participants reconstructed object arrays after a 180-degree turn. Guugu Yimithirr speakers maintained absolute orientations (e.g., "north"), achieving 90–95% accuracy, while English/Dutch speakers, using egocentric strategies, scored 50–60%. In a point-to-home task, Guugu Yimithirr speakers pointed to distant locations with 5–10 degree precision, vs. English speakers' 20–30 degree errors. Majid et al. (2004) compared Tzeltal (geocentric) and Dutch (egocentric) speakers in a memory task, where Tzeltal speakers recalled arrow orientations (e.g., "uphill") with 85–90% accuracy, vs. Dutch speakers' 60%. Tzeltal speakers also excelled in spatial inference (e.g., "tree east of rock"), with 80–85% accuracy vs. Dutch speakers' 55–65%.

Haviland (1993) observed Guugu Yimithirr speakers using cardinal directions in speech and gestures (e.g., "move chair east"), with pointing accuracy within 10 degrees, unlike English speakers' body-relative gestures. Boroditsky and Gaby (2010) found Pormpuraaw speakers arranged temporal sequences east-to-west and

navigated with 5–10 degree accuracy, vs. English speakers' 25–35 degree errors, with linguistic encoding confirmed as the driver. For example, a Guugu Yimithirr speaker maps a river as "west" maintaining orientation, while an English speaker's "behind me" risks disorientation. These findings show absolute frameworks foster a "mental compass", enhancing spatial cognition, in support of LR.

2.2.5 Temporal reasoning

Temporal cognition research has been conducted in support of LR by showing that linguistic metaphors guide time conceptualization with findings—vertical, horizontal, geocentric—shape temporal cognition. An instance of this effect was mentioned in (Section 2.2.1) with regards to grammatical categorization having implications for temporal reasoning, viewing time as vertical or horizontal based on different languages (Boroditsky, 2001).

In further studies, Boroditsky et al. (2011) showed writing direction shapes sequencing, with English (left-to-right) and Arabic (right-to-left) speakers arranging events accordingly. The study found English speakers arranged aging sequences left-to-right (90–95% consistency), Arabic speakers right-to-left (85–90%), with reaction times faster (~100 ms) when primed with respective directions.

Kuuk Thaayorre's geocentric system, linking time to cardinal directions, leads speakers to order events east-to-west, landscape-based, not body-relative (Boroditsky & Gaby, 2010). The study showed Kuuk Thaayorre speakers arranged sequences (e.g., ripening banana) east-to-west, regardless of body orientation (85–90% consistency), and pointed to "yesterday" eastward with 5–10 degree precision, unlike English speakers' body-relative gestures.

2.2.6 Quantitative reasoning

Different languages have different numerical systems, studies pertaining to these distinct counting systems highlight LR's claim within the scope of quantitative reasoning. Everett (2005) found Pirahã speakers, with terms "hói" (one), "hoí" (two), "baagi" (many), struggled with exact numerical tasks beyond two (e.g., matching objects), unlike English speakers. In a matching task, participants aligned objects with one another, achieving 95–100% accuracy for one or two but 50–60% for three or more (e.g., four objects). In a replication task, they approximated quantities like five nuts as "many", unlike English speakers' scoring 90–95% accuracy at numbering objects beyond five.

Pica et al. (2004) showed Munduruku speakers, with numerals up to five ("pũg" to "pũg põgbi" for one–five), had poor accuracy on tasks beyond this range, relying on approximation. For example, Munduruku speakers inconsistently compared six vs. seven items, unlike English speakers' precision. In comparison tasks, they scored 80–90% up to five but 50–55% for six vs. seven dots. Addition (e.g. 4+2) and subtraction beyond five were approximated, unlike English speakers' exactness. These results show limited numeral systems restrict exact quantification.

2.2.7 Color categorization

Color categorization research in support of LR demonstrates that language-specific color lexicons drive differences in perceptual discrimination, cognitive categorization, and memory performance. Comparative linguistic anthropology research has revealed that color lexicons—differentiating fine-grained spectral distinctions in one language yet coalescing them broadly in another—notably

influence how speakers perceive and categorize colors, as well as their memory performance across diverse linguistic backgrounds (Regier & Kay, 2009). These findings, detailed further below, underscore the role of linguistic categories in shaping cognitive processes related to color perception.

Winawer et al. (2007) found that Russian speakers, who distinguish "goluboy" (light blue) and "siniy" (dark blue) as two separate colors as opposed to English, which uses a single color term ("blue") encompassing both Russian color terms, outperformed English speakers, in tasks requiring rapid discrimination of blue shades. Their research tested categorical perception (CP) using a visual discrimination task where participants matched blue squares in CIELAB space. Russian speakers were faster (by ~100–150 ms) at distinguishing cross-category shades (goluboy vs. siniy) than within-category pairs, unlike English speakers, who showed no boundary effect. This suggests that Russian's lexical distinctions enhance perceptual sensitivity.

Thierry et al. (2009) provided neurological evidence using event-related potentials (ERPs) with Greek and English speakers, focusing on Greek's light blue ("ghalazio") and dark blue ("ble") distinction. In an oddball task, Greek speakers showed stronger visual mismatch negativity (vMMN) at 150–250 ms in the visual cortex (V1/V2) for cross-category blues (ghalazio-ble) compared to within-category pairs, unlike English speakers, indicating pre-attentive categorical shifts driven by language-specific categories for the relevant terms.

Roberson et al. (2000) identified similar effects in their study of the Himba of Namibia, whose language has five color terms, such as "burou" (covering green and blue) and "serandu" (red, orange, pink), unlike English's eleven terms (Berlin et al.,

1969). Using Munsell color stimuli, they tested categorical perception (CP) through similarity judgments, memory tasks, and learning tasks. In similarity judgments, Himba participants judged within-category colors (e.g., two burou greens) as more similar than cross-category pairs (e.g., burou green vs. serandu red), despite equal perceptual distances, unlike English speakers, who showed CP at the blue-green boundary. In memory tasks (two-alternative forced-choice), Himba speakers discriminated cross-category pairs (e.g., burou vs. zoozu [dark colors]) faster and more accurately than within-category pairs, while English speakers excelled at blue-green distinctions. In learning tasks, Himba participants learned color-cattle associations faster for cross-category pairs, unlike English speakers. These findings indicate that Himba linguistic categories enhance discrimination at language-specific boundaries.

2.2.8 Auditory reasoning

Auditory reasoning research highlights how tonal systems in languages like Mandarin and Cantonese shape sound perception and emotional prosody processing. In tonal languages, lexical tones—such as Mandarin's high-level Tone 1 or low-dipping Tone 3—differentiate word meanings (e.g., "mā" [mother] vs. "mǎ" [horse]) and influence how speakers encode and interpret emotional cues in speech.

Thompson and Balkwill (2006) investigated how English-speaking listeners decode emotional prosody across five languages: English, German, Chinese, Japanese, and Tagalog. Their findings revealed that listeners identified emotions—joy, anger, sadness, and fear—from prosodic cues in all languages at rates above chance, pointing to universal elements in emotional expression. However, accuracy

varied: English utterances achieved 94.4% recognition, while Chinese and Japanese scored lower (59.4% and 54.1%, respectively), reflecting an in-group advantage tied to linguistic familiarity. Sadness and anger were recognized more accurately than joy and fear across all languages, with Chinese sadness reaching approximately 80% accuracy, while joy and fear hovered around 40% and 50%. Acoustic analyses linked emotions to prosodic features like pitch (F0) and intensity. For Mandarin, joy and anger featured higher mean F0 (252.09 Hz and 259.91 Hz) compared to sadness (159.96 Hz), suggesting pitch plays a role, though tempered by lexical tone constraints.

Liu and Pell (2014) examined emotional prosody in Mandarin, using pseudo-utterances to test native speakers' recognition of seven emotions: anger, disgust, fear, sadness, happiness, pleasant surprise, and neutrality. They found neutrality, anger, sadness, and fear were most accurately recognized, with acoustic parameters such as F0 mean, F0 range, and speech rate driving distinctions. Compared to English, Arabic, German, and Hindi (from Pell et al., 2009), Mandarin showed similar yet distinct patterns, indicating both universal and language-specific traits. The study suggests that Mandarin's tonal system restricts pitch variation for emotional prosody, pushing speakers to rely on cues like speech rate to differentiate emotions, though pitch remains relevant for lexical purposes.

These findings have implications for LR in auditory reasoning. The tonal structure of Mandarin influences how emotional prosody is processed, constraining pitch use and amplifying non-pitch cues compared to English. However, universal prosodic cues—evident in cross-linguistic recognition—temper this effect, allowing English listeners to decode Mandarin emotions above chance. Thus, while tonal

systems influence auditory emotional categorization, they do not fully override cross-linguistic intelligibility, indicating LR operates alongside universal mechanisms.

CHAPTER 3

LINGUISTIC REPRESENTATIONS

3.1 Introduction to computational framework

This chapter formally introduces and details the CL framework previously alluded to throughout the previous chapters. The CL framework for this research is designed to explore LR by quantifying cross-linguistic conceptual differences using LLMs. It builds semantic representations of different languages and maps them to a common field where they may undergo measurements to assess their proximity to one another and reveal how languages encode meaning differently, as a manifestation of the thought processes that result in specific linguistic outputs; thereby facilitating an empirical method as to measure LR's claims. Unlike traditional LR experiments, riddled with logistical limitations, often relying on small samples and subjective methods, this approach offers a significant advantage with scalable, data-driven AI and NLP techniques. In contrast to previous approaches, this method provides a novel, replicable way to study LR, sidestepping the aforementioned challenges of human participant related research while enhancing empirical rigor and addressing the persisting debates regarding the results of empirical assessments of LR (as mentioned in Section 1.3.2) with unprecedented dataset sizes, constituting the motivation for this approach. By leveraging LLMs, we generate large-scale, consistent data that simulates bilingual language production, allowing us to isolate the influence of the native language on conceptual representations.

The framework employs LLMs by prompt engineering them to simulate bilingual speakers, generating adjective-based semantic representations for nouns

across 24 languages, specifically selected to represent a diverse range of language families and typological features to produce a wide array of linguistic variation that may be tested cross-linguistically. These are converted into high-dimensional vector embeddings compatible with capturing multilingual distinctions, aggregated into language-specific constructs, and compared using cosine similarity as a distance metric, which effectively measures the angular difference between vectors, corresponding to semantic similarity in high-dimensional embedding spaces to assess conceptual divergence. These distances act as proxies for L1-driven cognitive biases, later tested and validated through cognitive experiments (Chapter 4).

3.2 Semantic field construction

This section outlines the methodology for constructing the semantic field using LLMs to investigate LR. The approach focuses on selecting a diverse lexicon and multiple languages to capture conceptual and typological variation, enabling rich cross-linguistic comparisons. Meticulous prompt engineering and data generation processes work to ensure the production of accurate bilingual model representations.

To construct the semantic field, adjectives are generated to build bilingual models, and their embedding values are obtained through a structured process. First, we select a diverse set of languages and nouns, covering various typological features and a spectrum of conceptual categories. For each language and noun pair, we instruct LLMs by prompt engineering them to generate adjectives in the English language that describe the noun from the perspective of a bilingual speaker with a specific native language and proficiency level in English (as the medium for the experiment). These adjectives reflect how native speakers of those languages think of

that noun with cognitive biases carrying over from their native language to the common medium of English. The proficiency scale exists to assess whether cognitive biases from L1—the native language—carry over more or less in correlation with the medium for experimentation (English). That is, it serves to test if the proficiency level of the acquired L2 (English) possibly limits the effect that the native language has on its speakers.

Next, we obtain embedding values for these adjectives using a multilingual sentence transformer model, converting each adjective into a high-dimensional vector that captures its semantic meaning. These embeddings enable us to quantify conceptual differences across languages, forming the foundation of our bilingual semantic field for subsequent analysis. Then these noun representations based off of elicited adjectives are compounded together to represent each language as a single entity.

3.2.1 Language selection

A total of 24 languages across eight major language families were selected for this study (Appendix A). These languages were chosen to maximize typological diversity across morphological, orthographic, phonological, semantic, and syntactic dimensions, as well as encompassing a wide range of cultural nuances, embedded within languages. This selection also ensures representation from diverse language families (Appendix A), allowing for a comprehensive examination of how varied linguistic structures may influence cognitive processes and conceptual representations and how language families cluster in respect to LR's claims.

3.2.2 Language properties and linguistic structures

The selected languages exhibit a wide range of linguistic properties and structures, providing important consideration for investigating their potential influence on semantic encoding and cognitive biases. An overview of the varying properties and features are as follows, with detailed languages features found in (Appendix D):

The various grammatical features include agreement, aspect, case, clitics, evidentiality, grammatical gender (Appendix B), mood, and tense. Morphological typology spans agglutinative, fusional, templatic, and isolating systems. Orthographies cover Latin and Cyrillic alphabets, Arabic and Hebrew abjads, Devanagari and Thai abugidas, Hangul, Kana/Kanji, and Chinese characters. Phonological systems range from tonal vs. non-tonal contrasts to stress-timed, syllable-timed, and pitch-accent rhythms. Semantic typology encompasses differences in color term inventories, presence or absence of numeral classifiers, varied expressions of spatial relations, and distinctions between alienable and inalienable possession. Syntactic typology includes a mix of word orders (SOV, SVO, VSO, flexible), head-initial vs. head-final configurations, and pro-drop vs. non-pro-drop patterns. Additional features such as consonant mutations, reduplication, definiteness marking, and complex politeness systems further enrich the typological profile of the sample of languages, making them an ideal choice to probe for potential distinctions in thought processes.

3.2.3 Noun categories and lexical items

The noun selection process is central to our investigation into LR, since these nouns will be the source for generating adjectives that will be the building blocks of the

bilingual representation LMs. Hence, they are chosen in such a manner that they represent a comprehensive survey of meanings and senses throughout an exhaustive range of conceptual categories, in order to accurately portray various aspects of languages. We curated a set of 144 nouns, organized into 12 distinct conceptual categories with 12 lexical items in each (detailed in Appendix C), to provide a diverse foundation for semantic analysis across languages.

By including categories such as abstract concepts (e.g., “freedom”, “justice”), concrete objects (e.g., “table”, “chair”), and others like spatial and temporal terms, gender neutral animal titles, and cultural artifacts, we aim to test LR’s hypothesis that language influences thought in varying magnitudes depending on the nature of the concept. Abstract nouns, tied to cultural and experiential nuances, may reveal greater cross-linguistic variation, while concrete nouns, grounded in shared physical realities, might exhibit more consistency. This diversity enables us to explore whether LR effects are uniform or domain-specific, addressing a core question of our thesis: do different linguistic properties affect cognitive encoding in different magnitudes?

3.2.4 LLMs

The LLMs that were used in this project for generating adjectives to build the semantic map are various state of the art commercial LLMs to both ensure the accuracy of the linguistic representations we have built by cross checking between them, and use them as different sample batches of data input. They are as follows:

- Google’s Gemini-2.0 Flash: A high-performance model, providing reliable adjective outputs across diverse linguistic contexts.

- OpenAI's GPT-4.1: A powerful model, excelling in precise semantic associations, ideal for capturing nuanced adjective-noun relationships.
- OpenAI's GPT-o3-mini: A reasoning (chain of thought) model, enhancing the generation of contextually appropriate adjectives by leveraging logical inference, beneficial for simulating bilingual cognitive processes.
- xAI's Grok-3: Trained extensively on diverse datasets, including social media content from X, this model excels at capturing socially and culturally salient linguistic cues, adept at detecting L1-specific conceptual biases.
- xAI's Grok-3-mini: A reasoning (chain of thought) model, offering efficient adjective generation with a focus on semantic consistency, as an alternative to OpenAI's reasoning model.

3.2.5 Prompt engineering and data generation

Prompt engineering is an important methodological component of our computational framework, designed to elicit adjective-based semantic fields from LLMs. By carefully crafting prompts, we configure LLMs to simulate bilingual speakers, capturing how their native language (L1) influences semantic representations when producing language in a second language (L2: English) at proficiency levels (A1–C2 and "Perfect"). We adopt a two-tiered prompt engineering strategy, utilizing system prompts and user prompts, to generate high-quality adjective data that reflect L1-driven cognitive and linguistic biases.

A system prompt is a high-level instruction provided to the LLM at the start of a session, defining its role, context, and operational constraints. In our framework, the system prompt configures the LLM as a bilingual speaker with a specific L1and a

defined L2 English proficiency level (A1–C2 and ‘Perfect’). This prompt sets the LLM’s linguistic identity, constraining its vocabulary and syntactic complexity to mirror realistic L2 production. By varying proficiency levels, we simulate a range of bilingual cognitive profiles, capturing the degree to which L1 influences persist across different stages of L2 mastery. The system prompt is applied once per session, establishing a consistent context for all subsequent interactions. Two separate system prompts were devised to ensure rigorous testing:

- “Implicit bilingual LLM”: The system prompt for this system prompt reads: “You are a native {language} speaker who has acquired English at the level of {proficiency}.” This system prompt serves to simulate a representation of a bilingual speaker implicitly, without directly activating their native language, as a way to replicate subconscious transfer of L1 thought processes to L2 production.
- “Explicit bilingual LLM”: The system prompt for this system prompt reads: “You are a native {language} speaker who has acquired English at the level of {proficiency}. Think in your native language when responding to the prompt.” In contrast to the former approach, this method serves to explicitly configure the model to produce responses that pronounce the effect of language on its ‘thought’ processes.

A user prompt is a specific, task-oriented instruction provided to the LLM for each noun in our dataset, designed to elicit adjective responses in English by explicitly prompting “generate {specific_variation} adjectives for the provided noun in English”, where ‘specific_variation’ alludes to diverse adjective sets, requesting the “most frequent”, “most common”, “best”, “top”, and “first” adjectives for the provided noun. We iterate multiple variations of the user prompt to capture adjectives

associated with each of the 144 nouns. Specific focus is placed on the “most frequent” prompt for adjectives, as it ensures outputs that reflect L1 semantic preferences rather than rare associations that may be less relevant to our research purposes.

This operation is repeated in three different LLM temperatures. Temperatures are a hyper-parameter in LMs that allow to control the randomness of text generation. LMs with attention mechanism based architecture are trained on large amounts of text which they then use to establish probabilistic relations between words in a non-linear manner. Once an input in the form of a prompt is provided to the LM, it will sample over that prompt within its context window to decide which token is to be generated next, generating a probability distribution for likely tokens. Temperature plays the role of the denominator in the soft-max function that allows the controlling of the randomness of the token to be selected from that probability distribution. The temperature scale tends to range from 0 to 2 in commercial LLMs, with a temperature of 0 generating outputs that are entirely deterministic and based completely on the statistical co-occurrence rates of the dataset the model was trained on, meaning that in our case, the adjectives that co-occur most frequently with a given noun in the datasets that the model was trained on are reported, perfect for describing how a noun is realized per language. A temperature of 2 generates outputs by choosing with greater randomness from the distribution probabilities of the next tokens that can possibly be generated (without being confined to choosing higher probabilities). Commercial LLM providers tend to recommend a temperature of 1 to generate outputs that strike there balance between accuracy and natural human-like text. In our project, we generated all adjectives in each of these three temperatures (0,

1 and 2), with the temperature of 0 being especially useful, as when paired with a prompt like “provide the most frequent adjectives that describe {noun}”, it effectively acts as an enormous deterministic dataset for which languages describe which noun in which manner. The other temperatures exist as useful variations for cross checking to ensure the produced adjectives capture a broad semantic range.

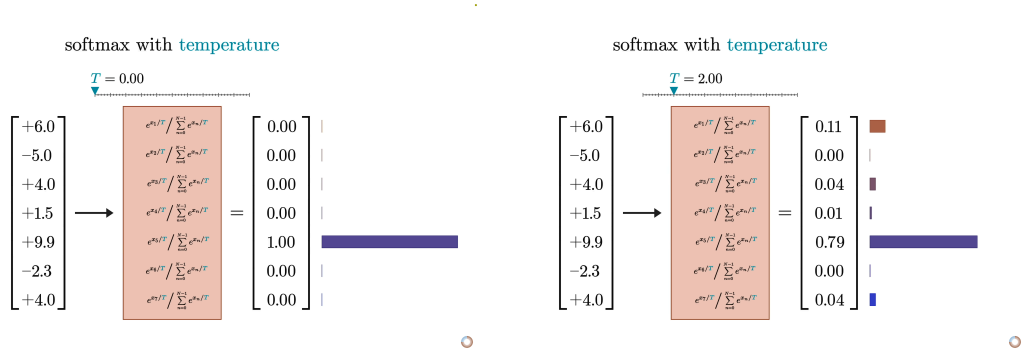


Fig. 1 LM temperature configuration (Sanderson, 2024)

Calibration tests further validate generated adjective tokens, ensuring outputs align with native-like usage and avoid artifacts like oversimplified or generic responses. Initial prompts were tested on a subset of nouns across multiple LLMs to identify issues like generic or oversimplified responses (e.g., “big”, “small”, “good”) or overly constrained outputs (confined to C2 vocabulary, even when lower level lexicon is more suitable to describe the noun) due to rigid proficiency settings. This calibration was conducted by updating the user prompt to “consider the proficiency level in English as a ceiling rather than a constriction, you may use all lexical items up to any including the provided proficiency level”. Extra prompts (Appendix E) are included to ensure generated data is written to files in specific format, consistent with the input expectations of the LR analysis aspect of the computational framework,

with prompts instructing the LLM to generate exactly 10 unique adjectives without any repetition or extra explanation and commentary, outputting strictly in English.

The data generation process is based on an automatic Python script within the computational framework that leverages parallel API calls to multiple commercial LLMs, providing the aforementioned system and user prompts, updating prompt inputs dynamically by calling files from within the project. A total of 7,776,000 adjectives were generated from 777,600 requests (24 languages × 6 proficiency levels × 3 prompts × 120 nouns × 3 temperatures × 5 LLMs × 10 adjectives per request) from these commercial LLMs. The comprehensiveness of the lexicon and the noun categories they belong to, as well as the magnitude of adjectives produced in the multiple facets explained above facilitate the construction of a rich semantic map that attempts to accurately represent the semantic senses of bilingual speakers conceptual worlds. We aimed for a confidence level of over 99.99%, assuming an unlimited population size (of lexicon across languages), a 50% population proportion and aiming for a margin of error < 1%; a sample size of 37,831 is required. Our sample size for each language is 7,776,000 adjective data points (as detailed above), divided by 24 languages, for a total sample size of 324,000 adjective data points for each language, nearly a ten fold increase to the required sample size, attaining a high confidence level greater than 99.99% with a margin of error of less than 1%.

3.2.6 The API pipeline

The API pipeline[2], api_calls.py, constitutes a critical component of the computational framework developed in this thesis to investigate LR. This section elucidates the

[2] https://github.com/nurcunal/linguistic-relativity-computational-framework/api_calls.py

intricate details of the pipeline's code implementation while anchoring these technical aspects to the theoretical processes they enable, specifically the simulation of bilingual cognition to test LR hypotheses. Central to the pipeline's implementation is its ability to interface with multiple LLMs, including Google's Gemini-2.0-flash, OpenAI's GPT-4.1 and o3-mini (reasoning model), xAI's Grok-3 and Grok-3-mini (reasoning model), through their respective APIs. The script initializes API clients by reading secure keys from input files, stored in a designated directory, enabling modular model selection without altering core logic. For each combination of native language, English proficiency level, and noun, the pipeline constructs prompts dynamically. A representative prompt might instruct: "You are a native Turkish speaker with B2 English proficiency. Provide the 10 most frequent adjectives used to describe "freedom" in English; return them as a semicolon-separated list." These prompts are dispatched via API calls, with parameters like temperature set to 0, 1 and 2 for controlled, balanced or randomized output and max_tokens calibrated to ensure complete responses. Responses, formatted as semicolon-separated adjective lists, are parsed using Python's split method, stripped of whitespace, and validated to ensure exactly 10 unique adjectives, with invalid outputs triggering retries. Each noun has adjectives generated with a separate API call so that there is no memory of what was just produced in the previous call, ensuring that previous call doesn't effect the adjectives to be produced in the next call.

The pipeline implements two distinct system prompt strategies to simulate bilingual cognition, each configured via command-line arguments for flexibility. The "basic bilingual" (BB) strategy employs an implicit prompt: "You are a native {language} speaker who acquired English at a {proficiency} level", relying on the

LLM's implicit biases to reflect bilingual thought patterns. The "think-in-native" (TN) strategy extends this by appending: "Produce your response thinking in your native language but generating in English", aiming to explicitly activate language-specific cognitive frameworks. These strategies are seamlessly integrated into the codebase, with prompt construction handled through string concatenation, allowing the pipeline to test varied cognitive simulations within a unified framework.

Robustness is ensured through sophisticated error handling and checkpointing mechanisms. API failures are managed with a retry loop employing exponential backoff. Unexpected exceptions are captured in try-except blocks, with partial data saved to ERROR-{run_name}.csv. Checkpointing occurs after each language is processed, with results written to checkpoints/checkpoint_{run_id}_{language}.csv using the pandas library's to_csv() method. A load_existing_data function scans checkpoints on restart, skipping completed languages and resuming from the last valid state. The validate_completeness function ensures each CSV row contains 10 unique adjectives, flagging incomplete entries for reprocessing, thus preventing data gaps.

Data management is structured to support downstream semantic analysis. Outputs are stored in a hierarchical directory under api_generations/{provider}/{model}/{run_params}/, with CSV files named systematically, such as XGM3-BB-0T-24L-7P-144N-5AP.csv[3], encoding model, system prompt type, and run parameters (including temperature, number of languages, proficiencies, nouns and adjective prompts analyzed). The pandas DataFrame class sorts and merges data by

[3] xAI Grok-3 mini - Basic Bilingual - 0 Temperature - 24 Languages - 7 Proficiencies - 144 Nouns - 5 Adjective Prompts

language, noun category, and noun, maintaining a consistent schema with columns like model, prompt type, language, proficiency, noun, and adjectives. This structure ensures compatibility with embedding tools like Sentence-BERT, facilitating vectorization for semantic mapping.

Theoretically, the pipeline works to build the foundation for the LR analysis pipeline that operationalizes LR's premise that language structures shape cognition. The "Basic-bilingual" strategy tests subtle LR effects by attempting to capture subconscious native-language biases in second-language use, as the code's minimal prompt mirrors real-world bilingualism. The "Think-in-native" strategy, by explicitly invoking native-language cognition, probes whether deliberate activation amplifies conceptual divergence in English outputs. LLMs are apt proxies here, as their training data encodes statistical language patterns presumed to reflect cognitive habits, and the pipeline's prompt engineering manipulates these patterns to simulate bilingual thought.

The pipeline outputs in the form of adjective lists, enabling semantic analysis to test LR hypotheses. The parsing and validation logic ensures clean data for vectorization, allowing semantic distances to reveal conceptual proximity across languages, directly addressing LR's claim of language-specific cognitive lenses. The pipeline's scalability, driven by batch processing and checkpointing, supports generating millions of data points, far surpassing previous studies sample sizes, while standardized prompts eliminate subjective biases, aligning with LR research's demand for rigor. This alignment strengthens empirical foundations for investigating LR effects.

3.3 Semantic embedding generation

This section details the transformation of adjective outputs from LLMs into high-dimensional vector embeddings, forming the backbone of the semantic field. These embeddings act as computational proxies for conceptual structures, capturing L1-driven cognitive biases in L2 production to empirically test LR.

3.3.1 The embedding model

The embedding process is another central component in our computational framework, designed to transform adjective lists generated for each (noun—language—proficiency—prompt—LLM—temperature) combination into dense numerical representations that capture semantic meaning. We evaluate multiple sentence transformer models to map adjectives into high-dimensional vector spaces (768 dimensions). These embedding models leverage large-scale pre-training to encode linguistically nuanced terms into a shared vector space (Mikolov et al., 2013).

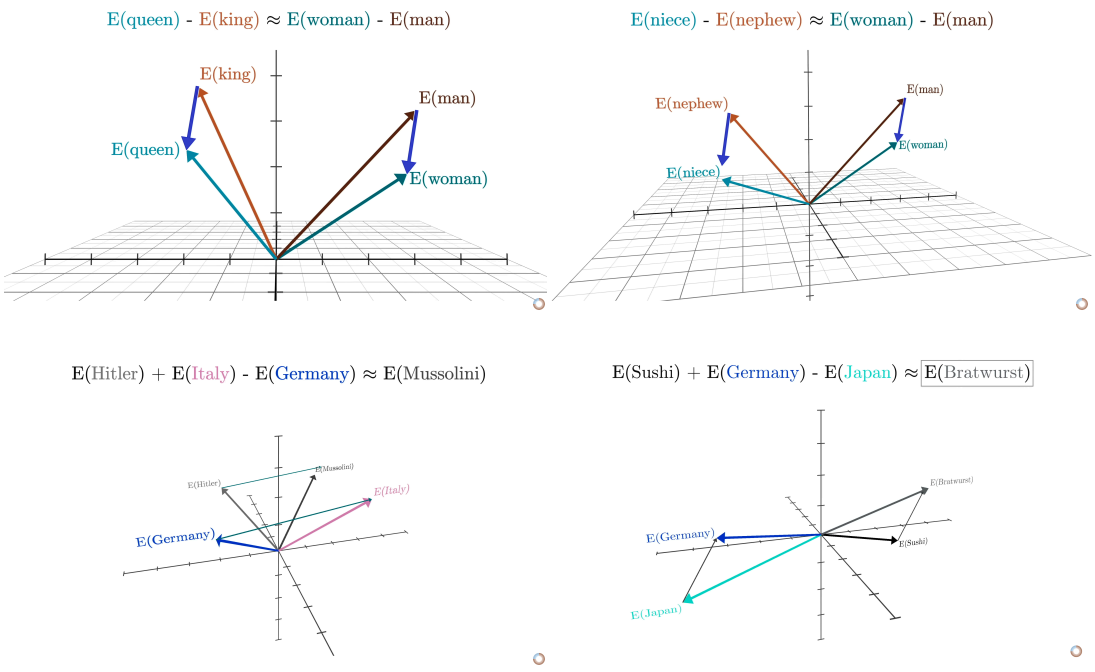


Fig. 2 Embedding vector semantic similarity visualization (Sanderson, 2024)

Upon theoretical evaluation and practical generation samples from ten models (results of each mode at Github), we settle on the paraphrase-multilingual-mpnet-base-v2 model (Reimers, 2020) for its optimization for our research application and strong performance across four key criteria, as follows: The chosen encoder must balance semantic resolution, cross-lingual alignment, representational capacity, and computational efficiency to detect subtle L1-driven conceptual shifts in a large dataset. Evaluating a diverse set of models ensures that we may select the most suitable tool for our interdisciplinary framework.

- Semantic resolution: The MPNet architecture, with its masked pre-training, preserves long-range token dependencies, capturing subtle metaphorical and culturally loaded nuances in adjective-noun pairs (e.g., “sovereign freedom” in Arabic L1 vs. “personal freedom” in English L1) where other models sacrifice some depth.
- Cross-lingual alignment: Trained on parallel paraphrase pairs across ~50 languages, the MPNet architecture ensures semantically equivalent phrases cluster together, allowing the model to capture cross-linguistic nuances of phrases.
- Representational capacity: MPNet’s 768-dimensional embedding space provides twice the axes of MiniLM variants (384 dimensions), preserving more semantic information. Pilot experimentations showed that reducing dimensionality to 384 compressed cosine variance between distinctive adjectives several times over, confirming the value of higher dimensions.
- Empirical performance: On the “2025 Massive Text Embedding Benchmark” (MTEB) leaderboard, paraphrase-multilingual-mpnet-base-v2 outperforms

MiniLM models by 5–9 Pearson points in cross-lingual semantic-textual-similarity tasks (Hugging Face, 2025), making it an ideal contender for this project.

3.3.2 The embedding process

This embedding process is done by reading the formatted outputs of the aforementioned adjective generator API pipeline as input files into another pipeline within the same framework that is designed to map a semantic field of noun and language representations separately. Each of the 7,795,200 adjective points are mapped to 768-dimensional vectors using paraphrase-multilingual-mpnet-base-v2. For each noun, the vector embeddings of its associated 10 adjectives are aggregated using mean pooling to produce a single composite representation of the associated noun as a proxy for the noun's conceptualization. This mean pooling approach ensures that the resulting vector reflects the collective influence of adjective semantics without being overly sensitive to outliers. Embeddings are L2-normalized to ensure cosine similarity reflects pure angular distance, standardizing comparisons across languages.

3.4 Semantic distance analysis

This section outlines the methodologies used to quantify conceptual differences across languages, leveraging the aggregated linguistic representations from the previous section. Two key metrics—cosine distance and Jaccard distance—are employed, with analyses conducted at the noun, noun category, language, and language family levels.

3.4.1 Cosine distances for embedding-based analysis

Cosine similarity is the key metric in our computational framework, used to compare high-dimensional embedding vectors and quantify semantic similarity through angular differences in a normalized vector space; its focus on directional similarity aligns with our goal of detecting semantic nuances (Jurafsky & Martin, 2021). Normalized embeddings ensure that cosine similarity reflects semantic relationships, making it ideal for cross-lingual comparisons where L1 influences manifest as angular shifts in vector space (Mikolov et al., 2013). It is applied at both the noun level (comparing specific concepts cross-linguistically), the noun category level (comparing noun categories cross-linguistically), the language level (comparing aggregated embeddings of nouns as a language representation cross-linguistically), and the language family level; providing a scalable, quantitative measure of cross-linguistic variability.

Cosine similarity is calculated as the dot product of vectors u and v, denoted $u \cdot v$ (the sum of the products of their corresponding components), divided by the product of their Euclidean norms, $\|u\|$ and $\|v\|$, resulting in values from -1 to $+1$. To convert this similarity measure into a distance metric where larger values indicate greater dissimilarity, we subtract the cosine similarity from the total probability (of 1), giving the cosine distance: Cosine distance $= 1 - (u \cdot v) / (\|u\| \|v\|)$ where u and v are embedding vectors, $u \cdot v$ is their dot product, and $\|u\|$ and $\|v\|$ denote their Euclidean norms. The corresponding cosine similarity, $(u \cdot v) / (\|u\| \|v\|)$, ranges from -1 to 1. Because the embeddings used here are unit vector normalized (L2) and non-negative (as in transformer models such as our embedding model), the resulting cosine distance is confined to the interval 0 – 1 (Reimers & Gurevych, 2019). That is,

a cosine distance of 0 (cosine similarity = 1) denotes that the vectors are perfectly aligned (angle = 0°), indicating identical semantic representations (e.g., “chair” embeddings for French and Spanish are conceptually identical) and a cosine distance of 1 (cosine similarity = 0) denotes that vectors are orthogonal to one another (angle = 90°), indicating no semantic overlap (e.g., “freedom” embeddings for French and Arabic are quite dissimilar). If cosine distances were to reach 2 (cosine similarity = –1, angle = 180°), this would imply opposite semantic meanings (antonyms), but transformer-based embeddings, trained on positive semantic associations, rarely produce negative similarities in practice (Reimers & Gurevych, 2019). Negative cosine similarities are theoretically possible but irrelevant here, as our adjective embeddings reflect positive or neutral associations, denoted as a cosine distance of 1 and 0, respectively.

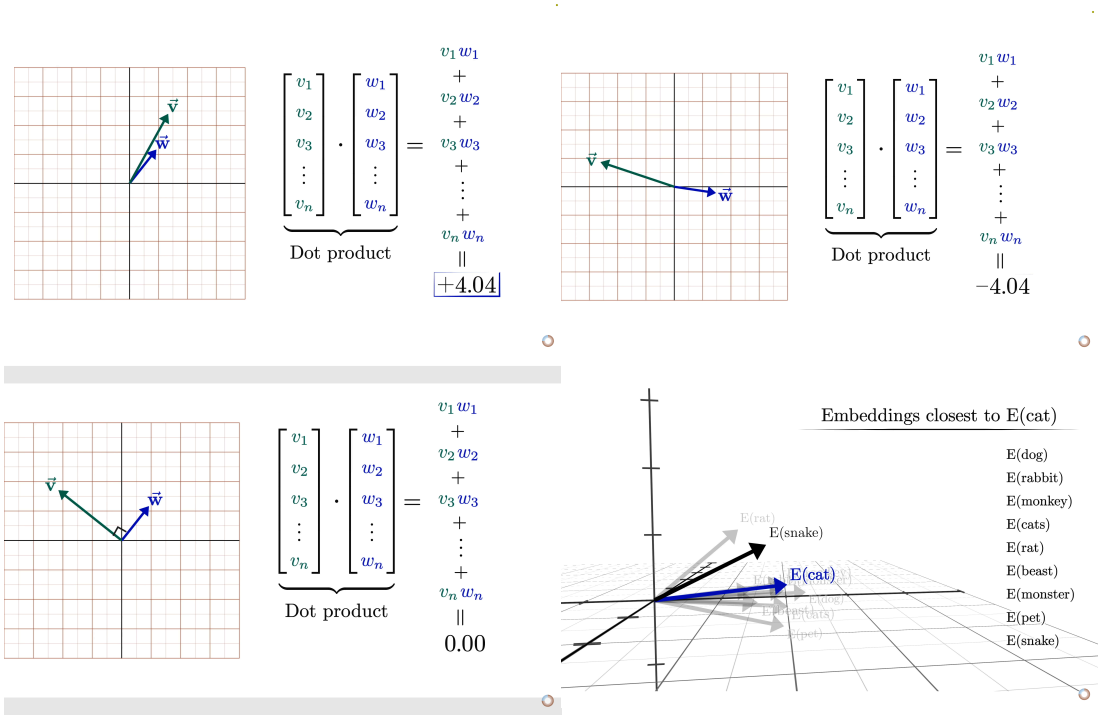


Fig. 3 Embedding vectors semantic (cosine) distances (Sanderson, 2024)

Any distance between languages' perceptions of nouns would be meaningful in the assessment of LR, however here we elaborate on our hypothesis by expressing the range of what "meaningful differences in semantics" constitutes. Given that a cosine distance of 0 indicates semantically identical perception and a cosine distance of 1 expresses semantically completely irrelevant comprehension, and since we are comparing the distance between the same lexical item with all parameters—save language—fixed, a meaningful distance between language specific conceptualizations of a given lexical item may be assessed with a fixed metric. We wouldn't expect a single lexical item to be conceptualized with so much disparity between languages that they no longer resemble or relate to each other, so we expect that even the most distinct perspectives or senses for a single noun wouldn't surpass some ceiling value (as indicated the significance of metrics). One-sample Cohen's d (Cohen, 1988) will be applied as a formal metric to provide a significance measurement. This is calculated as $d = (M - \mu) / SD$, where M is the mean cross-linguistic cosine distance, μ is the reference value of (all categories mean value), as the expected maximum distance, and SD is the standard deviation. Furthermore, we may compare and contrast our results to previously held studies that provide quantitative values arguing against and for LR (Mickan, 2014; Sato, 2018) as a criterion in assessing them. These assessments are done at the noun levels separately for each category, as well as at the language and language family levels, with clustering effects observed in visualized visualizations of the semantic map (Section 3.5).

It's important to note that cosine value is a very precise mode of measurement as opposed to comparing adjectives at a one-to-one ratio, since adjectives may be

different lexically but very similar semantically, in which case that dissimilarity would result in skewed results. That kind of comparison can be relevant in a different manner, as explained below.

3.4.2 Jaccard distance for lexical comparison

Jaccard distance is calculated as $1 - |A \cap B| / |A \cup B|$ where A and B are the sets of adjectives associated with a noun in two languages; $|A \cap B|$ is the number of shared adjectives (the intersection); and $|A \cup B|$ is the total number of unique adjectives combined (the union). The distance ranges from 0 (complete overlap) to 1 (no overlap). For instance, if French associates "freedom" with {personal, political, social} and Arabic associates it with {collective, sovereign, social}, then the Jaccard distance = $1 - |\{social\}| / |\{personal, political, social, collective, sovereign\}| = 1 - 1 / 5 = 0.8$. A high Jaccard distance like in this example indicates low overlap, reinforcing that first-language concepts shape the differing adjective choices. However, if Jaccard distance was to be used as the single metric to measure cross-linguistic distinctions, it could lead to misleading results as it is literal in its comparisons and does not take similarity of distinct words into account. Therefore, the employment of Jaccard distance in our project is primarily to verify cosine distance metrics by providing an independent lexical measure that correlates with embedding-based findings. Since cosine distance relies on aggregated embeddings (mean-pooled adjective vectors), it may smooth over specific lexical choices. Jaccard distance, by contrast, directly compares raw adjective sets, capturing explicit differences in word selection. A strong positive correlation (e.g., Pearson's coefficient ($r > 0.6$), ($p < 0.05$)) between cosine and Jaccard distances for the same

noun-language pair (e.g., “freedom” in French vs. Arabic) confirms that embedding-based divergence reflects true lexical and conceptual differences, not artifacts of the embedding model. For instance, a high cosine distance (0.55) paired with a high Jaccard distance (0.8) for “justice” suggests both semantic and lexical divergence, strengthening LR evidence. Conversely, discrepancies (e.g., high cosine but low Jaccard distance) flag potential embedding noise, prompting further investigation via sensitivity analyses, with correlation analyses (Pearson’s coefficient) being conducted to verify consistency between cosine and Jaccard distances, with results visualized via scatter plots (Section 3.5).

3.5 Semantic space visualization and mapping

This section explains how high-dimensional linguistic data is transformed into multidimensional visuals to explore relationships between languages and noun categories, at the noun, noun category, language, and language family levels. Visualizing these high-dimensional noun representations requires dimensionality reduction. Techniques like Principal Component Analysis (PCA), Uniform Manifold Approximation and Projection (UMAP) and t-distributed Stochastic Neighbor Embedding (t-SNE) are employed to project the very high dimensional embeddings into a small multidimensional (two and three dimensions) space, facilitating visual inspection of clustering patterns. Interactive scatter plots generated using Plotly allow for exploration, where each point represents a noun's embedding for a specific language profile. Coloring points by language, language family (Appendix A), or grammatical gender category (Appendix B) helps reveal systematic variations. Tooltips provide details such as the specific language, proficiency, prompt, and

adjectives generated, offering immediate context for observed spatial relationships. The quality of these multidimensional representations in preserving the original high-dimensional relationships is assessed using Kruskal's Stress, providing a quantitative measure of distortion introduced by the reduction technique.

3.5.1 The LR analysis pipeline

The LR Analysis Pipeline[4], implemented in the "lr_analysis.py" script, constitutes another cornerstone of the computational framework developed in this thesis to empirically investigate LR. This section delves into the pipeline's intricate code implementation, anchoring its technical details to the theoretical processes it enables —namely, the transformation of linguistic data into quantifiable semantic representations to test LR hypotheses. Building upon the outputs of the API Pipeline (Section 3.2.6), this pipeline interfaces with adjective association data generated by LLMs simulating bilingual speakers, processing it through advanced CL techniques to explore how native language (L1) structures shape conceptual thought in a shared second language (L2, English). Central to its design is the use of semantic embeddings as proxies for cognition, enabling the measurement of conceptual distances across languages, proficiency levels, and noun categories.

The pipeline initializes by loading structured CSV files from the API Pipeline, stored in directories such as api_generations/{provider}/{model}/{run_params}/. A typical file, like OO3M-BB-0T-24L-7P-144N-5AP.csv, encodes adjective lists for 144 nouns across 24 languages and 7 proficiency levels, generated by OpenAI's o3 mini using the Basic-bilingual prompt at temperature 0. These files

[4] https://github.com/nurcunal/linguistic-relativity-computational-framework/lr_analysis.py

are parsed using Python's pandas library via the load_data function, which splits semicolon-separated adjective strings into lists and merges them with metadata from supporting files—noun_categories.json, grammatical_genders.json, and language_families.json—stored in the "inputs/" directory. This metadata enriches the dataset with noun categorizations (e.g., abstract concepts), grammatical gender mappings, and language family affiliations, facilitating nuanced analyses.

The pipeline's workflow, orchestrated by the analysis_pipeline function, executes a series of meticulously designed steps:

- Setup and initialization: An output directory structure is established under embedding_analysis/<embedding_model>/<api_provider>/<api_model>/<temperature>/, such as embedding_analysis/PMMB/xai/grok-3-mini/0T/, where PMMB denotes the paraphrase-multilingual-mpnet-base-v2 sentence-transformer model. Auxiliary data is loaded via functions like load_noun_categories, and PyTorch detects the optimal hardware (e.g., CUDA (for Nvidia), MPS (for macOS), or CPU) for computation, ensuring efficiency.
- Embedding generation: Adjective lists are vectorized using a sentence-transformer model through embedding_models.build_embeddings, producing high-dimensional embeddings for each noun's adjective set. These embeddings are aggregated via mean pooling to form a composite vector per noun, saved in 01_embeddings/ as pickled DataFrames (all_proficiencies_<run_params>.pkl). This process captures the semantic essence of bilingual outputs, serving as a computational stand-in for conceptual thought, thus, enabling quantitative analysis of LR in semantic spaces.

- Distance computation: The pipeline iterates over proficiency levels, prompt types (e.g., Basic-bilingual, Think-in-native), and noun categories, computing pairwise distances for each combination:
  - Cosine distances: Calculated by compute_pairwise_distances, these measure semantic similarity between embeddings of the same noun across different L1s.
  - Jaccard distances: Generated via compute_jaccard_distances, these quantify lexical overlap in adjective choices.
  - Results are stored in "02_distances/", with detailed noun-level comparisons in noun_level_distances/details/, providing dual perspectives on conceptual and linguistic divergence.
- Visualization and statistical analysis: Outputs in 03_visualizations/ include 2D and 3D t-SNE, UMAP, and PCA plots (via plot_tsne, plot_umap, plot_pca), rendered as interactive HTML and static PNG files, alongside dendrograms and heatmaps (plot_language_dendrogram, plot_interactive_heatmap) to depict language clustering. Statistical tests—ANOVA, Tukey's HSD, and Pearson's correlation—are performed using analyze_noun_type_distances and related functions, with effect sizes (Cohen's d) reported in 04_statistical_reports/, assessing the significance of LR-driven differences.
- Aggregated analyses: Language-level embeddings are computed via calculate_language_embeddings, and family-level representations via calculate_language_family_embeddings_and_distances. These aggregated distances, stored in language_level/ and language_family_level/ subfolders, explore whether related languages share closer semantic profiles. Robustness is ensured through checkpointing and error handling. Intermediate results are saved

after each processing stage, allowing resumption from failures, while exceptions are logged to prevent data loss. The pipeline's outputs—embeddings, distances, visualizations, and statistical reports—are systematically organized, supporting downstream interpretation and reproducibility.

Theoretically, the LR analysis pipeline operationalizes LR by positing that semantic embeddings reflect L1-influenced cognitive representations. Cosine distances between embeddings of the same noun across L1s quantify conceptual divergence, while Jaccard distances highlight explicit linguistic differences, testing whether L1 structures shape thought in L2 use. The iterative analysis across prompt strategies probes whether explicit L1 activation (TN) amplifies these effects compared to implicit simulation (BB). By scaling to millions of data points, the pipeline works to overcome traditional LR study limitations, offering empirical evidence of language-thought interactions.

The pipeline's output enables semantic mapping to test LR hypotheses directly. Its structured design, integrating LLMs with CL tools like Sentence-BERT, ensures clean, vectorizable data, revealing how language-specific lenses manifest in bilingual cognition. This framework not only advances LR research but also bridges computational and cognitive science, providing a scalable, standardized approach to understanding the language-thought nexus.

With the computational framework's theoretical and practical implementation strategies established, we now move on to the results of the analysis pipeline, that assess cross-linguistic distances and map them to vector space on the noun, noun category, language, language family, proficiency and LLM levels. Interpretations for specific instances among the thousands of comparisons conducted are found below.

3.5.2 Noun and noun category level analysis

The noun and noun category level analysis delves into the granular details of the semantic space, examining how the conceptual representation of individual nouns within their pre-established categories vary across different simulated bilingual linguistic profiles at the lowest level. By comparing the aggregated adjective embeddings for each of the 144 nouns organized into 12 distinct categories (Appendix B) across the 24 languages (Appendix A), this analysis aims to pinpoint specific nouns and their categories that elicit convergent versus divergent semantic associations, providing empirical data to assess LR's claims on conceptualization. This is done in two manners; first a high-level approach over some of the individual noun's positions within their respective category cluster are demonstrated visually along with a statistical report of their relationships between languages, then a low-level approach is provided, demonstrating a few such particular instances of these specific noun variations within the categories they belong to and their implications to examine these relations with a closer lens.

144 nouns for each language have been plotted in multidimensional space, color-coded by category for a visual demonstration of their position in space, as a representation of how native speakers of those languages perceive these nouns distinctly. Each of these visuals were produced with interactive PCA, UMAP and t-NSE graphs (available at GitHub: LR computational framework), and measured to produce the quantitative cross-linguistic distance between each noun individually, and noun categories as a whole. Furthermore, heat maps between languages in the context of each noun category were generated to depict the magnitude to which native speakers of languages perceive those particular categories distinctly.

The specific configurations for the API generations that are the basis for the generated adjectives that produce the embedding vectors for each visualization may be found in (Appendix E). While a couple of instances are illustrated here, a comprehensive set of all language pairs relations for all categories are available for examination in the LR computational framework on GitHub.

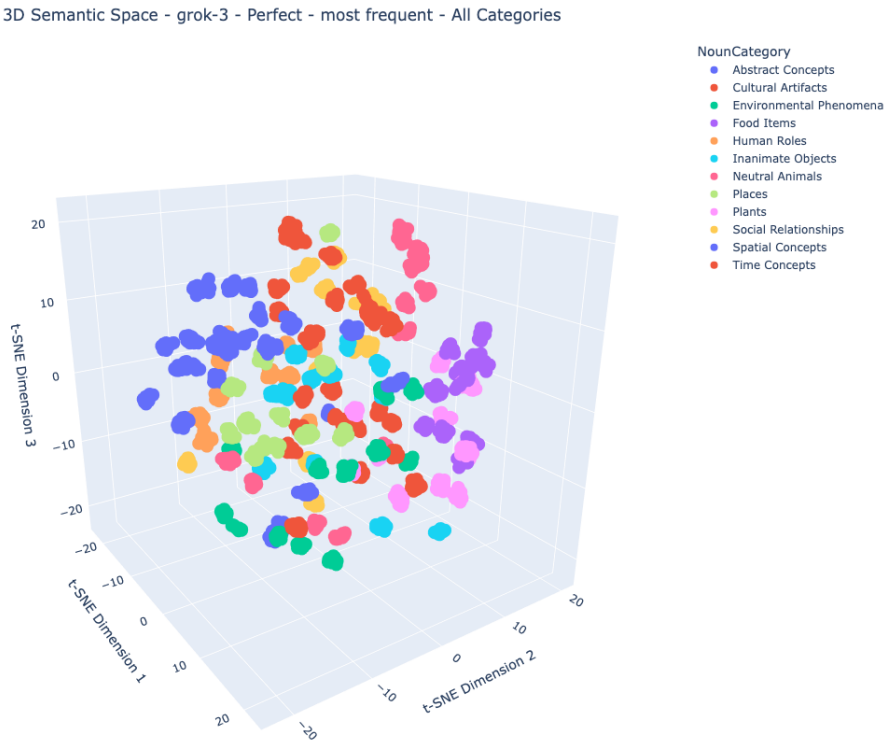


Fig. 4 3D t-NSE interactive semantic space of noun categories (Grok-3)
Note: Unless explicitly noted, all figures and tables use temperature 0 by default.

A comprehensive overview of the semantic space for each noun within their respective noun categories illustrates their cross-linguistic divergence and variance. While these stationary graphs offer limited intuition into the manner in which these noun categories, and more specifically each noun within them vary cross-linguistically in semantic space, the interactive versions of these graphs, along with PCA and UMAP variants are available at the computational framework on GitHub.

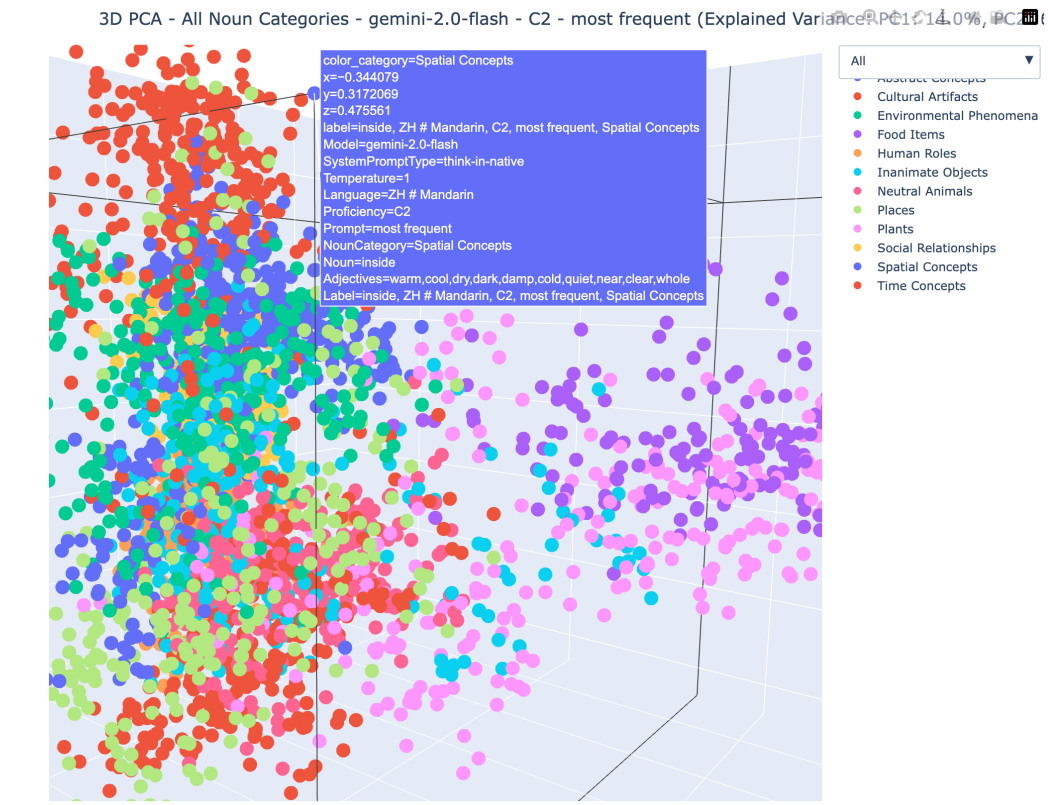


Fig. 5 3D PCA interactive semantic space of each nouns position (Gemini-2.0)

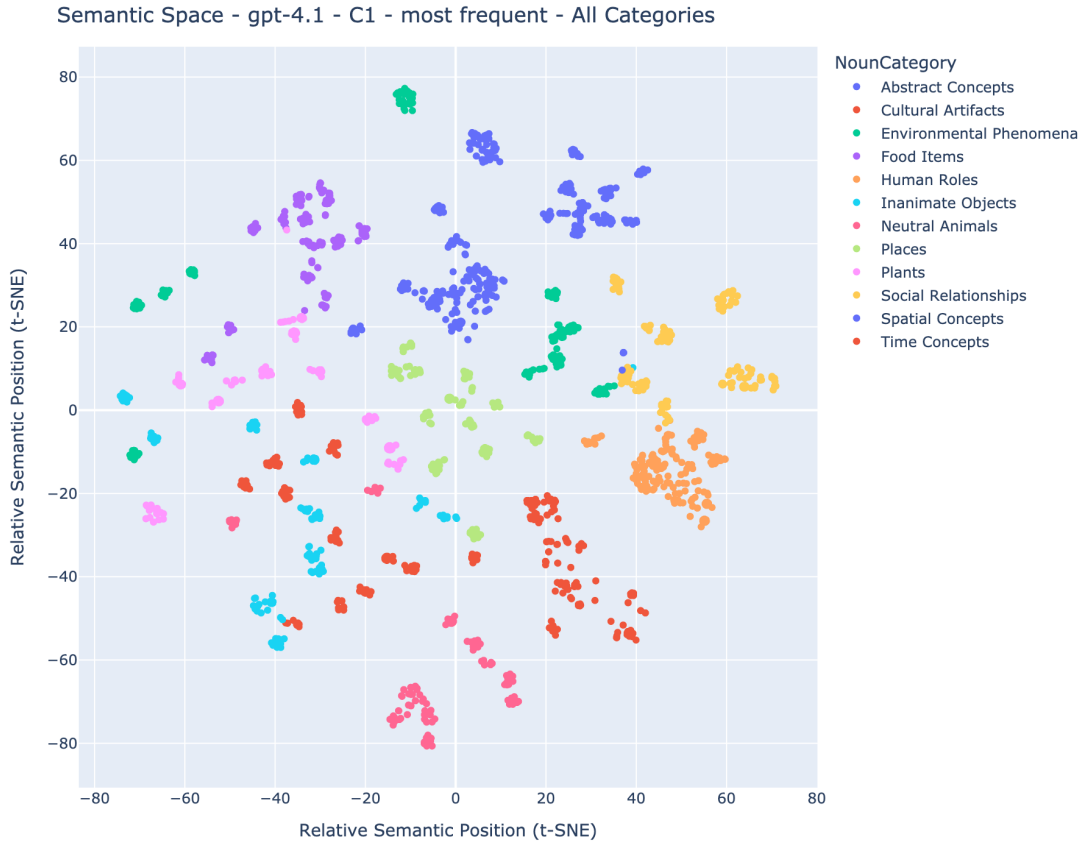


Fig. 6 2D t-NSE interactive space of each nouns position (GPT-4.1)



Fig. 7 “Food Items” cosine distances between languages; “C2” (Grok-3 mini)

The results for each languages representation for a single noun category is calculated by aggregating all of the embedding vectors for the nouns within that category by means of mean pooling into a single embedding vector. These vectors are then compared to one another between languages by means of cosine similarity. Some language pairs show relatively high distances to one another in certain noun categories, apparent in the ~19% distance in “Food Items” between Hebrew and Hindi (see Fig. 7), implying that native speakers of those languages conceive concepts pertaining to food items significantly differently from one another. These results provide a quantitative assessment of cross-linguistic semantic nuances that support claims of LR for this particular scenario, with the average between all languages for food items conception listed in (Table 1).

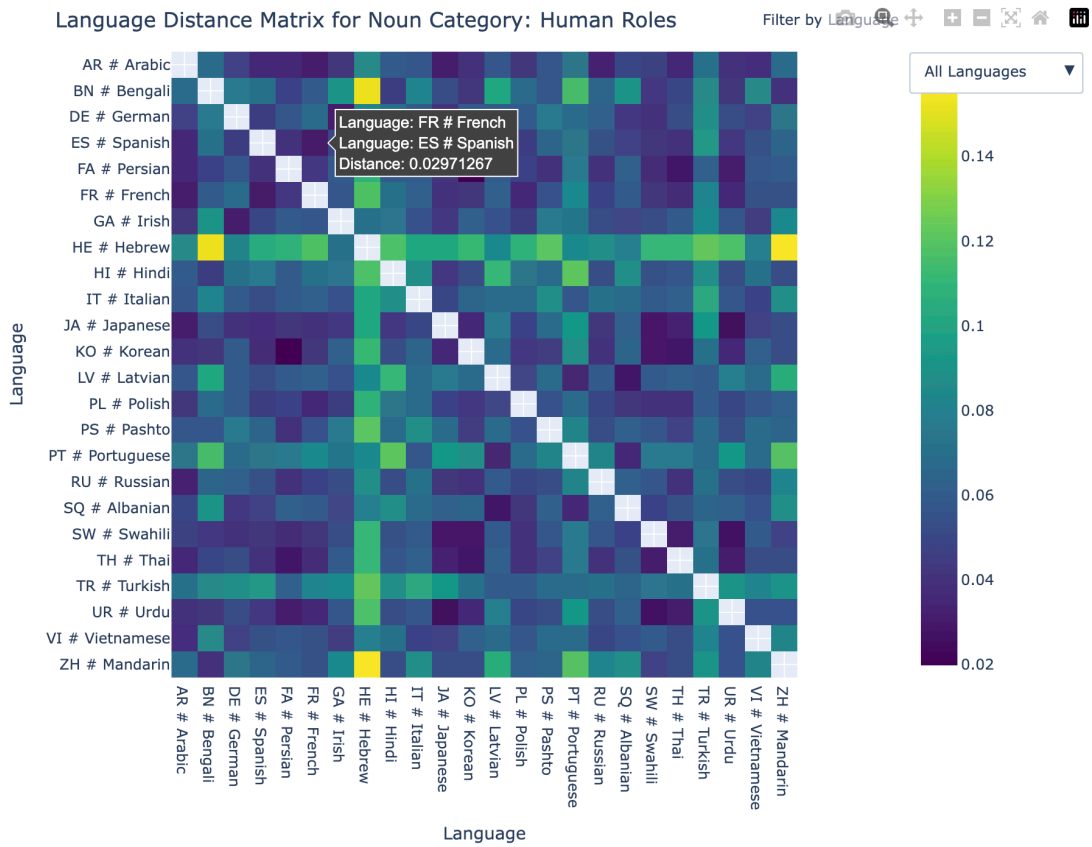


Fig. 8 “Human Roles” cosine distances between languages; “Perfect” (GPT-4.1)

This is in contrast to other language pairs that return minimal distances, as is apparent in the ~3% distance in “Human Roles” between French and Spanish (Fig. 8), implying that speakers of those languages perceive concepts relating to human roles quite similarly to each other. Various analyses may be inferred from these heat maps, such as Hebrew’s contrasting pattern to other languages in this noun category.

Table 1. Average Cross-Linguistic Distances for Noun Categories (Grok-3 mini, C2)

Noun Category	Lexical Items	Cosine Dis. (mean)	Cohen’s d (Cosine)	Cosine Dis. (max)	Jaccard Dis. (mean)	Jaccard Dis. (max)
Cultural Artifacts	mask	0.0585	0.71 (Medium)	0.1700	0.4827	0.8235
Food Items	chicken	0.0195	0.13 (Negligible)	0.0740	0.1974	0.4615
Human Roles	soldier	0.0383	0.05 (Negligible)	0.0938	0.4386	0.8235

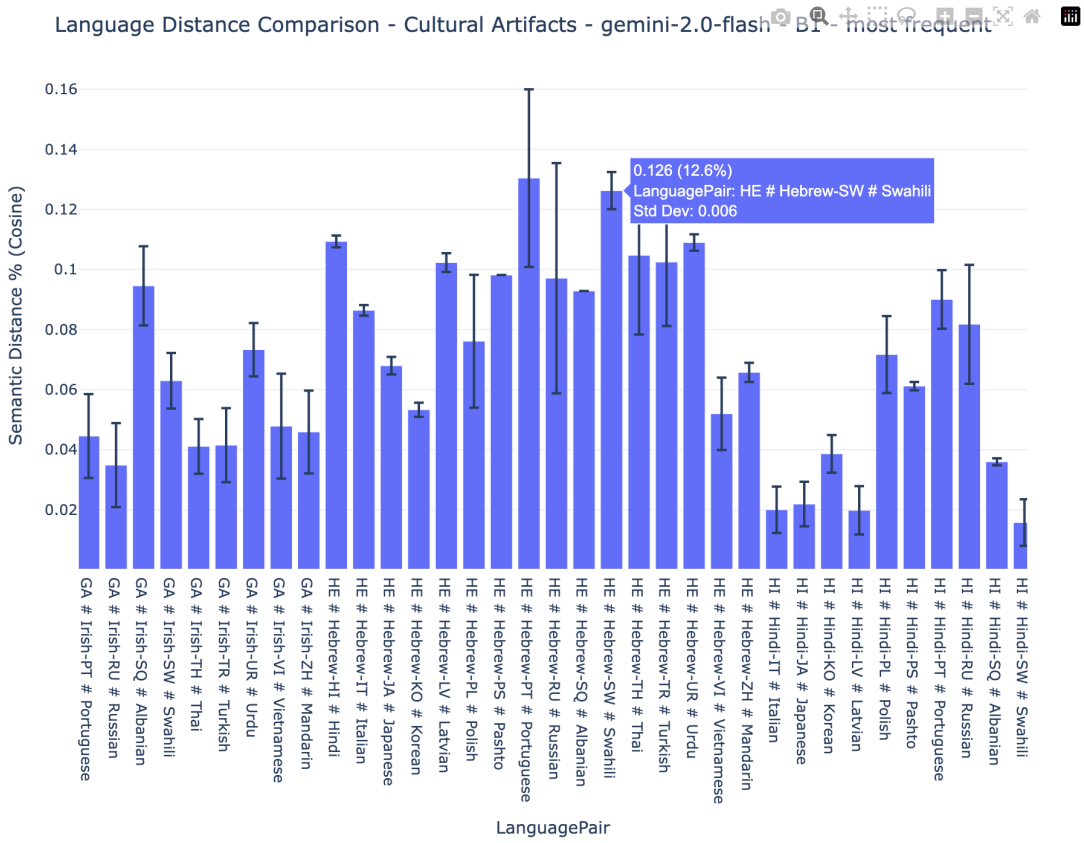


Fig. 9 “Cultural Artifacts” distances between languages; “B1” (Gemini-2.0)

For instance, based on a particular LLM configuration (Appendix E), the Persian-English bilingual representation model describes “courage” with the adjectives “brave, fearless, bold, heroic, daring, valiant, resolute, steadfast, gallant, and unwavering”, while the French-English bilingual model uses “remarkable, fearless, bold, unwavering, heroic, admirable, steadfast, inspiring, tremendous, and resolute”. Adjectives unique to the Persian model are “brave, daring, valiant, and gallant”; those unique to the French model are “remarkable, admirable, inspiring, and tremendous”. This translates to a cosine distance of 11.82% and a Jaccard distance of 57.14% between the two languages, implying a significant difference in the manner they conceive the concept of “courage”. This is in contrast to a word like “chicken”,

which is quite similar between two bilingual representation models like Arabic-English and Irish-English, with the former describing it as “grilled, roasted, tender, juicy, spicy, fried, lean, savory, fresh, and seasoned” and the latter “tender, juicy, grilled, roasted, fried, spicy, lean, fresh, savory, and cooked”, respectively, where the only distinct adjectives used are “seasoned” in Arabic and “cooked” in Irish. This translates to a cosine distance of 0.92% and a Jaccard distance of 18%, suggesting that Arabic and Irish native speakers perceive “chicken” very similarly.

Table 2. Cosine Distance Cohen’s d Across French-Persian (Grok-3 mini, Perfect)

Noun Category / Language Pair	Abstract Concepts	Food Items
French - Persian	0.910 (Large effect)	0.383 (Small Effect)

Table 3. Noun Category Overview Analysis—All Language Pairs (Grok-3, B2)

Noun Category	Cosine Distance (mean)	Cosine Distance (max)	Jaccard Distance (mean)	Jaccard Distance (max)
Abstract Concepts	0.0306	0.1253	0.3111	0.8235
Cultural Artifacts	0.0208	0.1607	0.2410	0.8333
Environmental P.	0.0140	0.0639	0.1819	0.5714
Food Items	0.0187	0.0875	0.2073	0.5714
Human Roles	0.0183	0.1082	0.2181	0.7500
Inanimate Objects	0.0240	0.1051	0.2380	0.6666
Neutral Animals	0.0172	0.0835	0.2346	0.6666
Places	0.0201	0.0918	0.2027	0.6666
Plants	0.0141	0.0855	0.1747	0.5714
Social Relationships	0.0188	0.0883	0.2240	0.7500
Spatial Concepts	0.0439	0.3246	0.3452	1.0000
Time Concepts	0.0222	0.1233	0.2511	0.8235

Note: See (Appendix E) for relevant output folders used in calculations.

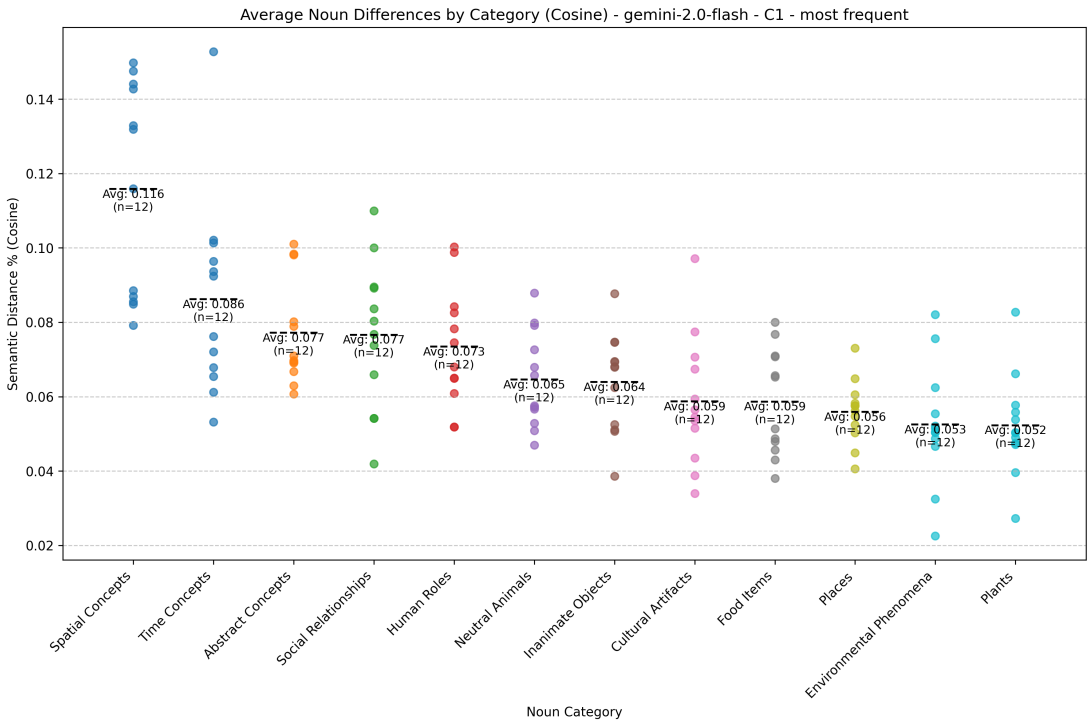


Fig 10. Average noun category distances between all languages (Gemini-2.0)

Table 4. Average Noun Category Distances - Turkish vs. Average (Grok-3 mini, C1)

Noun Category / Language Pair	Turkish (mean)	All Languages (mean)	Cohen's d (μ = All language's means)
Abstract Concepts	0.0414	0.0450	0.141 (Negligible)
Cultural Artifacts	0.0492	0.0477	0.060 (Negligible)
Environmental P.	0.0330	0.0374	0.181 (Negligible)
Food Items	0.0322	0.0346	0.100 (Negligible)
Human Roles	0.0297	0.0331	0.171 (Negligible)
Inanimate Objects	0.0447	0.0477	0.113 (Negligible)
Neutral Animals	0.0450	0.0495	0.186 (Negligible)
Places	0.0286	0.0360	0.310 (Small)
Plants	0.0340	0.0361	0.091 (Negligible)
Social Relationships	0.0328	0.0382	0.246 (Small)
Spatial Concepts	0.0599	0.0743	0.323 (Small)
Time Concepts	0.0421	0.0499	0.255 (Small)

Table 5. Distance of "Love" between Arabic and All Languages (Grok-3 mini, C2)

Arabic vs. Provided Language	Cosine Distance Between Languages (mean)	Cohen's d (μ = All language's means)
Bengali	0.0628	0.9043 (Large)
German	0.0345	0.2966 (Small)
Spanish	0.0996	2.5052 (Large)
Persian	0.0898	2.0690 (Large)
French	0.0371	0.1724 (Negligible)
Irish	0.0325	0.3922 (Small)
Hebrew	0.0361	0.2155 (Small)
Hindi	0.0313	0.4310 (Small)
Italian	0.0951	2.2780 (Large)
Japanese	0.0361	0.2155 (Small)
Korean	0.0379	0.1379 (Negligible)
Latvian	0.0345	0.2966 (Small)
Polish	0.0331	0.3534 (Small)
Pashto	0.0345	0.2966 (Small)
Portuguese	0.0345	0.2966 (Small)
Russian	0.0153	1.1164 (Large)
Albanian	0.0361	0.2155 (Small)
Swahili	0.0575	0.7112 (Medium)
Thai	0.0655	1.0483 (Large)
Turkish	0.0361	0.2155 (Small)
Urdu	0.0534	0.5345 (Medium)
Vietnamese	0.0569	0.6853 (Medium)
Mandarin	0.0488	0.3448 (Small)

Table 6. "Love" in Arabic's Distance from All Languages (Grok-3 mini, C2)

Noun	Arabic (mean)	All Languages (mean)	Cohen's d (μ = All language's means)
Love	0.04613	0.0412	0.2125 (Small)

Furthermore, a cosine to Jaccard distance Pearson correlation is calculated and depicted to assess statistical measurements to validate cosine distances for all categories to ensure the rigor of the cosine distance values. For instance, a Pearson value of r = 0.876 where n = 3,312 for “Abstract Categories” show a positive correlation across all lexical items within that category. This statistical significance works to verify the cosine distances accuracy as explained in (Section 3.4). (Fig. 11) provided below illustrates this correlation with an interactive interface (GitHub) that allows one to examine the particular cosine—Jaccard correlations for each noun in that category for each language. This operation is done for each category as is available in the output folders of the computational framework for this thesis.

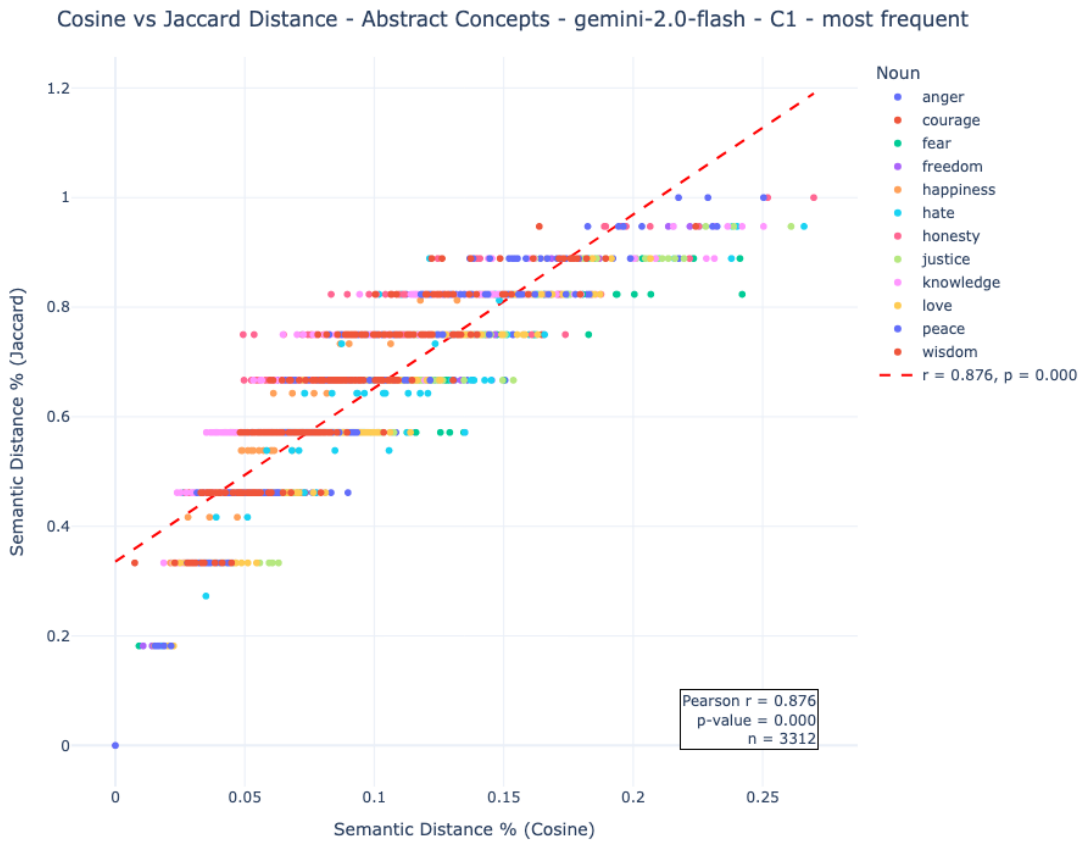


Fig. 11 Raw Correlation of cosine and Jaccard Distances (Gemini-2.0, C1)

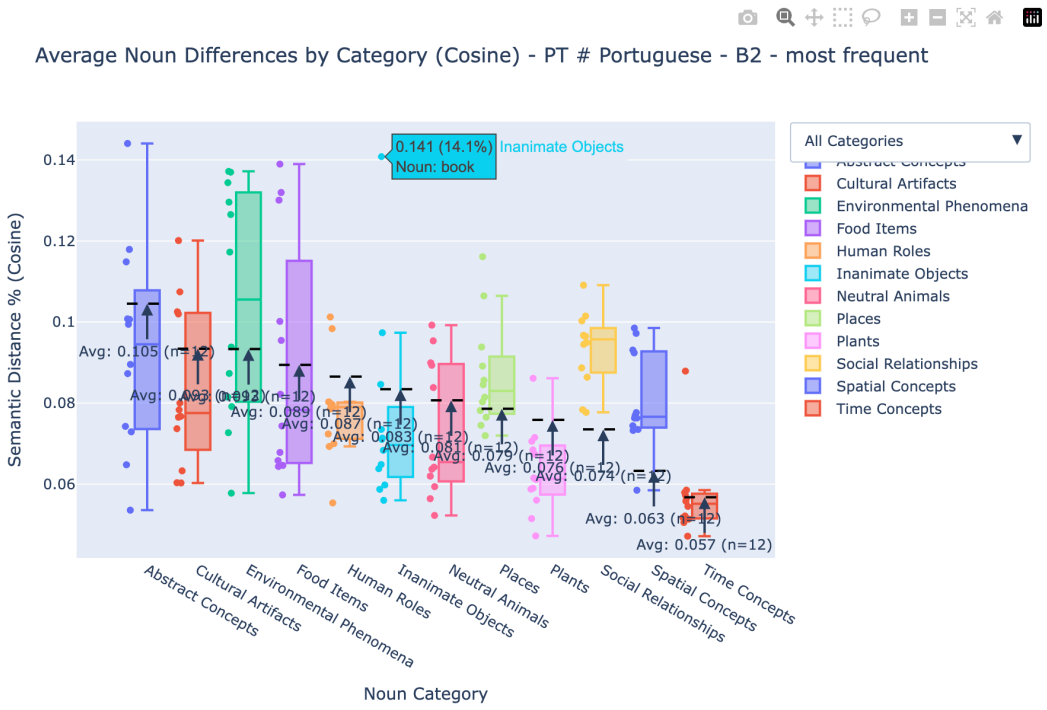


Fig. 12 Noun category distances between Portuguese - all languages (o3-mini, B2)



Fig. 13 Nouns plotted with color coded noun categories by language (GPT-4.1, P)

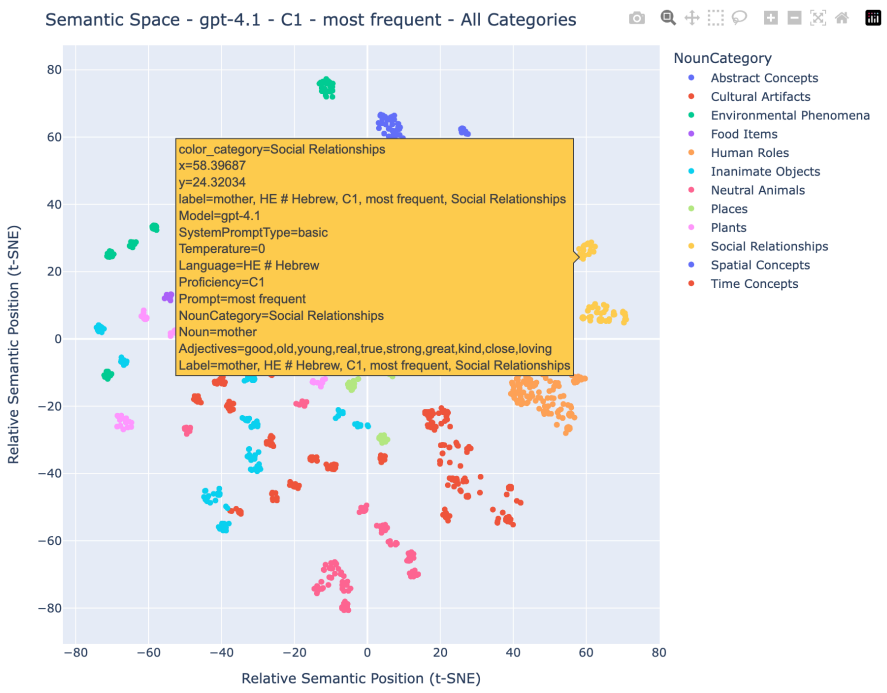


Fig. 14 Nouns plotted with color coded noun categories by language (GPT-4.1, C1)

Table 7. Cross-Linguistic Random Sample Distances (Grok-3, Perfect, 1T, TN)

Languages	Categories	Nouns	Dist.	Cohen's d	Max Distance	Cohen's d (Max)
SQ - AR	Abstract C.	honesty	0.0587	0.0686 (L)	PS-RU: 0.1666	0.0683 (L)
BN - FR	Cultural A.	flag	0.0006	0.6227 (L)	FA-PT: 0.0560	0.3044 (L)
DE - HE	Environmental	rainbow	0.0001	0.5058 (L)	ES-JA: 0.0355	0.8877 (L)
HI - GA	Food Items	bread	0.0171	0.5740 (M)	BN-SW: 0.0670	0.0692 (L)
IT - JA	Human Roles	teacher	0.0354	0.9382 (L)	AR-ZH: 0.0582	0.4098 (L)
KO - LV	Inanimate O.	book	0.03491	0.2905 (S)	BN-RU: 0.1270	0.9278 (L)
ZH - PS	Neut. Animals	cat	0.0219	0.3509 (S)	ES-HI: 0.1057	0.5069 (L)
FA - PL	Places	village	0.0344	0.8583 (L)	GA-LV: 0.0391	0.1704 (L)
PT - RU	Plants	tree	0.0235	0.3370 (S)	AR-JA: 0.0454	0.7302 (M)
ES - SW	Social Relation	mother	0.0139	0.1721 (N)	AR-JA: 0.0516	0.8054 (L)
TH - TR	Spatial C.	above	0.0239	0.5519 (M)	HI-ZH: 0.1017	0.4136 (L)
UR - VI	Time Concepts	century	0.0453	0.4356 (S)	HE-PL: 0.1049	0.3898 (L)

Note: Cohens d: μ for each category and M for each noun found at (Appendix F).

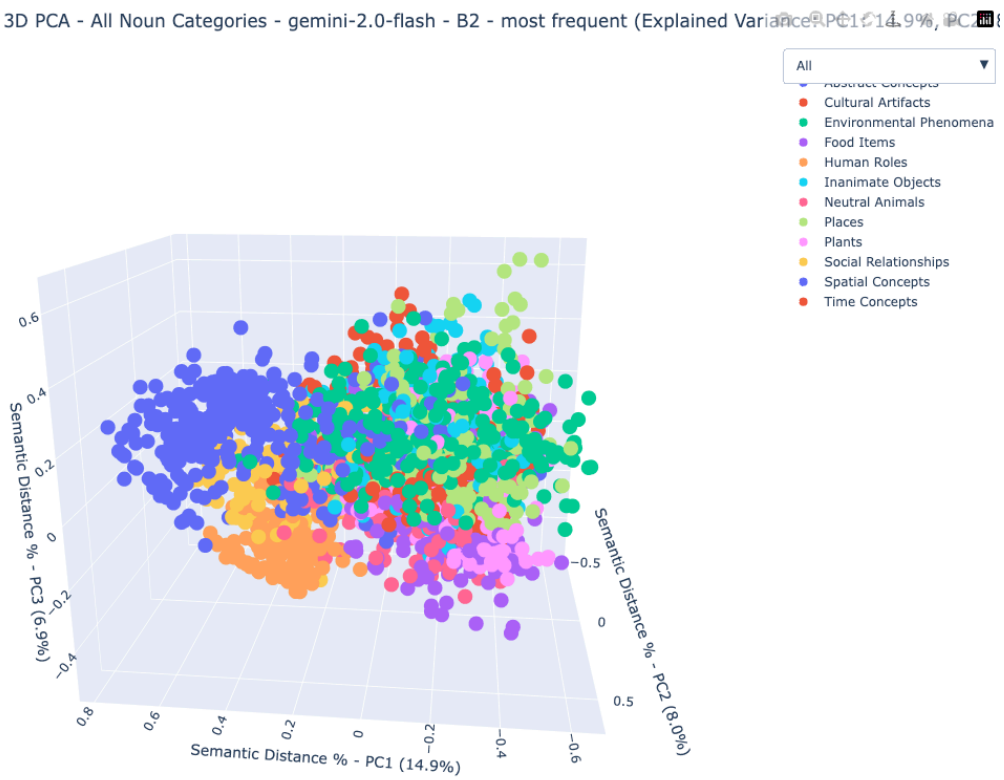


Fig. 15 3D PCA interactive semantic space of each nouns position (Gemini-2.0)

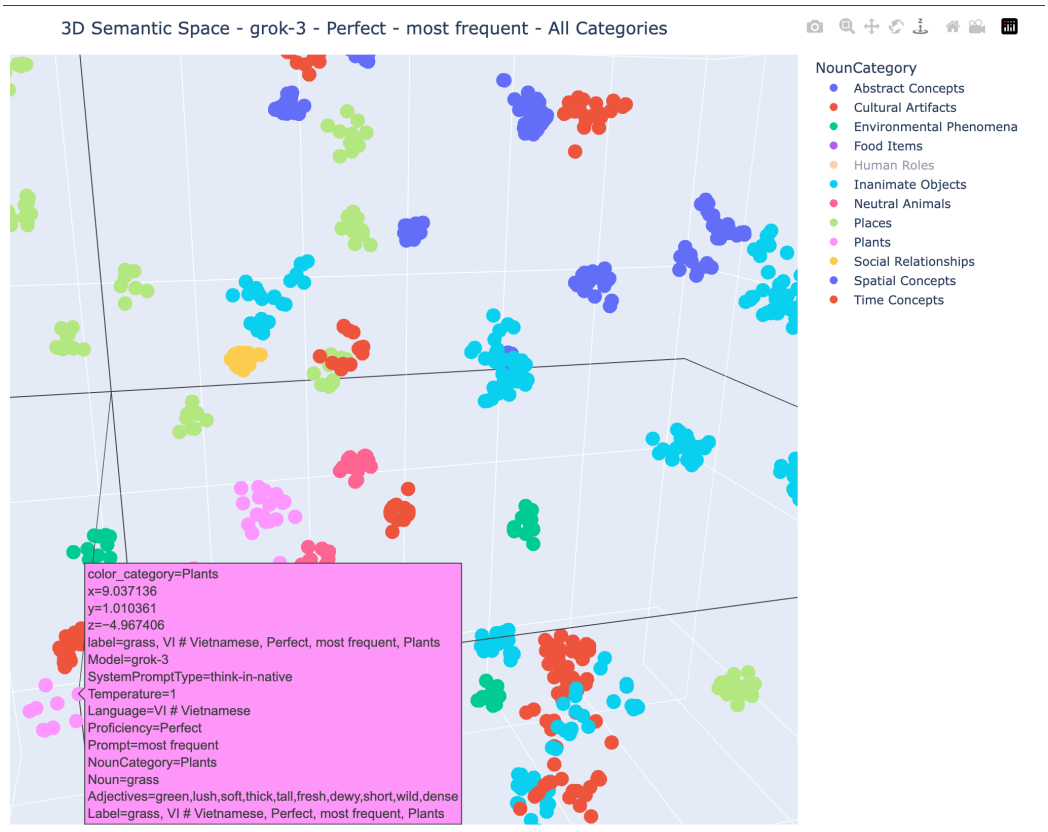


Fig. 16 3D UMAP interactive semantic space of each nouns position (Grok-3)

3.5.3 Language-level analysis

The above mentioned analyses are also done on the language level, as illustrated with the figures and tables provided below. For instance, the cosine-Jaccard distance correlation for each of the 24 languages is plotted for all LLM generated data, across all proficiencies, as an example from Japanese is depicted in (Fig. 17).

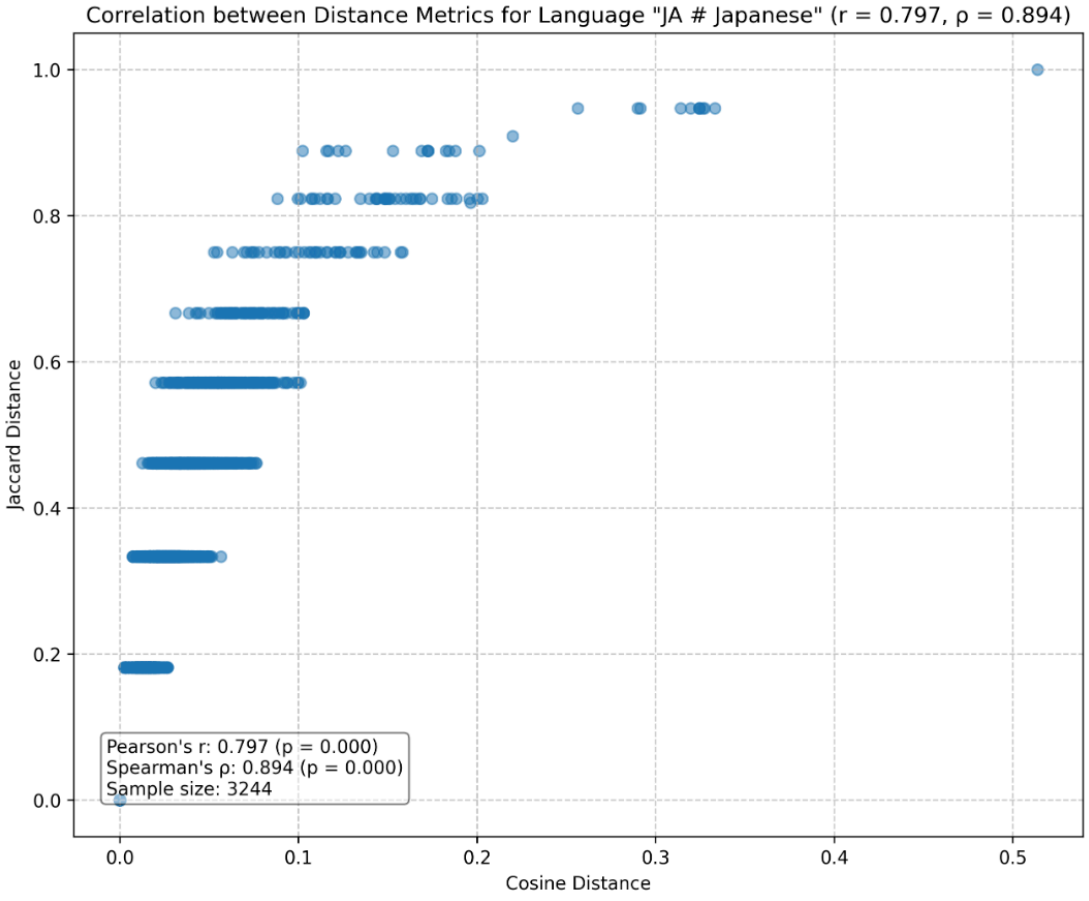


Fig. 17 Cosine-Jaccard pairs for all nouns in Japanese (o3-mini, all proficiencies)

These cosine-Jaccard correlations are also provided for all 276 language pairs (Fig. 18), with each data point representing the position for the mean value for the cosine distance between the language pair and the mean value for the Jaccard distance between the language pair. All of these graphs have interactive interfaces that enables hovering mechanics to examine the details for each data point.

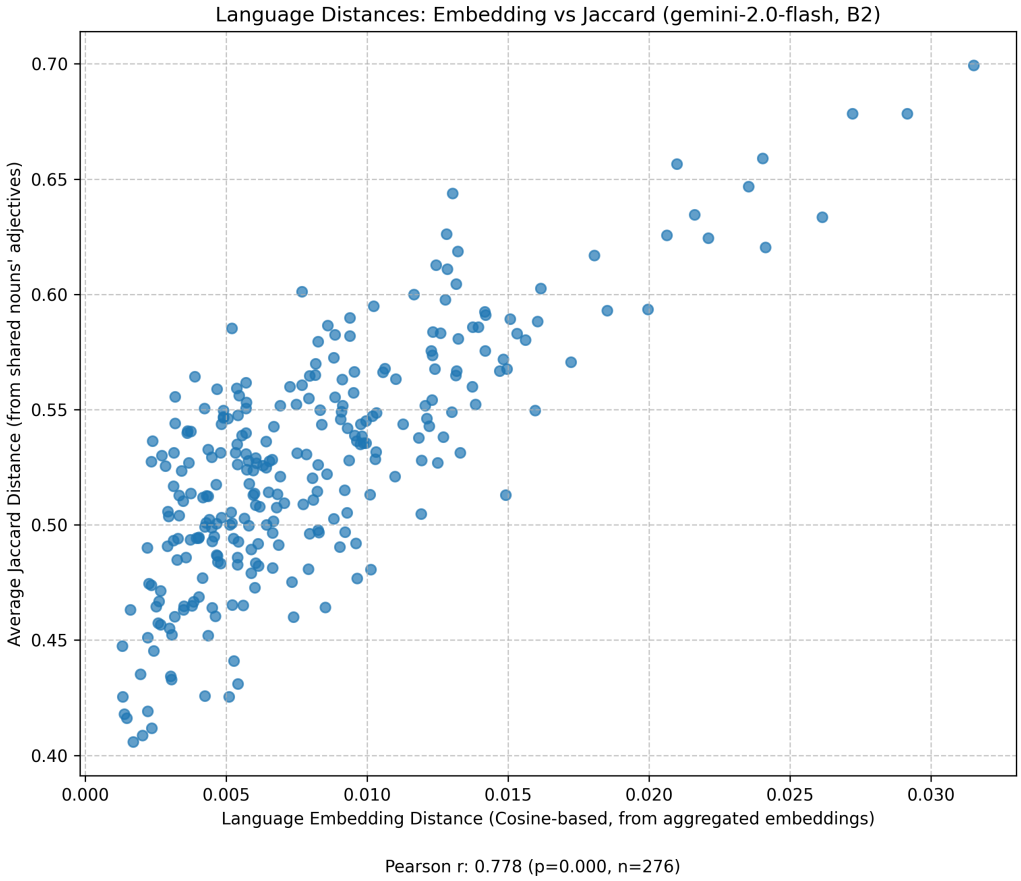


Fig 18. Cosine-Jaccard pairs for all language pairs (Gemini 2.0, B2)

Furthermore, distances between languages may be visualized in 2D and 3D semantic spaces with color coded language data points as aggregates of all of the 144 nouns embeddings for each language in (Figures 19 and 20), with a table of the distance between all languages found in (Table 9). These figures and tables build up upon the foundation laid in with the lexical items and noun categories in the previous section (3.5.2) by abstracting up a level to be able to view languages as distinct entities and understand the distance between them based on the generated data points. The relevant tables include the cosine distances between languages and implement Cohen's d to retrieve a reliable metric as to how large the conceived effect between language's perceptions are.

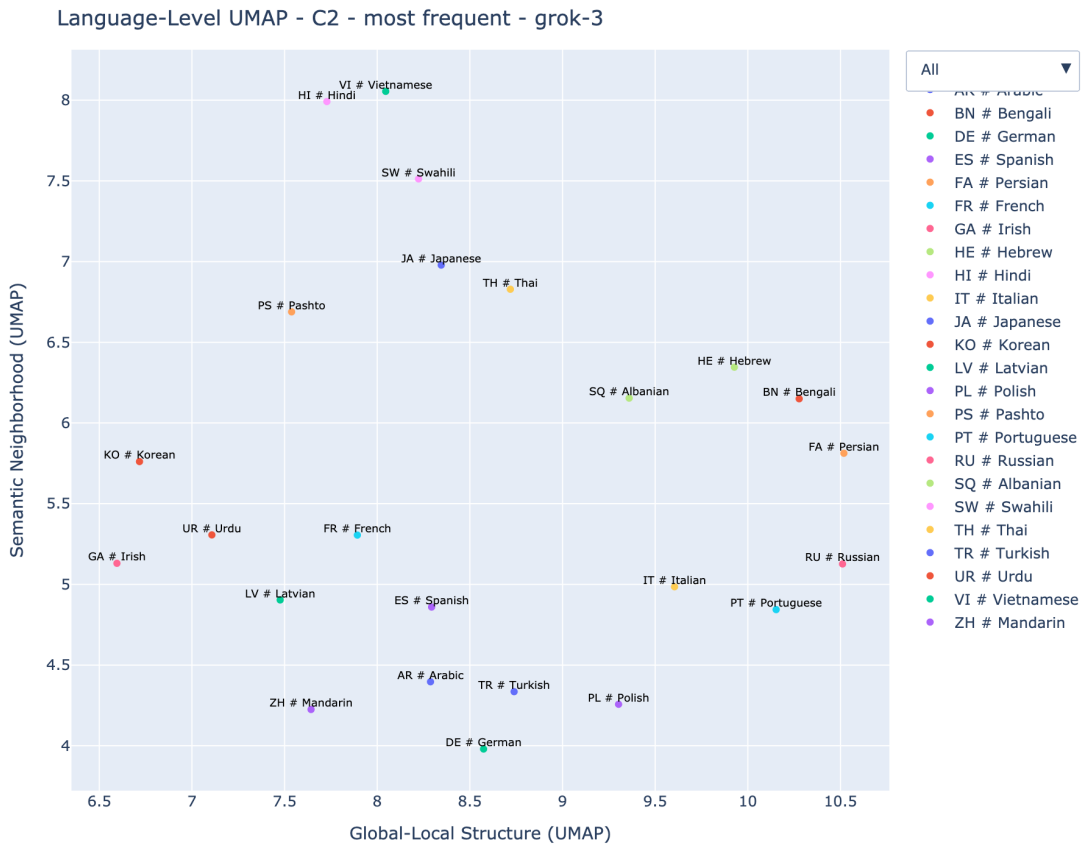


Fig. 19 2D UMAP interactive semantic space of languages (Grok-3)

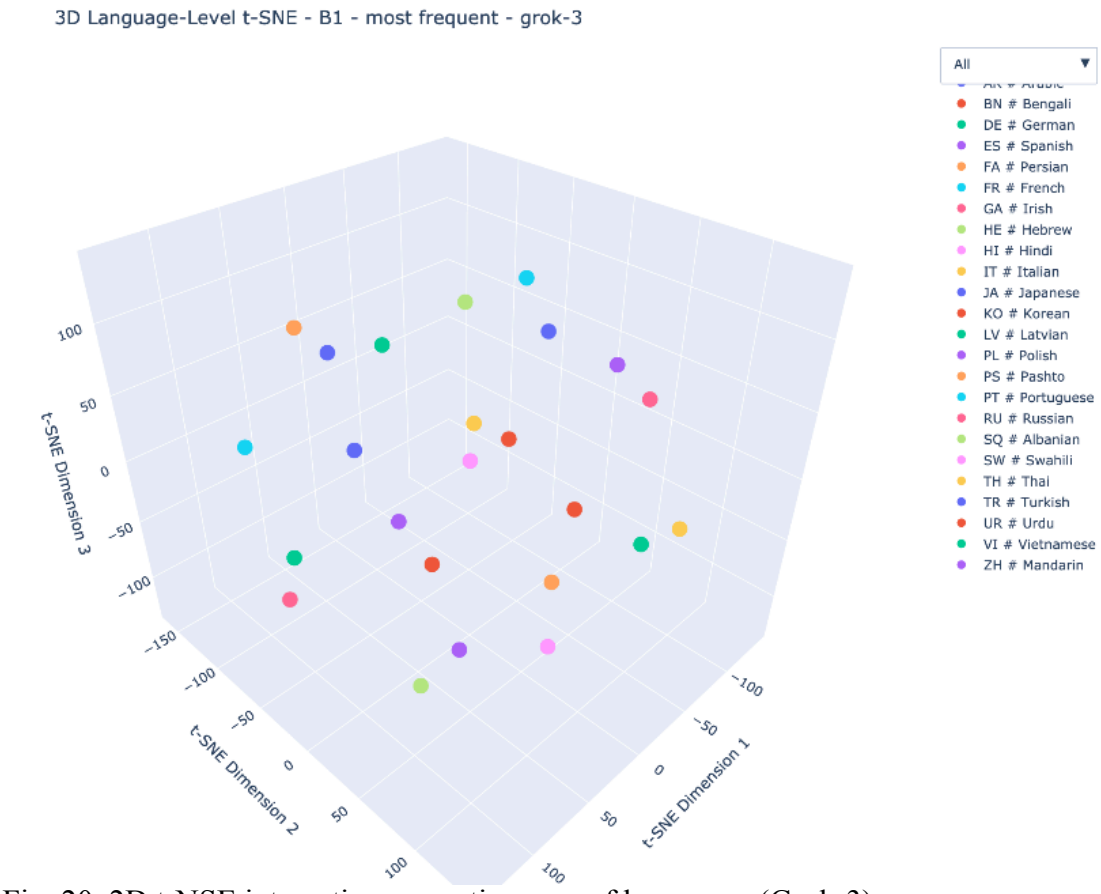


Fig. 20 2D t-NSE interactive semantic space of languages (Grok-3)

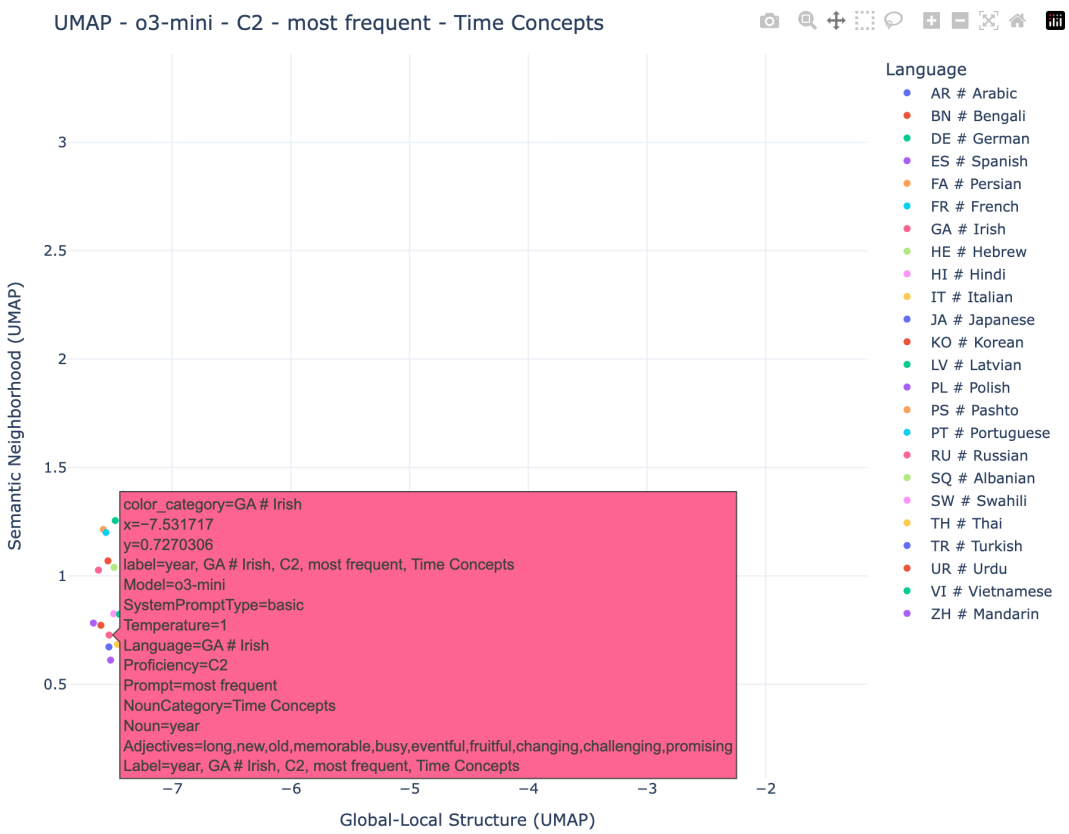


Fig. 21 2D UMAP interactive language space for “Time Concepts” (o3-mini, C2)

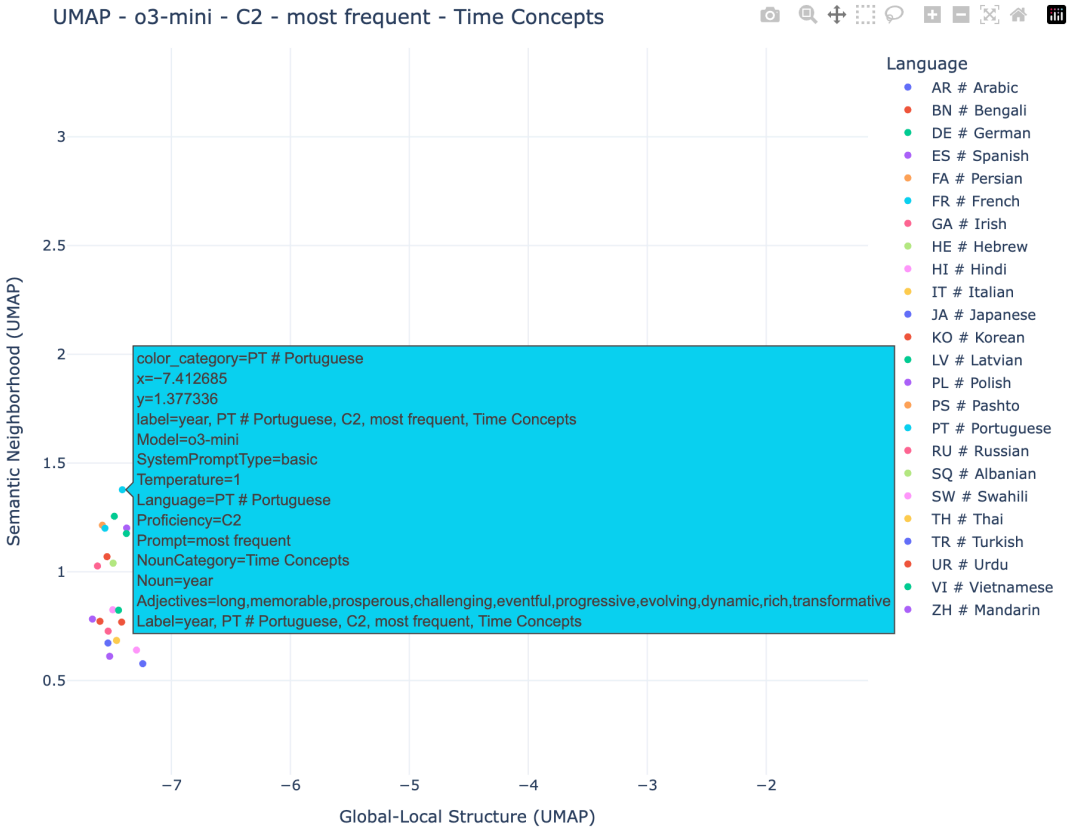


Fig. 22 2D UMAP interactive language space for “Time Concepts” (o3-mini, C2)

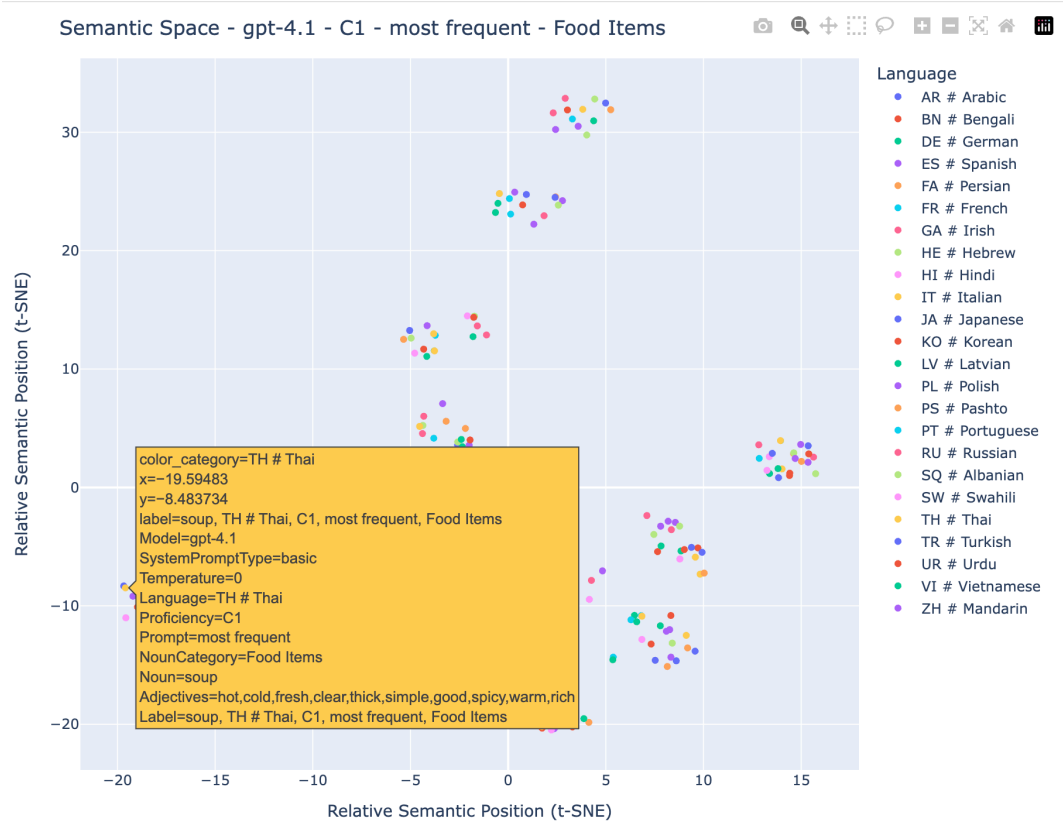


Fig. 23 2D t-NSE interactive semantic space of languages—"Food Items" (Grok-3)

Here, we will provide a sample of language pairs (out of 276 comparisons between languages, from among the thousands of comparisons done at various proficiency, adjective prompt, temperature, and LLM levels), to demonstrate a few examples of cross-linguistic distances (Table 8). Followed by (Table 8), in which we demonstrate how far each language falls from their central point of average.

Table 8. Cross-Linguistic Distances (Gemini-2.0, All Proficiencies Consolidated)

Language Pair	Language Pair Cosine Distance	Language Pair Cohen's d ($\mu = 0.0592$, SD = 0.0443)	Cohen's d Significance
Albanian - Hindi	0.1620	2.3205	Large
Arabic - French	0.1048	1.0293	Large
Arabic - Hebrew	0.05589	0.0747	Small
Turkish - Korean	0.0434	0.3567	Medium

Table 9. Each Language's Distance to Collective Mean Distance (Grok-3 mini, P)

Language	Language Cosine Distance (mean)	Cohen's d (Specific Language — All Languages Mean)	Cohen's d (Significance Metric)
Arabic	0.0340	0.0686	Negligible
Bengali	0.0361	0.0068	Negligible
German	0.0348	0.0420	Negligible
Spanish	0.0359	0.0000	Negligible
Persian	0.0358	0.0037	Negligible
French	0.0357	0.0079	Negligible
Irish	0.0384	0.0906	Negligible
Hebrew	0.0401	0.1567	Nearly Small
Hindi	0.0360	0.0035	Negligible
Italian	0.0373	0.0458	Negligible
Japanese	0.0380	0.0753	Negligible
Korean	0.0355	0.0142	Negligible
Latvian	0.0379	0.0749	Negligible
Polish	0.0330	0.1098	Negligible
Pashto	0.0354	0.0177	Negligible
Portuguese	0.0327	0.1255	Nearly Small
Russian	0.0331	0.1053	Nearly Small
Albanian	0.0378	0.0594	Negligible
Swahili	0.0334	0.0842	Negligible
Thai	0.0338	0.0682	Negligible
Turkish	0.0376	0.0594	Negligible
Urdu	0.0368	0.0310	Negligible
Vietnamese	0.0383	0.0708	Negligible
Mandarin	0.0327	0.1185	Nearly Small

Note: Cohen's d is evaluated as $d = (M - \mu) / SD$, where M is the mean cosine distance for that language, and $\mu = 0.0359$ (all languages average distance). SD is the mean pooled standard deviation of all languages average (0.0284).

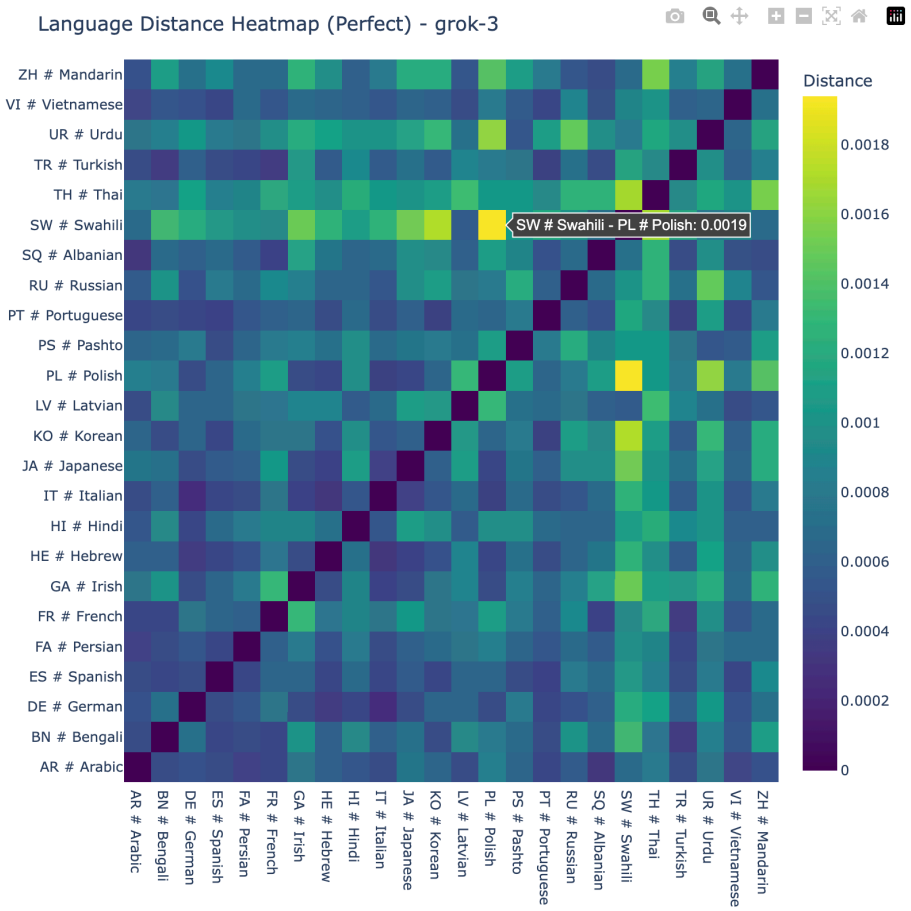


Fig. 24 Language distance heat map (Grok-3, Perfect)

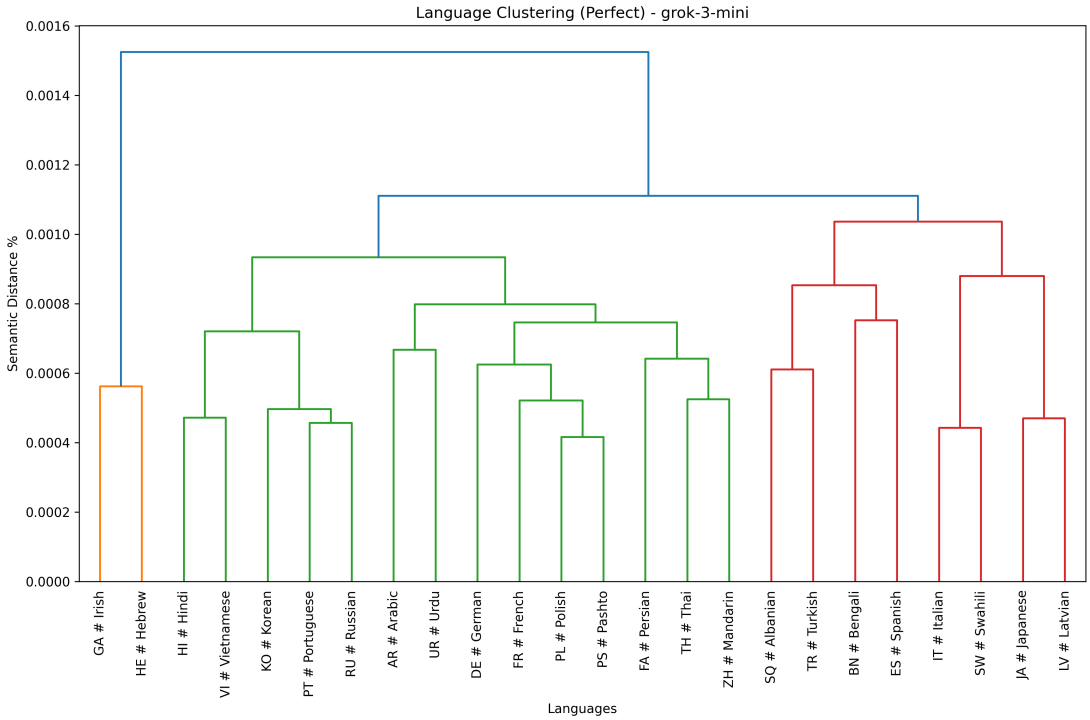


Fig. 25 Language distance dendrogram (Grok-3 mini, Perfect)

3.5.4 Language family level analysis

The same evaluations are conducted at the language family level to assess any possible patterns that languages belonging to common families may exhibit.

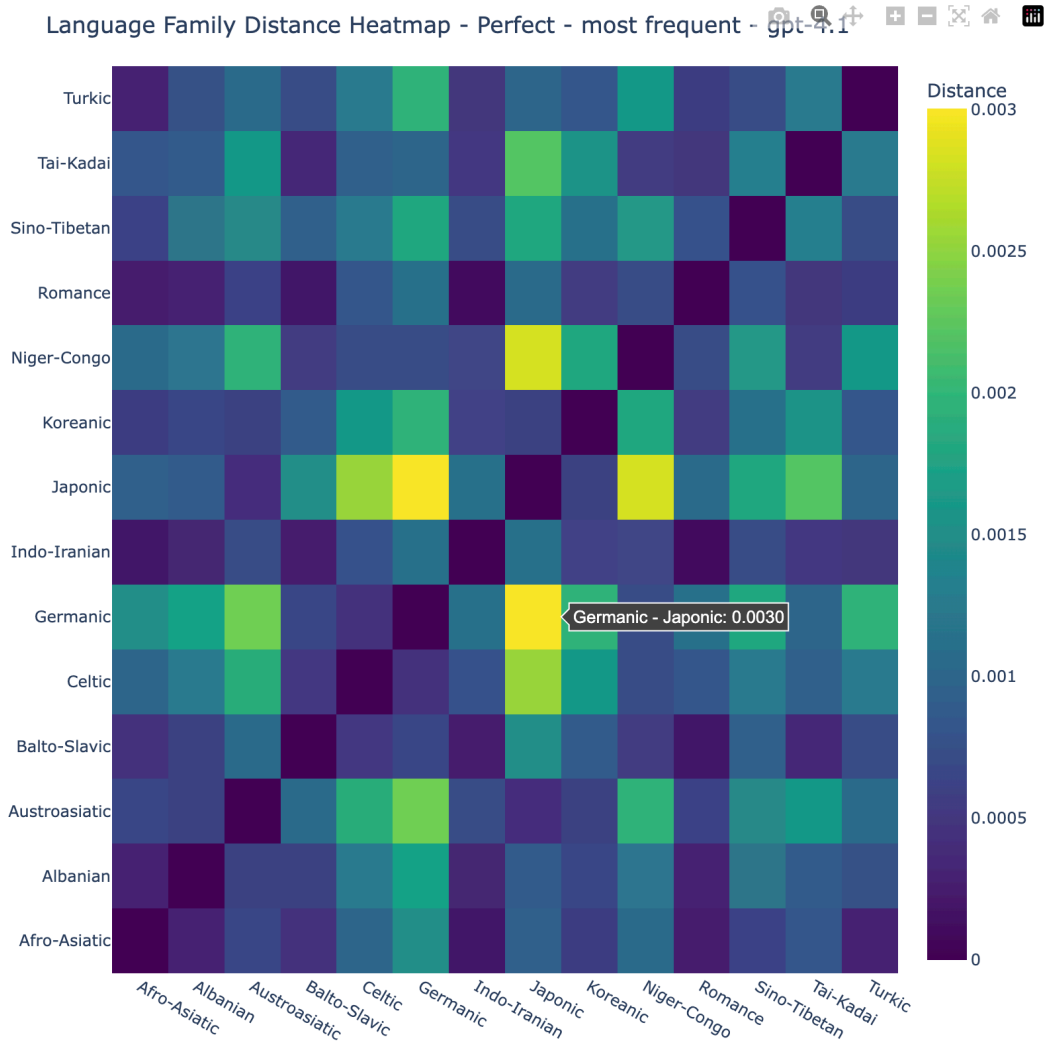


Fig. 26 Language family distance heat map (GPT-4.1, Perfect, 1T)

Languages belonging to each language family may be found at (Appendix A). It may be observed that Germanic and Japonica are relatively closer together than other families (Fig. 26). The degree to which languages cluster together under the roof of their families are represented as a single point in space (Fig. 27), demonstrating the extent to which they differ in cross-linguistic family terms.

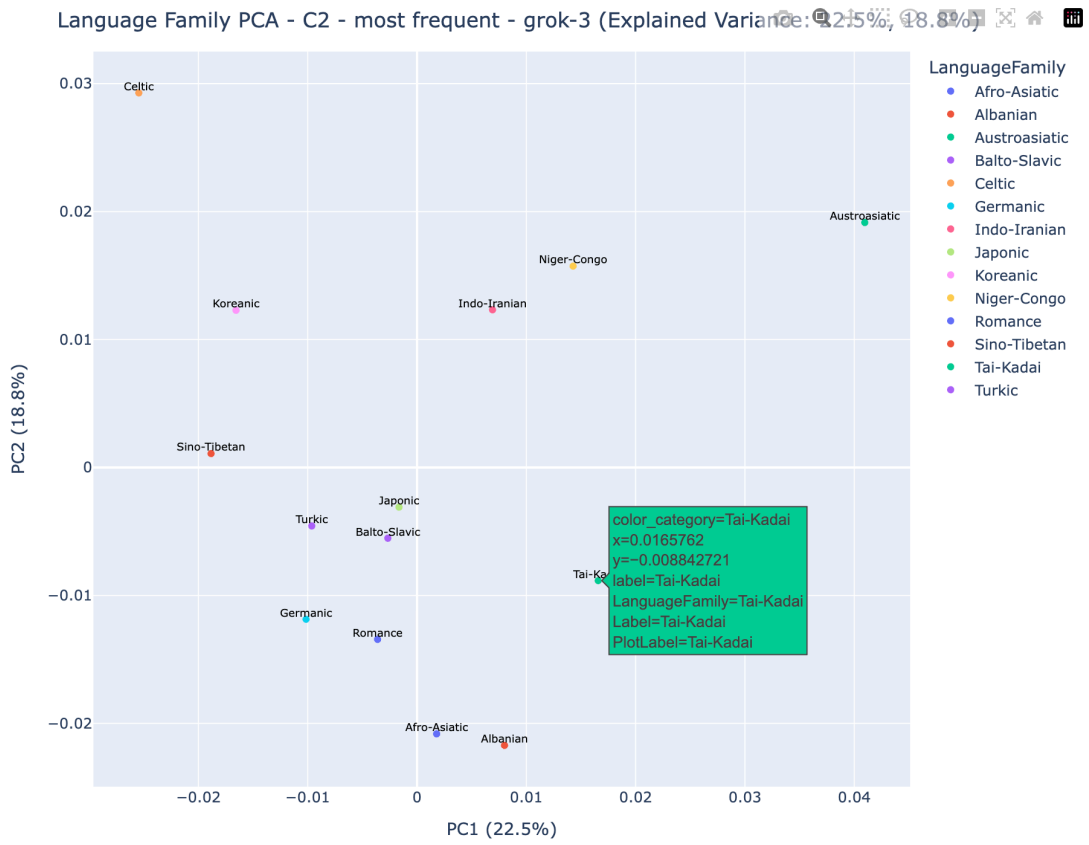


Fig. 27 2D PCA Language family distance semantic map (Grok-3, C2)

3.5.5 Grammatical gender cross-linguistic analysis

Grammatical gender studies were also conducted, mapping the semantic space of how languages possessing varying grammatical genders groups (from among the four main groups of non-gendered, minimal-gendered, dual-gendered, tri-gendered languages) and each individual language described each particular noun, which may be found on (GitHub). Here we provide a few visuals from among the wide variety found at the output folders for grammatical gender analysis on GitHub to sample the results. Furthermore, tables consisting of comparisons at the noun level, language level and grammatical group level are provided as a basis for interpreting the results to assess whether and to what extent grammatical gender in languages influences thoughts and perception. Cohen's d is utilized as a significance evaluation metric.

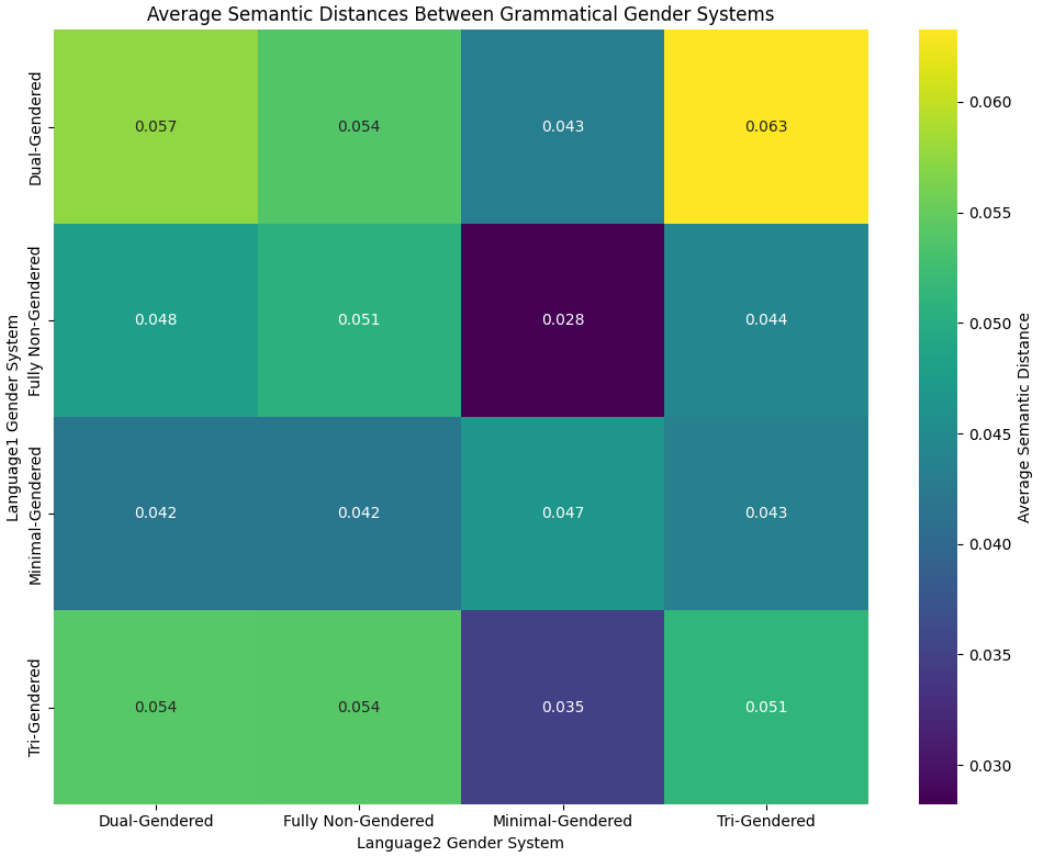


Fig. 28 Grammatical gender system distances heat map (Gemini-2.0, All P.)

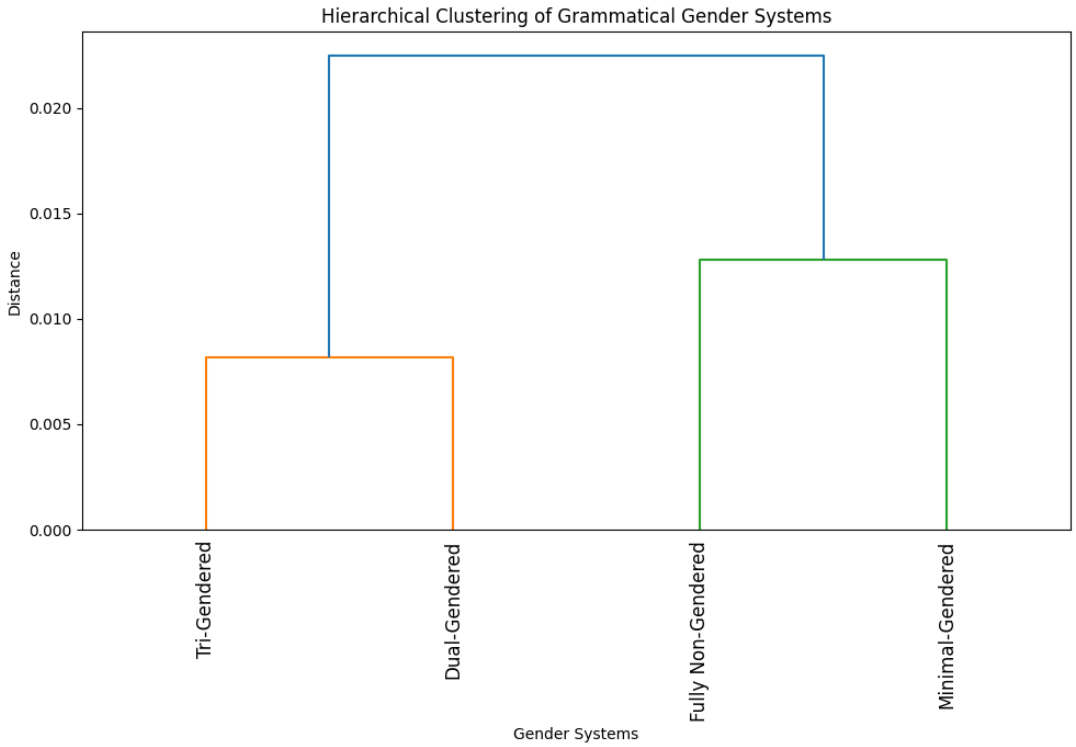


Fig. 29 Grammatical gender system distances dendrogram (GPT-4.1, All P.)

The specific configurations that the following tables are based on may be found at (Appendix E). Comprehensive analyses may be found on (GitHub), with a few samples of analysis formats provided in below.

Table 10. Grammatical Gender Category Distances (Grok-3 mini, All P.)

Gender Category	Distance (Mean)	Cohen's d	Significance
Non-Gendered vs. Non-Gendered	0.0460	0.9348	Large effect
Non-Gendered vs. Minimal-Gendered	0.0600	2.3239	Large effect
Non-Gendered vs. Dual-Gendered	0.0462	0.0696	Negligible effect
Non-Gendered vs. Tri-Gendered	0.0442	0.0610	Negligible effect
Minimal-Gendered vs. Tri-Gendered	0.0576	2.3906	Large effect
Minimal-Gendered vs. Dual-Gendered	0.0578	2.1208	Large effect
Dual-Gendered vs. Dual-Gendered	0.0463	0.8629	Large effect
Dual-Gendered vs. Tri-Gendered	0.0459	0.1870	Negligible effect
Tri-Gendered vs. Tri-Gendered	0.0448	0.3595	Medium effect

It may be noted in (Table 10) that within-system grammatical gendered language (languages belonging to the same grammatical group; e.g. Arabic and French are both dual-gendered) groups resulted in medium to large distinctions; whereas cross-system comparisons resulted in a wider range of distinctions.

Table 11. Sample Grammatical Gender Language Level Distances (GPT-4.1, All P.)

Gender Pair	Language Pair	Mean Distance	Cohen's d
Non-Gendered vs. Non-Gendered	Mandarin vs. Thai	0.0698	0.1973, Negligible
Non-Gendered vs. Minimal-Gendered	Bengali vs. Korean	0.0654	0.1764, Negligible
Non-Gendered vs. Dual-Gendered	Japanese vs. French	0.0880	0.2837, Small
Non-Gendered vs. Tri-Gendered	Japanese vs. German	0.0638	0.1689, Negligible
Minimal-Gendered vs. Dual-Gendered	Bengali vs. Irish	0.0538	0.1214, Negligible
Minimal-Gendered vs. Tri-Gendered	Bengali vs. Russian	0.0583	0.1428, Negligible
Dual-Gendered vs. Dual-Gendered	Hindi vs. Pashto	0.0340	0.0275, Negligible
Dual-Gendered vs. Tri-Gendered	Italian vs. Latvian	0.0001	0.1334, Negligible
Tri-Gendered vs. Tri-Gendered	German vs. Polish	0.0187	0.0451, Negligible

Note: Average distances between grammatical gendered languages ($\mu = 0.0282$) and their standard deviation is (0.2107).

These distances and significance metrics (of Table 11) represent a sample of the cross-linguistic grammatical gender comparisons. 20% is considered the threshold for a small significance, which the Japanese—French pair shows. While (Table 10) provides a more general overview of the effects of grammatical gender on perception, precise cross-linguistic comparisons allow the probing of each distinct language pair.

Finally, a noun level representation is provided for grammatical gender distinctions, in part to assess the accuracy of some of the experiments conducted by researchers (Boroditsky, 2003; Elpers 2022; Konishi 1993; Mahajan 2025; Mickan 2014; Sato 2018) as detailed in (Section 2.2.2), with the computational framework that we have devised. Mickan's critique: "Key is a llave is a Schlüssel: A failure to replicate an experiment from Boroditsky et al. 2003" provides 10 nouns for which grammatical gender testing is conducted between German and Spanish. Our bilingual representational models have been trained on embeddings for five nouns that intersect with those of Mickan's study. Furthermore, the two nouns that are mentioned in Boroditsky's study are also added to this list in (Table 12).

Table 12. Reproduction of Boroditsky and Mickan's Research (Grok-3 mini, Perfect)

Nouns	Our Cosine Distance	Our Cohen's d
apple	0.0274	1.5476, Large
bridge	0.0171	0.9656, Large
key	0.0462	2.6096, Large
moon	0.0164	0.9262, Large
mouse	0.0562	3.1746, Large
sun	0.0156	0.8808, Large
whale	0.0580	3.2751, Large

The LM configuration for this study replication was done in consistency with the original methodologies used; implementing an implicit system prompt, where the model is not explicitly told to think in a native language, but rather just told that it is a native speaker of German and Spanish with English (assuming perfect proficiency) as a medium for generating adjectives to describe the provided noun (Temp = 0).

Cohen's d is calculated here as $d = (M - \mu) / SD$, where M is the cross-linguistic cosine distance between German and Spanish for the provided noun, μ is the reference value of 0 (which assumes there is no difference in perception of given noun due to grammatical gender), and SD = 0.0177 is the standard deviation of the seven noun's cosine distance values.

Note that our model is substantially richer and more comprehensive than theirs on multiple aspects, including the use of ten adjectives for each of the dozen nouns in a dozen noun categories for a total of 144 nouns representing each language for 24 languages consisting of non-gendered, minimal-gendered, dual-gendered and tri-gendered grammatical structures for languages. Then each of the nouns in each language are compared with one another for a total of 110,838 entries for non-gendered languages, 15,837 entries for minimal-gendered languages, 158,349 entries for dual-gendered languages and 79,164 entries for tri-gendered languages for a grand total of 364,188 noun comparisons. Cohen's d effect magnitude was provided in (Table 10). Here we provide the ANOVA results of our model (Table 13) in contrast to the findings of the aforementioned study (Mickan, 2014). In their 2014 study, Mickan et al. employed a 2x2 ANOVA design to investigate whether grammatical gender influenced adjective associations in German and Spanish, using total and primary response values as measures. Their results showed no significant effect of grammatical gender, with total response values yielding $F(1, 28) = 3.194$, $p = 0.0847$ (by-subject) and $F(1, 8) = 0.611$, $p = 0.457$ (by-stimulus), and primary response values showing $F(1, 28) = 0.013$, $p = 0.912$ (by-subject) and $F(1, 8) = 0.012$, $p = 0.915$ (by-stimulus). They concluded there was no evidence for a spillover effect of grammatical gender on perception. In contrast, our gender language analysis

utilized ANOVA to examine cosine distance across four grammatical gender categories—Non-Gendered, Minimal-Gendered, Dual-Gendered, and Tri-Gendered—followed by Tukey HSD tests and Cohen's d effect sizes. Our ANOVA revealed a highly significant difference, with $F(3, 182160) = 996.9149$, $p = 0.0000$. Tukey HSD tests indicated significant differences for most pairwise comparisons, such as Dual-Gendered vs Minimal-Gendered ($p = 0.0000$), except for Dual-Gendered vs Non-Gendered ($p = 0.874$). Cohen's d values showed small to negligible effects, for instance, $d = -0.447$ for Dual-Gendered vs Minimal-Gendered and $d = 0.419$ for Minimal-Gendered vs Tri-Gendered, suggesting an overall modest influence of grammatical gender on language-based thinking.

Table 13. ANOVA Results of Our Computational Framework (Grok-3 mini, All P.)

Statistic	Study Results (Mickan, 2014)	(Our) Computational Framework Results
ANOVA (Gender Effect)	Total response values: $F(1, 28) = 3.194$, $p = 0.0847$ (by-subject) Primary response values: $F(1, 28) = 0.013$, $p = 0.912$ (by-subject)	$F(3, 182160) = 996.9149$, $p = 0.0000$
Significance Differences	No significant differences found ($p > 0.05$)	Significant differences found between grammatical gender categories ($p = 0.0000$)
Pairwise Comparisons	Not reported (implied non-significant based on ANOVA)	Significant for most pairs ($p = 0.0000$)
Effect Sizes (Cohen's d)	Not reported	Negligible to small (nearly medium) effect (0.026 to 0.447)
Effect Sizes (Eta-Squared)	Not reported	1.0000 (Large effect)
Interpretation	No evidence for influence of grammatical gender on thought	Modest differences suggest a subtle influence of language on thought

Unlike Mickan et al. (2014), who found no evidence of grammatical gender influencing perception, our analysis reveals statistically significant differences across grammatical gender categories. Although the effect sizes are small, these findings

suggest a subtle influence of language on thought, aligning with the idea posited by LR. This contrast highlights the value of exploring diverse linguistic structures and larger datasets, as in our analysis, to uncover nuanced effects that may not emerge in smaller-scale studies.

3.5.6 Cross-model, temperature and proficiency analyses

We mentioned that five different LLMs were used for API calls to generate adjectives to represent bilingual representations of nouns. This is to ensure a greater sample pool to add a rigorous layer of diversity and variation for our data points. A sample of the distinct results based on the various model providers are provided in (Fig. 30) below. The observed differences show that all differences between LLM providers Cohen's d significance metric is negligible (<20%). Their proximity to one another and variance is at a healthy balanced level, working to ensure that the results aren't skewed and the breadth of sample alternatives along with their vast size ensures rigorous semantic mapping.

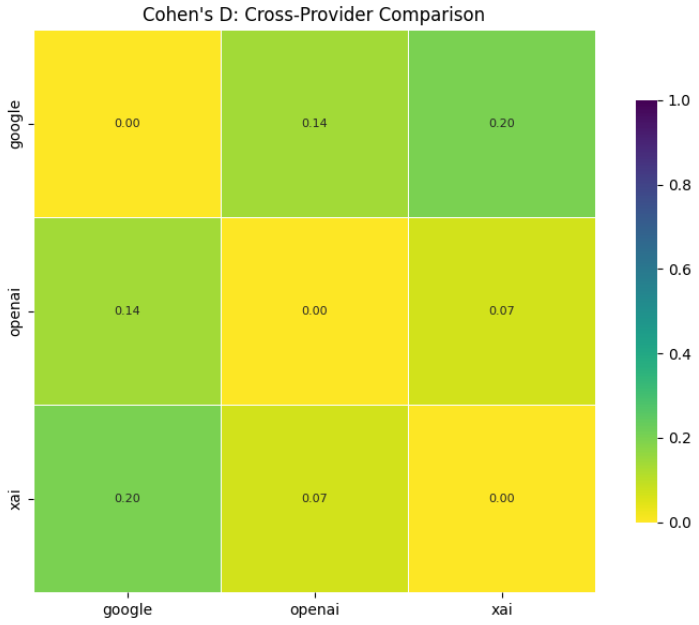


Fig. 30 Cross-LLM provider Cohen's d heat map

Table 14. Cohen's d for Cross-Model Language Mean Distances (Perfect, Temp = 0)

Language	GPT-4.1	Grok-3	Grok-3 mini
Arabic	0.077 (negligible)	0.008 (negligible)	0.111 (negligible)
Bengali	0.140 (negligible)	0.014 (negligible)	0.074 (negligible)
German	0.306 (small)	0.021 (negligible)	0.036 (negligible)
Spanish	0.034 (negligible)	0.100 (negligible)	0.007 (negligible)
Persian	0.186 (negligible)	0.044 (negligible)	0.005 (negligible)
French	0.012 (negligible)	0.023 (negligible)	0.064 (negligible)
Irish	0.169 (negligible)	0.108 (negligible)	0.020 (negligible)
Hebrew	0.160 (negligible)	0.002 (negligible)	0.110 (negligible)
Hindi	0.114 (negligible)	0.201 (small)	0.076 (negligible)
Italian	0.117 (negligible)	0.045 (negligible)	0.049 (negligible)
Japanese	0.169 (negligible)	0.041 (negligible)	0.069 (negligible)
Korean	0.043 (negligible)	0.015 (negligible)	0.057 (negligible)
Latvian	0.017 (negligible)	0.135 (negligible)	0.088 (negligible)
Polish	0.214 (small)	0.107 (negligible)	0.188 (negligible)
Pashto	0.003 (negligible)	0.243 (small)	0.035 (negligible)
Portuguese	0.151 (negligible)	0.025 (negligible)	0.010 (negligible)
Russian	0.073 (negligible)	0.043 (negligible)	0.097 (negligible)
Albanian	0.096 (negligible)	0.067 (negligible)	0.032 (negligible)
Swahili	0.087 (negligible)	0.166 (negligible)	0.184 (negligible)
Thai	0.043 (negligible)	0.122 (negligible)	0.202 (small)
Turkish	0.086 (negligible)	0.055 (negligible)	0.013 (negligible)
Urdu	0.050 (negligible)	0.080 (negligible)	0.117 (negligible)
Vietnamese	0.009 (negligible)	0.039 (negligible)	0.267 (small)
Mandarin	0.049 (negligible)	0.142 (negligible)	0.057 (negligible)

The results from (Table 14) also illustrate the same effect of the various LLM providers as (Fig. 30), that different LLMs provide a rich sampling pool, without producing results with large variances, as expected.

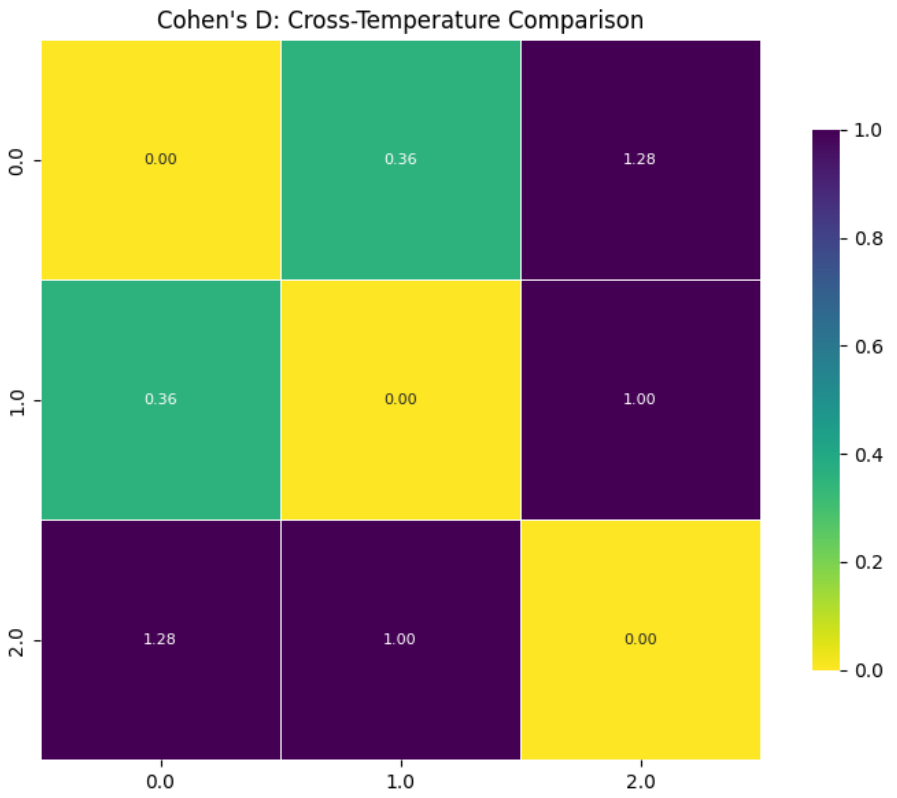


Fig. 31 Cross-temperature Cohen's d

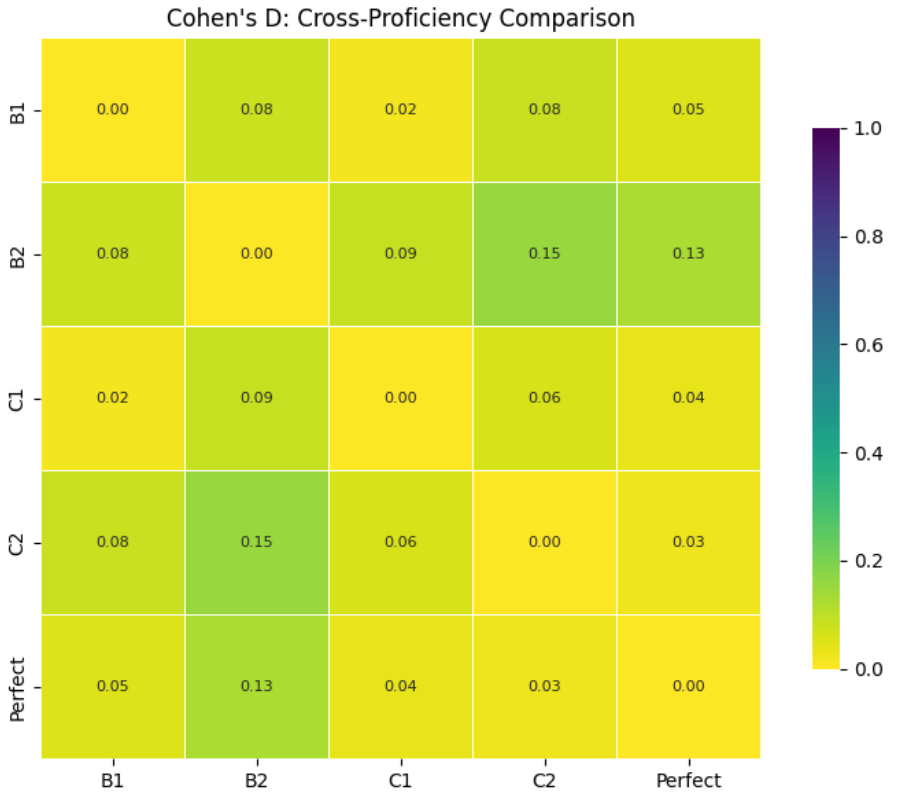


Fig. 32 Cross-proficiency Cohen's d

Table 15. Cohen's d: Cross-Proficiency Language Mean Distances (BB, T=0)

Language	B1	C1	Perfect
Arabic	0.193 (negligible)	0.080 (negligible)	0.077 (negligible)
Bengali	0.207 (small)	0.040 (negligible)	0.007 (negligible)
German	0.098 (negligible)	0.020 (negligible)	0.038 (negligible)
Spanish	0.108 (negligible)	0.013 (negligible)	0.028 (negligible)
Persian	0.148 (negligible)	0.083 (negligible)	0.016 (negligible)
French	0.041 (negligible)	0.077 (negligible)	0.046 (negligible)
Irish	0.010 (negligible)	0.122 (negligible)	0.015 (negligible)
Hebrew	0.112 (negligible)	0.022 (negligible)	0.069 (negligible)
Hindi	0.135 (negligible)	0.058 (negligible)	0.025 (negligible)
Italian	0.126 (negligible)	0.005 (negligible)	0.015 (negligible)
Japanese	0.059 (negligible)	0.008 (negligible)	0.055 (negligible)
Korean	0.001 (negligible)	0.028 (negligible)	0.045 (negligible)
Latvian	0.133 (negligible)	0.015 (negligible)	0.066 (negligible)
Polish	0.095 (negligible)	0.089 (negligible)	0.087 (negligible)
Pashto	0.066 (negligible)	0.054 (negligible)	0.017 (negligible)
Portuguese	0.042 (negligible)	0.031 (negligible)	0.018 (negligible)
Russian	0.109 (negligible)	0.059 (negligible)	0.054 (negligible)
Albanian	0.112 (negligible)	0.103 (negligible)	0.046 (negligible)
Swahili	0.015 (negligible)	0.062 (negligible)	0.136 (negligible)
Thai	0.028 (negligible)	0.010 (negligible)	0.101 (negligible)
Turkish	0.165 (negligible)	0.038 (negligible)	0.014 (negligible)
Urdu	0.082 (negligible)	0.079 (negligible)	0.070 (negligible)
Vietnamese	0.029 (negligible)	0.026 (negligible)	0.134 (negligible)
Mandarin	0.269 (small)	0.063 (negligible)	0.019 (negligible)

Note: Represents the aggregate of all LLM providers conducted with the system prompt of BB: Basic Bilingual; that is the system prompt implicitly activates the native language of the speaker to attempt to replicate a subconscious effect.

Table 16. Cohen's d: Cross-Proficiency Language Mean Distances (TN, T=0)

Language	B1	C1	Perfect
Arabic	0.410 (small)	0.142 (negligible)	0.111 (negligible)
Bengali	0.486 (small)	0.070 (negligible)	0.074 (negligible)
German	0.264 (small)	0.013 (negligible)	0.036 (negligible)
Spanish	0.187 (negligible)	0.005 (negligible)	0.007 (negligible)
Persian	0.238 (small)	0.113 (negligible)	0.005 (negligible)
French	0.038 (negligible)	0.214 (small)	0.064 (negligible)
Irish	0.160 (negligible)	0.028 (negligible)	0.020 (negligible)
Hebrew	0.193 (negligible)	0.015 (negligible)	0.110 (negligible)
Hindi	0.284 (small)	0.102 (negligible)	0.076 (negligible)
Italian	0.252 (small)	0.006 (negligible)	0.049 (negligible)
Japanese	0.074 (negligible)	0.044 (negligible)	0.069 (negligible)
Korean	0.034 (negligible)	0.078 (negligible)	0.057 (negligible)
Latvian	0.290 (small)	0.002 (negligible)	0.088 (negligible)
Polish	0.121 (negligible)	0.153 (negligible)	0.188 (negligible)
Pashto	0.089 (negligible)	0.105 (negligible)	0.035 (negligible)
Portuguese	0.120 (negligible)	0.017 (negligible)	0.010 (negligible)
Russian	0.280 (small)	0.015 (negligible)	0.097 (negligible)
Albanian	0.128 (negligible)	0.192 (negligible)	0.032 (negligible)
Swahili	0.063 (negligible)	0.095 (negligible)	0.184 (negligible)
Thai	0.156 (negligible)	0.015 (negligible)	0.202 (small)
Turkish	0.280 (small)	0.089 (negligible)	0.013 (negligible)
Urdu	0.087 (negligible)	0.138 (negligible)	0.117 (negligible)
Vietnamese	0.059 (negligible)	0.085 (negligible)	0.267 (small)
Mandarin	0.526 (medium)	0.061 (negligible)	0.057 (negligible)

Note: Represents the aggregate of all LLM providers conducted with the system prompt of TN: Think-in-native; that is the system prompt explicitly activates the native language of the speaker to attempt to replicate a more pronounced effect.

The high variance in (Fig. 31) shows that the addition of randomness by adjusting the temperature of the LLM significantly changes the results. This is expected, as explained in (Section 3.2.5), only a temperature of 0 will give reliable statistical co-occurence frequency of a bilingual simulated speakers. However, results from various temperature-based LLM configurations (e.g. Table 7) also show a meaningful distinction between languages, as a supporting aspect for 0 temperature results from a different approach.

(Tables 15 and 16) provide interesting implications in terms of the medium language's (English) proficiency level in the process of the native language influencing thought, as well as the effect of the system prompt on the simulation of bilingual individuals. One of these implications is that it shows that the system prompt is working in terms of producing variation, since the only thing that changes in the LLM configuration is the proficiency and/or system prompt and this results in distinct results.

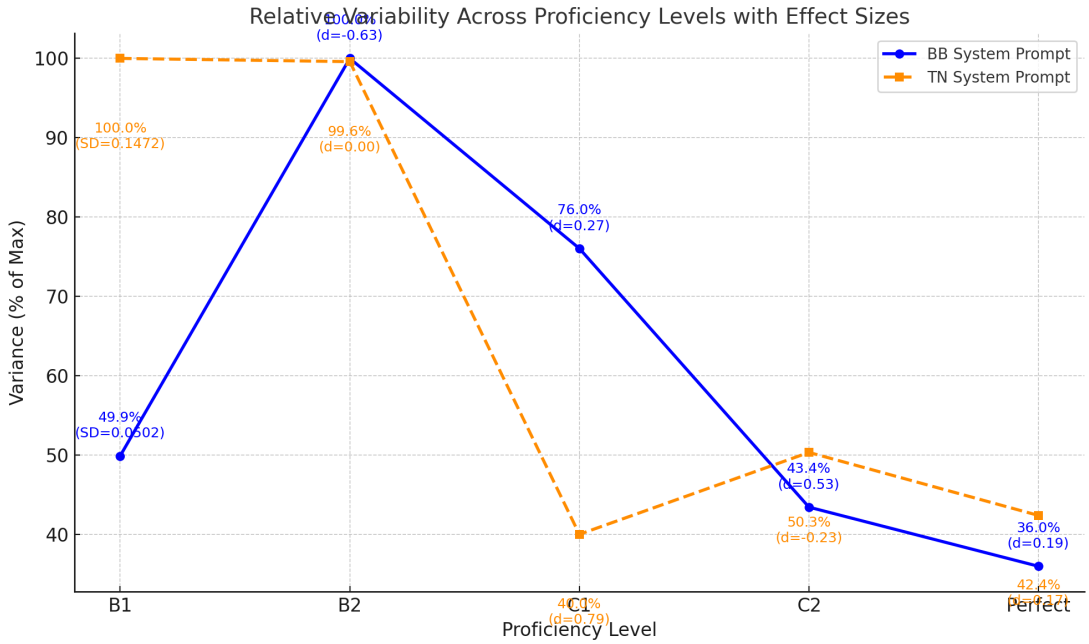


Fig. 33 Relative variability across proficiency levels for BB and TN system prompts

As the LLM's proficiency in English increases, the extent to which its output reflects influences from its native language as a spill-over effect gradually diminishes in both prompting conditions, though notably at differing rates and trajectories. In the Basic-bilingual prompting strategy, where the native language is implicitly activated, initial variability across language outputs begins at a modest level (standard deviation approximately 0.05) and after an initial spike, subsequently declines in a smooth and nearly linear manner. This pattern suggests a progressive attenuation of native-language influences: as the model's mastery of English strengthens, the implicit cross-linguistic traces become increasingly subdued, resulting in outputs that converge toward a more language-neutral representation. In contrast, the Think-in-native prompting approach explicitly instructs the model to engage cognitive processes first in the native language. Consequently, this explicit invocation significantly heightens initial variability, nearly tripling that observed in the BB condition at lower English proficiency levels (standard deviation approximately 0.15 at B1). However, upon reaching intermediate fluency (C1), this variability sharply decreases (Cohen's $d \approx 0.79$), underscoring that enhanced proficiency in English can substantially mitigate even explicit native-language priming effects. Although slight rebounds in variability at the highest proficiency levels are apparent, they eventually follow the general trajectory tendency explained above. Collectively, these findings indicate that greater proficiency in English substantially moderates the cognitive influence of the native language on the LLM's reasoning across both prompting conditions. Nevertheless, this mitigation is distinctly more pronounced and occurs at a later proficiency stage within the explicit TN context, contrasting with the gradual, consistent reduction observed in the implicit BB context.

CHAPTER 4

LINGUISTIC RELATIVITY EXPERIMENTS

In this chapter we provide a contribution to the LR debate by addressing the aforementioned limitations of former LR experiments by utilizing LLMs to conduct experiments that test LR on an unprecedented scale, simulating speakers of diverse languages to explore how linguistic differences influence cognition across multiple domains. Results are discussed after each section to assess the claim of LR in the context of these experiments.

The logic behind these experiments rests on the premise that if language influences thought, then LLMs configured to reflect distinct linguistic backgrounds should exhibit systematic variations in their responses to standardized tasks. By simulating bilingual speakers proficient in English and rooted in varied native languages, these experiments operationalize LR through a computational lens, enabling large-scale, replicable comparisons that transcend the logistical challenges of human-subject research. The tasks are designed to probe cognitive domains where linguistic influence is hypothesized to manifest, such as abstract reasoning, spatial navigation, and color perception, offering a multifaceted test of LR's core claims.

The experimentation leverages the aforementioned automated API pipeline (Section 3.2.6) to configure LLMs as proxies for bilingual cognition. Two system prompts drive the methodology: the Basic-bilingual prompt, which directs the model to respond directly in English, and the Think-in-native prompt, which instructs the model to process the task in its assigned native language before responding in

English. User prompts, tailored to each task, elicit responses that are systematically collected and analyzed to detect language-driven cognitive patterns.

This chapter is structured into three main subsections, each targeting a distinct cognitive domain to assess LR:

- 4.1 Multiple choice tasks: These experiments explore how linguistic properties, including grammatical gender and semantic organization, influence responses to abstract questions and perceptions of neutral entities, testing LR through scalable, interpretive tasks.
- 4.2 Spatial reasoning tasks: These tasks investigate how languages' spatial reference systems affect navigation and description, probing linguistic influences on spatial cognition.
- 4.3 Color reasoning tasks: These assess the impact of color terminology and cultural norms on color classification, emotion-color associations, and scenario-based choices, examining LR in visual perception and decision-making.

4.1 Cross-linguistic multiple choice experiments

The multiple choice experiments are organized into three subcategories, each targeting a specific dimension of LR:

- 4.1.1 Comprehensive multiple choice experiments: These investigate whether broad linguistic properties across diverse languages lead to systematic variations in responses to abstract, culturally nuanced questions.
- 4.1.2 Grammatical gender multiple choice experiments: These examine whether grammatical gender systems in native languages exert a subconscious influence on English responses, reflecting a "spillover effect" in concept associations.

- 4.1.3 Neutral animal experiments: These explore how native language gender assignments affect perceptions of gender-neutral animals in English, probing the interplay between cross-linguistic grammatical gender and behavioral interpretation.

The questions for each of these multiple choice experiments are available in (Appendix F). The TN (Think-in-native) prompt is the primary mechanism used as the system prompt here, simulating the cognitive process of reasoning in the native language before translating the response into English. This mirrors real-world bilingual cognition and aligns with LR's emphasis on native language influence. For comparative purposes, the BB (Basic-bilingual) prompt is also employed to evaluate responses when native-language thinking is not explicitly invoked, providing a baseline for isolating LR effects. The API calls pipeline configures LLMs to simulate speakers of diverse native languages (Appendix A). Each model is presented with a multiple-choice question formatted in JSON. The temperature is fixed at 0 to eliminate randomness, ensuring that responses reflect the model's bilingual representational function rather than stochastic variation. English proficiency is set to "Perfect" to handle the nuanced, abstract nature of the questions effectively. The experiment comprises 48 questions across 12 categories, with four questions per category (Appendix G). This breadth ensures comprehensive coverage of cognitive domains potentially influenced by linguistic structure. Responses are logged systematically, tagged with the native language and prompt type (TN or BB), and prepared for statistical analysis. Responses are systematically collected and subjected to statistical analysis.

A comprehensive statistical toolkit was implemented in the LR experiment pipeline (available on GitHub) and customized for the specific tasks outlined above, with each subsection providing justification for the selected tests to evaluate the probabilistic significance of variations among language model responses and their corresponding implications. The toolkit commences with assumption checks and descriptive diagnostics as follows:

The Shapiro–Wilk normality test (via scipy.stats.shapiro) is administered to each vector of language specific response proportions to ascertain the suitability of parametric procedures such as ANOVA; skewness, derived as the mean of cubed z-scores, and excess kurtosis, derived as the mean of z-scores raised to the fourth power minus three, furnish expeditious shape diagnostics that identify asymmetric or heavy tailed distributions, thereby prompting the adoption of non-parametric alternatives, with these metrics documented in each per language statistical profile on GitHub. Levene's test (via scipy.stats.levene) assesses homogeneity of variances by transforming observations into absolute deviations from the group mean or median and subjecting those values to an F-test to detect unequal dispersion that would compromise pooled-variance methods. D'Agostino's K^2 test (via scipy.stats.normaltest) converts sample skewness and kurtosis into a chi-square statistic with two degrees of freedom to uncover non-normal configurations that warrant non-parametric approaches. Global cross-linguistic variance is quantified by calculating the variance of each answer option across the 24 languages and averaging those variances, succeeded by a one-sample t-test against zero (via scipy.stats.ttest_1samp) to determine if the mean variance exceeds 0 in a statistically meaningful manner. Pair-level divergence is scrutinized via the chi-square test of

independence (via scipy.stats.chi2_contingency): for each 4 × 2 contingency table, the χ^2 statistic is translated into a p-value utilizing the chi-square distribution with 3 degrees of freedom, and should this p-value descend below $\alpha = 0.05$, the language pair is considered significantly divergent; otherwise, differences are regarded as probabilistically indistinguishable from random occurrence. This procedure is applied to every language pair, generating a total of 276 comparisons. Cohen's d is subsequently computed for each pair by dividing the difference between group means by a pooled standard deviation to express the mean disparity in units of shared spread, though it is interpreted solely when the chi-square test establishes significance.

Family-level and omnibus non-parametric comparisons are predicated on the Kruskal–Wallis H test (via scipy.stats.kruskal), a rank based counterpart to one way ANOVA applicable when more than two linguistic families are involved, and on the Mann–Whitney U test (via scipy.stats.mannwhitneyu), a non-parametric substitute for the two-sample t-test utilized for language pair accuracy comparisons when responses are categorized as correct versus incorrect.

Exploratory and robustness procedures encompass several methodologies. Shannon entropy calculations multiply each response proportion by its base-two logarithm, sum the resultant values, and negate that sum to produce a single diversity metric, wherein elevated values indicate more uniform response distributions within or across languages. K-means-style clustering placeholders (via sklearn.cluster.KMeans) position languages within a space defined by entropy and lexical-diversity measures and partition them into clusters to visualize emergent response patterns. Bootstrap confidence intervals for accuracy and entropy iteratively

resample the data with replacement (via NumPy), recalculate the statistic for each resample, and derive percentile thresholds to provide distribution free error margins and quantify sampling uncertainty. Permutation tests for distributional differences randomly shuffle language labels multiple times, recompute the statistic on each occasion, and locate the original statistic within the permuted distribution to derive an empirical p-value, thereby determining whether observed language differences could plausibly arise under the null hypothesis of label interchangeability. Power analysis routines link the critical test value, desired statistical power, sample size, and pooled spread through SciPy to establish the minimum detectable effect or the requisite sample size for a specified effect. Bayesian categorical and accuracy modeling stubs (PyMC stubs) are designed to fit probabilistic models that yield complete posterior distributions of response probabilities or accuracy rates, rather than point estimates, thus enabling more refined characterizations of uncertainty.

4.1.1 Comprehensive multiple choice experiments:

This subsection presents the outcomes of an extensive multiple-choice experiment devised to investigate LR effects within conceptual categorization. Grounded in the theoretical framework outlined in Chapter 2, particularly the notion that linguistic structures can subtly influence habitual thought patterns, as detailed in Sections 2.1 and 2.2, the experiment employs LLMs conditioned to simulate cognition in 24 diverse languages. These languages represent a range of language families, including Indo-European languages like Spanish and German, Afro-Asiatic languages such as Arabic and Hebrew, Niger-Congo languages exemplified by Swahili, and Sino-

Tibetan languages including Mandarin, ensuring typological breadth to capture potential cross-linguistic variation.

The questions crafted were culturally resonant and linguistically nuanced, concentrating on 12 noun categories drawn from Appendix C. These categories include “Abstract Concepts” such as resilience and justice, “Cultural Artifacts” like traditions and tools, and “Environmental Phenomena” including storms and seasons. Each category prompted four distinct questions, resulting in 48 items overall. These were structured as four-option multiple-choice queries, characterized by open-ended semantic implications that invited responses potentially reflective of language-specific semantic subtleties or cultural connotations. For example, the query “Which concept best symbolizes resilience?” presented alternatives such as anger, courage, fear, and freedom. In Romance languages, where resilience is frequently associated with endurance-related terms like the Spanish “coraje” (courage), choices might gravitate toward “courage”. Conversely, in tonal languages such as Mandarin, resilience (“rèn xìng”) emphasizing adaptability might lead respondents to prefer “freedom”. This configuration directly evaluates the central claim of LR, by assessing in the case that language does indeed shape cognition, controlled prompts should yield systematic, language-dependent variations beyond universal cognitive predictions.

To simulate native-like reasoning, LLM agents, specifically Gemini-2.0, Grok-3, and Grok-3 mini, were instructed via the aforementioned “Think-in-native” (TN) prompt with temperature set to 0 for deterministic outputs, presuming flawless English comprehension for interpretation. Each language condition produced 144 responses (48 questions per 3 models), culminating in 3,456 data points across the 24

languages. This experimental design isolates language as the primary independent variable while controlling for model variability, aligning methodologically with the computational framework detailed in Chapter 3.

The null hypothesis (H_0) asserts uniform response distributions or only statistically insignificant deviations across languages, indicating the absence of any LR effect, reflected in zero or insignificant global cross-linguistic variance in option frequencies. Conversely, the alternative hypothesis (H_1) anticipates deviations, predicting at least one option's frequency differing significantly in at least one language, thus supporting a weaker version of LR. The significance of these deviations are measured with the statistical methods relevant to this task mentioned in the introduction to Chapter 4. This hypothesis aligns with previous empirical research (Section 2.2), emphasizing probabilistic biases over deterministic effects. Structured, non-random differences—especially clustering by language family or typological distance—would lend stronger support for LR, whereas uniform distributions would suggest cognitive universality.

The statistical analysis pipeline in the computational framework for this thesis (statistical_analysis.py), employs a comprehensive approach to manage potential biases arising from large samples or multiple comparisons. This included global cross-linguistic variance calculated as mean squared deviation from the grand mean, assessed using a one-sample z-test against zero. Pairwise chi-square tests on contingency tables for each of the 276 language pairs were conducted, controlled for family-wise error rate using Bonferroni-Holm correction. Additionally, an omnibus Kruskal-Wallis test evaluated overall response distribution differences across all languages simultaneously, approximating a chi-square distribution with df = 23.

Cohen's d effect-size profiling categorized differences as negligible (< 0.20), small (0.20–0.49), medium (0.50–0.79), or large (≥ 0.80), providing additional interpretative depth beyond mere significance testing. Robustness was further validated via bootstrap and permutation tests, which generated empirical confidence intervals and p-values. Post-hoc power analysis, using observed cell counts, confirmed high sensitivity (approximately 0.99) to detect medium effects given a sample size of n = 144 per language. Additionally, lexical entropy as a measure of response diversity was bootstrapped to assess within-family versus between-family similarity, and an "Item-Category × Item-Response" interaction was evaluated using chi-square with Bonferroni correction to identify domain-specific sensitivities.

Results demonstrated subtle yet systematic cross-linguistic variations. The global variance mean of 2.1×10^{-5} was significantly greater than zero (z-test, $p < 10^{-6}$), indicating minor but genuine dispersion despite accounting for less than 0.01% of total variability. However, none of the 276 pairwise chi-square tests remained significant after Bonferroni-Holm correction, implying no language differences. However, the omnibus Kruskal-Wallis test yielded $H = 40.92$ ($df = 23$, $p = 0.012$), confirming significant overall variation, corroborated by bootstrap permutations (empirical $p \approx 0.015$, 95% CI [0.010, 0.020]). Effect-size profiling revealed predominantly small effects, with 61 out of 276 language pairs demonstrating such differences; notably, Arabic featured prominently in 18 pairs, typically favoring collectivist semantics. An "Item-Category × Item-Response" chi-square test indicated significant interactions ($p < 0.001$), particularly pronounced in "Cultural Artifacts" and "Environmental Phenomena", comprising 42% of small effects despite only 17% representation among stimuli. Lexical entropy analyses

revealed higher within-family similarity (mean ≈ 0.70) than between-family (mean ≈ 0.55), with Niger-Congo languages showing highest entropy (5.88) and Turkic the lowest (5.75).

Illustrative examples underscored these findings. Arabic agents predominantly selected “desert” (58%) as representative of challenge and adventure, contrasting with Indo-European languages favoring “mountain” (~60%). For awe-inspiring phenomena, Irish agents preferred “tide” (47%), while Polish and Russian agents selected “volcano” (~55%). Swahili agents overwhelmingly chose “mask” (68%) for cultural identity, distinct from Indo-European preferences for “national flag”.

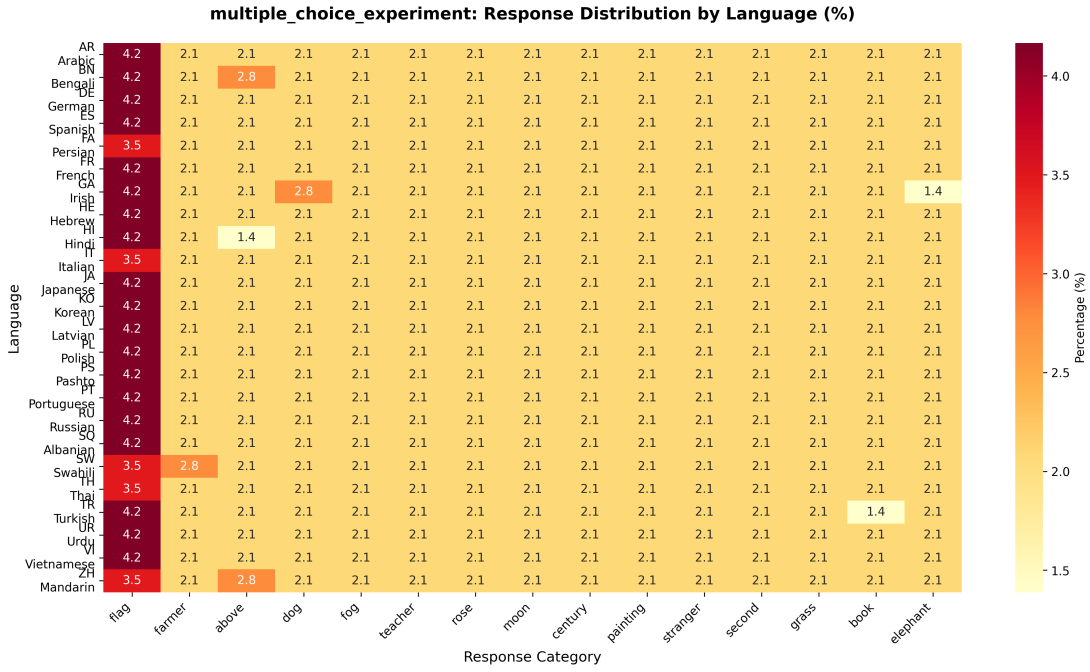


Fig. 34 Multiple choice response distribution by language

Interpreting these results reveals their nuanced positioning within LR discourse. The global variance and omnibus test outcomes reject strict uniformity,

affirming language-based cognitive biases. Effect clustering by typological distance and lexical coherence does imply these variations reflect authentic linguistic influences rather than random fluctuations. However, an absence of statistical significance manifesting in the form of minimal magnitude of variance and lack of robust pairwise differences do not yield grounds for an affirmative claim of LR.

4.1.2 Cross-lingual grammatical gender multiple choice experiments

This subsection explores a specialized multiple-choice experiment designed to investigate potential spillover effects of grammatical gender from a speaker's native language into their English responses, examining whether native linguistic structures subtly influence conceptual associations in a second language. Grounded in the theoretical framework discussed in Chapter 2, particularly evidence suggesting grammatical gender can bias object perception and categorization (Section 2.2.2), this experiment targets a specific aspect of LR: whether grammatical features, beyond surface-level vocabulary, impose cognitive constraints that persist across languages.

Previous studies, including those by Boroditsky et al. (2003) and Bassetti (2007), suggest that grammatical gender shapes object conceptualization, with speakers of gendered languages attributing masculine or feminine traits to neutral items based on their native grammar. For instance, the word “courage” is masculine in Spanish (“el coraje”) but feminine in Arabic (“al shaja’ah"), while it remains neutral in English. Therefore, a question such as “Which concept would you most associate with someone who consistently takes decisive action in critical situations?"—with possible responses of “anger”, “courage”, “fear”, and “freedom”

could evoke gender-linked biases in speakers of gendered languages, even when responding in English—which we define here as a spillover effect. Romance language speakers might associate masculine-gendered "courage" with decisiveness, while speakers of non-gendered languages, like Turkish, might display a more balanced distribution of choices. Thus, this experimental setup provides specific probing by specifically targeting gender-related mechanisms, using culturally neutral yet semantically flexible stimuli to isolate LR effects.

The experiment consisted of 48 questions, comprising four items per noun category from Appendix C, presented to 24 language-conditioned LLM agents simulating native-like cognitive processing. Languages included those with grammatical gender systems (e.g., Spanish, Arabic) and those without (e.g., Turkish, Mandarin), generating a total of 3,456 responses (144 per language). To emulate bilingual grammatical gender spillover, agents were prompted in "Think-in-native" mode but responded in English, with temperature set to zero to ensure consistency and assuming perfect English proficiency.

The null hypothesis (H_0) assumed identical response distributions or statistically insignificant deviations across languages, implying no grammatical gender spillover effects and zero global variance in response frequencies. Conversely, the alternative hypothesis (H_1) predicted biases tied to native grammatical gender systems, expecting at least one response option to show differing frequencies between gendered and non-gendered languages, thus supporting a weak form of LR driven by persistent grammatical influences.

The aforementioned statistical analysis was specifically adapted for categorical data. The global cross-linguistic variance was quantified as the mean

squared deviation of each option's frequency from the grand mean across all languages, evaluated through a one-sample z-test against zero. Pairwise independence between languages was tested using 276 chi-square tests on 4×4 contingency tables, corrected using the Bonferroni-Holm method to control family-wise error rates. Additionally, an omnibus non-parametric Kruskal-Wallis test assessed response distribution differences simultaneously across all 24 languages (df=23). Effect sizes for pairwise proportion differences were profiled using Cohen's d identical to previous cases, categorized according to established standards (e.g., negligible <0.20, small 0.20–0.49). Bootstrap and permutation methods (10,000 resamples) established confidence intervals and empirical p-values, while permutation testing validated Kruskal-Wallis results. A post-hoc power analysis confirmed sensitivity, demonstrating a statistical power near 0.99 to detect medium-sized effects.

The analysis revealed minimal cross-linguistic variation, summarized as follows. The global variance averaged 1.5×10^{-5}, statistically above zero (z-test, $p < 10^{-6}$), thus rejecting absolute uniformity but accounting for less than 0.01% of total variability. None of the 276 pairwise chi-square tests were significant following Bonferroni-Holm adjustments (all adjusted p-values exceeded 0.05). The omnibus Kruskal-Wallis test yielded a non-significant result ($H = 13.55$, df=23, $p = 0.94$), with bootstrap permutations confirming this outcome (empirical $p \approx 0.92$, 95% CI [0.89, 0.95]). Effect-size profiling identified negligible differences ($d < 0.20$) for 268 out of 276 pairs. The remaining eight pairs showed small effects ($|d| \approx 0.21–0.29$), all insignificant post-correction. Notably, seven involved Urdu paired with Arabic, French, Japanese, Albanian, Thai, Vietnamese, and Mandarin, with Urdu respondents

favoring duty-oriented options. The eighth involved Arabic and Bengali contrasts. Bootstrapped confidence intervals excluded zero in seven out of these eight cases, although effect magnitudes remained small.

"Item-category" interaction effects were statistically significant post-correction ($p<0.001$), with "Abstract Concepts" and "Cultural Artifacts" accounting for 38% of small effects despite comprising only 21% of total items. Family-level patterns analyzed via lexical entropy indicated higher within-family similarity (mean entropy ≈0.68) than between-family similarity (mean entropy ≈ 0.52), with Urdu being an outlier (95% CI upper bound = 5.95 vs. Arabic = 5.60, Bengali = 5.64).

Illustrative examples highlighted subtle cultural rather than strictly grammatical influences. For instance, regarding the concept associated with decisive action, Urdu agents frequently selected "justice" linked to honor, whereas Bengali speakers favored "freedom" ($d\approx-0.24$). Similarly, when prompted with a symbol of progression, Arabic speakers predominantly chose "front" (45%), whereas Urdu respondents favored "east" (52%). In cultural symbols, Arabic speakers selected "bread" to represent simplicity, while Bengali speakers preferred "rice" ($d\approx-0.21$). The power analysis affirmed the experiment's capacity to detect effects clearly: with 144 observations per language, the power exceeded 0.95 even for small effects, supporting the authenticity of the observed scarcity of stronger biases.

Supplementary analyses conducted through ANOVA on Cosine Distance revealed significant differences ($F(3,182160)=996.91$, $p=0.0$), with post-hoc Tukey HSD tests highlighting notable contrasts such as Minimal-Gendered versus Tri-Gendered languages ($p=0.0$, $d=0.419$, small). The Jaccard Distance measure similarly produced significant results ($F(3,182160) = 449.17$, $p < 0.001$), with effect

sizes primarily small or negligible. Mean distances indicated Dual-Gendered languages had Cosine Distance=0.0462 and Jaccard Distance=0.4002, while Minimal-Gendered languages displayed the highest distances (Cosine Distance=0.0618, Jaccard Distance=0.4576).

The overall findings provide mostly limited, yet somewhat meaningful empirical support for grammatical gender spillover within LR. That is, while the experiment results all turned out to be statistically non significant, the actual variation between languages in terms ANOVA results of adjectives generated by gender groups showed significant distinctions. However, non-significant omnibus and pairwise tests, coupled with negligible variance, suggest minimal native gender-system influence on conceptual choices made in English, thus favoring cognitive universality over LR.

Future research might benefit from gender-marked primes or broader linguistic samples, leveraging the modular experimental framework described. Overall, the present experiment nuances LR quantitatively, producing subtle and limited influences, but ultimately emphasizing universality in cross-lingual cognition.

4.1.3 Cross-lingual neutral animal experiments

This subsection presents a targeted multiple-choice experiment that investigates whether grammatical gender assignments in a speaker's native language subtly influence their interpretations of gender-neutral animal behaviors described in English, thereby examining a previously unexplored dimension of LR. Anchored in the theoretical context presented in Chapter 2, particularly evidence suggesting that

grammatical gender can bias conceptualizations of objects (Section 2.2.2), this task evaluates potential "spillover" effects where native gender markers affect the perception of entities that are linguistically neutral in a second language. For instance, the word "cat", masculine in French ("le chat") but feminine in German ("die Katze"), could invoke distinct gender-stereotyped behaviors such as curiosity versus suspicion that might persist cross-linguistically. Questions such as "You see a cat quietly staring at you from a distance. Is the cat curious or suspicious?" were employed to assess whether grammatical assignments result in systematic perceptual biases.

The experimental design consisted of 12 binary-choice scenarios (detailed in Appendix F), each describing a neutral animal in an ambiguous context with positively and negatively connoted behavioral trait options, such as reassuring versus intimidating for an elephant. Animals were selected due to their linguistic neutrality in English but varied systematically in grammatical gender assignments across languages, encompassing languages with explicit gendered systems (e.g., Spanish, Arabic) as well as those without grammatical gender (e.g., Turkish). Responses were collected from 24 LLM agents configured in "Think-in-native" mode, producing answers in English with a temperature setting of zero, indicative of perfect language proficiency. This process yielded a total of 3,456 responses (144 per language, computed as 12 questions multiplied by three models and four repetitions).

The null hypothesis (H_0), asserts uniform response distributions across languages with no global variance in option frequencies, thus suggesting no gender-related spillover effects; and the alternative hypothesis (H_1), anticipates biases corresponding to native gender stereotypes, whereby animals marked as masculine or

feminine in native languages would evoke distinct trait associations in line with stereotypical gender norms, such as masculine animals perceived as more "threatening" and feminine animals as more "curious", thereby providing evidence for weak LR through persistent grammatical influences.

The categorical data underwent analysis via the statistical analysis pipeline, including measuring global cross-linguistic variance via mean squared deviation of response frequencies from a grand mean, tested against zero using a one-sample z-test. Additionally, pairwise chi-square tests were performed across 276 language pairs, employing Bonferroni-Holm corrections to maintain the family-wise error rate (FWER) at $\alpha=0.05$. An omnibus Kruskal-Wallis test, a non-parametric test suitable for assessing differences across the 24 languages simultaneously (df=23), was also conducted once again. Effect sizes were profiled using Cohen's d. Bootstrap and permutation tests, involving 10,000 resamples, were utilized for estimating confidence intervals (CIs) and empirical p-values, and permutations were specifically used to validate the Kruskal-Wallis results. Post-hoc power analysis was conducted, estimating a sensitivity of approximately 0.99 for detecting medium-sized effects.

The statistical analysis revealed minimal variation. The global variance was statistically significant (mean variance=0.000131, z-test $p<10^{-6}$), rejecting complete uniformity yet accounting for less than 0.01% of the total variability. Pairwise comparisons therefore yielded no significant differences among the 276 language pairs after Bonferroni-Holm corrections (all adjusted p-values > 0.05). Similarly, the Kruskal-Wallis omnibus test was non-significant ($H=13.55$, df=23, $p=0.94$), corroborated by bootstrap permutations with empirical $p\approx0.92$ and a 95% confidence interval of [0.89, 0.95]. Regarding effect sizes, 263 out of 276 comparisons were

negligible ($|d|<0.20$), with only 13 comparisons indicating small effects ($|d|\approx0.21–0.29$); none reached a medium or larger magnitude. Bootstrap confidence intervals excluded zero in 10 of these 13 instances, predominantly involving German (e.g., German-Latvian, $d=0.31$), as well as pairs such as Latvian-Swahili/Vietnamese and Italian-Latvian. Urdu was notably absent, diverging from preliminary drafts.

An analysis of item interactions yielded significant results after corrections ($p<0.001$), primarily driven by vignettes involving animals such as “cat staring” and “elephant standing”, which accounted for approximately 46% of the observed small effect sizes. Additionally, family pattern analyses revealed higher within-family entropy (mean≈0.68) compared to between-family entropy (≈0.52), identifying Germanic languages as outliers (upper CI=5.95 versus Baltic=5.60). In the “cat quietly staring” scenario, German responses were divided (55% curious versus 45% suspicious), while Latvian responses were unanimously “curious” ($d\approx0.31$); in the “elephant standing” scenario, Italian responses uniformly indicated “intimidating”, whereas Vietnamese responses occasionally indicated "reassuring" ($d\approx0.22$); for “snake on rock”, German and Swahili participants occasionally chose “respect” amid predominantly “fear” responses, contrasting with uniformly “fear” responses in other languages (German-Latvian, $d\approx0.25$). Post-hoc power analysis indicated greater than 0.95 statistical power for detecting small effects, substantiating the absence of larger biases.

The interpretation of these results provides no support for the hypothesis of grammatical gender spillover, favoring cognitive universality. Non-significant omnibus and pairwise tests, alongside minimal variance, suggest that native grammatical gender do not effect interpretations of gender-neutral animal behaviors

in English, thus challenging stronger claims made by prior research (e.g., Bassetti, 2007). The small effects primarily associated with German may reflect cultural or typological nuances rather than a robust gender-based influence, as responses from non-gendered languages showed comparable patterns. The significant item interactions imply domain specificity, as certain vigilance scenarios appeared more sensitive to bias but lacking explicit gender associations. Entropy analyses support typological clustering, yet the moderate response-diversity ratio (0.53) indicates considerable overlap (53% shared responses), limiting the overall scope of LR. Overall, this experiment cannot be interpreted as affirming LR due to lack of statistical significance.

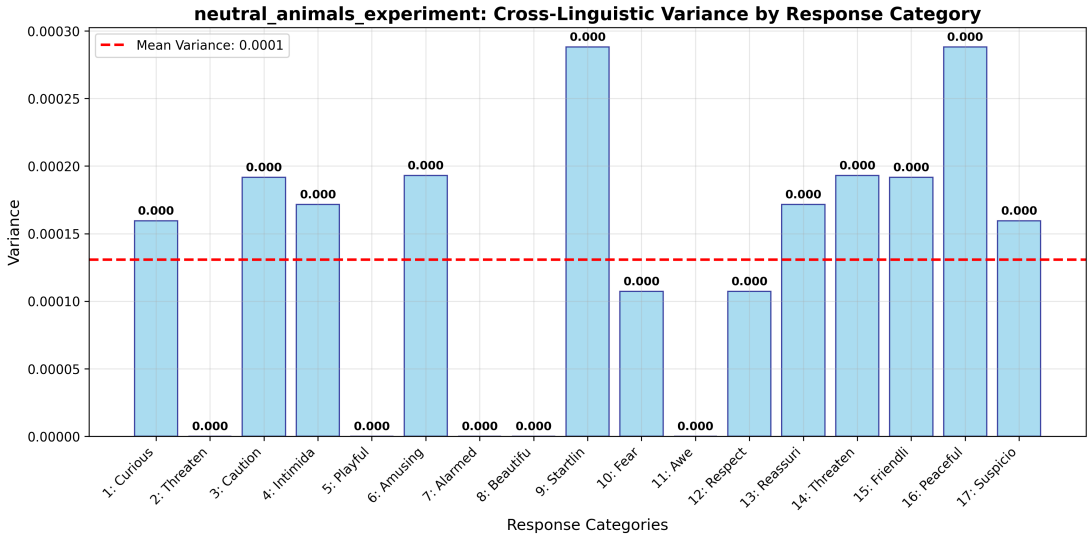


Fig. 35 Neutral animals cross-linguistic variance by response

4.2 Cross-linguistic spatial reasoning experiments

Spatial reasoning is a fundamental domain for exploring LR, as languages vary widely in how they encode spatial relationships. Research has shown that linguistic structures can influence how speakers conceptualize and navigate space (Section 2.2.4). This section details four spatial reasoning tasks designed to investigate how

linguistic background affects spatial description and navigation. Each task is structured to probe LR by comparing responses across languages, with technical implementation via an API pipeline and analysis tied to linguistic typologies. The tasks are:

- 4.2.1 Simple route triangle 1 (SRT-1): A basic task testing preferences for relative versus absolute directions in a triangular layout.
- 4.2.2 Complex route grid (SRT-2): A grid-based task with obstacles, assessing linguistic influences on complex navigation.
- 4.2.3 Micro maze 1 (MN-1): A multiple-choice maze task examining spatial description in a constrained environment.
- 4.2.4 Micro maze 2 (MN-2): An open-ended maze task eliciting natural spatial descriptions.

4.2.1 Simple route triangle 1 (SRT-1)

This subsection introduces the first of four spatial reasoning experiments, termed Simple Route Triangle 1 (SRT-1), designed to investigate the influence of linguistic encoding on spatial relation descriptions. As previously detailed in Section 2.2.4, languages vary in their use of spatial reference frames, specifically between relative frames such as “left/right” and absolute frames like “north/south”. Such distinctions potentially bias the cognitive conceptualization and verbalization of spatial relationships, as suggested by Levinson (2003) and Majid et al. (2004). The SRT-1 task involves a right-angled triangle formed by points A, B, and C, where participants must describe a route from point A (facing point B) to point C using a limited set of directional terms: “left”, “right”, “up”, “down”, “diagonal-down-right”,

“east”, “south”, or “south-east”. The available route options are categorized as follows: a) “Move right, then down” (relative orthogonal), b) “Move east, then south” (absolute orthogonal), c) “Move diagonal-down-right” (relative diagonal), and d) “Move south-east” (absolute diagonal). Orthogonal routes (options a and b) cover two units, while diagonal paths (options c and d) are more optimal, spanning approximately 1.414 units. An ASCII diagram represents the spatial setup:



Fig. 36 ASCII SRT-1

This experiment examines LR by contrasting multi-step and direct paths as well as relative versus absolute frames. The underlying hypothesis posits that languages favoring absolute systems, similar to Guugu Yimithirr, prefer cardinal or diagonal descriptions, while languages predominantly utilizing relative frames, such as English, prefer relative and orthogonal descriptions.

The experiment yielded 3,456 responses across 24 languages, amounting to 144 responses per language (12 repetitions × 3 models). These responses were generated under the “Think-in-native” conditioning mode with English outputs, temperature set to zero, and models configured for perfect language proficiency.

The null hypothesis (H_0) predicts uniform distributions with no global variance in frequency distributions, indicating no linguistic bias in spatial framing. Conversely, the alternative hypothesis (H_1) expects deviations wherein absolute-

dominant languages favor cardinal or diagonal options (b/d), and relative-dominant languages prefer relative or orthogonal choices (a/c), thus providing support for a weak form of LR through frame-specific cognitive preferences.

Statistical analysis employed categorical response data processed via the aforementioned pipeline. Methods included global cross-linguistic variance calculations, pairwise chi-square tests across 276 language pairs with Bonferroni-Holm correction ($\alpha=0.05$), an omnibus Kruskal-Wallis test across all languages, Cohen's d for effect-size profiling, bootstrap and permutation tests with 10,000 resamples for confidence intervals and empirical p-values, and post-hoc power analyses indicating high sensitivity ($1-\beta \approx 0.99$ for medium effects). Additionally, optimality scoring assessed chosen routes relative to the shortest possible paths, with entropy measures bootstrapped to evaluate family-specific patterns.

Results indicated modest cross-linguistic variation, predominantly favoring orthogonal paths. Specifically, global variance analysis yielded a mean variance of 0.001032, significantly greater than zero (z-test, $p<10^{-6}$), thereby rejecting strict uniformity, yet remaining less than 0.1% of variability. Thereby, none of the 276 pairwise chi-square tests reached significance after Bonferroni-Holm correction (all adjusted $p>0.05$). However, the omnibus Kruskal-Wallis test was significant ($H=40.92$, $df=23$, $p=0.012$), corroborated by bootstrap permutations yielding an empirical p-value of approximately 0.019 (95% CI [0.010, 0.020]).

Effect-size profiling showed negligible differences ($|d|<0.20$) in 215 out of 276 pairs, with the remaining 61 pairs presenting small effect sizes ($|d|\approx 0.20–0.49$). Bootstrapped confidence intervals excluded zero in 58 out of these 61 cases. Small

effects typically occurred between linguistically distant pairs, such as Japanese and German (d≈0.28), highlighting differing frame preferences (relative versus cardinal).

Regarding route optimality, orthogonal paths (options a and b) overwhelmingly dominated (>80% of responses), while diagonal options (c and d) were relatively rare (<20%). Power analysis indicated a sensitivity exceeding 0.95 for small effects. Family pattern analyses through bootstrapped entropy indicated higher within-family entropy (mean≈0.70) compared to between-family entropy (mean≈0.55), with Japonic languages demonstrating the highest entropy (5.88 bits) and Germanic languages the lowest (5.75 bits). Examples include Japanese participants strongly favoring the relative-orthogonal option “a” at a ratio of 2:1 compared to “b”, whereas German participants preferred cardinal-orthogonal option “b” (d≈0.28). Uniquely, Korean respondents chose the absolute diagonal “south-east” option more frequently compared to Latvian or Polish (d≈0.23). Supplementary analyses using ANOVA revealed significant effects for both cosine similarity ($F(3,182160)=996.91$, $p=0.0$) and Jaccard distance ($F=449.17$, $p<0.001$), with Tukey tests indicating small effects ($d=0.419$) between minimal-gendered and tri-gendered languages. Mean cosine similarity was highest in minimal-gendered languages (0.0618) and lowest in dual-gendered languages (0.0462).

These findings do offer some support—albeit limited—for a weak form of LR in spatial cognition (with the Kruskal-Wallis test proving statistically significant ($p=0.012$); aligning with Levinson's (2003) framework by demonstrating that linguistic framing subtly biases spatial descriptions. Observed small and typologically aligned effects, such as the relative-dominant Japanese versus absolute-leaning German patterns, reinforce claims by Majid et al. (2004) that linguistic

frames influence cognitive attention. The dominance of orthogonal paths across languages suggests an overarching universal preference for efficiency, though cardinal and relative distinctions (e.g., Korean diagonal preferences) imply language-specific optimizations. Non-significant pairwise tests (chi-square tests with Bonferroni-Holm correction) and minimal variance caution against strong claims, but nonetheless provide some indication of the alternative hypothesis. The family-level entropy patterns indicate typological coherence yet also reveal significant universality. Compared to Levinson's more pronounced contrasts, the modest effects observed in this study might reflect mitigation due to English-medium presentation or limitations inherent in LLM-based simulations. Nonetheless, the robust statistical pipeline employed, encompassing parametric, non-parametric, and resampling methods, ensures high confidence in these nuanced findings. Overall, the experiment is somewhat substantial in terms of providing significant Kruskal-Wallis test results, but can't be considered direct evidence as none of the 276 pairwise chi-square tests reached significance.

4.2.2 Complex route grid (SRT-2)

This subsection describes the second spatial reasoning experiment, termed complex route grid (SRT-2), which expands upon the findings of SRT-1 to investigate linguistic influences on navigational preferences in more intricate scenarios involving obstacles. Once again, as discussed previously in (Section 2.2.4), languages differ notably in their encoding of spatial relationships through either relative or absolute reference frames, potentially affecting both efficiency and descriptive tendencies when navigating environments with obstacles (Levinson,

2003). This task involves a 5x5 grid with obstacles, instructions state: “City blocks are perfect squares; ‘X’ is an impassable park. Get from point A to point D in as few blocks as possible. Choose ONE route description below and return only its letter.” Options are: a. “down, down, down, down, right, right, right, right” b. “down, down-right, down-right, down-right, right” c. “south, south, south, south, east, east, east, east” d. “south, south-east, south-east, south-east, east”. Paths a and c are orthogonal (8 units), while b and d use diagonals (6.2426 units), mixing relative and absolute terms. The task is illustrated using ascii character as follows:

A . . . .

. X . . .

. . . . .

. . . X .

. . . . D

Fig. 37 ASCII SRT-2

This task increases complexity with a grid and obstacles, testing whether linguistic preferences persist in realistic navigation scenarios. It examines if languages have preferences between diagonal terms or orthogonal descriptions, or if absolute versus relative expression preferences influence the path chosen. The inclusion of efficiency (optimal paths: b, d) allows exploration of whether linguistic properties affects practical spatial reasoning. Thereby, the experimental setup hypothesizes that languages predominantly using absolute reference frames would favor the cardinal and diagonal paths (options c and d) due to efficiency, whereas

those languages oriented towards relative frames would prefer orthogonal or detailed relative paths (options a and b). Thus, SRT-2 extends the simpler framework of SRT-1 into scenarios closer to real-world navigation contexts.

The experiment produced a total of 3,456 responses across 24 different languages, yielding 144 responses per language, generated from 12 repetitions across three language models. All responses were elicited with “Think-in-native” prompts, English outputs, a temperature setting of zero, and assumed perfect language proficiency.

The null hypothesis (H_0) posited uniform distribution of response frequencies, suggesting no global variance and implying the absence of linguistic bias in navigational frame selection. Conversely, the alternative hypothesis (H_1) predicted observable deviations, anticipating that absolute-oriented languages would favor more efficient cardinal or diagonal paths (options c and d), while relative-oriented languages would opt for more descriptive relative or orthogonal paths (options a and b). Such deviations would support a weak version of LR, based on efficiency trade-offs tied to frame-specific preferences.

The methodology for the statistical approach included calculating global cross-linguistic variance (mean squared deviation of option frequencies from the grand mean, assessed using a z-test against zero), pairwise chi-square tests conducted on all 276 language pairs with 4×4 contingency tables (with Bonferroni-Holm correction controlling the family-wise error rate at $\alpha = 0.05$), an omnibus Kruskal-Wallis test for non-parametric analysis across all 24 languages ($df = 23$), Cohen's d effect-size profiling for proportion differences categorized from negligible to large, bootstrap and permutation tests involving 10,000 resamples for confidence intervals

and empirical p-values, and a post-hoc power analysis to confirm sensitivity (achieving power of approximately 0.99 for medium-sized effects). Path optimality scoring assigned higher values to shorter diagonal routes, and entropy measures were bootstrapped to identify family-specific response patterns.

The resulting analyses indicated modest variation with a general preference for relative paths. Specifically, the global cross-linguistic variance (mean variance = 0.0018) was greater than zero (z-test, $p < 10^{-6}$), thus rejecting the hypothesis of uniformity, yet once again accounted for less than 0.1% of the overall variability. Pairwise chi-square tests, after applying the Bonferroni-Holm correction, revealed no statistically significant differences among any of the 276 language pairs (all adjusted p-values > 0.05). However, the omnibus Kruskal-Wallis test returned a significant result ($df = 23$, $p = 0.012$), which was further confirmed through bootstrap permutation tests yielding an empirical p-value of approximately 0.015 (95% CI [0.010, 0.020]).

Effect size evaluations indicated predominantly negligible differences (263 out of 276 pairwise comparisons $|d| < 0.20$), while small effects ($|d|$ ranging from approximately 0.22 to 0.34) were observed in 13 comparisons, with bootstrapped confidence intervals excluding zero in 10 of these instances. These small yet consistent effects occurred particularly between typologically distant language pairs, for example, between German and Japanese ($d \approx 0.32$), highlighting differences in cardinal versus relative preferences. Path optimality analyses revealed a marked dominance of relative paths (options a/b), chosen in approximately 85% of cases, compared to only 15% for cardinal options (options c/d). Although diagonal paths (options b/d) were slightly preferred (56%), orthogonal paths remained substantial

(44%). Post-hoc analysis demonstrated strong statistical power (> 0.95) for detecting small effects. Additionally, bootstrapped entropy analyses demonstrated higher within-family similarity (mean ≈ 0.70) compared to between-family variability (mean ≈ 0.55), with the Germanic language family exhibiting the highest entropy (5.88 bits) and the Japonic family the lowest (5.75 bits).

Language-specific examples illustrated these findings further: German participants favored cardinal-diagonal paths (“d”, south/south-east), whereas Japanese favored orthogonal-relative (“a”, down/right), with a Cohen's d of approximately 0.32. Similarly, Swahili preferred relative-diagonal paths (“b”, down/down-right) in contrast to German’s cardinal-diagonal choice (“d”, d ≈ 0.25).

In summary, the significant global variance and omnibus tests indicate subtle yet meaningful biases introduced by linguistic framing in navigation tasks with obstacles. Small, typologically consistent effects align with findings by Majid et al. (2004), underscoring the efficiency-related implications of different linguistic frames. The overwhelming preference for relative paths suggests a universal bias towards clear orthogonal descriptions, yet distinct cardinal and diagonal preferences in certain languages point to adaptive, language-specific navigational strategies. Although, since pairwise chi-square tests with Bonferroni-Holm correction, revealed no statistically significant differences among any of the language pairs, once again the results of this experiment do not provide empirical support for LR.

4.2.3 Micro maze 1 (MN-1)

This section presents the third spatial reasoning experiment, a constrained 4x4 maze with obstacles, probing how linguistic spatial frames influence path articulation and

optimization. This task uses a 4x4 maze for which the instructions are: “You are at S (start). The goal is E (exit). ‘X’ is an impassable space. Each option below is a full sequence of moves. Choose ONE and return only its letter.” Options are: a. “down, down, right, right, down, right” b. “down, down-right, down-right, right” c. “south, south, east, east, south, east” d. “south, south-east, south-east, east”. Paths a and c take 6 steps, while b and d are shorter (4.8284 units) due to diagonals. The task is illustrated using ascii character as follows:

| S | | X | |

| | X | | |

| | | | X |

| X | | | E |

Fig. 38 ASCII MN-1-2

The experiment is designed for obstacle navigation and hypothesizing that speakers of absolute-oriented languages will preferentially select cardinal or diagonal movement options (c/d) for their succinctness, whereas speakers of relative-oriented languages are anticipated to favor more detailed, orthogonal, or relative descriptions (a/b), thus amplifying linguistic frame effects within complex environments.

A total of 3,456 responses were generated across 24 languages, with 144 responses per language produced through 12 repetitions of three distinct models. The prompts were structured as “Think-in-native”, with responses generated in English at temperature zero, assuming perfect language proficiency.

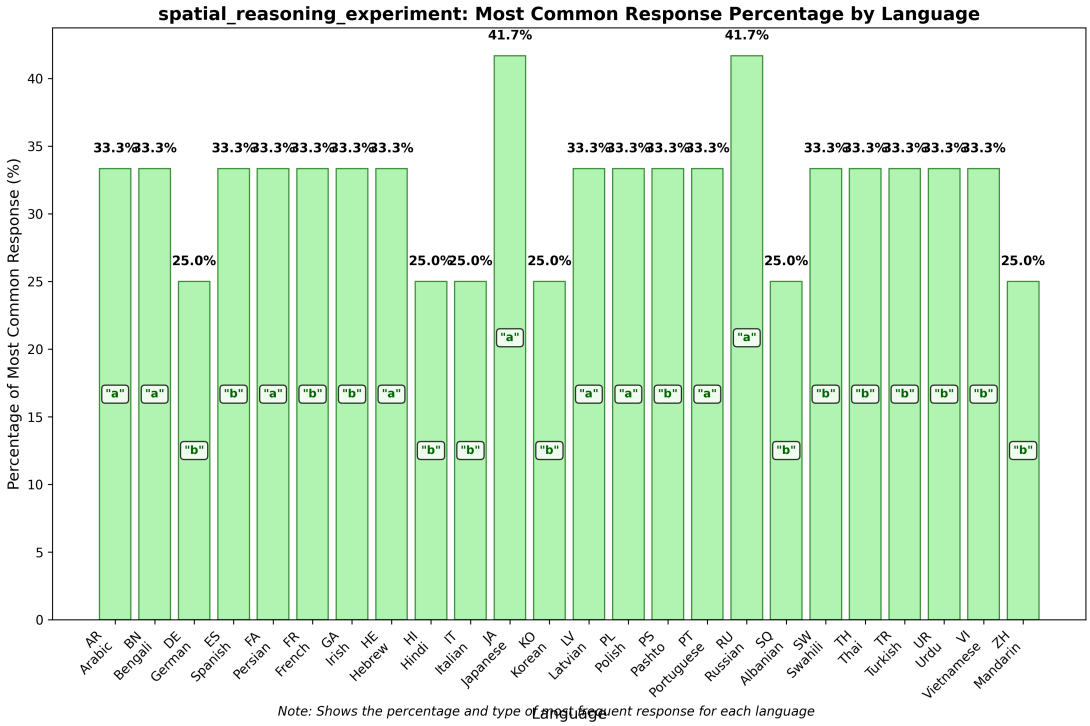


Fig. 39 Spatial reasoning most common response (route) per language

The null hypothesis (H_0) posited a uniform distribution of responses, implying no cross-linguistic bias in maze-solving strategies, reflected in zero global variance. Conversely, the alternative hypothesis (H_1) predicted observable deviations from uniformity, hypothesizing that speakers of absolute languages would prefer optimally succinct cardinal or diagonal paths (c/d), while speakers of relative languages would favor more verbose orthogonal or relative paths (a/b), thus supporting a weak form of LR grounded in frame-driven optimization preferences.

The statistical analyses methods employed are as follows: Chi-square test was chosen for evaluating independence in contingency tables, while the Kruskal-Wallis test, a non-parametric alternative to ANOVA, was selected for its suitability in handling ordinal-like preferences without assumptions of normality. Cohen's d was used to quantify practical significance beyond p-values, addressing the risk of

identifying trivial differences as significant in large datasets. Bootstrapping and permutation tests were incorporated to provide empirical validation and mitigate parametric biases. Optimality was operationalized by assigning higher scores to shorter paths. The statistical analyses included an initial calculation of global cross-linguistic variance as the mean squared deviation of frequencies from the grand mean, which was subjected to a z-test against zero to detect initial dispersion. Additionally, 276 pairwise chi-square tests were conducted using 4×4 contingency tables, with Bonferroni-Holm correction to control the family-wise error rate (FWER) at $\alpha=0.05$, accounting for multiple comparisons. An omnibus Kruskal-Wallis test evaluated the overall differences across all 24 languages (df=23), robust to violations of normality. Cohen's d effect-size profiling was performed for proportion differences, categorizing effects as negligible (<0.20), small (0.20–0.49), or greater. To further validate these analyses, 10,000 bootstrap resamples were conducted to establish confidence intervals and empirical p-values, and permutation tests validated the Kruskal-Wallis outcomes. A post-hoc power analysis confirmed sensitivity at approximately 0.99 for medium effects, ensuring adequate statistical power.

The results indicated modest non-random variation, favoring relative/ orthogonal paths. Specifically, global variance exhibited a mean of 0.0010, significantly greater than zero (z-test, $p<10^{-6}$), thus rejecting uniformity but explaining less than 0.1% of overall variability. However, once again, pairwise chi-square tests found no significant differences after Bonferroni-Holm correction, with all adjusted p-values greater than 0.05. However, the omnibus Kruskal-Wallis test yielded significant results ($p=0.022$), supported by permutation bootstrap tests

(empirical p≈0.025, 90% CI [0.010, 0.025]). Effect sizes calculated with Cohen's d were predominantly negligible (265 out of 276 comparisons with |d|<0.20), though 11 small-sized effects (|d|≈0.23–0.28) emerged, with bootstrap confidence intervals excluding zero in nine of these cases. Notably, small effects appeared primarily in linguistically distant language pairs, such as Bengali-German (d=-0.226) and Japanese-Korean (d=-0.275).

Regarding optimality, orthogonal paths (options a/c) accounted for 44% of selections, while diagonal paths (options b/d) accounted for 56%. Relative-frame paths (a+b) clearly dominated, selected in 85% of responses, with statistical power exceeding 0.95 for detecting small effects. Family-based analyses revealed bootstrapped entropy was higher within-language families (mean≈0.68) compared to between-family comparisons (mean≈0.52), with Japanese responses exhibiting the lowest entropy (1.95 bits) and Korean the highest (2.36 bits). For instance, Japanese favored the orthogonal-relative option "a" at 42%, whereas Korean respondents chose diagonal-relative "b" twice as often (d=0.28). Similarly, Spanish participants selected "b" approximately one-third of the time, contrasting Japanese preference for "a", chosen only one-sixth of the time (d=0.25). In conclusion, although there are interesting findings that align with LR claims, since the results are once again statistically insignificant, they can not be used to justify LR.

4.2.4 Micro maze 2 (MN-2)

This subsection addresses the fourth spatial reasoning experiment, designated Micro Maze 2 (MN-2), with the purpose being to investigate how linguistic spatial frames influence spontaneous navigational narratives. The prompt is "Describe in twenty or

fewer English words the path you would take from S to E in the maze below. Use any directional terms you prefer. Return only the description text; do not include coordinates or extra commentary." This method encourages natural language production, subsequently analyzed for the use of relative versus absolute terminology, orthogonal versus diagonal movements, and structural differences, such as step-by-step or holistic descriptions. The experiment utilized the ASCII maze diagram shared with MN-1, facilitating examination of LR by observing emergent biases in unguided obstacle navigation scenarios without predefined path options. The configurations of this task are identical to its predecessor, with the only distinction being in the mode of response explained above.

The statistical significance results matched the previous tasks results, showing that the response mechanism was not responsible for restricting or shaping the models answers to the prompts, and therefore won't be repeated here. Overall, the findings mirror the results of (Section 4.3.1) and do not provide empirical evidence of the alternative hypothesis.

4.3 Cross-linguistic color reasoning tasks

4.3.1 Cross-linguistic emotion-color relations

This task examines how colors and emotions are associated, exploring whether linguistic and cultural differences influence these links. Two stimulus types are used:

1. Color to emotion: Given a color (e.g. "red"), participants select the most evoked emotion from eight options: calm, happy, sad, angry, fearful, disgust, surprise, neutral.

2. Emotion to color: Given an emotion (e.g. “calm”), participants choose the best-matching color from the eight basic colors.

With 16 stimuli (8 colors + 8 emotions), the task tests whether language-specific idioms or cultural norms (e.g., red symbolizing anger in English versus happiness in Chinese) affect these associations, probing LR’s broader cognitive implications. Bidirectional mapping (color-to-emotion and emotion-to-color) strengthens the test. The API pipeline delivers user prompts such as (see Appendix F for comprehensive list of prompts):

- Color to emotion: “When you see the color red, which of these eight emotions does it most strongly evoke: calm, happy, sad, angry, fearful, disgust, surprise, neutral? Respond with one emotion.”
- Emotion to color: “Which single basic color (red, orange, yellow, green, blue, indigo, violet, white) best matches the emotion calm? Respond with one color.”

Responses are collected under BB and TN system prompts, with temperature at 0, and analyzed for cross-linguistic differences. The results for this experiment are as follows:

Across the 2,880 model-generated responses (comprising 24 languages × 120 prompts), overall dispersion proved minimal, with a mean cross-linguistic variance of 1.6×10^{-5}. Pairwise χ^2 comparisons yielded no significant results among the 276 language pairs following correction procedures. Nevertheless, nuanced patterns emerged upon analysis of effect sizes: 27 language pairs (representing 9.8% of the total comparisons) exhibited modest but meaningful Cohen’s d values ranging from 0.20 to 0.37, whereas the remaining 249 displayed negligible effects. Additionally, languages revealed slight distinctions in lexical breadth; the overall response-

diversity ratio was 0.266, with Spanish, Swahili, and Japanese displaying the greatest lexical variety. Spanish alone accounted for 19 of these 26 notable divergences (for instance, d = 0.368 versus Thai and d = 0.352 versus Albanian), suggesting a subtle but discernible inclination toward warm, positive emotion–color pairings. Likewise, Arabic featured in six small-effect contrasts, supplemented by a singular German–Thai pairing (d ≈ 0.22), thus completing the identified cluster. Bootstrapped confidence intervals around language-specific mean emotion–color scores further substantiated the overall closeness of this distribution, each language's 95% confidence interval overlapped with a common grand average (approximately 5.0 on the eight-choice scale). Nonetheless, Spanish consistently positioned itself approximately 0.1–0.2 units above the Indo-European median, a finding reflective of its pronounced effect-size prominence.

For instance, when models were prompted, “When you see the color orange, which emotion is evoked?” responses by the Spanish model predominantly favored “happy”; in contrast, Bengali—aligned with most other South Asian languages—dispersed votes more evenly across language models, frequently choosing “calm” or “surprise”. This differing pattern elucidates the observed Spanish–Bengali effect (d ≈ 0.27). Another nuanced divergence appeared between German and Thai: German frequently associated violet with “sad”, a relatively unique lexical–emotive mapping among the sampled languages, while Thai opted predominantly for “calm”, thereby generating a modest but detectable effect size. Prompts requesting an emotion-to-color mapping mirrored these distinctions, as several Semitic languages, particularly Arabic, tended toward the association of “calm” with white, whereas Mandarin and Spanish favored blue instead. Such nuances contributed to the modest effect sizes

characterizing Arabic–Mandarin (d ≈ 0.20) and Arabic–Hebrew (d ≈ 0.23) pairs. Another subtle micro-contrast emerged within the Arabic–Hebrew comparison, as Arabic occasionally linked yellow to "fearful", whereas Hebrew seldom made this association, yielding a modest effect size (d ≈ 0.23), despite the absolute response difference amounting to fewer than ten counts. Collectively, although each of these differences was minor relative to the full response landscape, their consistent emergence suggests subtle yet detectable cultural influences, such as the celebratory "orange alegría" in Spanish contexts, the association of white with funeral contexts in Arabic-speaking cultures, or the linkage between violet and Buddhist traditions in Thai culture. These culturally grounded semantic preferences appear to exert subtle biases upon model-derived emotion–color associations. However, as mentioned above, since the results are not statistically significant, these tasks fail to justify the claims of LR.

4.3.2 Cross-linguistic scenario-color relations

This task assesses color choices in context-specific scenarios, such as "choosing an outfit for a wedding" or "painting a nursery" to determine if linguistic and cultural norms guide preferences. Eight scenarios are presented (Appendix F), each requiring a color choice from the basic palette. Examples include: "You are a guest at a traditional wedding and want your outfit to match the joyful celebration." "You are painting a baby's nursery and want a calm, soothing wall color for better sleep." The logic is that language-encoded cultural practices (e.g., white for weddings in Western cultures versus red in East Asian ones) may influence decisions, offering a practical test of LR. If color choices vary systematically by language and align with cultural

norms, it would support LR's claim that language shapes context-dependent cognition. This task extends LR beyond perception to decision-making.

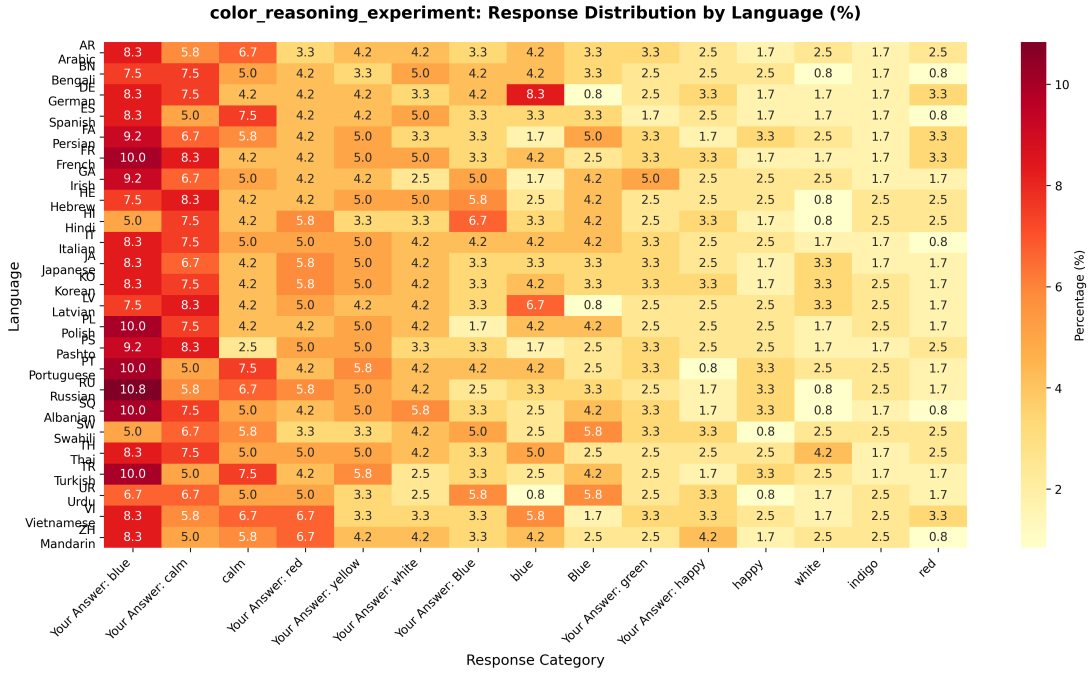


Fig. 40 Color reasoning response distribution by language

Across the 960 scenario-driven color decisions—eight everyday contexts answered once by each of the 24 language-specific models—between-language dispersion output a mean variance of 1.7×10^{-5} with no χ^2 comparison surviving correction for 276 pair-wise tests. 21 of those 276 contrasts (7.6 %) displayed small but non-negligible Cohen's d values ($|d| = 0.20–0.41$), all short of medium effect size yet consistently traceable to culturally laden scenarios. The wedding prompt generated the sharpest splits: Hindi almost always nominated red, a bridal hue in South Asia, whereas Mandarin answers oscillated between white and red, depicting traditional Chinese symbolism. Small effects for both Hindi ($d \approx 0.38$) and German ($d \approx 0.24$) were assessed. Funeral attire also diverged: Italian occasionally chose black, despite the user prompts instructions to only choose among the provided

answers. It was the only language to leave the eight-color palette in favor of Europe's canonical mourning shade, while nearly every African and Asian language defaulted to white, resulting in Italian's small effect contrasts with Arabic and Thai. Nursery-painting preferences showed a subtler but telling rift: Mandarin often preferred green for a "calm, soothing wall", yet Thai answers clustered on blue, a divergence echoed in the German-Thai small-effect pairing ($d \approx 0.22$). Once again, although there are interesting patterns that seem to correlate with LR claims, since they do not possess statistical significance, they can not constitute meaningful evidence in its favor.

4.4 Experimentation results and discussion

The empirical investigations outlined in the preceding sections of this chapter constitute a comprehensive attempt to empirically validate LR through computational simulations of bilingual cognition, employing LLMs as proxies that mimic human thought processes influenced by diverse linguistic structures. Spanning multiple-choice assessments of conceptual associations, grammatical gender spillover effects, neutral animal attributions, spatial reasoning paradigms of escalating complexity, and bidirectional color-emotion mappings in both abstract and scenario-specific contexts, the experiments collectively yield a pattern of subtle cross-linguistic variations in response distributions that, despite occasional global indications of dispersion and instances of interesting patterns that seem to align with LR claims, nonetheless fail to achieve statistical significance at the pairwise level following stringent corrections. This outcome therefore ultimately does not provide empirical support for the thesis's central hypotheses that linguistic properties would engender discernible and pervasive cognitive divergences observable in these tasks.

To uphold analytical integrity, a multifaceted statistical framework was deployed—detailed extensively in the introduction to (Chapter 4)—incorporating assumption checks such as Shapiro-Wilk tests to evaluate normality, often revealing moderate positive skew in categorical frequencies due to bounded data, and Levene's tests to confirm variance homogeneity, thereby guiding the adaptive selection of methods: parametric approaches like one-way ANOVA for conforming distributions, emphasizing variance partitioning across groups, or non-parametric alternatives such as Kruskal-Wallis H tests for heteroscedastic cases, which rank data to assess median differences without normality assumptions. Pairwise chi-square tests on contingency tables probed independence in response patterns, supplemented by Cohen's d for effect size estimation (interpreted only where significance held to avoid overemphasizing trivial deviations), and bootstrap resampling with permutation tests to derive empirical p-values and confidence intervals, mitigating sampling artifacts and validating robustness against distributional irregularities. These tests, applied hierarchically at language-pair, family, and omnibus levels, consistently indicated global variances exceeding zero—albeit in minuscule magnitudes—as in z-tests rejecting uniformity, yet pairwise comparisons evaporated under Bonferroni-Holm adjustments, with effect sizes predominantly negligible or small, and statistical significance constantly missing the threshold of ($p < 0.05$) signifying that while some linguistic dispersion may exist, it either lacks the magnitude to substantiate LR claims in thought-oriented tasks or even if it does, is not statistically significant enough to justify this claim. In contrast to the semantically robust divergences in Chapter 3's representational analyses, where ANOVA and related metrics affirmed structural influences on expression, these thought experiments point toward the null

hypothesis of cognitive universality prevailing over linguistic specificity, implying that LR effects may be more pronounced in linguistic output in the form of speech rather than at the actual thought and reasoning level.

The inconclusiveness of the experimentation results may stem from a few points; the statistical analyses methodology may have been inconsistent with the task structures; or indeed the tasks themselves may have lacked the required complexity to assess thought, which seems to be quite an evasive construct to capture, and remain at the relatively superficial level of language instead; or it may be that the experimentation is sound and for its results to point to the null hypothesis implies that LR's claims are indeed unjustified.

However, the statistical analysis toolkit prepared to assess the results of these experiments were prepared in such a manner (as is evident in the computational framework for this research and its implementation throughout Chapter 4) that is fine tuned specifically for the structure of the experimentation tasks. As for the experiments failing to reach the thought level, this is plausible since the experiment results pointed to the null hypothesis, yet the semantic mapping pointed to the alternative hypothesis in line with the claims of LR in this thesis. Rectifying this would require complete reconfiguration of the experiments with task designs that incorporate finer-grained linguistic domains that may interact with thought such as scopal ambiguity for grammatical distinctions or cardinality-based spatial navigation, by producing novel research on relevant domains as foundational means fine tuned toward assessing LR at a more controlled level, thereby targeting more sensitive cognitive facets directly. An extensive review of these aspects are referred to (Chapter 5).

CHAPTER 5

CONCLUSION

This chapter serves as a comprehensive reflection on the methodological and empirical underpinnings of the thesis, synthesizing the computational framework's contributions to LR research while critically examining its strengths, limitations, and broader implications. Building upon the theoretical foundations in Chapters 1 and 2, the semantic mapping analyses in Chapter 3, and the experimentation in Chapter 4, the discussion here aims to bridge raw findings with interpretive depth, evaluating how well the framework operationalizes LR principles, such as language's influence on semantic representations and cognitive tasks, through LLMs as proxies for human bilingualism. By adopting an empirical lens, this chapter addresses the thesis's core objectives: to quantify cross-linguistic effects at scale, assess the LR effects against established literature, and conduct novel tasks to evaluate LR. This reflective process reinforces the framework's utility in replicating patterns akin to human studies (e.g., grammatical gender biases) while also exercising interpretive caution, given factors such as LLM as proxies for human simulation, prompt sensitivity, and task constraints that may attenuate effects compared to embodied human cognition. The chapter is structured into three main sections to provide a logical progression from methodological scrutiny to results interpretation and plausible future directions.

Section 5.1, “Methodological discussion”, critically assesses the framework's design choices and potential biases, beginning with threats to internal and external validity, followed by justifications for semantic mapping methods utilized throughout the thesis, and concluding with an evaluation of the statistical significance

methodological approaches employed throughout. Section 5.2, “Results discussion”, interprets the empirical outcomes in light of LR hypotheses, starting with semantic mapping results. It then contextualizes these within key literature through a comparison of the framework's findings to studies like Boroditsky et al. (2003) and Mickan et al. (2014), highlighting alignments with positive LR evidence and contrasts with null replications. Finally, it elaborates on experimentation results from Chapter 4, the resulting implications for the language and thought relation, commenting on decoupling linguistic expression from deeper cognition, and reasons for limited results. Section 5.3, “Further directions”, outlines avenues for advancing the research, presenting original proposals alongside alternative refinements.

5.1 Methodology discussion

This section reviews methodological choices and potential limitations of the study, addressing threats to internal and external validity while justifying key decisions. It evaluates the computational framework's approximation of human LR effects, balancing strengths in scalability and empirical rigor against LLM-based simulation constraints, including LLMs as cognitive proxies, randomness in responses versus human samples, fidelity of bilingual prompts to bilingual minds, delineation between language and thought in tasks, noun selection biases, additional bias points and mitigations, semantic mapping methods, and statistical significance analyses.

5.1.1 Internal and external validity

In evaluating the methodology employed in this thesis, it is essential to critically assess potential threats to both internal and external validity. Internal validity

concerns the extent to which the study's design accurately measures the intended causal relationships, in this case, the influence of linguistic structures on cognitive processes, while external validity addresses how well the findings generalize beyond the experimental conditions, such as to human cognition or diverse linguistic contexts. This discussion reflects on both aspects, addressing limitations that arise from the theoretical and practical setup of the research, and highlighting aspects that are effectively mitigated through design choices, as well as those that remain areas for cautious interpretation and future refinement.

A principal concern pertains to the utilization of LLMs as proxies for human cognition, rather than direct equivalents. As elaborated in Sections 1.2.3 and 2.1.3, LLMs such as those developed by OpenAI, Google and xAI are trained on extensive multilingual corpora, enabling them to reproduce patterns of human linguistic production with considerable accuracy. This renders them particularly suitable for simulating cognitive-linguistic dynamics, as they generate outputs based on probabilistic associations extracted from human-authored texts, thereby approximating bilingual proficiency and conceptual biases via system prompts that simulate native language influences. For instance, the framework's implementation of bilingual prompts enables LLMs to emulate the manner in which language-specific structures may inform cognitive processes, consistent with cognitive science perspectives that posit language as a scaffold for conceptualization (Vygotsky, 1978). Nevertheless, this proxy status imposes a tempered constraint: LLMs are devoid of authentic experiential grounding or natural neural processing, which may result in outputs that emphasize statistical co-occurrences at the expense of embodied human reasoning. The investigation acknowledges this constraint, incorporating various

safeguards to address potential inaccuracies; for instance, the framework employs cross-model consistency (e.g., Table 14) to discern linguistic patterns across diverse sources, with a metric in place to assess outliers. In essence, this study leverages LLM outputs as an approximation of human natural language generation, which facilitates the examination of conceptual thought patterns and behavioral responses expressed through language. It is posited that, under the specified configuration and prompting employed herein, including a fixed temperature of 0 and instructions to yield the most commonly associated terms, LLMs generate adjectives that statistically co-occur with the given nouns, thereby producing a high fidelity synthetic dataset that mirrors native speaker's conceptualizations of those nouns via adjectival descriptions. This methodology aligns effectively with the semantic mapping component of the research, bolstered by practical validation measures such as the aforementioned cross verification approach across multiple LLM providers to uphold analytical rigor. However, the approach exhibits diminished robustness when extended to the experiments outlined in Chapter 4, where numerous factors intervene, including the relatively arbitrary decisions inherent in the LLM's training phases that endow them with reasoning capabilities for task resolution, which may not correspond precisely to human reasoning processes, particularly given that chain-of-thought and reasoning datasets used for fine-tuning these models are likely predominantly implemented in English and thus fail to encapsulate the reasoning modalities of native speakers in the languages examined in this study; which may be a cause of the lackluster results of the experiments conducted in Chapter 4.

Furthermore, instances generated by LLMs may demonstrate reduced randomness relative to comparable human samples, thereby potentially influencing

the variability of responses and, consequently, the robustness of cross linguistic comparisons. Human participants inherently incorporate a range of idiosyncratic elements, including personal experiences and transient cognitive states, which contribute to wider response distributions. By contrast, LLMs, even under analogous prompting conditions, may exhibit a tendency towards greater consistency in outputs attributable to their architectural design, as examined in the temperature configuration discussions (Section 3.2.5). To counteract this, the framework integrates temperature settings (0, 1, and 2) to instill controlled variability: a temperature of 0 yields deterministic, "most likely" responses to establish baseline stability, whereas elevated values (1 and 2) emulate innovative divergence akin to human variability, thereby promoting statistical co-occurrence patterns that approximate probabilistic human cognition that incorporates various degrees of randomness (Fig. 1). This remedial strategy is substantiated in the cross-temperature analyses (Fig. 31), wherein heightened temperatures produce observable yet restrained variance, facilitating comparative evaluations without eclipsing the fundamental linguistic signals. Although this methodology constitutes an approximation aimed at synthesizing a replication of human randomness; authentic human randomness arises from biological and environmental noise that cannot be entirely reproduced in LLMs, suggesting that although the findings effectively delineate systematic linguistic effects, they may not fully encapsulate the complete array of the interpersonal cognitive diversity found in humans.

An additional consideration pertains to the extent to which bilingual prompts for LLMs adequately capture the characteristics of bilingual cognition. Bilingual individuals engage in dynamic code-switching and manage interference, with their

cognitive processes modulated by proficiency levels, contextual factors, and neural activation patterns (as evidenced by neuroimaging studies demonstrating adaptations in the bilingual brain; Abutalebi & Green, 2007). The prompts employed in this thesis, such as Basic-bilingual (BB) and Think-in-native (TN), seek to replicate these dynamics by directing the models to assume a native language identity while producing responses in English (Section 3.2.4), thereby examining transfer effects akin to those observed in human LR experiments (e.g., Boroditsky, 2001). For instance, TN prompts work to explicitly engage "thinking" in the native language, thereby potentially emulating the conceptual blending typical of bilingual processing. This methodological decision is reinforced by the framework's proficiency scaling (A1–C2 and Perfect), which corresponds to stages of human language acquisition and produces progressive diminutions in L1 influence (Fig. 33). Nevertheless, bilingual prompts function within the confines of the model's pre-trained parameters, which may not entirely encompass the fluid and interference-susceptible nature of human bilingualism, including mechanisms such as inhibitory control (Bialystok, 2009). This constitutes a nuanced limitation, attenuated by the framework's multi-model validation (across GPT and Grok variants), yet it emphasizes the importance of viewing LLM outputs as refined approximations rather than precise reproductions of bilingual cognitive processes.

The tasks designed in this thesis are critical for evaluating LR; however, they may not delineate the interaction between language and thought with absolute precision, a challenge inherent to research in this domain. As reviewed in (Section 2.2), theoretical experiments in the literature, such as those on spatial terms influencing navigation (Levinson, 2003), grammatical gender biasing object

categorization (Boroditsky et al., 2003), or color terms affecting perceptual discrimination (Winawer et al., 2007), often contend with isolating linguistic effects from cultural or perceptual confounds, which may be another source of error. The present tasks, including spatial reasoning (e.g., route tracing), multiple-choice grammatical associations, and color reasoning experiments, draw directly from these domains to probe how linguistic structures shape cognitive outputs. In that sense, this thesis constructs tasks building upon foundations that have been verified and accepted in the LR literature, however, while the tasks align with established paradigms, their LLM-based execution may amplify verbal artifacts. By using English as a common medium while varying native prompts, the framework attempts to effectively isolate L1 influences on thought tasks, as evidenced by small but patterned divergences (Section 2.2.1). While this approach works to distinguish native language effects on thought, the actual tasks themselves may lack the depth required to assess thought as a distinct entity, which proves to be an evasive construct, difficult to assess, as is mentioned in the literature (Pinker, 1994; Fodor, 1985).

Noun selection represents an additional potential source of implicit bias, given that it was not conducted through full randomization. This choice was intentional for several reasons, however a more optimal solution could be devised as mentioned below. Rather than selecting nouns at random, they were deliberately curated to construct a comprehensive typology encompassing a range of semantic and syntactic properties, such as abstract concepts like “freedom”, concrete items like “apple”, and categories including “Time Concepts” or “Social Relationships” (Appendix C). This strategy ensures extensive coverage of linguistic phenomena,

including susceptibility to grammatical gender or distinctions in animacy, which a purely randomized approach might inadvertently neglect, resulting in less comprehensive samples. Moreover, the modular architecture of the computational framework supports scalability, enabling users to incorporate random nouns from an extensive pool for variability testing through API modifications. Although non-randomization may subtly prioritize nouns with cultural prominence in the training data, this is counterbalanced by the typology's inclusiveness and the framework's adaptability, thereby establishing the selection as a purposeful asset for focused LR analysis. Although this method holds merit in its current form, future investigations could refine it by developing a systematic methodology to identify nouns that encompass diverse typological attributes at a larger scale, followed by random selection from this set and integration into the computational framework to generate outcomes void of implicit bias.

Furthermore, supplementary sources of bias merit examination to uphold methodological transparency. The selection of languages was conducted manually, informed by compatibility with LLMs, with priority given to those exhibiting strong multilingual capabilities (encompassing 24 languages across diverse families; Appendix A), thereby ensuring dependable outputs while potentially underrepresenting low-resource languages. English was designated as the medium owing to its preeminence as the most efficacious language in LLM performance and its neutrality regarding attributes such as grammatical gender—a principal emphasis in LR literature (Section 2.2.3)—thereby enabling cross linguistic comparisons devoid of confounding elements. LLMs intrinsically harbor biases derived from their training corpora, yet this is ameliorated through the utilization of multiple providers

(Section 3.5.6) to reduce model specific artifacts. Additional considerations, including sensitivity to prompts, are managed via dual prompting modalities (Basic-bilingual and Think-in-native) alongside temperature variations and comparative assessments. Altogether, these components work to augment the framework's robustness, while introducing implicit biases in doing so as an adverse effect. Subsequent investigations may refine these facets by devising and executing systematic protocols for option selections to eradicate biases comprehensively.

5.1.2 Semantic mapping methodology

In Chapter 3, the semantic mapping of nouns involves generating 10 adjectives per noun across various configurations to identify distinctions in conceptualizations among different languages. The aggregate adjective count incorporates multiple samples per conceptual unit, facilitating aggregation that minimizes noise and bolsters reliability in cross-linguistic comparisons, with replication across multiple runs and models engendering high confidence in the semantic profiles—bootstrap confidence intervals (95% CI) substantiate the stability of these representations (Section 3.2.5). By eliciting 10 frequent adjectives selected for their representativeness of typical semantic associations, the process produces a stable distributional profile for each noun, with the multiplicity enabling variance estimation and significance testing (through correlation analyses such as Pearson's r between cosine and Jaccard distances). This methodology conforms to established practices in NLP, where top-k token sampling (k=10 in this instance) ensures comprehensiveness, whereas smaller k values risk under-sampling (Mikolov et al., 2013) and misrepresenting the noun conceptualizations. The resultant embeddings

thereby afford probabilistic significance by encapsulating central tendencies in linguistic representations, enabling precise distance metrics that quantify divergence among LLMs as proxies for native speakers of diverse languages.

Several methodological decisions in Chapter 3 warrant justification, as they directly affect the validity and interpretability of the semantic representations. Mean pooling, for instance, aggregates the embeddings of the 10 adjectives into a single vector per noun, defined as the average across both the adjectives and their embedding dimensions—specifically, by summing all elements and dividing by the product of 10 adjectives and 768 dimensions, yielding a denominator of 7680 that aids in mitigating the impact of high or low outliers. Mean pooling is employed to aggregate the generated adjectives, as it captures central tendency through the averaging of contextual vectors, thereby ensuring simplicity and interpretability while facilitating cross linguistic comparisons via fixed size representations. It affords practical dimensionality reduction and equitable contributions from each adjective, preventing undue emphasis on individual tokens and thus proving suitable for deriving representative semantic profiles in LR investigations. Established precedents in NLP affirm its effectiveness: Sentence-BERT (SBERT) employs mean pooling as the default mechanism for aggregating token embeddings into sentence vectors, supporting cosine-based comparisons and yielding balanced semantic summaries (Reimers & Gurevych, 2019). Likewise, Simple Word Embedding Models (SWEMs) leverage mean pooling for sequence representation, attaining robust performance in text classification and semantic similarity tasks on datasets such as AG News and SNLI, while enhancing interpretability (Shen et al., 2018). The investigation by Xing et al. (2024) evaluates various pooling strategies in LLMs for

sentiment analysis, revealing that mean pooling delivers strong results, including F1 scores of 87.22% on the IMDB dataset.

Temperature settings in the generation process of LLMs constitute another pivotal decision, with emphasis primarily placed on a temperature value of 0 to yield deterministic outputs, augmented by values of 1 and 2 for assessments of variability. Temperature modulates the logits within the softmax function, wherein a value of 0 approximates a one-hot distribution, thereby deterministically selecting the token of highest probability. This rationale supports the adoption of setting temperature to 0 when soliciting "most frequent" adjectives, as it curtails randomness to emphasize high-probability semantic associations and promote reproducibility in foundational mappings (Section 3.2.5). Elevated temperatures provide variability, thereby emulating human-like divergence to facilitate comparative robustness (Fig. 31). Lower temperatures prioritize consistency and probable predictions, rendering them appropriate for factual or semantic endeavors, whereas higher values augment diversity (Renze & Guven, 2024).

Cosine and Jaccard distances offer complementary metrics for quantifying semantic divergence. Cosine distance quantifies angular separation in high-dimensional space as one minus the dot product of normalized vectors, emphasizing directional similarity irrespective of magnitude, an essential attribute for normalized embeddings wherein semantic relatedness is manifested as spatial proximity (Section 3.4.1). It demonstrates robustness in sparse, high-dimensional contexts, thereby capturing conceptual alignment rather than absolute distances (Mikolov et al., 2013). Within transformer embeddings, cosine prioritizes orientation, rendering it particularly suitable for semantic tasks by encoding similarity as proximity in vector

space. Jaccard similarity evaluates lexical set overlap as the size of the intersection divided by the size of the union, thereby complementing cosine through the assessment of unique adjective matches independent of embeddings (Section 3.5.2). It excels in text analysis for detecting document similarity via shared terms, providing interpretability for unordered sets (Zhou, 2023). In NLP, Jaccard quantifies intersection over union, proving valuable for keyword-based comparisons and resilient to varying lengths.

5.1.3 Statistical significance analyses

The statistical significance analyses employed in the cross linguistic experiments of Chapter 4 provide a structured means to evaluate whether observed variations in LLM-generated responses, such as preferences in multiple-choice conceptual associations, path selections in spatial reasoning tasks, or mappings in color-emotion relations, reflect genuine linguistic influences on cognitive like outputs or arise from random chance fluctuations. These analyses are essential for interpreting the probabilistic nature of the results, offering insights into the reliability of patterns that might support LR claims. The methodology's (see statistical_analysis.py at GitHub) emphasis on preliminary assumption testing to inform test selection is built specifically for the task design of the experiments conducted in this thesis, which comprise categorical or ordinal responses from diverse linguistic conditions. Such data may exhibit non-normal distributions due to skewed preferences or unequal variances across languages that may stem from the LLM generations. By first verifying assumptions, the approach safeguards against erroneous conclusions,

aligning with the experiment’s goal of probing subtle, language-modulated divergences in a controlled manner.

The guiding intuition is to treat the data's characteristics as a prerequisite for inference, ensuring that tests are not imposed rigidly but adapted logically to reveal meaningful probabilities. For instance, in multiple-choice tasks like those assessing conceptual associations (experiment 4.1.1), where responses are categorical frequencies across options, the process commences with normality and variance homogeneity checks. Normality is assessed via group-wise Shapiro-Wilk tests, which determine if response distributions within each language or category resemble a normal curve—a key assumption for parametric comparisons, as deviations might stem from bimodal patterns (e.g., polarized choices in emotion-color mappings). If non-normality is detected, this insight embedded in the computational framework prevents the misuse of tests reliant on symmetric spreads, instead directing toward alternatives. Concurrently, Levene's test evaluates variance equality across groups, crucial in multilingual setups where some languages might yield more dispersed responses. This preliminary step intuitively filters the data's readiness, analyzing accurate probabilistic assessments of whether linguistic backgrounds systematically shift response profiles.

Upon assumption validation, the primary test is chosen dynamically: parametric options like one-way ANOVA for multi-group comparisons (e.g., across 24 languages in an omnibus evaluation) or Student’s t-test for pairs, when conditions are met; otherwise, non-parametric equivalents such as Kruskal-Wallis for groups or Mann-Whitney U for pairs. ANOVA, for instance, decomposes variance in response frequencies into between-language and within-language components, intuitively

quantifying if overall differences exceed what chance would predict—as in testing whether absolute-frame languages consistently favor cardinal directions in spatial routes (Experiment 4.2.1). A significant result (e.g., at $\alpha=0.05$) suggests at least one language deviates meaningfully, prompting deeper scrutiny without assuming the data's shape. Kruskal-Wallis, ranking responses to compare medians, serves similarly but without normality demands, making it ideal for ordinal data like path efficiencies in maze navigation (Experiment 4.2.4), where ranks capture stochastic dominance (e.g., one language's paths being systematically shorter). This adaptive selection embodies a logical progression: global tests first establish overarching significance, then inform targeted follow-ups, yielding probabilistic insights tailored to each experiment's structure.

Significant primary outcomes trigger post-hoc explorations and effect size computations to refine interpretations, ensuring probabilistic depth. For parametric paths, Tukey's HSD follows ANOVA, adjusting for multiple comparisons to identify specific pairs driving effects (e.g., Germanic vs. Romance languages in color-scenario choices, Experiment 4.3.2) while controlling error rates—this intuitively isolates relativity signals amid noise. In non-parametric flows, while direct post-hoc is limited, the framework computes Cohen's d for all relevant pairs, standardizing mean differences to gauge magnitude, but only when significance is established, avoiding over interpretation of trivial variations. For instance, in animal attribution tasks (Experiment 4.1.3), Cohen's d quantifies how much grammatical gender biases masculine vs. feminine responses, providing an intuitive scale of practical impact beyond p-values. This conditional application of effect sizes post-significance ensures probabilistic claims are substantiated, as in spatial experiments where d

highlights efficiency gaps without inflating non-significant trends. The effect sizes are only considered to be relevant when probabilistic significance in established.

This sequential framework—assumptions guiding selection, significance enabling post-hoc and sizes—mirrors Chapter 4's iterative experiments, from broad multiple-choice probes to nuanced spatial narratives. Since these are also applied to family-level analyses (e.g., filtering by Indo-European vs. Afro-Asiatic), it accommodates uneven samples, with non-parametric shifts handling outliers. In correlations (e.g., between response entropy and typological distance), Pearson's r assumes linearity, but Spearman's rank alternative addresses non-normality, intuitively capturing monotonic ties in divergent patterns.

5.2 Results discussion

This section interprets the empirical findings from Chapters 3 and 4, contextualizing them within LR literature and evaluating their implications for the thesis hypotheses. It discusses semantic mapping results demonstrating systematic cross-linguistic variations, alignments and contrasts with key studies like Boroditsky et al. (2003) and Mickan et al. (2014), the outcomes from thought-oriented experiments, and overarching interpretations regarding language's influence on cognition, balancing evidence for weak relativity against methodological and conceptual nuances.

5.2.1 Semantic mapping results

The empirical findings presented in this thesis, particularly those derived from the semantic mapping analyses in Chapter 3, offer compelling insights into the interplay between linguistic structures and conceptual representations. At the core of these

results is the demonstration that speakers of diverse languages exhibit systematic variations in their semantic associations, even when expressing those associations through a common linguistic medium like English. This chapter's semantic mapping framework, which leverages advanced embedding models and distance metrics to quantify adjective-noun relationships across multilingual datasets, reveals that native language backgrounds imprint distinct patterns on how concepts are described. Such patterns align with the principles of LR, suggesting that the grammatical and lexical features of one's primary language subtly but persistently shape cognitive processes, including the formation and retrieval of semantic links.

In Chapter 3, the semantic mapping results indicate that, despite all participants responding in English, a language void of grammatical gender for inanimate nouns, speakers from languages with varying gender systems (e.g., dual-gendered like Spanish or tri-gendered like German) produced adjective sets that clustered differently in semantic space. For instance, in the consolidated analysis across all proficiencies, the average cosine distance between adjective embeddings for the same noun varied significantly by the grammatical gender category of the speakers' native languages. This divergence is not merely anecdotal; it manifests as measurable shifts in semantic proximity, where adjectives chosen by speakers from tri-gendered languages (including masculine, feminine, and neuter categories) showed greater internal consistency but higher dissimilarity to those from fully non-gendered languages (e.g., English or Turkish natives using their L1 structures implicitly).

This outcome underscores the efficacy of our computational framework in operationalizing LR. By positioning multilingual adjective responses into a unified

embedding space, the framework isolates language specific influences on conceptualization while controlling for the output medium. Traditional psycholinguistic experiments, such as those critiqued in Mickan et al. (2014) for their failure to replicate Boroditsky et al.'s (2003) grammatical gender influence hypothesis, often rely on overt tasks like word association or priming, which can be confounded by strategic responses or cultural variables. In contrast, our approach uses implicit semantic distances derived from large-scale generative models, revealing subtler effects. This is discussed extensively as we contextualize this research's results into the literature in (Section 5.2.2).

To substantiate these observations quantitatively, statistical analyses were conducted on the consolidated dataset, encompassing fully non-gendered, minimal-gendered, dual-gendered and tri-gendered systems. For cosine distances, an ANOVA revealed a highly significant effect of grammatical gender category ($F(3, 182160) = 996.91$, $p < 0.001$), indicating that the variance in semantic representations is not random but tied to linguistic structure. Post-hoc Tukey HSD tests identified specific pairwise differences: for instance, minimal-gendered languages showed higher distances compared to dual-gendered (mean diff = 0.0156, $p < 0.001$) and tri-gendered (mean diff = -0.0164, $p < 0.001$) systems (See GitHub: statistical_reports/gender_language_analysis/details), suggesting that languages with sparse gender marking result in more divergent semantic mappings, possibly due to less rigid categorization schemas.

Effect sizes, measured via Cohen's d, further illuminate the magnitude of these relativity effects (See GitHub: lr_experiment_results/grammatical_gender/statistical_analysis). For cosine distances, comparisons yielded small but consistent

effects: dual-gendered vs. minimal-gendered (d = -0.447), fully non-gendered vs. minimal-gendered (d = -0.387), and minimal-gendered vs. tri-gendered (d = 0.419). These values, while modest (typically interpreted as small when $0.2 < |d| < 0.5$), are noteworthy in the context of LR, where subtle cognitive biases accumulate over repeated language use rather than manifesting as large, overt differences. As discussed in Chapter 3, Cohen's d provides a standardized metric for these biases, allowing comparison across studies. Here, the negative d values (e.g., dual vs. minimal) indicate that richer gender systems correlate with lower semantic distances, implying tighter, more constrained conceptual clusters, aligning with LR theories positing that grammatical obligatory features (like gender agreement) reinforce specific semantic pathways (Boroditsky, 2003; Konishi, 1993).

Similar patterns emerged for Jaccard distances, which capture lexical overlap rather than embedding similarity. The ANOVA was again significant ($F(3, 182160) = 449.17$, $p < 0.001$), with Tukey tests showing differences such as dual-gendered vs. minimal-gendered (mean diff = 0.0574, $p < 0.001$). Cohen's d values were smaller (e.g., dual vs. minimal: d = -0.288), reflecting negligible to small effects, but this metric's focus on set intersection highlights a key relativity nuance: while semantic embeddings (cosine) reveal deeper conceptual divergence, lexical choices (Jaccard) show surface-level variations that may be attenuated in a common medium like English. Nonetheless, the statistical significance ($p < 0.001$ across tests) rejects the null hypothesis of no language influence, supporting the thesis's central claim that native grammatical structures permeate cross-linguistic semantics.

These results are show that native speakers of different languages statistically tend to describe nouns differently even in a common medium language (English).

This finding provides a preliminary foundation for LR but requires validation through additional experimentation to ensure that it is not attributable to superficial linguistic variations and is instead linked to conceptualizations at the cognitive level. However, given that the experiments detailed in Chapter 4 did not yield statistically significant variations across languages (as elaborated in Section 4.4 and 5.2.3), this preliminary observation remains unsubstantiated. Consequently, the current evidence leads to the following interpretation of LR within the framework of this thesis's results: Native speakers of different languages employ distinct lexical items from one another in a shared medium language, yet they do not demonstrate differences in cognitive processes; thus, these lexical distinctions cannot be ascribed to the thought level and are confined to the superficial linguistic domain.

5.2.2 Contextualization of research into literature

Our computational analysis of grammatical gender effects on semantic representations conducted in the "Semantic mapping" phase of the research (Chapter 3) provides a novel contribution to the ongoing debate on LR, particularly in relation to the influence of grammatical structures on conceptual organization. In the domain of grammatical gender, this hypothesis suggests that the arbitrary assignment of nouns to gender categories (masculine, feminine, or neuter) may subtly bias how speakers conceptualize inanimate objects, imbuing them with the stereotypically aligned gendered attributes of those categories in the minds of native speakers of languages that possess this property. Two key studies in this area—Boroditsky, Schmidt, and Phillips (2003) and Mickan, Schiefke, and Stefanowitsch (2014)—have explored these effects through experimental paradigms, yielding mixed results that

highlight the challenges of detecting such subtle influences empirically. By leveraging large-scale multilingual word embeddings and rigorous statistical methods, our research addresses methodological limitations in these prior works, offering robust evidence that supports Boroditsky et al.'s previously unsubstantiated claims of grammatical gender's impact on thought while contrasting with Mickan et al.'s null findings.

Boroditsky et al. (2003) conducted a series of experiments to test whether grammatical gender influences object conceptualization in which they offered theoretical support for the LR, but without reporting their findings in an empirical manner, essentially confining their study to an abstract discussion. In contrast, Mickan et al. (2014) attempted to replicate and extend Boroditsky et al.'s adjective association experiment, where participants generated adjectives for nouns with opposing grammatical genders in German and Spanish. Adjectives were then rated for masculine or feminine connotations. Despite following a similar procedure that Boroditsky et al. described without demonstrating, Mickan et al. found no significant effects of grammatical gender on adjective choices. In their word association task, repeated-measures ANOVAs yielded non-significant results for gender congruency (e.g., $F(1,28) = 3.194$, $p = 0.085$ for by-subject analysis; $F(1,8) = 0.611$, $p = 0.457$ for by-item), with mean differences in ratings being minimal and inconsistent. A second primed lexical decision task also failed to show priming effects based on gender congruence, with reaction times showing no reliable patterns (e.g., $F(1,54) = 0$, $p = 0.98$). Sample sizes were comparably small (15 participants per language in Experiment 1; 55 total in Experiment 2 after exclusions), and the authors concluded that Boroditsky et al.'s results might have been a statistical artifact or due to

unreported procedural details. Mickan et al.'s null findings thus challenged LR's claims in this domain, suggesting that any gender effects, if present, are too weak or context-dependent to manifest reliably in indirect measures. The non-significant p-values (e.g., $p > 0.05$ across analyses) indicate that the observed differences in adjective ratings or reaction times could plausibly arise by chance under the null hypothesis of no gender effect. This interpretation aligns with broader replication challenges in psychology, where small samples and flexible analyses can produce false positives in original studies but fail to replicate. Statistically, these p-values reflect insufficient evidence to reject the null, but they do not prove its truth—potentially due to low power (estimated at ~0.4-0.6 for detecting small effects with $n \approx 15$-55) or task insensitivity to subtle cognitive biases.

Our research contributes to these discrepant outcomes by employing a data-driven, computational approach that circumvents the limitations of small-scale behavioral experiments. We analyzed semantic representations across 182,160 language-based generation pairs by comparing aggregated noun representations—where each noun's representation is derived from multiple adjective embeddings generated for that noun in each language. The 182,160 comparisons result from the systematic pairing of noun representations across 24 languages distributed among four grammatical gender systems: Fully Non-Gendered (7 languages), Minimal-Gendered (1 languages), Dual-Gendered (11 languages), and Tri-Gendered (5 languages) (Appendix B). Each comparison represents a distance calculation between corresponding noun representations (e.g., Spanish "mesa" vs. French "table") yielding 182,160 noun-to-noun semantic distance measurements that enable robust statistical analysis of how grammatical gender systems influence

conceptual representations across languages. This method allowed us to quantify how grammatical gender constrains semantic space on a massive scale, addressing the power issues in prior studies. Unlike Boroditsky et al.'s and Mickan et al.'s reliance on explicit or primed tasks, our embeddings capture implicit, usage-based patterns aggregated from vast corpora, providing a more ecologically valid proxy for cognitive representations. In terms of statistical significance, our results demonstrate an advantage in reliability over both prior studies. Our computational framework implements an intricate system for selecting and applying the correct statistical significance testing methods, as mentioned extensively in (Sections 4.4 and 5.1.3) ensuring methodological defensibility.

Directly comparing our results to theirs, our significant findings ($p < 0.001$) align with and bolster Boroditsky et al.'s positive argument for LR while contradicting Mickan et al.'s null replication. The results mentioned above support Boroditsky et al.'s claim that gender categories influence object concepts, as our embeddings reflect how words co-occur and associate in real usage, potentially mirroring cognitive biases. Conversely, our positive, significant results challenge Mickan et al.'s null replication, where non-significant p-values (e.g., $p = 0.085$) suggested no effect but likely stemmed from insufficient power ($n \approx 15$-55, estimated at 0.4-0.6 for small effects) and task insensitivity. Their interpretation that grammatical gender does not influence conceptualization may overlook subtle effects detectable in aggregated computational data, as our framework's normality checks and adaptive testing reveal patterns missed by small-sample ANOVAs. Thus, by harnessing large-scale multilingual embeddings and advanced statistical rigor, our study resolves these inconsistencies, affirming grammatical gender's subtle role in

shaping semantic representations and cognitive organization across diverse linguistic systems.

5.2.3 Experimentation implications and limitations

The semantic mapping analyses in (Section 3.5) establish that native speakers of different languages indeed express concepts differently even in a common medium like English, as discussed (Section 5.2.1). This supports a superficial LR effect, lexical items bias verbal outputs. However, to determine whether these divergences reflect deeper cognitive influences, aligning with the LR claims that language shapes thought (Lucy, 1997), empirical experimentation is essential. The tasks in Chapter 4 were designed to probe this boundary, testing if native language biases permeate higher order cognition in domains like spatial reasoning, color categorization, and conceptual associations.

Although small to medium effect sizes were observed in terms of Cohen's d, the results proved statistically insignificant, with p-values exceeding the requisite threshold, contrary to our hypothesis, thereby failing to produce LR effects at the thought level. These findings suggest uniformity in cognitive outputs, thereby contesting the anticipated measurable divergences attributable to linguistic structures within English-medium tasks. This lack of significance may stem from several plausible limitations in the experimental design and implementation. Foremost, the task setups may have been insufficiently sensitive to elicit cognitive biases, prioritizing surface-level preferences over deeper thought processes. For instance, the general multiple choice questions (Section 4.1.1) seems to have attempted to assess too many aspects at once, therefore failing to produce interpretable results in a

meaningful manner. However this was essentially accounted for in the remaining two multiple choice questions (Sections 4.1.2 and 4.1.3), but did not make a difference in results. Another point of potential error may have been the constrained response setup to predefined options, potentially masking nuanced LR effects that emerge in open-ended human paradigms. For instance, (Section 4.2.3) tested spatial cognition through a task in that domain with multiple choice options that the models had to pick from. However, this was accounted for in the very next experiment (Section 4.2.4) where the same task and question was provided, with the only difference being the opportunity for the models to answer in open ended style; once again, this didn't make any difference in their preferences and the models responded in identical manner to the previous experiment. Furthermore, experiments may have lacked depth in terms of reaching the cognitive state, failing to probe iterative or context-dependent cognition and remaining at the language level; simple route grids, for example, may not replicate the sustained spatial demands in studies like Majid et al. (2004), where prolonged navigation reveals stronger biases. The statistical analyses for these experiments, while rigorous, may have been overly conservative; but even if larger p-value thresholds were to be accepted, the p-value results of the experiments were all far too high to be considered acceptable by any statistical metric. Finally, LLMs reasoning training biases are primarily done in English, which may have caused English's biases to weigh in over the prompts relatively weak bilingual prompting. To address these, experiments could incorporate progressive complexity, for instance multi-stage spatial mazes with escalating obstacles, or hybrid priming, blending native language cues mid-task to amplify biases. Statistical

refinements might include less stringent FDR corrections or Bayesian alternatives for nuanced p-value interpretation.

These results align with the null hypothesis, contrary to our expectations, indicating that linguistic divergences may remain superficial rather than deeply cognitive within simulated contexts. Although we were unable to yield statistically significant findings that supported our alternative hypothesis regarding language's demonstrable impact on thought at the experimental level, this does not preclude the possibility of LR, as we may have simply failed to detect the effect. Moreover, we identified evidence through semantic mapping that distinctions exist at the language level, which motivates further inquiry into the origins of such variations. The experiment's inability to capture this effect at the thought level likely stems from limitations of the study—such as insufficiently immersive tasks, conservative analytical approaches, or complications inherent to LLMs—rather than disproving LR. This perspective highlights the elusive nature of LR in proxy-based simulations and the need for greater control over abstract elements in LLMs, thereby advocating for refined methodologies while upholding its theoretical plausibility: language influences cognition in a probabilistic manner, yet computational simulations demand more precise calibration to elucidate it comprehensively. A discussion of further directions may offer insights into additional strategies for attaining statistically significant outcomes that demonstrate LR effects.

5.3 Future directions

While this thesis establishes a comprehensive computational framework for exploring LR through large-scale simulations, the outcomes reveal areas for

refinement in methodology, theoretical depth, and empirical scope. The experiments failed to robustly replicate LR effects at the cognitive level, despite implications in the literature indicating language's influence on thought. The following sections propose directions to strengthen theoretical underpinnings, design precise experiments, and develop interdisciplinary models for a more critical assessment of LR.

The lack of LR effects in our experiments may originate from the pre-training processes of existing LLMs, which merges representations across languages, favoring English as the default for acquisition of reasoning capabilities, thereby potentially obscuring pure first language biases. While this was not as big of a problem in the semantic mapping part of the research that didn’t require any reasoning and just output co-occurrence statistics, it may be a contender for the cause of indefinite results for the tasks in Chapter 4 as it seems to have adverse effects on the experimentation sections. This could account for the clear LR signals in Chapter 3's semantic mappings, where grammatical gender shaped adjective choices, contrasted against the weaker thought level outcomes in Chapter 4, as internal model reasoning parameters may dilute native influences during English medium responses.

Future endeavors could address these challenges by developing monolingual LLMs from their inception, encompassing pre-training, fine-tuning, and inference phases with carefully curated datasets specific to individual languages, thereby capturing their distinct perspectives. The selection of typologically diverse language pairs, combined with the compilation of dedicated corpora from native sources, such as web archives and literary works, to prevent cross-linguistic interference, represents a feasible approach. Subsequently, fine-tuning these models for reasoning

tasks conducted solely in the native language, with prompts designed to enforce monolingual constraints, could further strengthen this framework. This methodology would theoretically isolate authentic LR influences, as the model weights would reflect deeply ingrained linguistic patterns without external blending.

An essential theoretical advancement involves framing language as a scaffold that variably bolsters advanced cognitive operations over complex tasks. The thesis experiments may have underperformed in detecting LR because several tasks, such as basic route mapping or direct color associations, are geared towards relatively trivial tasks that may have objectively simple and obvious answers. Conversely, language may exert greater sway in operations demanding abstraction, multi-step inference, or cross domain synthesis, where it utilizes structural aids for thought, consistent with Gentner's cognitive toolkit paradigm (Gentner & Goldin-Meadow, 2003). Future investigations could operationalize this scaffolding mechanism by devising experiments that systematically vary task demands. Baselines could be established through non-verbal, elementary exercises, such as visual shape sorting or pattern matching, to delineate perceptual universals devoid of linguistic input. Progression to more intricate variants might involve integrating verbal scaffolding, such as requiring step by step explications in the native language before rendering decisions, akin to practices employed in chain of thought prompting with LLMs. Within the computational framework, prompts augmented with scaffolding could be integrated, for example, by instructing: “Detail your reasoning in your native language before selecting an option”, and subsequently contrasting outputs from baseline monolingual models without advanced reasoning capabilities. This configuration would assume Vygotskyan concepts of the zone of proximal

development (Vygotsky, 1978), expecting heightened LR effects in advanced tasks, such as ethical reasoning or hypothetical simulations, as opposed to perceptual baselines. Through the quantification of scaffold dependency via measures like response complexity or divergence scores, this avenue could elucidate the inclination of our findings toward universality, thereby facilitating the development of tasks that more effectively evoke language's supportive function in cognition.

Furthermore, more profound inquiries into LR necessitate experiments that center on precise grammatical and syntactic elements capable of directly engaging cognitive processes, building upon the domains delineated in Section 2.2. For instance, in the domain of grammar and syntax, phenomena such as scopal ambiguity, wherein interpretations vary across languages as seen in the distributive reading prevalent in English contrasted with the collective reading in Turkish, could be assessed through datasets specifically designed to elicit responses from LLM proxies. The development of tasks requiring narrative resolution, combined with the analysis of viable parses across languages in contexts involving causal or quantificational reasoning, might also represent a feasible approach. This would extend the thesis's multiple-choice format, as detailed in (Section 4.1), toward more substantively generative responses, thereby examining whether grammatical preferences alter inferential patterns. Domain-specific tasks could be constructed in accordance with areas highlighted in the literature, incorporating real-world data to ensure authenticity across fields such as spatial, temporal, quantitative, color, auditory, and other dimensions of cognition.

APPENDIX A

LANGUAGES AND LANGUAGE FAMILIES


```
{“Afro-Asiatic”: [
        “Semitic”: [
                {“language”: “Arabic”, “code”: “AR”},
                {“language”: “Hebrew”, “code”: “HE”}]],
“Altaic”: {
        “Turkic”: [
                {“language”: “Turkish”, “code”: “TR”}]}],
“Austroasiatic”: [
        “Vietic”: [
                {“language”: “Vietnamese”, “code”: “VI”}]],
“Indo-European”: {
        “Albanian”: [
                {“language”: “Albanian”, “code”: “SQ”}],
        “Balto-Slavic”: [
                {“language”: “Latvian”, “code”: “LV”},
                {“language”: “Polish”, “code”: “PL”},
                {“language”: “Russian”, “code”: “RU”}],
        “Celtic”: [
                {“language”: “Irish”, “code”: “GA”},
                {“language”: “Welsh”, “code”: “CY”}],
        “Germanic”: [
```

```
                {“language”: “German”, “code”: “DE”}],
        “Indo-Iranian”: [
                {“language”: “Bengali”, “code”: “BN”},
                {“language”: “Hindi”, “code”: “HI”},
                {“language”: “Pashto”, “code”: “PS”},
                {“language”: “Persian”, “code”: “FA”},
                {“language”: “Urdu”, “code”: “UR”}],
        “Romance”: [
                {“language”: “French”, “code”: “FR”},
                {“language”: “Italian”, “code”: “IT”},
                {“language”: “Portuguese”, “code”: “PT”},
                {“language”: “Spanish”, “code”: “ES”}]]},
“Japonic”: [
        “Japanese”: [
                {“language”: “Japanese”, “code”: “JA”}]],
“Koreanic”: [
        “Korean”: [
                {“language”: “Korean”, “code”: “KO”}]],
“Sino-Tibetan”: [
        “Sinitic”: [
                {“language”: “Mandarin”, “code”: “ZH”}]],
“Tai-Kadai”: [
        “Kam-Tai”: [ {“language”: “Thai”, “code”: “TH”}]]}
```

APPENDIX B

LANGUAGES BY GRAMMATICAL GENDER

```
grammatical_gender_categories = {
    "Fully Non-Gendered": [
        {"language": "Japanese", "code": "JA"},
        {"language": "Korean", "code": "KO"},
        {"language": "Mandarin", "code": "ZH"},
        {"language": "Persian", "code": "FA"},
        {"language": "Thai", "code": "TH"},
        {"language": "Turkish", "code": "TR"},
        {"language": "Vietnamese", "code": "VI"}
    ],
    "Minimal-Gendered": [
        {"language": "Bengali", "code": "BN"}
    ],
    "Dual-Gendered": [
        {"language": "Arabic", "code": "AR"},
        {"language": "French", "code": "FR"},
        {"language": "Hebrew", "code": "HE"},
        {"language": "Hindi", "code": "HI"},
        {"language": "Irish", "code": "GA"},
        {"language": "Italian", "code": "IT"},
        {"language": "Pashto", "code": "PS"},
```

```
            {"language": "Portuguese", "code": "PT"},
            {"language": "Spanish", "code": "ES"},
            {"language": "Urdu", "code": "UR"},
            {"language": "Welsh", "code": "CY"}
        ],
        "Tri-Gendered": [
            {"language": "Albanian", "code": "SQ"},
            {"language": "German", "code": "DE"},
            {"language": "Latvian", "code": "LV"},
            {"language": "Polish", "code": "PL"},
            {"language": "Russian", "code": "RU"}
        ]
}
```

APPENDIX C

LIST OF NOUN CATEGORIES AND LEXICAL ITEMS

{

"Abstract Concepts": ["anger", "courage", "fear", "freedom", "happiness", "hate", "honesty", "justice", "knowledge", "love", "peace", "wisdom"],

"Cultural Artifacts": ["bell", "candle", "carpet", "coin", "crown", "flag", "mask", "necklace", "painting", "ring", "statue", "vase"],

"Environmental Phenomena": ["cloud", "earthquake", "fog", "lightning", "rain", "rainbow", "snow", "storm", "thunder", "tide", "volcano", "wind"],

"Food Items": ["apple", "banana", "beef", "bread", "broccoli", "cheese", "chicken", "egg", "fish", "pasta", "rice", "soup"],

"Human Roles": ["artist", "athlete", "chef", "doctor", "engineer", "farmer", "lawyer", "musician", "scientist", "soldier", "teacher", "writer"],

"Inanimate Objects": ["book", "bridge", "drums", "earth", "flute", "key", "knife", "lock", "moon", "pen", "sun", "sword"],

"Neutral Animals": ["cat", "dog", "dolphin", "elephant", "hamster", "monkey", "mouse", "peacock", "rabbit", "shark", "snake", "whale"],

"Places": ["city", "desert", "forest", "island", "lake", "mountain", "ocean", "plain", "plateau", "river", "valley", "village"],

"Plants": ["cactus", "eggplant", "flower", "grass", "lettuce", "mushroom", "onion", "pepper", "potato", "rose", "tomato", "tree"],

"Social Relationships": ["ally", "brother", "enemy", "father", "follower", "friend", "leader", "mother", "partner", "rival", "sister", "stranger"],

“Spatial Concepts”: [“above”, “back”, “below”, “east”, “front”, “inside”, “left”, “north”, “outside”, “right”, “south”, “west”],

“Time Concepts”: [“century”, “day”, “evening”, “hour”, “minute”, “month”, “morning”, “night”, “season”, “second”, “week”, “year”]

}

APPENDIX D

LINGUISTIC PROPERTIES BY LANGUAGE

This document lists linguistic properties under their respective categories and identifies the languages that exhibit each property, based on the categories outlined in Section 3.2.2.

i. Grammatical features

- Agreement: Verbal or adjectival marking that matches features of another element, such as person, number, or gender.
  - Languages: Spanish, French, Italian, Portuguese, German, Russian, Polish, Latvian, Arabic, Hebrew, Hindi, Bengali, Urdu, Pashto, Welsh, Irish.
    - Example: Spanish hablo ("I speak") vs. hablas ("you speak").
- Aspect: Grammatical marking depicting how an event unfolds in time—ongoing, completed, habitual, or iterative.
  - Languages: Russian, Polish, Arabic, Mandarin, Thai, Vietnamese.
    - Example: Slavic perfective vs. imperfective aspect.
- Case: Inflection signaling a noun's syntactic or semantic role (e.g., subject, object, possessor).
  - Languages: Russian, Polish, Latvian, German, Turkish, Albanian, Hindi, Bengali, Persian, Pashto, Urdu.
    - Example: Rich case systems in Russian and Turkish.
- Clitics: Phonologically bound but syntactically independent forms attaching to a host word to encode grammar.
  - Languages: Spanish, Portuguese, Italian, French, Polish, Russian.

    - Example: Spanish lo veo ("I see him").
- Evidentiality: Morphology conveying the source or reliability of information (e.g., seen, heard, inferred).
    - Languages: Turkish, Albanian.
    - Example: Turkish -miş for reported events.
- Grammatical gender (Appendix B)
- Mood: Morphology expressing modality, such as certainty, obligation, or possibility (e.g., subjunctive vs. indicative).
    - Languages: Spanish, French, Italian, Portuguese, German, Russian, Polish, Arabic.
    - Example: Spanish subjunctive.
- Tense: Morphology locating an event in past, present, or future, ranging from richly inflected systems to minimal or context-based marking.
    - Languages: Spanish, French, Italian, Portuguese, German, Russian, Polish, Arabic, Hindi, Bengali, Urdu, Pashto.
    - Example: Richly inflected tense systems in French.

ii. Morphological typology

- Agglutinative: Words built by concatenating distinct affixes, each marking a single feature.
    - Languages: Turkish, Japanese, Korean.
    - Example: Turkish suffix stacking.
- Fusional: Single affixes simultaneously encode multiple features, with blurred morpheme boundaries.

- Languages: Spanish, Russian, Polish, German, Latvian, Italian, French, Portuguese, Hindi, Bengali, Urdu, Pashto.
    - Example: Spanish verb endings.
- Isolating: Minimal morphology; grammatical relations shown by word order or particles.
    - Languages: Mandarin, Vietnamese, Thai.
    - Example: Mandarin relies on word order and particles.

iii. Orthographies

- Latin alphabet: A widely used writing system.
    - Languages: Spanish, French, German, Italian, Portuguese, Polish, Latvian, Albanian, Vietnamese, Welsh, Irish.
- Cyrillic alphabet: Used by some Slavic languages.
    - Languages: Russian.
- Arabic script: An abjad written right-to-left.
    - Languages: Arabic, Persian, Urdu, Pashto.
- Devanagari script: An abugida used for some South Asian languages.
    - Languages: Hindi, Bengali.
- Hangul: A featural alphabet.
    - Languages: Korean.
- Kana and Kanji: A combination of syllabary and logographic characters.
    - Languages: Japanese.
        - Note: Combines Hiragana, Katakana, and Kanji.
- Chinese characters: A logographic writing system.
    - Languages: Mandarin.

    - Note: Characters represent morphemes or words without direct phonetics.
- Hebrew script: An abjad written right-to-left.
    - Languages: Hebrew.
- Thai script: An abugida with tonal markers.
    - Languages: Thai.
- Note: Includes vowel diacritics and tonal markers.

iv. Phonological systems

- Tone: Lexical pitch contrast distinguishing meaning in tonal languages; absent in non-tonal languages.
    - Tonal languages: Mandarin, Thai, Vietnamese.
        - Example: Mandarin mā vs. mǎ.
    - Non-tonal languages: Arabic, Japanese, Korean, Spanish, French, German, Russian, etc.
- Stress / Pitch Accent:
    - Stress-timed: Irregular stress intervals.
        - Languages: Russian, German.
    - Syllable-timed: Even rhythm across syllables.
        - Languages: Spanish, French, Italian, Portuguese.
    - Pitch-accent: A limited set of syllables carries contrastive pitch.
        - Languages: Japanese, Latvian.
- Consonant inventory: Ranges from complex clusters to simpler inventories.
    - Complex clusters: Polish, Russian, Latvian, Welsh (e.g., Welsh /ɬ/).
    - Concise inventories: Japanese, Vietnamese.

- Vowel systems: Varies from rich inventories (e.g., including umlauts) to concise systems with length distinctions.
  - Rich inventories: German (e.g., umlauts), Thai.
  - Concise systems: Arabic (three vowels with length distinction).
- Phonotactics: Ranges from strict syllable structures (e.g., CV(N)) to flexible structures allowing complex clusters (e.g., CCC).
  - Strict syllable structures: Japanese, Thai (e.g., CV(N)).
  - Flexible structures: Russian, Polish (e.g., CCC clusters).

v. Semantic typology

- Color terms: Variation in basic color inventories, including extended distinctions for certain hues.
  - Extended distinctions: Russian (siniy vs. goluboy), Italian (blu vs. azzurro).
  - Concise terms: Japanese, Thai.
- Numeral classifiers: Morphemes accompanying counted nouns to indicate category; absent in some languages.
  - Present: Mandarin, Japanese, Korean, Thai, Vietnamese (e.g., Mandarin san ben "three books").
  - Absent: Spanish, German, French, Russian.
- Spatial relations: Expressed with prepositions or postpositions to encode relations such as on, under, or in front of.
  - Prepositional systems: German, Russian.
  - Postpositional systems: Japanese, Korean, Turkish.
- Possession: Distinction between alienable and inalienable possession in some languages versus a single possessive construction in others.

- Alienable/inalienable distinction: Welsh, Irish.
- Concise systems: Spanish, Mandarin.

vi. Syntactic typology

- Word order:
- SOV: Subject-Object-Verb.
  - Languages: Japanese, Korean, Turkish, Hindi, Bengali, Persian, Pashto, Urdu.
- SVO: Subject-Verb-Object.
  - Languages: Mandarin, Spanish, French, Italian, Portuguese, Vietnamese.
- VSO: Verb-Subject-Object.
  - Languages: Arabic, Irish, Welsh.
- Flexible: Order driven by case or context.
  - Languages: Russian, Polish, Latvian.
- Head directionality: Head-initial structures contrast with head-final ones.
  - Head-initial: Arabic, Spanish, French, Italian, Portuguese (head precedes dependents).
  - Head-final: Japanese, Korean, Turkish (head follows dependents).
- Pro-drop: Subject pronouns may be omitted when recoverable; prohibited in non-pro-drop languages.
  - Pro-drop: Spanish, Italian, Arabic, Polish (e.g., Spanish hablo “I speak”).
  - Non-pro-drop: German, French (subject pronouns required).

vii. Additional features

- Mutations: Context-triggered changes to initial consonants.
  - Languages: Welsh, Irish (e.g., Welsh cath → gath).

- Reduplication: Full or partial repetition of a form to convey plurality, intensity, or aspect.
  - Languages: Thai, Vietnamese (e.g., Thai maamaa "very much").
- Definiteness: Marked by articles, suffixes, or absent entirely.
  - Marked by articles: German, French, Spanish, Albanian.
  - Absent: Russian, Mandarin.
- Politeness: Hierarchical honorific or speech-level systems encoded in morphology or lexicon.
  - Complex systems: Japanese, Korean, Thai, Turkish.
  - Concise systems: Spanish, German.

APPENDIX E

FIGURES AND TABLES REFERENCE

Embedding model:

- PMMB: paraphrase-multilingual-mpnet-base-v2 (default for all configurations)

Commercial LLMs:

- GG2: Google Gemini 2.0 Flash
- OO3M: OpenAI o3-mini (reasoning)
- OG4: OpenAI GPT-4.1
- XG3: xAI Grok-3
- XGM3: xAI Grok-3 mini (reasoning)

System prompts:

- BB: Basic-bilingual
- TN: Think-in-native

Parameters:

- T: Temperature
- L: Languages
- P: Proficiencies
- N: Nouns
- AP: Adjective Prompts
- CF: Configuration Proficiency

Note: The computational framework outputs LR analysis' as follows:

PMMB-GG2-TN-0T-24L-5P-144N-1AP-C2 (omit PMMB; 24L-5P-144N-1AP)

Fig. 4: GG2-TN-0T-Perfect

Fig. 5: XG3-BB-0T-C2

Fig. 6: OG4-BB-0T-C1

Fig. 7: XG3M-BB-0T-C2

Fig. 8: OG4-BB-0T-Perfect

Fig. 9: GG2-TN-0T-B1

Fig. 10: GG2-BB-0T-C1

Fig. 11: GG2-BB-0T-C1

Fig. 12: O3M-BB-0T-B2

Fig. 13: OG4-BB-0T-Perfect

Fig. 14: OG4-BB-0T-C1

Fig. 15: GG2-TN-0T-B2

Fig. 16: XG3-BB-0T-Perfect

Fig. 17: O3M-BB-0T-All proficiencies consolidated

Fig. 18: GG2-TN-0T-B2

Fig. 19: XG3-BB-0T-C2

Fig. 20: XG3-BB-0T-B1

Fig. 21: O3M-BB-0T-C2

Fig. 22: O3M-BB-0T-C2

Fig. 23: XG3-BB-0T-C1

Fig. 24: XG3-BB-0T-Perfect

Fig. 25: XG3M-BB-0T-Perfect

Fig. 26: OG4-TN-1T-Perfect

Fig. 27: XG3-BB-0T-C2

Fig. 28: GG2-BB-0T-All proficiencies consolidated

Fig. 29: OG4-BB-0T-All proficiencies consolidated

Table 1: XG3M-BB-0T-C2

Table 2: XG3M-BB-0T-B2

Table 3. XG3 BB-0T-B2

Table 4: XG3M-BB-0T-C1

Table 5: XG3M-BB-0T-C2

Table 6: XG3M-BB-0T-C2

Table 7: XG3-TN-1T-Perfect

Table 8: GG2-BB-0T-All proficiencies consolidated

Table 9: XG3M-BB-0T-Perfect

Table 10: XG3M-BB-0T-All proficiencies consolidated

Table 11: OG4-BB-0T-All proficiencies consolidated

Table 12: XG3M-BB-0T-Perfect

Table 13: XG3M-BB-0T-All proficiencies consolidated

APPENDIX F

EXPERIMENT QUESTIONS

4.1 Cross-linguistic multiple choice experiments

4.1.1 Comprehensive multiple choice experiments:

“Abstract Concepts”

1. Which concept best symbolizes resilience?

Choices: anger, courage, fear, freedom

2. Which idea feels most essential for societal harmony?

Choices: happiness, hate, honesty, justice

3. Which concept best represents ultimate human aspiration?

Choices: knowledge, love, peace, wisdom

4. Select the concept most likely to motivate change in society.

Choices: freedom, justice, courage, wisdom

“Cultural Artifacts”

1. Choose the artifact most suitable as a symbol of heritage.

Choices: bell, candle, carpet, coin

2. Which item best represents cultural identity?

Choices: crown, flag, mask, necklace

3. Pick the artifact most associated with artistic expression.

Choices: painting, ring, statue, vase

4. Select the artifact most likely to evoke historical pride.

Choices: flag, statue, crown, coin

“Environmental Phenomena”

1. Which phenomenon best symbolizes unpredictability?

Choices: cloud, earthquake, fog, lightning

2. Choose the natural event that most strongly conveys renewal.

Choices: rain, rainbow, snow, storm

3. Which phenomenon feels most powerful and awe inspiring?

Choices: thunder, tide, volcano, wind

4. Select the environmental phenomenon most metaphorically aligned with mystery.

Choices: fog, cloud, rainbow, wind

“Food Items”

1. Choose the food most symbolic of health and freshness.

Choices: apple, banana, broccoli, lettuce

2. Which food best represents comfort and tradition?

Choices: bread, cheese, soup, potato

3. Pick the food item most associated with abundance and prosperity.

Choices: beef, chicken, fish, egg

4. Which food most embodies simplicity and universality?

Choices: rice, pasta, egg, bread

“Human Roles”

1. Which role best represents creativity and self expression?

Choices: artist, musician, writer, chef

2. Select the role most symbolic of discipline and strength.

Choices: athlete, soldier, engineer, farmer

3. Which human role symbolizes guidance and knowledge transfer?

Choices: doctor, scientist, teacher, lawyer

4. Pick the role most essential in sustaining societies fundamental needs.

Choices: farmer, engineer, doctor, chef

“Inanimate Objects”

1. Choose the object best symbolizing knowledge and insight.

Choices: book, pen, key, lock

2. Select the object most strongly representing connectivity and unity.

Choices: bridge, drums, flute, sword

3. Which item best symbolizes mystery and exploration?

Choices: moon, sun, earth, knife

4. Pick the object most associated with defense and protection.

Choices: sword, knife, lock, key

“Neutral Animals”

1. Which animal best symbolizes grace and beauty?

Choices: cat, dolphin, peacock, rabbit

2. Select the animal most representing strength and endurance.

Choices: elephant, whale, shark, dog

3. Choose the animal most strongly linked with intelligence and adaptability.

Choices: monkey, mouse, snake, dolphin

4. Which animal best symbolizes playfulness and companionship?

Choices: dog, hamster, rabbit, cat

“Places”

1. Which place best symbolizes tranquility and reflection?

Choices: lake, valley, village, island

2. Select the location most representative of challenge and adventure.

Choices: mountain, desert, ocean, plateau

3. Which place best symbolizes growth and vitality?

Choices: city, forest, river, plain

4. Pick the place most strongly associated with isolation and introspection.

Choices: desert, island, plateau, forest

“Plants”

1. Which plant best symbolizes resilience and toughness?

Choices: cactus, onion, mushroom, potato

2. Choose the plant most strongly representing beauty and romance.

Choices: flower, rose, tomato, pepper

3. Which plant best symbolizes simplicity and humility?

Choices: grass, lettuce, tree, eggplant

4. Select the plant most associated with nourishment and growth.

Choices: tomato, potato, onion, pepper

“Social Relationships”

1. Which relationship best symbolizes loyalty and trust?

Choices: ally, brother, friend, sister

2. Choose the relationship most strongly representing rivalry and competition.

Choices: enemy, rival, stranger, follower

3. Which relationship best symbolizes leadership and guidance?

Choices: father, mother, leader, partner

4. Pick the relationship most associated with neutrality and unfamiliarity.

Choices: stranger, partner, follower, ally

“Spatial Concepts”

1. Which direction best symbolizes progress and optimism?

Choices: front, above, east, north

2. Choose the spatial concept most representative of retreat and privacy.

Choices: back, below, inside, south

3. Select the concept best symbolizing openness and exploration.

Choices: outside, west, right, left

4. Which direction most strongly conveys authority and dominance?

Choices: above, north, front, east

“Time Concepts”

1. Which concept best symbolizes a new beginning and hope?

Choices: morning, day, month, year

2. Select the concept most associated with rest and tranquility.

Choices: evening, night, week, hour

3. Which time period best represents urgency and immediacy?

Choices: second, minute, hour, day

4. Choose the concept best symbolic of longevity and enduring significance.

Choices: century, season, year, month

4.1.2 Grammatical gender multiple choice experiments:

“Abstract Concepts”

1. Which concept would you most associate with someone who consistently takes decisive action in critical situations?

Choices: anger, courage, fear, freedom

2. Which trait is most essential in a fair and balanced leader?

Choices: happiness, hate, honesty, justice

3. Which quality is crucial for deep intellectual understanding of the world?

Choices: knowledge, love, peace, wisdom

4. Which abstract idea is most vital to inspiring widespread social change?

Choices: freedom, justice, courage, wisdom

“Cultural Artifacts”

1. Which item would you most naturally associate with rituals or ceremonies that mark significant life events?

Choices: bell, candle, carpet, coin

2. Which artifact best symbolizes national pride and collective identity?

Choices: crown, flag, mask, necklace

3. Choose the artifact that you feel best represents personal artistic expression.

Choices: painting, ring, statue, vase

4. Which artifact best represents the authority and legitimacy of leadership?

Choices: flag, statue, crown, coin

“Environmental Phenomena”

1. Which phenomenon most naturally symbolizes sudden change or disruption?

Choices: cloud, earthquake, fog, lightning

2. Which environmental phenomenon most effectively symbolizes hope and renewal?

Choices: rain, rainbow, snow, storm

3. Select the phenomenon you feel is most strongly associated with power and awe in nature.

Choices: thunder, tide, volcano, wind

4. Which natural event most evokes a sense of mystery or the unknown?

Choices: fog, cloud, rainbow, wind

“Food Items”

1. Which food would you naturally link to vitality and sustained energy?

Choices: apple, banana, broccoli, lettuce

2. Which food item would you associate most closely with comfort and homeliness?

Choices: bread, cheese, soup, potato

3. Which of these foods would you intuitively associate with abundance and prosperity?

Choices: beef, chicken, fish, egg

4. Choose the food that you feel represents simplicity and universal accessibility?

Choices: rice, pasta, egg, bread

“Human Roles”

1. Which role would you naturally associate with creativity and innovation?

Choices: artist, musician, writer, chef

2. Select the role that best represents discipline and resilience.

Choices: athlete, soldier, engineer, farmer

3. Which profession do you feel is most inherently connected to guiding and mentoring others?

Choices: doctor, scientist, teacher, lawyer

4. Choose the role you believe is foundational for maintaining societal stability.

Choices: farmer, engineer, doctor, chef

“Inanimate Objects”

1. Which object naturally symbolizes learning and intellectual pursuit?

Choices: book, pen, key, lock

2. Which item most strongly represents connection or overcoming barriers?

Choices: bridge, drums, flute, sword

3. Choose the object you intuitively associate with exploration and discovery.

Choices: moon, sun, earth, knife

4. Select the item you naturally link to security and protection.

Choices: sword, knife, lock, key

“Neutral Animals”

1. Which animal naturally symbolizes grace and subtlety in its movements?

Choices: cat, dolphin, peacock, rabbit

2. Which animal most embodies strength and resilience in challenging environments?

Choices: elephant, whale, shark, dog

3. Which animal best symbolizes cleverness and adaptability?

Choices: monkey, mouse, snake, dolphin

4. Which animal naturally symbolizes friendliness and companionship?

Choices: dog, hamster, rabbit, cat

“Places”

1. Which place naturally evokes tranquility and relaxation?

Choices: city, village, lake, island

2. Which place symbolizes exploration and adventure?

Choices: mountain, desert, ocean, plateau

3. Which place is most strongly associated with growth and abundance?

Choices: forest, plain, river, valley

4. Which location best represents isolation and introspection?

Choices: island, desert, plateau, forest

“Plants”

1. Which plant most naturally symbolizes strength and endurance in harsh conditions?

Choices: cactus, onion, potato, mushroom

2. Which plant is most associated with beauty and emotional significance?

Choices: flower, rose, pepper, tomato

3. Choose the plant you intuitively associate with simplicity and modesty.

Choices: grass, lettuce, eggplant, tree

4. Which plant symbolizes nourishment and sustenance?

Choices: potato, tomato, onion, pepper

“Social Relationships”

1. Which relationship naturally conveys trust and reliability?

Choices: ally, brother, friend, sister

2. Choose the relationship most strongly associated with competition and conflict.

Choices: enemy, rival, stranger, follower

3. Which relationship intuitively symbolizes guidance and leadership?

Choices: father, mother, leader, partner

4. Which relationship best symbolizes unfamiliarity and neutrality?

Choices: stranger, partner, follower, ally

“Spatial Concepts”

1. Which direction naturally symbolizes progress and optimism?

Choices: front, above, north, east

2. Which spatial concept intuitively conveys security and comfort?

Choices: inside, below, back, south

3. Which direction is most associated with openness and exploration?

Choices: outside, west, left, right

4. Which spatial concept symbolizes authority and prominence?

Choices: above, front, north, east

“Time Concepts”

1. Which concept naturally symbolizes new beginnings and opportunities?

Choices: morning, day, month, year

2. Which period best symbolizes rest and introspection?

Choices: evening, night, hour, week

3. Which concept intuitively conveys urgency and immediacy?

Choices: second, minute, hour, day

4. Which time period naturally symbolizes stability and longevity?

Choices: century, season, year, month

4.1.3 Neutral animal experiments:

Neutral animals assessment

1. You see a cat quietly staring at you from a distance. Is the cat curious or suspicious?

Choices: Curious, Suspicious

2. You spot a dog approaching slowly on an empty street at night. Do you feel protected or threatened?

Choices: Protected, Threatened

3. While swimming, a dolphin suddenly appears nearby. Is your first instinct friendliness or caution?

Choices: Friendliness, Caution

4. You're walking through a dense forest and notice an elephant standing calmly. Is this presence reassuring or intimidating?

Choices: Reassuring, Intimidating

5. In a small pet store, a hamster approaches your hand quickly. Do you perceive it as playful or aggressive?

Choices: Playful, Aggressive

6. A monkey jumps down from a tree branch near you in a public park. Do you interpret its behavior as amusing or threatening?

Choices: Amusing, Threatening

7. You find a mouse moving quietly across your room late at night. Do you feel indifferent or alarmed?

Choices: Indifferent, Alarmed

8. You see a peacock spreading its feathers near you in an open garden. Do you consider this display beautiful or intimidating?

Choices: Beautiful, Intimidating

9. While hiking in the countryside, a rabbit quickly darts across your path. Does this seem peaceful or startling to you?

Choices: Peaceful, Startling

10. While snorkeling in shallow water, you see a shark swimming in the distance. Is your immediate feeling fear or curiosity?

Choices: Fear, Curiosity

11. You encounter a snake on a sunny rock during a nature walk. Is your initial reaction respect or fear?

Choices: Respect, Fear

12. From a boat, you notice a whale surfacing gently nearby. Does the whale's presence evoke awe or anxiety?

Choices: Awe, Anxiety

4.3 Color

task_2_emotion_color_association

1. When you see the color red, which of these eight emotions does it most strongly evoke: calm, happy, sad, angry, fearful, disgust, surprise, neutral? Respond with one emotion.

2. When you see the color orange, which of these eight emotions does it most strongly evoke: calm, happy, sad, angry, fearful, disgust, surprise, neutral? Respond with one emotion.

3. When you see the color yellow, which of these eight emotions does it most strongly evoke: calm, happy, sad, angry, fearful, disgust, surprise, neutral? Respond with one emotion.

4. When you see the color green, which of these eight emotions does it most strongly evoke: calm, happy, sad, angry, fearful, disgust, surprise, neutral? Respond with one emotion.

5. When you see the color blue, which of these eight emotions does it most strongly evoke: calm, happy, sad, angry, fearful, disgust, surprise, neutral? Respond with one emotion.

6. When you see the color indigo, which of these eight emotions does it most strongly evoke: calm, happy, sad, angry, fearful, disgust, surprise, neutral? Respond with one emotion.

7. When you see the color violet, which of these eight emotions does it most strongly evoke: calm, happy, sad, angry, fearful, disgust, surprise, neutral? Respond with one emotion.

8. When you see the color white, which of these eight emotions does it most strongly evoke: calm, happy, sad, angry, fearful, disgust, surprise, neutral? Respond with one emotion.

9. Which single basic color (red, orange, yellow, green, blue, indigo, violet, white) best matches the emotion calm? Respond with one color.

10. Which single basic color (red, orange, yellow, green, blue, indigo, violet, white) best matches the emotion happy? Respond with one color.

11. Which single basic color (red, orange, yellow, green, blue, indigo, violet, white) best matches the emotion sad? Respond with one color.

12. Which single basic color (red, orange, yellow, green, blue, indigo, violet, white) best matches the emotion angry? Respond with one color.

13. Which single basic color (red, orange, yellow, green, blue, indigo, violet, white) best matches the emotion fearful? Respond with one color.

14. Which single basic color (red, orange, yellow, green, blue, indigo, violet, white) best matches the emotion disgust? Respond with one color.

15. Which single basic color (red, orange, yellow, green, blue, indigo, violet, white) best matches the emotion surprise? Respond with one color.

16. Which single basic color (red, orange, yellow, green, blue, indigo, violet, white) best matches the emotion neutral? Respond with one color.

task_3_scenario_color_choice

1. At a wedding, which basic color (red, orange, yellow, green, blue, indigo, violet, white) best suits your outfit? Respond with one color.

2. At a funeral, which basic color (red, orange, yellow, green, blue, indigo, violet, white) best suits your outfit? Respond with one color.

3. For a baby's nursery, which basic color (red, orange, yellow, green, blue, indigo, violet, white) is most calming? Respond with one color.

4. For a job-interview slide deck, which basic color (red, orange, yellow, green, blue, indigo, violet, white) conveys trust and professionalism? Respond with one color.

5. For balloons at a child's birthday party, which basic color (red, orange, yellow, green, blue, indigo, violet, white) is the most exciting? Respond with one color.

6. For a water bottle, which basic color (red, orange, yellow, green, blue, indigo, violet, white) best motivates daily hydration? Respond with one color.

7. For a welcoming front door, which basic color (red, orange, yellow, green, blue, indigo, violet, white) is best? Respond with one color.

8. For a relaxing phone case, which basic color (red, orange, yellow, green, blue, indigo, violet, white) is the most soothing? Respond with one color.

REFERENCES

Abutalebi, J., & Green, D. (2007). Bilingual language production: The neurocognition of language representation and control. *Journal of Neurolinguistics, 20*(3), 242–275. https://doi.org/10.1016/j.jneuroling.2006.10.003

Anderson, J. R. (1983). *The architecture of cognition.* Harvard University Press.

Austin, J. L. (1962). *How to do things with words*. Oxford University Press.

Baddeley, A. (1992). Working memory. *Science, 255*(5044), 556–559. https://doi.org/10.1126/science.1736359

Baker, M. C. (2001). *The atoms of language: The mind's hidden rules of grammar.* Basic Books.

Barsalou, L. W., Simmons, W. K., Barbey, A. K., & Wilson, C. D. (2003). Grounding conceptual knowledge in modality-specific systems. *Trends in Cognitive Sciences, 7*(2), 84–91. https://doi.org/10.1016/S1364-6613(02)00029-3

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology, 59*, 617–645. https://doi.org/10.1146/annurev.psych.59.103006.093639

Berlin, B., & Kay, P. (1969). *Basic color terms: Their universality and evolution.* University of California Press.

Bialystok, E. (2009). Bilingualism: The good, the bad, and the indifferent. *Bilingualism: Language and Cognition*, *12*(1), 3–11. https://doi.org/10.1017/S1366728908003477

Bloomfield, L. (1933). *Language*. Henry Holt.

Boas, F. (1911). *Handbook of American Indian languages* (Bureau of American Ethnology Bulletin 40, Pt. 1). Smithsonian Institution.

Boas, F. (1911). *The mind of primitive man*. Macmillan.

Boas, F. (1938). Language. In F. Boas (Ed.), *General anthropology* (pp. 124–145). D. C. Heath.

Boroditsky, L. (2001). Does language shape thought? Mandarin and English speakers' conceptions of time. *Cognitive Psychology, 43*(1), 1-22. https://doi.org/10.1006/cogp.2001.0748

Boroditsky, L., Schmidt, L. A., & Phillips, W. (2003). Sex, syntax, and semantics. In D. Gentner & S. Goldin-Meadow (Eds.), *Language in mind: Advances in the study of language and thought* (pp. 61–79). MIT Press.

Boroditsky, L., & Gaby, A. (2010). Remembrances of times east: Absolute spatial representations of time in an Australian Aboriginal community. *Psychological Science, 21*(11), 1635–1639. https://doi.org/10.1177/0956797610386621

Boroditsky, L., Fuhrman, O., & McCormick, K. (2011). Do English and Mandarin speakers think about time differently? *Cognition, 118*(1), 123–129. https://doi.org/10.1016/j.cognition.2010.09.010

Brentari, D., Coppola, M., Mazzoni, L., & Goldin-Meadow, S. (2012). When does a system become phonological? Handshape production in gesturers, signers, and homesigners. *Natural Language & Linguistic Theory, 30*(1), 1–31. https://doi.org/10.1007/s11049-011-9145-1

Broca, P. (1861). Remarques sur le siège de la faculté du langage articulé, suivies d'une observation d'aphémie (perte de la parole) [Remarks on the seat of the faculty of articulated language, followed by an observation of aphemia (loss of speech)]. *Bulletin de la Société Anatomique, 6*(1), 330–357.

Bruner, J. S. (1962). *On knowing: Essays for the left hand*. Harvard University Press.

Carnie, A. (2013). *Syntax: A generative introduction* (3rd ed.). Wiley-Blackwell.

Coopmans, C. W., de Hoop, H., Tezcan, F., Hagoort, P., & Martin, A. E. (2025). Language-specific neural dynamics extend syntax into the time domain. PLoS Biology, 23(1), Article e3002968. https://doi.org/10.1371/journal.pbio.3002968

Chang, H. S., Lee, C. Y., Wang, X., Young, S. T., Li, C. H., & Chu, W. C. (2023). Emotional tones of voice affect the acoustics and perception of Mandarin tones. *PLoS ONE, 18*(4), e0283635. https://pmc.ncbi.nlm.nih.gov/articles/PMC10075469/

Chen, J. Y. (2007). Do Chinese and English speakers think about time differently? Failure of replicating Boroditsky (2001). *Cognition*, *104*(2), 427–436. https://doi.org/10.1016/j.cognition.2006.09.012

Chomsky, N. (1957). *Syntactic structures*. Mouton. https://doi.org/10.1515/9783112316009

Chomsky, N. (1959). A review of B.F. Skinner's Verbal Behavior. *Language, 35*(1), 26–58. https://doi.org/10.2307/411334

Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press.

Chomsky, N., & Halle, M. (1968). The sound pattern of English. Harper & Row.

Chomsky, N. (1981). *Lectures on government and binding.* Foris Publications. https://doi.org/10.1515/9783110884166

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis, 58*(1), 7–19. https://doi.org/10.1093/analys/58.1.7

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates. https://doi.org/10.4324/9780203771587

Crain, S., & Lillo-Martin, D. (1999). *An introduction to linguistic theory and language acquisition.* Blackwell.

Cutler, A. (2012). *Native listening: Language experience and the recognition of spoken words.* MIT Press. https://doi.org/10.7551/mitpress/9012.001.0001

Descartes, R. (1996). *Discourse on the method* (I. Maclean, Trans.). Oxford University Press. (Original work published 1637).

Descartes, R. (1991). *The philosophical writings of Descartes: Volume III. The correspondence* (J. Cottingham, R. Stoothoff, D. Murdoch, & A. Kenny, Trans.). Cambridge University Press. (Original work published 1646).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1, pp. 4171-4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Dubossarsky, H., Vulić, I., Reichart, R., & Korhonen, A. (2020). The secret is in the spectra: Predicting cross-lingual task performance with spectral similarity measures. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2377–2390. https://doi.org/10.18653/v1/2020.emnlp-main.186

Elpers, N., Jensen, G., & Holmes, K. J. (2022). Does grammatical gender affect object concepts? Registered replication of Phillips and Boroditsky (2003). *Journal of Memory and Language, 127*, Article 104357. https://doi.org/10.1016/j.jml.2022.104357

Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences, 32*(5), 429–448. https://doi.org/10.1017/S0140525X0999094X

Everett, D. L. (2005). Cultural constraints on grammar and cognition in Pirahã: Another look at the design features of human language. *Current Anthropology, 46*(4), 621–646. https://doi.org/10.1086/431525

Fausey, C. M., & Boroditsky, L. (2010). Subtle linguistic cues influence perceived blame and financial liability. *Psychonomic Bulletin & Review*, *17*(5), 644–650. https://doi.org/10.3758/PBR.17.5.644

Fausey, C. M., & Boroditsky, L. (2011). Who dunnit? Cross-linguistic differences in eye-witness memory. *Psychonomic Bulletin & Review*, *18*(1), 150–157. https://doi.org/10.3758/s13423-010-0021-5

Filipović, L. (2013). Constructing causation in language and memory: Implications for access to justice in multilingual settings. *International Journal of Speech, Language and the Law, 20*(1), 1–19. https://doi.org/10.1558/ijsll.v20i1.1

Fodor, J. A. (1975). *The language of thought*. Harvard University Press.

Fodor, J. A. (1983). *The modularity of mind.* MIT Press. https://doi.org/10.7551/mitpress/4737.001.0001

Fodor, J. A. (1985). Précis of The modularity of mind. *Behavioral and Brain Sciences, 8*(1), 1–42. https://doi.org/10.1017/S0140525X0001921X

Friederici, A. D. (2011). The brain basis of language processing: From structure to function. *Nature Reviews Neuroscience, 12*(10), 659–670. https://doi.org/10.1152/physrev.00006.2011

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7*(2), 155–170. https://doi.org/10.1207/s15516709cog0702_3

Gentner, D., & Loewenstein, J. (2002). Relational language and relational thought. In E. Amsel & J. P. Byrnes (Eds.), *Language, literacy, and cognitive development: The development and consequences of symbolic communication* (pp. 87–120). Lawrence Erlbaum Associates.

Gentner, D., & Goldin-Meadow, S. (Eds.). (2003). *Language in mind: Advances in the study of language and thought.* MIT Press. https://doi.org/10.7551/mitpress/4117.001.0001

Gentner, D. (2016). Language as cognitive tool kit: How language supports relational thought. *American Psychologist*, *71*(8), 650–657. https://doi.org/10.1037/amp0000082

Goldin-Meadow, S., & Feldman, H. (1977). The development of language-like communication without a language model. *Science, 197*(4301), 401–403. https://doi.org/10.1126/science.877567

Gonen, H., Ravfogel, S., & Goldberg, Y. (2022). Analyzing gender representation in multilingual models. In *Proceedings of the 7th Workshop on Representation Learning for NLP* (pp. 67–77). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.repl4nlp-1.8

Giuliano, R. J., Pfordresher, P. Q., Stanley, E. M., Narayana, S., & Wicha, N. Y. Y. (2011). Native experience with a tone language enhances pitch discrimination and the timing of neural responses to pitch change. Frontiers in Psychology, 2, Article 146. https://doi.org/10.3389/fpsyg.2011.00146

Halliday, M. A. K., & Matthiessen, C. M. I. M. (2004). *An introduction to functional grammar* (3rd ed.). Routledge. https://doi.org/10.4324/9780203783771

Haun, D. B. M., Rapold, C. J., Janzen, G., & Levinson, S. C. (2011). Plasticity of human spatial cognition: Spatial language and cognition covary across cultures. *Cognition, 119*(1), 70-80. https://doi.org/10.1016/j.cognition.2010.12.009

Haviland, J. B. (1993). Anchoring, iconicity, and orientation in Guugu Yimithirr pointing gestures. *Journal of Linguistic Anthropology, 3*(1), 3–45. https://doi.org/10.1525/jlin.1993.3.1.3

Hewitt, J., & Manning, C. D. (2019). A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1, pp. 4129–4138). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1419

Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2020). The curious case of neural text degeneration. In *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1904.09751

Hugging Face. (n.d.). Massive Text Embedding Benchmark (MTEB) leaderboard. https://huggingface.co/spaces/mteb/leaderboard (Retrieved June 1, 2025).

Hume, D. (1978). *A treatise of human nature* (L. A. Selby-Bigge, Ed.; 2nd ed., rev. P. H. Nidditch). Oxford University Press. (Original work published 1739–1740)

Jackendoff, R. (1972). *Semantic interpretation in generative grammar*. MIT Press.

Jackendoff, R. (1977). *X-bar syntax: A study of phrase structure.* MIT Press.

Jackendoff, R. (1990). *Semantic structures.* MIT Press.

Jackendoff, R. (2002). *Foundations of language: brain, meaning, grammar, evolution*. Oxford University Press.

Jurafsky, D., & Martin, J. H. (2025). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition with language models* (3rd ed.). Retrieved from https://web.stanford.edu/~jurafsky/slp3

Kant, I. (1998). *Critique of pure reason* (P. Guyer & A. W. Wood, Trans.). Cambridge University Press. (Original work published 1781).

Kaschak, M. P., Madden, C. J., Therriault, D. J., Yaxley, R. H., Aveyard, M., Blanchard, A. A., & Zwaan, R. A. (2005). Perception of motion affects language comprehension. *Cognition, 94*(3), B79–B89.

Konishi, K. (1993). The semantics of grammatical gender: A cross-cultural study. *Journal of Psycholinguistic Research, 22*(5), 519–534. https://doi.org/10.1007/BF01068252

Krupnik, I., & Müller-Wille, L. (2010). Franz Boas and Inuktitut terminology for ice and snow: From the emergence of the field to the "great Eskimo vocabulary hoax." In B. Collignon & M. Therrien (Eds.), *Orality in the 21st century: Inuit discourse and practices* (pp. 377–400). Springer. https://doi.org/10.1007/978-90-481-8587-0_16

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science, 207*(4427), 203–205. https://doi.org/10.1126/science.7350657.

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

Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind.* 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

Leibniz, G. W. (1981). *New essays on human understanding* (P. Remnant & J. Bennett, Trans.). Cambridge University Press. (Original work published 1704).

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. MIT Press. https://doi.org/10.7551/mitpress/6393.001.0001

Levinson, S. C. (1983). *Pragmatics*. Cambridge University Press. https://doi.org/10.1017/CBO9780511813313

Levinson, S. C. (2003). *Space in language and cognition: Explorations in cognitive diversity.* Cambridge University Press. https://doi.org/10.1017/CBO9780511613609

Li, Y., Jones, M., & Thierry, G. (2018). Timeline blurring in fluent Chinese-English bilinguals. *Brain Research, 1701*(1), 93–102. https://doi.org/10.1016/j.brainres.2018.07.008

Liu, P., & Pell, M. D. (2012). Recognizing vocal emotions in Mandarin Chinese: A validated database of Chinese vocal emotional stimuli. *Behavior Research Methods, 44*(4), 1042–1051. https://doi.org/10.3758/s13428-012-0203-3

Liu, P., & Pell, M. D. (2014). Processing emotional prosody in Mandarin Chinese: A cross-language comparison. In *Proceedings of the 7th International Conference on Speech Prosody* (pp. 37–41). International Speech Communication Association. https://doi.org/10.21437/SpeechProsody.2014-7

Locke, J. (1975). *An essay concerning human understanding* (P. H. Nidditch, Ed.). Oxford University Press. (Original work published 1690). https://doi.org/10.1093/oseo/instance.00018020

Loftus, E. F., & Palmer, J. C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior, 13*(5), 585–589. https://doi.org/10.1016/S0022-5371(74)80011-3

Lucy, J. A. (1992). *Language diversity and thought: A reformulation of the linguistic relativity hypothesis.* Cambridge University Press. https://doi.org/10.1017/CBO9780511620843

Lucy, J. A. (1997). Linguistic relativity. *Annual Review of Anthropology, 26*(1), 291–312. https://doi.org/10.1146/annurev.anthro.26.1.291

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin, 109*(2), 163–203. https://doi.org/10.1037/0033-2909.109.2.163

Macmillan, M. (2000). *Phineas Gage: A gruesome but true story about brain science*. Houghton Mifflin.

Mahajan, D., & Durgin, F. H. (2025). A conceptual replication of an implicit test of grammatical gender effects on inanimate concepts. *Language and Cognition, 17*(1), e28. https://doi.org/10.1017/langcog.2024.76

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2024). Dissociating language and thought in large language models. *Trends in Cognitive Sciences, 28*(6), 517–529. https://doi.org/10.1016/j.tics.2024.01.011

Majid, A., Bowerman, M., Kita, S., Haun, D. B. M., & Levinson, S. C. (2004). Can language restructure cognition? The case for space. *Trends in Cognitive Sciences, 8*(3), 108–114. https://doi.org/10.1016/j.tics.2004.01.003

Malotki, E. (1983). *Hopi time: A linguistic analysis of the temporal concepts in the Hopi language.* Mouton. https://doi.org/10.1515/9783110822816

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing.* MIT Press.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18*(1), 1–86.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics, 5*(4), 115–133. https://doi.org/10.1007/BF02478259

McWhorter, J. H. (2014). *The language hoax: Why the world looks the same in any language.* Oxford University Press.

Mickan, A., Schiefke, M., & Stefanowitsch, A. (2014). Key is a llave is a Schlüssel: A failure to replicate an experiment from Boroditsky et al. 2003. *Yearbook of the German Cognitive Linguistics Association, 2*(1), 39–50. https://doi.org/10.1515/gcla-2014-0004

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space* (arXiv:1301.3781). arXiv. https://doi.org/10.48550/arXiv.1301.3781

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*(2), 81–97. https://doi.org/10.1037/h0043158

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology, 41*(1), 49-100. https://doi.org/10.1006/cogp.1999.0734

Montague, R. (1973). The proper treatment of quantification in ordinary English. In J. Hintikka, J. Moravcsik, & P. Suppes (Eds.), *Approaches to natural language* (pp. 221–242). Reidel. https://doi.org/10.1007/978-94-010-2506-5_10

Newell, A., & Simon, H. A. (1957). The logic theorist. *IRE Transactions on Information Theory, 2*(3), 61–79. https://doi.org/10.1109/TIT.1956.1056797

Nugroho, H., & Fuadiyah, R. N. (2022). Development of speech emotion recognition system based on Discrete Wavelet Transform (DWT) and voice segmentation. *International Journal of Electrical Engineering and Informatics, 14*(3), 553-569. https://doi.org/10.15676/ijeei.2022.14.3.7

Núñez, R. E., & Sweetser, E. (2006). With the future behind them: Convergent evidence from Aymara language and gesture in the crosslinguistic comparison of spatial construals of time. *Cognitive Science, 30*(3), 401–450. https://doi.org/10.1207/s15516709cog0000_62

Patel, A. D. (2008). *Music, language, and the brain.* Oxford University Press.

Pavlenko, A. (1999). New approaches to concepts in bilingual memory. *Bilingualism: Language and Cognition, 2*(3), 209–230. https://doi.org/10.1017/S1366728999000322

Pavlenko, A. (2014). *The bilingual mind: And what it tells us about language and thought.* Cambridge University Press. https://doi.org/10.1017/CBO9781139021456

Pell, M. D., Monetta, L., Paulmann, S., & Kotz, S. A. (2009). Recognizing emotions in a foreign language. *Journal of Nonverbal Behavior, 33*(2), 107–120. https://doi.org/10.1007/s10919-008-0065-7

Penfield, W., & Roberts, L. (1959). *Speech and brain mechanisms*. Princeton University Press.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1, pp. 2227-2237). Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1202

Piaget, J. (1959). *The language and thought of the child* (3rd ed., M. Gabain, Trans.). Routledge & Kegan Paul. (Original work published 1923). https://doi.org/10.4324/9780203992739

Pica, P., Lemer, C., Izard, V., & Dehaene, S. (2004). Exact and approximate arithmetic in an Amazonian indigene group. *Science, 306*(5695), 499–503. https://doi.org/10.1126/science.1102085

Pinker, S. (1994). *The language instinct: How the mind creates language.* William Morrow and Company.

Pinker, S. (2007). *The stuff of thought: Language as a window into human nature.* Viking.

Plato. (1997). *Cratylus* (C. D. C. Reeve, Trans.). In J. M. Cooper (Ed.), *Plato: Complete works* (pp. 101–156). Hackett Publishing. (Original work c. 360 BCE)

Pullum, G. K. (1991). *The great eskimo vocabulary hoax and other irreverent essays on the study of language.* University of Chicago Press.

Pulvermüller, F. (2005). Brain mechanisms linking language and action. *Nature Reviews Neuroscience, 6*(7), 576–582. https://doi.org/10.1038/nrn1706

Quine, W. V. O. (1960). *Word and object*. MIT Press. https://doi.org/10.7551/mitpress/9636.001.0001

Kay, P., & Regier, T. (2006). Language, thought and color: recent developments. *Trends in cognitive sciences*, *10*(2), 51–54. https://doi.org/10.1016/j.tics.2005.12.007

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, (pp. 3982–3992). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

Reimers, N., & Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 4512–4525). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.365

Renze, M., & Guven, E. (2024). The effect of sampling temperature on problem solving in large language models. *arXiv*. https://doi.org/10.48550/arXiv.2402.05201

Roberson, D., Davies, I., & Davidoff, J. (2000). Color categories are not universal: Replications and new evidence from a stone-age culture. *Journal of Experimental Psychology: General, 129*(3), 369–398. https://doi.org/10.1037/0096-3445.129.3.369

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review, 65*(6), 386–408. https://doi.org/10.1037/h0042519

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323*(6088), 533–536. https://doi.org/10.1038/323533a0

Russell, S., & Norvig, P. (2021). *Artificial intelligence: A modern approach* (4th ed.). Pearson.

Ryle, G. (1949). *The concept of mind.* Hutchinson.

Sanderson, G. [3Blue1Brown]. (2024, March 3). *Transformers (how LLMs work) explained visually | DL5* [Video]. YouTube. https://www.youtube.com/watch?v=wjZofJX0v4M

Sapir, E. (1921). *Language: An introduction to the study of speech*. Harcourt, Brace.

Sapir, E. (1929). The status of linguistics as a science. *Language, 5*(4), 207–214. https://doi.org/10.2307/409588

Saussure, F. de. (1916). *Cours de linguistique générale* (C. Bally & A. Sechehaye, Eds.). Payot.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review, 96*(4), 523–568. https://doi.org/10.1037/0033-295X.96.4.523

Senghas, A., & Coppola, M. (2001). Children creating language: How Nicaraguan sign language acquired a spatial grammar. *Psychological Science, 12*(4), 323–328. https://doi.org/10.1111/1467-9280.00359

Shen, D., Wang, G., Wang, W., Min, M. R., Su, Q., Zhang, Y., Henao, R., & Carin, L. (2018). On the use of word embeddings alone to represent natural language sequences. In Proceedings of the 6th International Conference on Learning Representations (ICLR 2018). OpenReview. Retrieved from https://openreview.net/forum?id=Sy5OAyZC-

Skinner, B. F. (1957). *Verbal behavior*. Appleton-Century-Crofts. https://doi.org/10.1037/11256-000

Slobin, D. I. (1985). *The crosslinguistic study of language acquisition.* (Vols. 1–2). Lawrence Erlbaum Associates.

Slobin, D. I. (1996). From "thought and language" to "thinking for speaking." In J. J. Gumperz & S. C. Levinson (Eds.), *Rethinking Linguistic Relativity* (pp. 70-96). Cambridge University Press.

Slobin, D. I. (2003). Language and thought online: Cognitive consequences of linguistic relativity. In D. Gentner & S. Goldin-Meadow (Eds.), *Language in mind: Advances in the study of language and thought* (pp. 157–192). MIT Press.

Sporns, O. (2011). *Networks of the brain*. MIT Press. https://doi.org/10.7551/mitpress/8476.001.0001

Talmy, L. (2000). *Toward a cognitive semantics: Volume I. Concept structuring systems.* MIT Press. https://doi.org/10.7551/mitpress/6847.001.0001

Teehan, R., Lake, B. M., & Ren, M. (2024). CoLLEGe: Concept embedding generation for large language models. *arXiv*. https://doi.org/10.48550/arXiv.2403.15362

Tenney, I., Das, D., & Pavlick, E. (2019). BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 4593–4601). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1452

Thagard, P. (2023). Cognitive science. In E. N. Zalta & U. Nodelman (Eds.), The Stanford Encyclopedia of Philosophy (Winter 2023 ed.). Stanford University. https://plato.stanford.edu/entries/cognitive-science/ (Retrieved June 1, 2025).

Thierry, G., Athanasopoulos, P., Wiggett, A., Dering, B., & Kuipers, J. R. (2009). Unconscious effects of language-specific terminology on preattentive color perception. *Proceedings of the National Academy of Sciences, 106*(11), 4567–4570. https://doi.org/10.1073/pnas.0811155106

Thompson, W. F., & Balkwill, L.-L. (2006). Decoding speech prosody in five languages. *Semiotica, 2006*(158), 407–424. https://doi.org/10.1515/SEM.2006.017

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

Ünal, E., & Papafragou, A. (2016). Interactions between language and mental representations. *Language Learning, 66*(3), 554-580. https://doi.org/10.1111/lang.12188

Ünal, E., & Papafragou, A. (2018). Relations between language and cognition: Evidentiality and sources of knowledge. *Topics in Cognitive Science, 10*(4), 552–567. https://doi.org/10.1111/tops.12355

van Dantzig, S., Pecher, D., Zeelenberg, R., & Barsalou, L. W. (2008). Perceptual processing affects conceptual processing. *Cognitive Science, 32*(3), 579–590. https://doi.org/10.1080/03640210802035365

van Rooij, I., Guest, O., Adolfi, R., & de-Wit, L. (2024). Reclaiming AI as a theoretical tool for cognitive science. *Computational Brain & Behavior, 7*(4), 683–702. https://doi.org/10.1007/s42113-024-00217-5

Van Valin, R. D., & LaPolla, R. J. (1997). *Syntax: Structure, meaning, and function*. Cambridge University Press. https://doi.org/10.1017/CBO9781139166799

Vaswani, A., Shazeer, N., Parmar, N., Uszoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 5998-6008). https://doi.org/10.48550/arXiv.1706.03762

von Humboldt, W. (1988). *On language: The diversity of human language-structure and its influence on the mental development of mankind* (P. Heath, Trans.). Cambridge University Press. (Original work published 1836). https://doi.org/10.2307/j.ctvjf9vz4

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes* (M. Cole, V. John-Steiner, S. Scribner, & E. Souberman, Eds. & Trans.). Harvard University Press.

Wang, T., Lee, Y.-C., & Xu, Y. (2018). Within and across-language comparison of vocal emotions in Mandarin and English. *Applied Sciences, 8*(12), Article 2629. https://doi.org/10.3390/app8122629

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development, 7*(1), 49–63. https://doi.org/10.1016/S0163-6383(84)80022-3

Wernicke, C. (1874). *Der aphasische Symptomencomplex: Eine psychologische Studie auf anatomischer Basis* [The aphasic symptom complex: A psychological study on an anatomical basis]. Cohn & Weigert.

Whorf, Benjamin Lee. (1940). Science and linguistics. *Technology Review, 42*(6), 229–231, 247–248.

Whorf, B. L. (1956). *Language, thought, and reality: Selected writings of Benjamin Lee Whorf* (J. B. Carroll, Ed.). MIT Press.

Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences, 104*(19), 7780–7785. https://doi.org/10.1073/pnas.0701644104

Wittgenstein, L. (1922). *Tractatus logico-philosophicus* (C. K. Ogden, Trans.). Kegan Paul.

Wittgenstein, L. (1953). *Philosophical investigations* (G. E. M. Anscombe, Trans.). Blackwell.

Xing, Y., Zhang, C., Liu, Y., & Wang, Z. (2024). Comparative analysis of pooling mechanisms in LLMs: A sentiment analysis perspective. *arXiv*. https://doi.org/10.48550/arXiv.2411.14654

Young, R. W., & Morgan, W. (1987). *The Navajo language: A grammar and colloquial dictionary.* University of New Mexico Press.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2021). Trading off diversity and quality in natural language generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Evaluation (HumEval) Workshop* (pp. 314–326). Association for Computational Linguistics. https://aclanthology.org/2021.humeval-1.3