



**The structure of the visual world in the human mind:
measuring object statistics from large sets of
real-world images and their impact on behaviour**

Dissertation

for the degree of

Doctor of Natural Science

submitted to the

Department of Psychology and Sports Sciences of

Goethe University, Frankfurt am Main

by

Jacopo Turini

from Cremona, Italy

Frankfurt am Main 2022

(D30)

Vom Fachbereich Psychologie und Sportwissenschaften der
Johann Wolfgang Goethe Universität als Dissertation angenommen.

Dekanin:

Prof. Dr. Sonja Rohrmann

Gutachter:

Prof. Dr. Melissa Le-Hoa Võ

Prof. Dr. Yee Lee Shing

Datum der Disputation: 19.04.2023

“Diceva che gli spiriti esistevano, ma non nei palazzi, nei vicoli e vicino alle porte antiche del Vasto. Esistevano nelle orecchie delle persone, negli occhi quando gli occhi guardavano dentro e non fuori, nella voce appena si comincia a parlare, nella testa quando si pensa, perché le parole ma anche le immagini sono zeppe di fantasmi”

E. F.

Contents

- **Zusammenfassung**.....p.11
- **Summary**.....p.17
- **Theoretical background**.....p.23
 - Mind and World.....p.24
 - Implicit and statistical learning: bringing the world into the mind...p.29
 - Implicit Statistical Learning in visual cognition: from contextual cueing to visual scenes.....p.31
 - Scene Grammar: learning the structure of *our* visual world.....p.34
 - Using large datasets to capture the structure of the visual world.....p.37
 - The current investigation.....p.41
- **Summary of empirical studies**.....p.43
 - Study 1: “Access to meaning from visual input: Object and word frequency effects in categorization behavior.”p.44
 - Study 2: “Hierarchical organization of objects in scenes is reflected in mental representations of objects”p.52
 - Study 3: “Scene hierarchy structures mental object representations while flexibly adapting to varying task demands”p.58
- **General discussion**..... p.63
 - Advantages and problems in using image datasets to extract objects statistics.....p.64
 - The future of Scene Grammar.....p.68
 - Scene Grammar and applications to Artificial Intelligence.....p.72
 - Our mind and our world.....p.74

- Original manuscripts.....	p.77 -
Bibliography.....	-
Erklärungen.....	

List of figures

• Figure 1 – Study 1: Experimental procedures.....	p.45
• Figure 2 – Study 1: Results Experiment 1.....	p.46
• Figure 3 – Study 1: Results Experiment 2.....	p.47
• Figure 4 – Study 1: Example of interaction between object frequency and conceptual distinctiveness.....	p.49
• Figure 5 – Study 1: Results interaction between object frequency and conceptual distinctiveness.....	p.50
• Figure 6 – Study 1: Results Experiment 3.....	p.51
• Figure 7 – Study 2: Presentation of the study.....	p.53
• Figure 8 – Study 2: Representation (Dis)similarity Matrices (RDMs) for the hierarchical measures.....	p.54
• Figure 9 – Study 2: Behavioural Representational (Dis)similarity Matrices (RDMs).....	p.55
• Figure 10 – Study 2: Results.....	p.57
• Figure 11 – Study 3: Behavioural Representational (Dis)similarity Matrices (RDMs).....	p.59
• Figure 12 – Study 3: Results.....	p.60

Zusammenfassung

Unser Geist hat die Aufgabe, die physische und soziale Welt, in der wir uns befinden, zu repräsentieren, damit wir effizient mit ihr interagieren können. Dies führt zu einer ständigen und dynamischen Interaktion zwischen Geist und Welt, die ein Gleichgewicht herstellt, wenn die Repräsentationen gleichzeitig genau sind in Bezug auf das, was die Welt unserem Organismus mitteilt, aber auch vereinbar mit der Funktionsweise unseres Geistes.

Ein paradigmatisches Beispiel für diese Interaktion ist die Wahrnehmung, die die mentale Funktion darstellt, die kontingente Aspekte der Welt aus dem, was unsere Sinne erfassen, abbildet. Die vorherrschende philosophische Sichtweise in der Kognitionswissenschaft ist, dass unsere Wahrnehmungszustände Repräsentationen der Welt sind und keinen direkten Zugang zu dieser Welt darstellen. Diese repräsentativen Wahrnehmungszustände haben daher einen Inhalt, nämlich die Aspekte der Welt, die sie repräsentieren und die die Wahrnehmung durch die Stimulierung unserer Sinnesorgane auslösen.

Wahrnehmungsrepräsentationen werden mit Hilfe von Informationen aus dem sensorischen System aufgebaut, d. h. Bottom-up-Informationen, sie werden aber auch durch zuvor erworbene Informationen integriert, d. h. Top-down-Informationen, so dass die Wahrnehmung mit dem Gedächtnis, aber auch mit der Sprache und anderen mentalen Funktionen interagiert. Man geht davon aus, dass eine solche Organisation einen allgemeinen Mechanismus unseres Geistes/Gehirns widerspiegelt, der darin besteht, Informationen zu erwerben und zu nutzen, um effiziente Vorhersagen für die Zukunft zu treffen, wobei ältere Informationen kontinuierlich durch aktuelle Informationen aktualisiert werden.

Diese vorausschauende Verarbeitung funktioniert, weil die Welt nicht zufällig ist, sondern eine regelmäßige Struktur aufweist, aus der sich zuverlässige Erwartungen ableiten lassen. Ein Weg, den unser Verstand nutzt, um diese Vorhersagen zu treffen, ist die Anpassung an die Struktur der Welt auf implizite, automatische und unbewusste Weise, ein Prozess, der als implizites statistisches Lernen (ISL) bezeichnet wird. ISL ist ein Lernprozess, der kein Bewusstsein erfordert und zufällig und spontan abläuft, wenn wir statistischen Regelmäßigkeiten in der Welt ausgesetzt sind. Dies geschieht beim Erlernen einer Sprache in der frühen Kindheit und ermöglicht es uns, implizit auf die phonologische Struktur von Sprache zu reagieren oder Sprachmuster mit Objekten und Ereignissen zu assoziieren, um die Wortbedeutung zu lernen.

Ein spezieller Fall von ISL ist das Erlernen der räumlichen Konfiguration in der visuellen Welt, das wir auf abstrakte Anordnungen von Gegenständen anwenden, aber vor allem auch auf natürlichere Umgebungen wie die visuellen Szenen, in die wir im Alltag eintauchen. Das Wissen, das wir uns über die Struktur von visuellen Szenen aneignen, wird als "Szenengrammatik" bezeichnet, weil es uns über das Vorhandensein und die Position von Objekten auf ähnliche Weise informiert wie die sprachliche Grammatik über das Vorhandensein und die Position von Wörtern. Wir erwerben also implizit eine Semantik von Szenen, indem wir lernen, welche Objekte zu einer bestimmten Szene gehören, und wir erwerben eine Syntax von Szenen, indem wir lernen, wo Objekte innerhalb einer bestimmten Szene auf konsistente Weise positioniert sind.

Neuere Entwicklungen haben vorgeschlagen, dass das Wissen über die Szenengrammatik in einem hierarchischen System organisiert sein könnte: Die Objekte sind in der Szene angeordnet, die den allgemeineren Kontext bietet, aber innerhalb einer Szene können wir verschiedene räumliche und funktionale Gruppen von Objekten identifizieren, die als "Phrasen" bezeichnet werden und eine zweite Kontextebene bieten; innerhalb jeder Phrase haben die Objekte dann einen unterschiedlichen Status, wobei in der Regel ein Objekt ("Ankerobjekt") eine starke Vorhersage darüber bietet, wo und welches die anderen Objekte innerhalb der Phrase sind ("lokale Objekte"). Diese weiteren Aspekte der Organisation von Objekten in Szenen sind jedoch noch wenig bekannt.

Ein weiteres Problem betrifft die Art und Weise, wie wir die Struktur von Szenen messen, um die Organisation der visuellen Welt mit der Organisation in unserem Kopf zu vergleichen. Um zu entscheiden, ob ein Objekt in einer bestimmten Szene vorkommt oder nicht, und ob es an einer bestimmten Position innerhalb einer Szene vorkommt oder nicht, stützen sich die Forscher in der Regel auf ihre Intuition und ihren gesunden Menschenverstand, wobei sie diese Entscheidungen zum Teil durch unabhängige Beurteiler validieren lassen. Es hat sich jedoch gezeigt, dass diese Entscheidungen oft begrenzt sind und komplexere Informationen über die Anordnung von Objekten in Szenen verloren gehen können.

Eine mögliche Lösung für dieses Problem könnte darin bestehen, große Mengen von Bildern aus der realen Welt zu verwenden, die Anmerkungen und Segmentierungen von Objekten enthalten, um Statistiken darüber zu messen, wie Objekte in der Umgebung angeordnet sind. Diese Idee nutzt die heutzutage größere Verfügbarkeit dieser Art von Datensätzen aufgrund der fortschreitenden Entwicklung von Computer-Vision-Algorithmen und weist auch Parallelen zur etablierten Verwendung großer Textkorpora in der Sprachforschung auf.

Ziel der vorliegenden Arbeit war es, Objektstatistiken aus diesen Bilddatensätzen zu extrahieren und zu testen, ob diese zuverlässig Verhaltensreaktionen während der Objektverarbeitung vorhersagen, sowie diese Statistiken zu nutzen, um komplexere Aspekte der Szenengrammatik zu untersuchen, wie z.B. ihre hierarchische Organisation, um zu sehen, ob sich diese Organisation in unserer mentalen Organisation von Objekten widerspiegelt.

In der ersten Studie, die wir vorstellen, haben wir neuartige Maße für die Objekthäufigkeit (object frequency, OF) berechnet, die angeben, wie oft ein Objekt in großen Datensätzen von realen Bildern vorkommt. Wir haben diese Maße zur Vorhersage von Verhaltensreaktionen bei der Objekterkennung verwendet und die OFs mit Worthäufigkeitsmaßen (word frequency, WF) verglichen, die sich zur Vorhersage von Verhaltensreaktionen bei der Worterkennung bewährt haben. Im ersten Experiment dieser Reihe (Exp. 1), in dem die Teilnehmer Objektbilder oder das dazugehörige geschriebene Wort als natürliche oder künstliche Konzepte kategorisieren mussten, fanden wir nur einen signifikanten Effekt einer aus Filmuntertiteln berechneten WF mit einer besseren Leistung für häufigere Konzepte. Überraschenderweise war der Effekt sowohl bei der Worterkennung als auch bei der Objekterkennung vorhanden.

In einem zweiten Experiment (Exp. 2) und dessen Replikation (Exp. 3) führten die Teilnehmer eine Priming-Aufgabe durch, bei der sie beurteilen mussten, ob Prime und Target die gleiche Bedeutung hatten. Die entscheidende Manipulation bestand darin, dass Priming und Target entweder aus der gleichen Modalität stammten (uni-modales Priming: Wort-Prime/Wort-Target, Objekt-Prime/Objekt-Target), was hauptsächlich perzeptuelle Verarbeitung beinhaltet, oder aus verschiedenen Modalitäten (cross-modales Priming: Wort-Prime/Objekt-Target, Objekt-Prime/Wort-Target), was hauptsächlich semantische Verarbeitung beinhaltet. Wir fanden erneut einen auf Untertiteln basierenden WF-Effekt, ohne Unterschied zwischen den Modalitäten, in den Durchgängen mit cross-modalem Priming, bei denen die Identität des Targets durch den Zugriff auf die Bedeutung des Primes vorhergesagt werden musste. In der gleichen cross-modalen Priming-Bedingung, und wieder ohne Unterschiede zwischen den Modalitäten, fanden wir einen signifikanten bildbasierten OF-Effekt. Die Richtung dieses OF-Effekts war jedoch dem WF-Effekt entgegengesetzt, mit einer besseren Leistung für seltenere Konzepte.

Wir interpretierten den WF-Effekt so, dass er die Stärke der Erfahrung mit einem Konzept widerspiegelt, das während der Sprache erworben und in das semantische Gedächtnis aufgenommen wurde. Der OF-Effekt scheint stattdessen einen Interferenzprozess im Gedächtnis widerzuspiegeln, der durch mehr Erfahrung mit Exemplaren einer Objektkategorie entsteht. Dieser Interferenzprozess wird jedoch bei Objektkategorien, die eine große Vielfalt an

Erscheinungsformen aufweisen, wieder ausgeglichen. Entscheidend ist, dass sowohl der WF- als auch der OF-Effekt eng mit dem semantischen Prozess verbunden sind und sich zumindest teilweise als unabhängig voneinander erweisen.

In der zweiten Studie, die wir vorstellen, untersuchten wir den Einfluss der oben erwähnten Szenenhierarchie auf die mentale Repräsentation von Objekten. Wir haben zum ersten Mal die verschiedenen Ebenen der Hierarchie zusammen betrachtet und zwei Modelle vorgeschlagen, die eine solche Organisation operationalisieren: erstens eine a priori Hierarchie, die eine Reihe von Objekten verschiedenen Szenen, Phrasen und Objekttypen (Anker oder lokal) zuordnet, unter Verwendung von Intuition und gesundem Menschenverstand; zweitens eine datengesteuerte Hierarchie, die wir unter Verwendung von Objektannotationen aus demselben Bilddatensatz, der in Studie 1 verwendet wurde, aufgebaut haben und aus der wir Maße für das gemeinsame Auftreten und die Clusterung derselben Objekte, die für die a priori Hierarchie verwendet wurde, extrahiert haben.

Diese hierarchischen Modelle wurden mit Verhaltensurteilen über die Ähnlichkeit der Objekte verglichen, die von zwei Gruppen von Teilnehmern erhoben wurden, von denen eine die Ähnlichkeit der Objektbilder und die andere die Ähnlichkeit der Objektwörter beurteilte. Die Ergebnisse zeigten einen signifikanten Effekt aller drei Ebenen der a priori-Hierarchie, wobei Objektpaare, die der selben Szene, der selben Phrase oder demselben Objekttyp zugeordnet waren, als ähnlicher beurteilt wurden als Objektpaare, die verschiedenen Szenen, verschiedenen Phrasen oder verschiedenen Objekttypen zugeordnet waren; außerdem fanden wir in der datengesteuerten Hierarchie auch einen signifikanten Effekt des gemeinsamen Auftretens von Objekten in der gleichen Szene oder der gleichen Phrase.

Dies deutet darauf hin, dass die vorgeschlagene Hierarchie von Objekten in Szenen tatsächlich erlernt und genutzt wird, um die Repräsentation von Objekten im Gedächtnis zu gestalten. Entscheidend ist, dass diese Effekte modalitätsübergreifend (Wörter und Objekte) stabil waren, was darauf hindeutet, dass diese Organisation der mentalen Repräsentation von Objekten abstrakt ist und in das semantische Gedächtnis aufgenommen wird.

Die dritte und letzte Studie, die wir vorstellen, ist eine Fortsetzung und Erweiterung dessen, was wir in Studie 2 untersucht haben. Es blieb unklar, ob die hierarchische mentale Organisation von Objekten, die die Hierarchie in der Umgebung widerspiegelt, stabil oder aufgabenabhängig ist, und was ihr Hauptzweck ist.

Wir haben den selben Datensatz von Objekten sowie die selben Maße der hierarchischen Organisation verwendet, um Verhaltensurteile über Ähnlichkeit zu modellieren, aber wir haben

in diesem Fall Verhaltensantworten von drei Gruppen von Teilnehmern gesammelt: Eine Gruppe musste die Ähnlichkeit von Objektbildern auf der Grundlage der Aktion beurteilen, die man mit den Objekten ausführen kann („Handlungsaufgabe“); eine andere Gruppe musste die Ähnlichkeit von Objektbildern auf der Grundlage des visuellen Erscheinungsbildes beurteilen („visuelle Aufgabe“); schließlich verwendeten wir die Daten der Gruppe von Teilnehmern, die die Ähnlichkeit von Objektbildern in Studie 2 beurteilten, in der sie keine expliziten Instruktionen erhielten („keine Aufgabe“).

Die Ergebnisse replizierten größtenteils die in Studie 2 gefundenen Haupteffekte und zeigten, dass die hierarchische Organisation der mentalen Repräsentationen von Objekten bei unterschiedlichen Aufgabenanforderungen insgesamt stabil ist. Gleichzeitig wurden die Effekte zwischen den Aufgaben verglichen, wobei sich herausstellte, dass die hierarchischen Maße bei der „Handlungsaufgabe“ und der Bedingung „keine Aufgabe“ größtenteils ähnlich sind, während ihre Auswirkungen bei der „visuellen Aufgabe“ geringer sind. Gleichzeitig werden einige der feineren hierarchischen Maße durch die „Handlungsaufgabe“ im Vergleich zur Bedingung „keine Aufgabe“ verstärkt; wenn man schließlich Maße der visuellen Ähnlichkeit betrachtet, wurden diese durch die „visuelle Aufgabe“ im Vergleich zur Bedingung „keine Aufgabe“ und „Handlungsaufgabe“ verstärkt.

Insgesamt deuten diese Ergebnisse darauf hin, dass die hierarchische Organisation von Objekten erlernt und genutzt wird, um die mentale Repräsentation für eine effiziente Interaktion mit Objekten zur Durchführung von Handlungen in der Umwelt zu organisieren.

Die Ergebnisse der drei Studien werden im Folgenden in einer allgemeineren Weise diskutiert. Wir reflektieren die positiven Ergebnisse, die mit Hilfe von Messungen aus kommentierten Bilddatensätzen erzielt wurden, um die vielen Herausforderungen hervorzuheben, die diese Art von Daten mit sich bringt: Annäherungen, Verzerrungen und individuelle Unterschiede.

Danach haben wir im Lichte der Ergebnisse von Studie 3 künftige Forschungsrichtungen im Bereich des szenengrammatischen Wissens erörtert und den Schwerpunkt darauf gelegt, wie szenengrammatisches Wissen komplexes Verhalten unterstützt, das in Handlungssequenzen aufgegliedert werden kann, die mehrere Prozesse der Objekterkennung, der Suche und des Erreichens von Objekten umfassen; darüber hinaus haben wir Unterschiede in Bezug auf Maßstab und Navigation zwischen verschiedenen Arten von Szenenkategorien (Innen- und Außenbereich) sowie die Rolle des Raums bei Fehlen einer klaren semantischen Beziehung zwischen Objekten und Szenen diskutiert.

Summary

Our mind has the function of representing the physical and social world we are in, so that we can efficiently interact with it. This results in a constant and dynamic interaction between mind and world that produces a balance when representations are at the same time accurate with respect to what the world is communicating to our organism, but also compatible with how our mind works.

A paradigmatic case of this interaction is offered by perception, which is the mental function that represents contingent aspects of the world built from what is captured by our senses. Indeed, the dominant philosophical view in cognitive science is that our perceptual states are representations of the world and not direct access to that world. These representational perceptual states therefore include the aspects of the world they represent and that initiate the perception by stimulating our sensory organs.

Perceptual representations are built using information from the sensory system, i.e., bottom-up information, but are also integrated with information previously acquired, i.e., top-down information, so that perception interacts with memory through language and other mental functions. Such organization is believed to reflect a general mechanism of our mind/brain, which is to acquire and use information to make efficient predictions about the future, continuously updating older information with present information.

This predictive processing works because the world is not random, but shows a regular structure from which reliable expectations can be built. One way that our minds make these predictions is by adapting to the structure of the world in an implicit, automatic and unconscious way, a process that has been called Implicit Statistical Learning (ISL). ISL is a learning process that does not require awareness and happens in an incidental and spontaneous way, with mere exposure to statistical regularities of the world. It is what happens when we learn a language during early childhood, and that allows us to be implicitly sensitive to the phonological structure of speech, or to associate speech patterns with objects and events to learn word meaning.

A specific case of ISL is the learning of spatial configuration in the visual world, which we apply to abstract arrays of items, but most importantly, also to more ecological settings such as the visual scenes we are immersed in during our everyday life. The knowledge we acquire about the structure of visual scenes has been called “Scene Grammar”, because it informs about presence and position of objects in a similar way to what linguistic grammar tells us about the presence and position of words. So, we implicitly acquire the semantics of scenes, learning which objects are consistent with a certain scene, as well as the syntax of scenes, learning where objects are positioned in a consistent way within a certain scene.

More recent developments have proposed that scene grammar knowledge might be organized based on a hierarchical system: objects are arranged in the scene, which offers the more general context, but within a scene we can identify different spatial and functional clusters of objects, called “phrases”, that offer a second level of context; within every phrase, then, objects have different status, with usually one object (“anchor object”) offering strong prediction of where and which are the other objects within the phrase (“local objects”). However, these further aspects of the organization of objects in scenes remain poorly understood.

Another problem relates to the way we measure the structure of scenes to compare the organization of the visual world with the organization in the mind. Typically, to decide if an object appears or not in a certain scene, and whether or not it appears in a certain position within a scene, researchers based their decision on intuition and common-sense, maybe validating those decisions with independent raters. But it has been shown that often these decisions can be limited and more complex information about objects’ arrangement in scenes can be lost.

A potential solution to this problem might be using large set of real-world images, that have annotations and segmentations of objects, to measure statistics about how objects are arranged in the environment. This idea exploits the nowadays larger availability of this kind of datasets due to increasing developments of computer vision algorithms, and also parallels with the established usage of large text corpora in language research.

The goals of the current investigation were to extract object statistics from this image datasets and test if they reliably predict behavioural responses during object processing, as well as to use these statistics to investigate more complex aspects of scene grammar, such as its

hierarchical organization, to see if this organization is reflected in the organization of objects in our mind

In the first study we present, we have computed novel measures of object frequency (OF) representing how many times an object occurs in large datasets of real-world images. We have used these measures to predict behavioural responses during object recognition, and compare OFs with word frequency (WF) measures that have been found to predict behavioural responses during word recognition. In a first experiment (Exp 1), where participants had to categorize object images or their corresponding written word as denoting an either natural or man-made concepts, we only found a significant effect of a subtitle-based WF with a facilitation for more frequent concepts. Surprisingly, the effect was present in both word recognition and object recognition.

In a second experiment (Exp 2), and its replication (Exp 3), participants performed a priming task where they had to judge if prime and target had the same meaning. The crucial manipulation was that prime and target were either from the same modality (uni-modal priming: word-priming-word, object-priming-object), which involves mainly perceptual processing, or from different modalities (cross-modal priming: word-priming-object, object-priming-word), which involves mainly semantic processing. We found again a subtitle-based WF effect, with no difference between modalities, in the cross-modal matching trials, where target identity needed to be predicted by accessing the meaning of the prime. In the same cross-modal matching condition, and again without differences between modalities, we found a significant image-based OF effect. However, this OF effect had an opposite direction than the WF effect, with behavioural facilitation for more rare concepts.

We interpreted the WF effect as reflecting the strength of experience with a concept acquired during language and incorporated into semantic memory. The OF effect, instead, seems to reflect a memory interference process produced by more experience with exemplars of an object category. This interference process is however counterbalanced for object categories that have a strong variety in the exemplars. Crucially, both WF and OF effect were strongly linked to semantic process and showed to be, at least partially, independent from each other.

In the second study we present, we investigated the impact of the above-mentioned scene hierarchy on mental representation of objects. We have for the first time considered together the different levels of the hierarchy, and we have proposed two models operationalizing such organization: first, an a priori hierarchy, built by assigning a set of objects to different scenes, phrases and object types (anchor or local), using intuition and common-sense; second, a data-driven hierarchy, that we built using object annotations in the same image dataset used in Study 1 and from which we extracted measures of co-occurrence and clustering of the same set of objects used for the a priori hierarchy.

These hierarchical models were compared to behavioural similarity judgements of the objects, collected from two groups of participants, one judging similarity of object pictures, one judging similarity of object words. Results showed significant effect of all the three levels of the a priori hierarchy, with pairs of objects assigned to the same scene, to the same phrase or to the same object type being judged as more similar than pairs of objects assigned to different scenes, different phrases or different object types; besides, we also found significant effects of co-occurrence of objects in scenes and co-occurrence of objects in phrases, measured from the data-driven hierarchy.

This suggests that the proposed hierarchy of objects in scenes is indeed learnt and used to shape how objects are represented in the mind. Crucially, these effects were stable across modalities (words and objects), suggesting that this organization of mental representation of objects is abstract and incorporated into semantic memory.

The third and final study we present is a follow-up and extension of what we have investigated in Study 2. It remained unclear whether the hierarchical organization of objects in the mind that reflects the hierarchy in the environment is stable or task-dependent, and which is its main goal.

We used the same set of objects as well as the same measures of hierarchical organization to model behavioural similarity judgements, but we have in this case collected behavioural responses from three groups of participants: one group judged similarity of object pictures based on the action you can perform with the objects (“action task”); another group judged the similarity of object pictures based on visual appearance (“visual task”); finally, we used the data

from the third group of participants who judged similarity of object pictures in Study 2 and who did not receive any explicit indication (“no task”).

Results mostly replicated the main effects found in Study 2, showing that the hierarchical organization of mental representations of objects is overall stable across different task demands. At the same time, the effects were compared across tasks revealing that hierarchical measures are mostly similar between “action task” and “no task”, while their effects are reduced during the “visual task”. At the same time, some of the more fine-grained hierarchical measures are further enhanced by the “action task” compared to the “no task”; finally, when considering measures of visual similarity, these were enhanced by the “visual task” compared to the “no task” and “action task”.

Altogether, these results suggest that hierarchical organization of objects is learned and used to organize mental representation for efficient interaction with objects to perform action in the environment.

The results from the three studies are further discussed in a more general way. We reflect on the positive results obtained from measures extracted from annotated image datasets to highlight the many challenges posed by this type of data: approximations, biases and individual differences.

After that, in light of the results of Study 3, we consider future directions of investigation on scene grammar knowledge, shifting the focus to how scene grammar supports complex behaviour that can be broken down into action sequences that involves multiple processes of object recognition, search and reaching; additionally, we discuss differences of scale and navigation between different types of scene categories (indoor and outdoor), and the role of space in the absence of a clear semantic relationship between objects and scenes.

Theoretical background

Mind and World

According to the intentionality-based definition (Brentano, 1874; Kim, 2018), the mind is a collection of processes that represent entities, properties and events of the world, with the purpose of allowing the organism where these processes take place to efficiently interact with the world, and survive. The entities, properties and events that are represented can be either actual (as the *computer* we are writing these words) or potential (as the *interest* we hope some of you might experience in reading this thesis). The world whose aspects are represented is physical (like coldness and colour green, materials and tools, hurricanes and photosynthesis, plants and animals, your own body and other people's bodies, etc.) but also non-physical (like socio-cultural norms and roles, your sense of *Self*, other people's aspirations, emotional and aesthetic values, etc.).

In this view, the importance of the relationship between mind and world emerges clearly: the mind, to function properly, needs to represent aspects of the world that enable the organism to act on and interact with that world in a fruitful way. Consider the extreme case of psychotic symptoms such as hallucinations and delusions, in which the mind is representing percepts and thoughts as actual state of the world, although they have no direct link to entities and events actually present in the world (Telles-Correia et al., 2015; Garety & Freeman, 2013). At the same time, the mind needs to organize and simplify the huge amount of information coming from the world into structures that are compatible with mental functioning itself. Therefore, for example, a group of atoms hit by a light is perceived not as a collection of individual sensations, but as a whole unique object (Koffka, 1935), and its appearance remains stable even when the movements of eyes and head change its reflection on the retina (O'Regan, 1992; Melcher, 2011). This constant dynamics of adaptation between mind and world, which accompanies phylogenic and ontogenetic development (Suddendorf & Whiten, 2001), is best exemplified by the case of perception.

Perception is the set of mental processes that represent the world in its contingent aspects starting from the information captured and processed by the sensory systems (Matthen, 2015). The relationship between the mental states of perception and the external world has long

been a topic of philosophical and psychological debates. As for the definitions of mind and perception we offered above, the dominant view in contemporary psychology and cognitive science is *representationalism* (Nanay, 2015), which conceives that mental states of perception are representations of the contingent world, in a similar way as other types of mental state are representations of aspects of the world (e.g., believes, desires, memories). According to this view, mental states are representations because they refer to something, which is defined as the content of the mental representations. For example, the belief that snow is cold is about a property of snow, its coldness. Similarly, the perception of a car in front of you is about that entity being in that specific position and being a car. A perception can therefore be veridical, if it is based on a correct representation of contingent aspects of the world, or it can be an illusion / hallucination if it is based on a misrepresentation. In this view, the relationship between the organism having a mind and the external world is not direct, but it is mediated by mental representations (a position that authors refer as *indirect* or *critical realism*).

Other theoretical approaches have instead opposed the notion of mental perceptual representation (and therefore have been named *anti-representationalisms*). One of the most relevant of these approaches is *enactivism*, for which the relationship between perceiver and the world is a dynamic and active process that does not require any intermediate and static entity as is the case for mental representations. For example, during vision, not all aspects of the visual environment are processed, and some of them are processed in more detail than others. Besides, through movements of the eyes, head and body, the same aspects of the visual environment that before were processed in detail could now be unattended or processed in less detail. Despite this dynamic process, our perception of the visual world remains stable and holistic, all the details of the environment seem available to us in a direct way (O'Regan, 1992). Enactivism has the merit to have highlighted in the theoretical debate the need to acknowledge the dynamic nature of mental processes, as well as the interplay of perception and action, reconnecting with the ecological approach to vision proposed by J. J. Gibson (Gibson, 1986). At the same time, acknowledging the dynamic and constructive nature of perceptual processes does not exclude *per se* the presence of perceptual representations or the plausibility of representationalist views, since those properties could be incorporated into the notion of representation.

Another anti-representational approach is the so-called *relationalism* (Nanay, 2014). According to this view, perception is a direct relationship between an organism and an object, rather than the content of a representation. Perceptual states are based on the actual perceived object. In this sense, both *enactivism* and *relationalism* relate to a *direct realism*. One of the main arguments that supports the relationalist proposal is that it considers the particularity of perception where *representationalism* would fail in doing so (Soteriou, 2000). The “particularity” of perception refers to the fact that when we perceive one object, we perceive a specific individual (a “token” of a kind). If a perceptually identical object replaced the previous object in the same position without our knowing, the perception of this new individual would be a different state, according to *relationalism*, but would be conceived as the same perceptual state by representationist views, because the content / representation would be the same. For *representationalism*, different perceptual states derive from different contents / representations (i.e., their perceptual properties), while, for *relationalism*, different perceptual states derive from different objects. This divergence seems to be reconcilable, though, since some forms of *representationalism* entail perceptual representational content as being “object-involving”, in the sense that it relates to a specific token (Siegel, 2006), therefore incorporating aspects of the relationalist view into the representational framework. One further argument relates to perceptual experience in the more phenomenological sense: as it seems to us, in first person, perception is direct and not mediated. One way to further reconcile the two positions is to consider the representational nature of perceptual states as better suited for explaining the mechanism of perceptual processes (and accounting also for unconscious perception), while the relationalist nature of perception might be more prominent when considering the phenomenology of perception and the conscious experience.

Beyond an attempt to reconcile and integrate representationalist and anti-representationalist views, it is important to acknowledge that the absence of perceptual representations would fail in explaining some empirical evidence (Nanay, 2015): for example, multisensory perception (Spence & Driver, 2004) indeed requires integration and matching of two distinct representation; similarly, the dissociation between ventral vision (for recognition) and dorsal vision (for visuo-motor control, Goodale & Milner, 1992) is explained by the presence

of two distinct representation, one veridical and one non-veridical, which produce two contradictory pattern of behaviour.

A second issue in perception, that we believe reconnects the philosophical tradition and theoretical discussion with current approaches in cognitive science, is concerned with the structure of perceptual representations and the way they come to be formed (Gordon & Slater, 1998). To simplify the discussion, on one side of the debate there is a position that reconnects with the tradition of *Empiricism* (e.g., Locke), where the information that structures perceptual representations is purely the information captured and processed by the sensory systems (“bottom-up” information), focusing on the role of experience in shaping perceptual processes and representations. We could say that according to this view, the world shapes the mind. On the opposite side, there is the tradition of *Nativism / Rationalism* (e.g., Descartes), where the structure of perceptual representations is given by innate and/or a priori mental processes (“top-down” information), and the focus is on the mind shaping the world. We could ascribe to the first group the contribution of theorists and experimentalists such as Helmholtz, Wundt and Piaget, while in the second group we can acknowledge the work of Gestalt psychologists (Koffka, 1935) and J. J. Gibson (Gibson, 1986). This categorization is actually much more complex, as most of the above-mentioned authors would have not denied some basic a priori/ innate abilities (for the empiricist) or the relevance of sensory information (for the nativist/ rationalist). Besides, for example, the unconscious inference proposed by the “empiricist” Helmholtz emphasizes the role of knowledge (we would say the “top-down effects”) in supporting the organization of sensory information; while the ecological approach of the “nativist” Gibson proposes that the visual signal already incorporated all complex information to give rise to a three-dimensional action-oriented perception.

A reconciliation of these two extreme approaches on the question of what is the structure of perceptual representations is nowadays a mainstream and prominent framework in cognitive science: *predictive processing* (Rao & Ballard, 1999; Bar, 2009; Clark, 2013). According to this view, the cognitive system, and its implementation in the brain, functions as a Bayesian probabilistic system that uses previously acquired knowledge to make prediction about upcoming sensory input. This system is organized in a hierarchy: bottom-up sensory input first

meets lower-level top-down predictions; then, only the residual bottom-up signal that was not met by the top-down expectation (the “prediction error”) moves to the next level of the hierarchy where it will be confronted with another top-down signal, and so on. While doing so, the bottom-up prediction error updates the top-down expectation, in order to make them more precise in predicting new inputs. Perceptual representations are the result of this balanced combination of sensory input and previous knowledge. Interestingly, the predictive processing framework that seems to “solve” the debates between empiricism and nativism-rationalism of perceptual representations has been identified to have roots in the philosophical proposal of Immanuel Kant (Swanson, 2016), whose aim was to offer a resolution to the empiricist vs. rationalist debate of his time. Aspects of the predictive processing proposal resonate with his notion of an active and generative role of mind in cognition and perception, for his ideas of a priori structure that informs the sensory experience and his notion of schemata.

We would like to emphasize two important ideas that come with the predictive processing framework. The first one is the *penetrability of perception* (Lupyan & Clark, 2015; but see a critical review from Firestone & Scholl, 2016): perception is not isolated (or “encapsulated”) from other cognitive functions, but is in interaction with thought, memory and language. And it is exactly because of this that perceptual processing is possible: previously acquired information represents the source from which to build goal-directed predictions. The second idea regards memory representations called “frames” or “schemata” (Bar, 2004; 2009), that are fairly abstract representation built from previous experience and used as top-down prediction to “test” or “shape” the upcoming sensory input in order to perform efficient perceptual recognition.

The relationship between mind and world, the intentional and representational nature of perception, the predictive processing with its combination of top-down and bottom-up signals, the penetrability of perception and the role of memory schemata, are all important concepts that will guide us through all this investigation.

Implicit and statistical learning: bringing the world into the mind

One reason the mind makes prediction about the world is that the world has some structure, some regularities, from which and for which expectations can be built. One fundamental way that the mind uses to build these expectations is by registering the regularities of the world during experiences in an implicit way, a phenomenon called *implicit learning* (IL) or *statistical learning* (SL). The two terms appeared in the experimental literature in different times, consolidating the use of different behavioural paradigm and emphasizing different aspects of the phenomenon, but for the current scope, we are going to use them interchangeably or simply as *implicit statistical learning* (ISL; about this definition issue, see Perruchet & Pachton, 2006; Batterink et al., 2019).

The term “Implicit learning” was first used by Reber (1967), who employed an artificial grammar to generate artificial verbal stimuli that followed a structure, unknown to participants, in comparison with randomly generated strings. Participants were asked to memorize the grammatical or the random strings in several training phases and then to reproduce them in a written way in test phases; crucially, no explicit information about an underlying structure were given. As a results of implicit learning, the group exposed to grammatical strings showed a decrease in error rate across different sets of materials, while the group exposed to random strings did not show this reduction (Reber, 1967). The term “Statistical learning” was proposed much later in a study by Saffran and colleagues (1996a), showing that 8-months-old infants were able to segment continuous speech input generated from an artificial grammar into words based on statistical regularities in neighbouring sounds, and that they were able to do so after short experience (Saffran et al., 1996a). The idea is that, in a structured acoustic input such as speech (or artificial speech), the probability of a sound following another one within a word is higher than the probability between two neighbouring sounds from different words (“transitional probability”, Saffran et al., 1996b). Humans seem to be sensitive to such structures and exploit these regularities starting at the early stages of development, segmenting continuous input into meaningful units or chunks (in this case words from speech sound).

In general, we can think about ISL as a set of learning mechanisms that function in a way that has been defined as incidental, unsupervised, uninstructed, automatic, unintentional, unaware, unconscious, and extracting from mere exposure, without analytical strategies, the regularities and statistical relationship between aspects of the world (Perruchet & Pachton, 2006; Schapiro & Turk-Brownie, 2015; Sherman et al., 2020). This ability of adaptation to environmental structure seems pervasive across many species and across different stages of the human lifespan, although researchers have often studied ISL's most rapid effect on behaviour, namely producing changes in performance after brief exposure to structured materials (Schapiro & Turk-Brownie, 2015). Beyond the clear advantage offered by such mechanisms during early stages of development, it has been proposed that ISL produces behavioural effects that might reflect important functional roles even during adult life: it seems to support guidance of attention in the environment and subsequent formation of efficient predictions in order to facilitate perceptual processing (this is why we discuss the role of ISL in predictive processing and the relationship between mind and world); but also, it has been proposed to help in the creation of unified representations by associating different features co-occurring in space and/or time, with the consequence of compressing information and reducing cognitive load; finally, it might influence how we structure memories in schemas and maps using the information extracted from the environment to create an internal "model" of the world, and support reward-oriented decision-making (Sherman et al., 2015).

Despite the implicit and unconscious nature of ISL, which would ascribe it to the domain of non-declarative memory systems (e.g., perceptual learning, motor skills acquisition), it has been proposed that, as for other types of learning, ISL might result from a complex relationship between non-declarative and declarative memory systems, even when one of the two is predominant (Batterink et al., 2019; Sherman et al., 2020). Encoding of episodic memories, for example, is strongly based on specific spatio-temporal coordinates that make that memory indeed unique. At the same time, the spatio-temporal structure of the environment where the episode takes place is, at least in some ways, familiar to us, in the sense that we are able, through ISL, to make prediction and consequently to process information in a more efficient way. Similar consideration can be made for encoding of space for navigation. Episodic event coding and

navigational coding, as well as ISL, have all been linked to activation of different but close neural circuitry in the Middle Temporal Lobe (MTL), and more specifically involving the structures of the hippocampus, parahippocampal cortex, perirhinal and entorhinal cortices (Batterink et al., 2019; Sherman et al., 2020; Arminoff et al., 2013).

Although ISL has been initially studied in relation to acoustic and/or linguistic materials and language acquisition, it has been revealed to be a domain-general mechanism that has been investigated using different types of stimuli, such as linguistic and non-linguistic sounds, tactile stimuli, visual shapes and complex visual configuration, and even multi-modal combination. The behavioural measures influenced by this phenomenon are also different, such as higher recognition accuracy, faster response times and eye-movements (Schapiro & Turk-Brownie, 2015). The domain-generality of ISL also becomes evident in the fact that it can transfer and generalize acquired knowledge between space and time, between different modalities and between levels of categorical specificity (Sherman et al., 2020).

Implicit Statistical Learning in visual cognition: from contextual cueing to visual scenes

When thinking about ISL as presented so far (and how it was investigated initially), the learning mechanism seems pretty effective in extracting meaningful representation from a continuous stream of sensory input unfolding over time, i.e., in detecting sequential regularities. However, our environment is highly structured in terms of spatial arrangement as well, especially when considering vision and processing of visual stimuli (Schapiro & Turk-Brown, 2015).

Since the seminal study of Biederman and colleagues (1982), it has been shown how the arrangement of objects in the environment follows a set of rules reflecting physical constraints or typical usage; when these rules are violated, participants' performance significantly deteriorated, suggesting that these rules are learnt through experience and used to guide perceptual processing (Biederman et al., 1982). One way this phenomenon was investigated in terms of ISL was by studying the so-called *contextual cueing* process, where structured visual arrays were shown to influence search performance (Chun & Jiang, 1998).

Chun & Jiang (1998) aimed to study the deployment of attention during search in complex visual displays that serve as a proxy of complex visual input coming from the environment during real-world search. They believed that what was missing from existing explanations of attentional selection was the global context given by configuration of multiple items, i.e., objects, that are often stable and predictable (Chun & Jiang, 1998; Biederman et al., 1982). Through ILS, these structured global context configurations are learnt and used to guide attention during search within the array. Besides, they proposed that these implicit memories are instance-based, i.e., they are implicit episodic memories (see discussion on interaction between declarative and non-declarative memory systems in Batterink et al., 2019; Sherman et al., 2020; Arminoff et al., 2013).

For their experiments, they employed visual arrays made of several two-lines items that can have different shapes (either more T-like or more L-like), different colours and different texture. They generated, randomly, configurations of these items and then kept them constant throughout the whole experiment (i.e., structured configurations), while other configurations were generated randomly in each block (i.e., unstructured configurations). Participants were asked to look for a T-shaped stimulus among L-shaped stimuli for several trials in several blocks. After 5 blocks out of 30 (i.e., 5 repetitions of the same configuration), search times were already faster in structured configuration than in unstructured configurations, indicating implicit episodic memory guidance of attention. Additionally, they showed that the memories for learnt configurations are abstract and generalize when changing identity of individual distractors, that they are formed even when arrays are much more visually homogeneous and tested that participants were not aware of or explicitly encoded the arrays as a search strategy (Chun & Jiang, 1998).

The theorization and investigation of contextual cueing has revealed the importance of the complexity of environment information, especially in its global configuration, but at the same time remained in line with traditional research in visual search that used highly artificial visual displays. After Chun & Jiang's study (1998), during the 2000s, it became more clear that visual search needed to be investigated using more ecological stimuli such as real-world images of visual scenes (Wolfe et al., 2011).

The classic body of research has shown that attention in visual search is guided by a target selection mechanism based on a set of attributes that simplify the number of items that need to be scrutinized (and recognized), given the limited capacity of object recognition that requires binding of different stimulus features (Wolfe, 1994; Treisman, 1996). This account worked quite well with artificial displays of items or isolated objects, but was found to be insufficient in accounting for search in visual scene images that seemed efficient despite the cluttered and complex nature of such visual stimuli. This is, again, due to the non-random but rather structured nature of the environment, that is learnt through experience and formation of a set of memories that offer scene-based guidance (Wolfe et al., 2011). One type of guidance is offered by episodic memory from previous exposure to a specific scene (as seen in contextual cueing, Chun & Jiang, 1998): for example, if you had to look for your pen on your desk, after returning from another room, you would benefit from the episodic memory that encodes the position of the pen when you left your office. But even more importantly, search in visual scenes is strongly guided by semantic memory (Brockmole & Võ, 2010), which is the set of general knowledge about the world: if you were invited to your friends' new house and asked to help cooking, you would look for knives only in the kitchen, not in the bathroom, and you would look for them in drawers not in the fridge. This is because we have learnt that an object is most likely to appear in a specific context, in a specific position and in specific relation to other objects (Bar, 2004; Oliva & Torralba, 2007). This semantic guidance is possible thanks to our ability to extract the "gist" of a scene, i.e., the basic information about its category and spatial layout, just with a glance, without the need of limited capacity attentional processing. Gist representation seems to be built quickly during early visual processing using specific spatial frequency information (low spatial frequencies), which offers a coarse global representation of the visual context (Oliva & Torralba, 2001) and that preactivates information in semantic memory associated with that visual context (i.e., typical objects, typical positions, etc.; Bar, 2004).

As a result, it has been proposed that visual search in real-world scenes is based on two types of processing pathways: one with limited capacity that works in a selective way to bind stimulus features for recognition based on attribute-based guidance, and another one that has

no constraint in capacity and exploits fast global gist representation to produce scene-based episodic and semantic guidance.

Scene Grammar: learning the structure of *our* visual world

The perceptual and cognitive analysis of scenes are processes that go beyond the recognition of an environment as belonging to a certain category (e.g., a kitchen). Scenes, in this sense, are visual input with a unique level of complexity and ecological validity, since we are always immersed in a visual scene. Scenes are stimuli that encompass most of our visual field, that have a spatial layout and contain several objects. Within a scene, we recognize and search for these objects in that spatial layout, we use the objects to perform certain actions and navigate through them and that space (Malcom et al., 2016).

The complexity of scenes is also reflected in the complexity of visual (and non-visual) features that can be extracted during rapid visual processing (Groen et al., 2017): scenes have low-level properties such as patterns of colour, texture and orientation; mid-level properties such as the spatial layout and global “shape”; and high-level properties such as the objects they contains or the action you can perform in there (their “functions”). This complexity posits an additional problem because these features are often intercorrelated and, although many individual features are sufficient to categorize a scene (e.g., consider “gist” perception), it is not clear which ones contribute primarily and fundamentally. Empirical evidence proposes that high-level features, in particular scene “functions”, are a strong guiding principle in scene categorization (Greene et al., 2016; Greene & Hansen, 2020), but task demand can prioritize the relevance of other task-related features (Greene & Hansen, 2020).

Beyond understanding how a scene is recognized, many researchers have focused on describing the details from which the extraction of scene gist produces an implicit scene-based semantic memory facilitation in tasks that involve object processing, such as recognition, search and interaction. One of these approaches has been called “Scene Grammar” (for a recent review, Vö, 2021).

The idea of Scene Grammar is based on an analogy first proposed by Biederman and colleagues (1982): the rules that govern the arrangement of objects in visual scenes are similar to the rules that govern the arrangement of words in a sentence; therefore, we can identify regularities that form a “scene semantics” (Biederman et al., 1982), which involves rules based on typical usage of an object, such as probability (“is this object typically appearing in this scene?”), position (“is this object typically appearing in this position within this scene?”) and size (“is this object typically appearing with this size compared to other objects within the scene?”); similarly, we can identify regularities that form a “scene syntax”, which are rules based on physical constraints of gravity on objects, such as support (“is an object resting on a surface?”) and interposition (“is an object interposed with another one?”). Whenever an object is violating one of these learnt rules, the processing of that object is degraded.

Scene Grammar proposal has reformulated Biederman and colleagues’ classification focusing on semantic violations as the ones that involve inconsistent object-scene pairing (e.g., a toilet paper roll appearing on a kitchen counter) and syntactic violations as the ones that involve inconsistent positioning of an object within a scene (e.g., a toilet paper roll appearing inside a toilet), with extreme syntactic violations involving placement of objects in an inconsistent way in relation to physical laws such as gravity (e.g., a toilet paper roll floating in the middle of a bathroom). Therefore, scene semantics connects an object with scene meaning, while scene syntax connects an object with scene structure (Vö & Henderson, 2009; Vö & Wolfe, 2013a). To further support this analogy and the reformulation of scene semantic and scene syntax offered by Scene Grammar, Vö and Wolfe (2013a) found that semantic and syntactic violations of object arrangement in scenes are reflected in similar electrophysiological signals to the ones elicited by semantic and syntactic violations in language, i.e., the N400 (for semantics; Kutas & Hillyard, 1980) and P600 (for syntax; Osterhout, & Holcomb, 1992) ERP components (Vö & Wolfe, 2013a; Ganis & Kutas, 2003).

The study of Scene Grammar violations have shown that these rules are learnt during visual experience and exploited to guide perception and attention, supporting object processing during complex behaviour such as recognition (Cornelissen & Vö, 2017), search (Vö & Wolfe, 2013b) and interaction (Draschkow & Vö, 2017).

Further investigations using this approach have focused on the components of a scene that are extracted to activate a contextual representation (Bar, 2004; Aminoff et al., 2013) that support object recognition in context. These “scene ingredients” (Lauer & Vö, 2022) have been identified as colored texture information (Lauer et al., 2018), orientation (Lauer et al., 2020), spatial layout (Brady et al., 2017), materials (Lauer et al., 2021) and neighbouring objects (Lauer et al., 2022). Other investigations have shown that Scene Grammar is developed quite early during childhood, with indications of behavioural and neural advantages for consistent object-scene pairing over inconsistent ones already in 2-years-old children and further strengthening of such scene knowledge in 4-years-old children (Öhlschläger & Vö, 2020; Maffongelli et al., 2020).

More recent developments of Scene Grammar have proposed that these object-scene regularities are organized according to a hierarchy (Vö et al., 2019). A scene (e.g., a kitchen), as it has been investigated in its global properties (for a review, see Lauer & Vö, 2022), represents only the most general contextual level; within a scene, we could think about spatial clusters of objects as a further and more specific context, which has been termed “phrase” continuing the language analogy (e.g., a cluster around a stove top); finally, within each phrase, it has been proposed that objects hold different status (Draschkow & Vö, 2017): one object (named “anchor object”, e.g., a stove), which is fairly typical of the scene, bigger and more static, offers strong predictions about identity and position of other objects within the phrase (named “local objects”, e.g., a pot). The idea is that such hierarchical structure organizes the set of expectations built through experience which forms Scene Grammar knowledge, so that complex behaviour within the environment proceeds following these contextual levels: if you were asked to check the boiling water in your friends’ new place, you will first look for a pot within the “scene” kitchen, narrow your search to the “stove top phrase”, and use the stove top “anchor” to identify and find the “local object” pot on the stove. Clearly, in a more familiar environment this general knowledge interacts with the specific knowledge of where things are kept or are left in that specific setting (“episodic guidance”, Vö & Wolfe, 2013b), but nonetheless, Scene Grammar knowledge is helpful to guide our behaviour whether it is in highly familiar or unfamiliar places.

While the role of Scene Grammar knowledge on object processing within the “scene level” has informed this approach from the beginning (and see also the seminal study of Daventport &

Potter, 2004), less attention has been given so far to the “phrase level” and to the “anchor vs local” distinction. Regarding the former, no explicit idea of a “phrase level” has been investigated, but there have been studies investigating the role of objects or group of objects in facilitating the perceptual processing of other objects. For example, both objects co-occurrence (a toilet and a toilet paper roll appearing together) and spatial dependency (a toilet paper roll appearing next to a toilet) result in facilitating visual search (Mack & Eckstein, 2011, Hwang et al., 2011) and object recognition (Auckland et al., 2007; Gronau & Shachar, 2014). Additionally, multiple objects are processed in a more efficient way if arranged according to typical semantic and spatial rules (Kaiser et al., 2014; Quek & Peelen, 2020) supporting a grouping mechanism that reduces the complexity of visual input. Regarding the anchor-local relationship, two eye-tracking studies (one in a typical lab setting and one in Virtual Reality, VR) have shown that when looking for a target local object within a scene (e.g., laundry basket), replacing the associated anchor object (e.g., washing machine) with another one (e.g., a cabinet) impaired search time and produced wider and less specific exploration of the scene (Boettcher et al., 2018; Helbing et al., 2022). In particular, they have shown that two key elements of the anchor-local relationship are the closeness in space and the predictable relative position along the vertical axis (Boettcher et al., 2018).

It remains to be investigated how the three different levels impact object processing at the same time. Future investigations need to explicitly address this hierarchical organization as a whole, to reveal if Scene Grammar knowledge is indeed structured this way.

Using large datasets to capture the structure of the visual world

The process of ISL, in general and for the specific case of Scene Grammar, manifests as a sensitivity to the structure of the world, as we have seen. Then, one central point in the empirical study of these phenomena is to find a way to measure the structure of the world and to see if neural signals and behaviour match this measured structure. One fairly common way that has been used in the study of ISL in visual scenes is based on an intuition-driven (or a priori) approach: researchers used their own experience, intuition and common-sense to decide which

configurations match typical aspects of the world (“consistent conditions”), and which ones violate these aspects of the world (“inconsistent conditions”). For example, as we have seen from the study of Scene Grammar, placing a sofa in the bathroom would make a semantically inconsistent object-scene pairing and placing a sofa floating in the living room would make a syntactically inconsistent object-scene pairing, while placing a sofa in the living room in the correct position would make a consistent object-scene pairing (Vö & Wolfe, 2013a). This operationalization of object arrangement in scenes has worked quite well in highlighting general principles of Scene Grammar, but limitations need to be acknowledged.

First, considering scene semantics, the distribution of objects in real-world settings is frequently overestimated in people’s judgments (Greene, 2016). As such, researchers claim strong contextual relationships between certain objects and a scenes (e.g., sandcastle and beach) despite the low frequency of occurrence of that object in the scene in the real world (e.g., sandcastle are not that common in a beach compared to other objects; Greene, 2013; Daventport & Potter, 2004). Secondly, contextual associations between object and scene might vary continuously or with intermediate degrees between “inconsistent” and “consistent”, so that using intuition-driven dichotomous distinction might mean losing a lot of the complexity of how objects are distributed in scenes. Similarly, not all positions might be syntactically “inconsistent” in the same way. Third, the relationship between objects and scenes is much more complex than can be captured by reducing it to the frequency of their occurrence or typical positioning. Other interesting aspects might include how specific an object is (does it appear in many different scenes or just one category?), how diagnostic an object is of a given scene (is the presence of an object a good index of which environment are we?), what the relationship is with other objects within a scene (how often they co-occur? How distant are they? What’s their typical spatial configuration?; Greene, 2013). These considerations seem to suggest that in order to extract good measures of statistical regularities of the world we need to find source materials that offer a truthful representation of the world, otherwise we would lose too much information and we would build a “model” of the world that is too imprecise and unable to match our mental representations.

One alternative approach to intuition-driven operationalization might come from the behavioural and neural study of language, where usage of large set of real-world linguistic information have been used to model how the mind tunes to the structure of the world. One fundamental example of this is represented by the study of the so-called *word frequency effect*, that is the behavioural facilitation in processing words that appear more frequently in our language (i.e., “the”) compared to words that are more rare (i.e., “platypus”; for reviews, see Brysbaert et al., 2011; 2018). This effect has first been reported in the 1930s (Preston, 1935) and for many years psychologists have relied on few datasets that were compiled using occurrence of words in non-fiction written text materials (e.g., newspapers, scientific essays, magazines; Kucera & Francis, 1965). However, the increasing availability of digital text materials have brought the creation of larger datasets (millions of words) called *corpora* that contain not only frequency information, but also other orthographic, phonological and syntactic information and that have been validated using behavioural experiments on large sets of participants (i.e., the megastudies, as e.g., Balota et al., 2007).

This availability of large digital text corpora, together with technological advancement in computer science and the rise of machine learning have transformed the usage of linguistic corpora, making them the valuable source for training algorithms in performing complex tasks, with many developments in the domain of Natural Language Processing (NLP, i.e., application of machine learning principles to model human language). One important case regards the applications of distributional semantics theories, which propose a model of word meaning where words that appear in similar linguistic contexts (i.e., they have a similar distribution in text) have similar meaning (for a review, Lenci, 2018). NLP have created algorithms that use distributional measures from text corpora to build representations of word meaning and perform operations on it (e.g., analogies, conceptual constructions). One typical way of representing word meaning by these algorithms is through *Word embeddings* which are multi-dimensional vectors where proximity indicates similarity in their meaning.

This revolution of computer science has involved the domain of vision as well, starting from the development of artificial neural networks (ANNs) that were able to categorize images of objects with a human-like accuracy (Khrizevsky et al., 2017). Similar achievements have been

made in the domain of scene recognition (Zhou et al. 2014), and have brought the need for more large datasets of real-world images that are necessary for training these algorithms (e.g., SUN database, Xiao et al., 2010; Place database, Zhou et al., 2014). Beyond scene recognition, the idea emerged to train machine to interact with the environment in a more complex way, which would require not just categorizing a scene, but also being able to segment it into objects and assigning these objects an identity. For this purpose, datasets were created that contained real-world images with segmented and annotated objects (e.g., COCO, Lin et al., 2014; ADE20K, Zhou et al., 2019). These kinds of datasets are annotated by human workers (e.g., via platform like LabelMe, Russel et al., 2008) and are based on verbal labels assigned to the portion of pixels associated with an object.

This type of data has a lot of potential to be exploited beyond the field of artificial intelligence, offering a unique opportunity to build a more reliable estimation of object statistics in real-world scenes to model behaviour and neural processing (Greene, 2013). From an annotated image, it is possible to measure the following: (a) something similar to word frequency but for object (“object frequency”); (b) object frequency in specific scene categories (a continuous measure of “object-scene consistency”); (c) object pairs co-occurrence, typical distance and typical arrangement on the vertical and horizontal axis; and (d) how objects clustered together within an image and many other aspects of scene structure. There has also been a recent attempt to use annotated image datasets to build a model based on principles of NLP and distributional semantics (Bonner & Epstein, 2021). This model is supposed to represent an object meaning based on the complex relationship it has with other objects in scenes, and first results seem to show a good match between this model and neural activity patterns in Parahippocampal Cortex (PHC), which has shown sensitivity to process of implicit statistical learning and contextual regularities in vision (Bonner & Epstein, 2021; Aminoff et al., 2013; Batterink et al., 2019; Sherman et al., 2020).

More experimental effort is needed to describe the potential and limitations of this annotated image datasets, as well as to validate object statistics extracted from them, but they seem to launch an important challenge on the way we have measured the world and investigated visual perception so far.

The current investigation

In the current investigation our aim was to use this kind of segmented and annotated real-world image datasets to compute objects statistics that allow us to further describe the Scene Grammar knowledge that is built through implicit visual experience with the world and used to make reliable predictions that make perceptual processing more efficient.

For this purpose, we focused on two datasets, that differed in quantity and variety of images, as well as on the quality of annotations (Greene dataset, Greene, 2013; ADE20K dataset, Zhou et al., 2019). Considerations on how the different features of these two datasets impact the choice and reliability of measures are presented.

Using measures extracted from the datasets, we have modeled behavioural responses of healthy adult human participants collected in three online experiments. The first experiment (“Study 1”, Gregorová et al., 2021) investigated the effect of newly computed variables that we named “object frequencies”, which measure the occurrence of an object in the world, on the process of recognition of both object pictures and written word denoting the very same objects. Besides, we investigated commonalities and differences between these new measures of object frequency and well-established measures of word frequency computed from large text corpora. The impact of frequency of occurrence offers insight into how the amount of experience with the world shapes our mental representations.

In a second study (“Study 2”, Turini & Vö, 2022a), we investigated for the first time in an explicit way the proposal of a hierarchical organization of Scene Grammar knowledge, measuring the impact of three different levels on mental organization of objects: organization at scene level, organization at phrase level, and organization at object type level (i.e., anchor vs local objects). To do so, we have confronted a hierarchy built using intuition and common-sense, together with a hierarchy that was built in a data-driven way from an annotated image datasets, taking into account statistics such as objects co-occurrence, distance and size. Additionally, the hierarchical organization was tested across different stimulus modalities, as object pictures and written

words denoting the objects. The impact of a hierarchical organization offers insight into the complexity of information built from the experience with a structured world.

Finally, in a third study ("Study 3", Turini & Vö, 2022b) the hierarchical organization proposed in Study 2 was further investigated in terms of stability and flexibility across different behavioural goals. The impact of the hierarchy on mental organization was measured when objects are judged without explicit task demands, as well as when objects are judged based on their action goals and based on their visual appearance. This offers insight into which purpose this mental organization might primarily serve.

Summary of empirical studies

Study 1: “Access to meaning from visual input: Object and word frequency effects in categorization behavior.”

Introduction

During visual recognition, very different visual inputs, like a picture of an apple and the written word “apple”, are transformed into the same semantic representation. For words, this process is influenced by the frequency of occurrence of the word in natural language, a learning phenomenon called “word frequency effect”: words that appear more often are processed preferentially, leading to faster response times during recognition, while more rare words are recognized with slower response times (Brysbaert et al., 2011; 2018). Word frequency (WF) measures are now typically computed from collection of digitally available text materials, called corpora. Using annotated image datasets, in this study we have computed measures of object frequency (OF) using similar logic and methodology.

The aims of this study were to 1) investigate the role of OF measures in influencing response times during object recognition, as it happens with WF measures during word recognition; 2) to compare the impact of OF measures and WF measures within their domain (WF for words, OF for objects) and across their domain (OF for words, WF for objects); 3) to understand the extent for which frequency effects involve semantic processing rather than perceptual processing.

Methods

We considered two measures of WF (one from movie/tv-shows subtitles, SUBTLEX WF, Brysbaert et al., 2011; one from books and magazine, dlexDB WF, Heister et al., 2011), as well as two measures of OF (Greene OF, Greene, 2013; ADE20K OF, Zhou et al., 2019).

These measures were tested in 3 experiments: Experiment 1, where participants (N=42) performed a semantic categorization (natural vs man-made) for 100 object concepts presented as either written (German) words or object pictures; Experiment 2, where the same participants of Experiment 1 had to judge if prime and target match or not, presenting participants with both uni-modal priming trials (word-priming-word and object-priming-object, which primarily involve

perceptual processing) and cross-modal priming trials (word-priming-object and object-priming-word, which primarily involve semantic processing); finally, Experiment 3, where we tried to replicate results of Experiment 2 but with two new and independent groups of participants performing either only uni-modal priming trials (N=53) or only cross-modal priming trials (N=53; see Figure 1 for schema of the procedures).

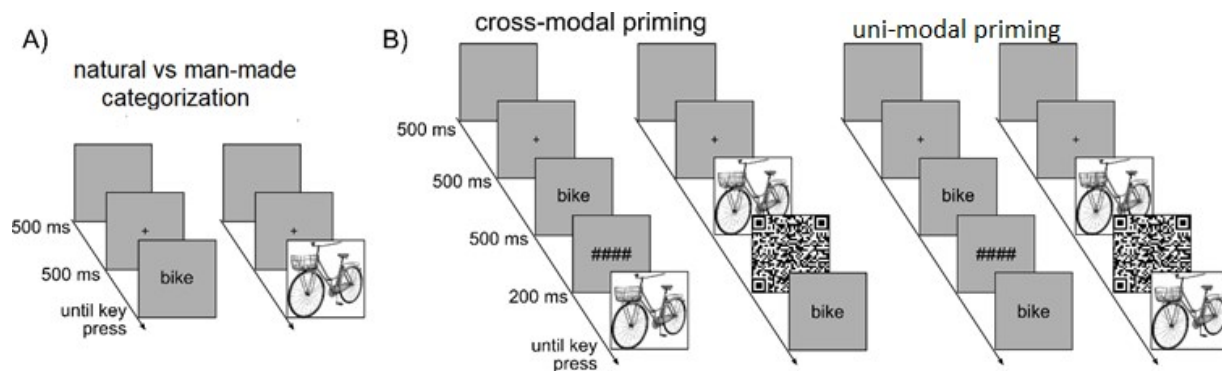


Figure 1 – Study 1: Experimental procedures – A) sequences for word (left) and object (right) trials of semantic categorization task (Experiment 1); B) sequences for matching prime-target trials of word-to-object and object-to-word (left) cross-modal priming task, and for word-to-word and object-to-object uni-modal priming task (Experiment 2 and 3).

In addition to the frequency measures, we computed several covariates to control for visual and orthographic confounds, but also to assess subjective familiarity and typicality of the employed key stimuli. Response times were analyzed using linear mixed-effect models (LMMs; Bates et al., 2014), first selecting the frequency measures that best fit the data, then inspecting the results and additional significant interactions.

Results and Discussion

In Experiment 1 (natural vs man-made semantic categorization), we found that only SUBTLEX WF frequency had a significant fit in both word recognition and object recognition trials, while dlexDB WF only fit with word recognition trials, and no OF measures was found to improve the fit in either modalities. Inspecting the results from the model including SUBTLEX WF, we found a main effect of SUBTLEX WF replicating typical WF effect, with faster response times for more frequent

concepts. Additionally, we found a significant interaction between SUBTLEX WF and Concept modality (Words vs Objects), showing a stronger SUBTLEX WF effect in words than in objects (see Figure 2). Running a post-hoc, we found that despite this difference, SUBTLEX WF effect was significant and had a similar direction in both word recognition and object recognition trials, when considered separately.

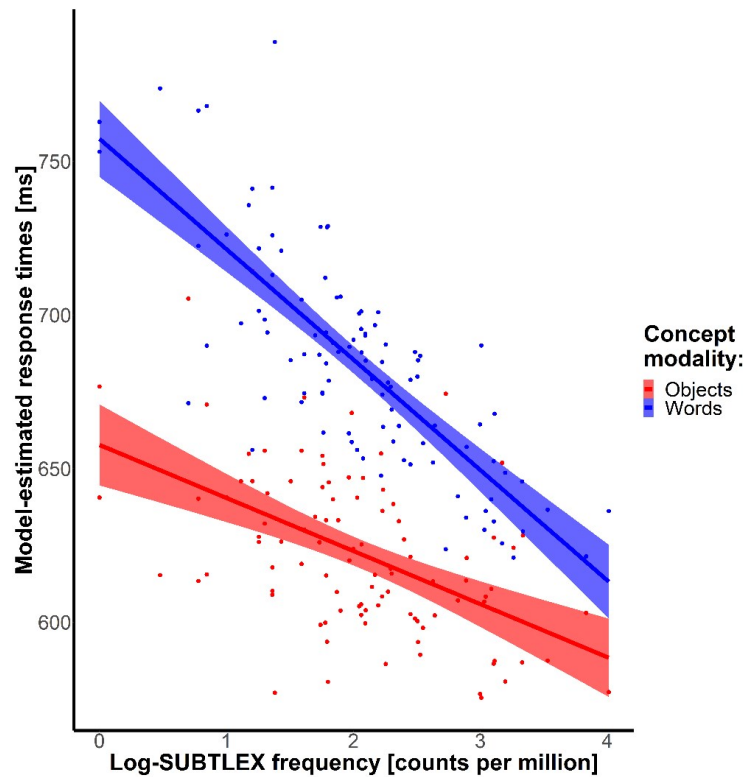


Figure 2 – Study 1: Results Experiment 1 – Participants’ response times (y-axis) estimated from the model, as a function of SUBTLEX WF frequency (x-axis), for words (blue) and objects (red). Points represent response times for individual concepts averaged across participants, lines represent linear fitting of the points; shaded areas around the lines represent 95 % confidence interval.

In Experiment 2 (prime-target matching task), we found that the frequency measures best fitting the data were SUBTLEX WF and Greene OF, in an independent way. Surprisingly, when inspecting the results of the model including both measures, we found a significant main effect of SUBTLEX WF, and a significant main effect of Greene OF going in the opposite direction. Similarly, we found significant interactions of SUBTLEX WF x Matching condition (Match vs Mismatch) x Priming condition (Uni-modal vs Cross-modal), and of Greene OF x Matching

condition (Match vs Mismatch) x Priming condition (Uni-modal vs Cross-modal), again going in opposite directions. Post-hoc analyses have revealed that both frequency effects were primarily significant in Cross-modal matching trials, where prime and target stimuli match and are from different modalities, therefore requiring more semantic processing.

In cross-modal matching trials, SUBTLEX WF had the same facilitatory effect for more frequent concepts, as in Experiment 1, while Greene OF had an opposite facilitatory effect for more rare concepts (see Figure 3). These frequency effects did not differ between modalities, whether the target was an object or a word.

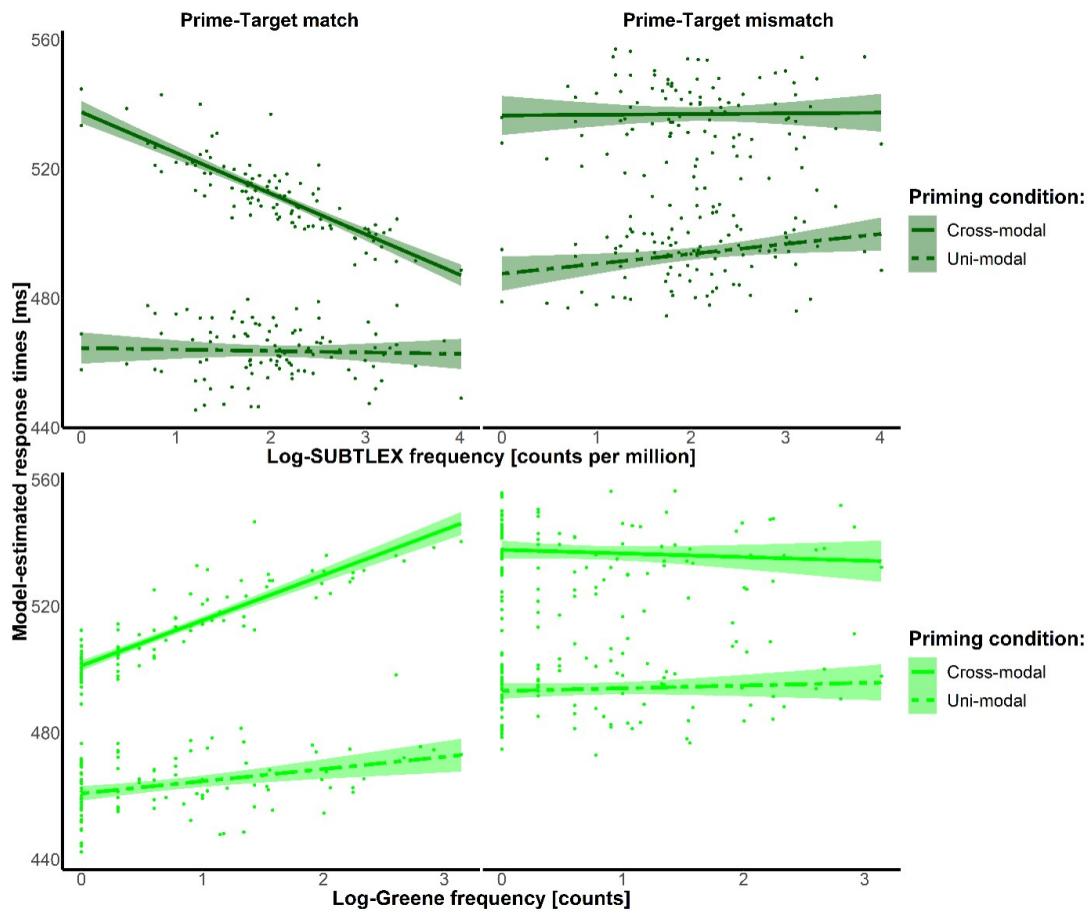


Figure 3 – Study 1: Results Experiment 2 – Participants’ response times (y-axis) estimated from the post-hoc models, as a function of SUBTLEX WF (x-axis up, darker green) and Greene OF (x-axis down, lighter green) in the different conditions of priming (Cross-modal = solid lines; Uni-modal = dashed-dotted lines), and in the different conditions of prime-target matching (Matching = left; Mismatching = right). Points represent response times for individual

concepts averaged across participants, lines represent linear fitting of the points; shaded areas around the lines represent 95 % confidence interval.

The facilitatory effect of SUBTLEX WF for more frequent words and objects in both Experiment 1 and 2 replicates the many findings of WF effect in word recognition literature (Brysbaert et al., 2011; 2018), but additionally suggest that the learning mechanism underlying this effect, at least the one captured by subtitle-based WF as SUBTLEX, reflects not simply the strength of experience with a word built during language (WF effect for words only), but the strength of experience with the meaning of that word (WF effect for words and objects).

The unexpected effect of Greene OF for both words and objects, with a facilitation for more rare concepts, was interpreted here as reflecting a memory interference process (Konkle et al., 2010). When we perceive a lot of exemplars of an object category, the higher occurrence weakens the memory process, while in the case of fewer encounters, the memory process does not show such interference. This interference given by high frequency objects can be counterbalanced if the many exemplars of the category are more diverse (Konkle et al., 2010).

To test this interpretation, we have collected ratings of Conceptual Distinctiveness (CD) from an independent group of participants (N=16). Object categories with high CD are categories whose exemplars are more diverse, not just at perceptual level, but in terms of different kinds (e.g., the category “tree” has high CD because it includes palms, pine tree, oaks, cypress, etc.; see Figure 4).

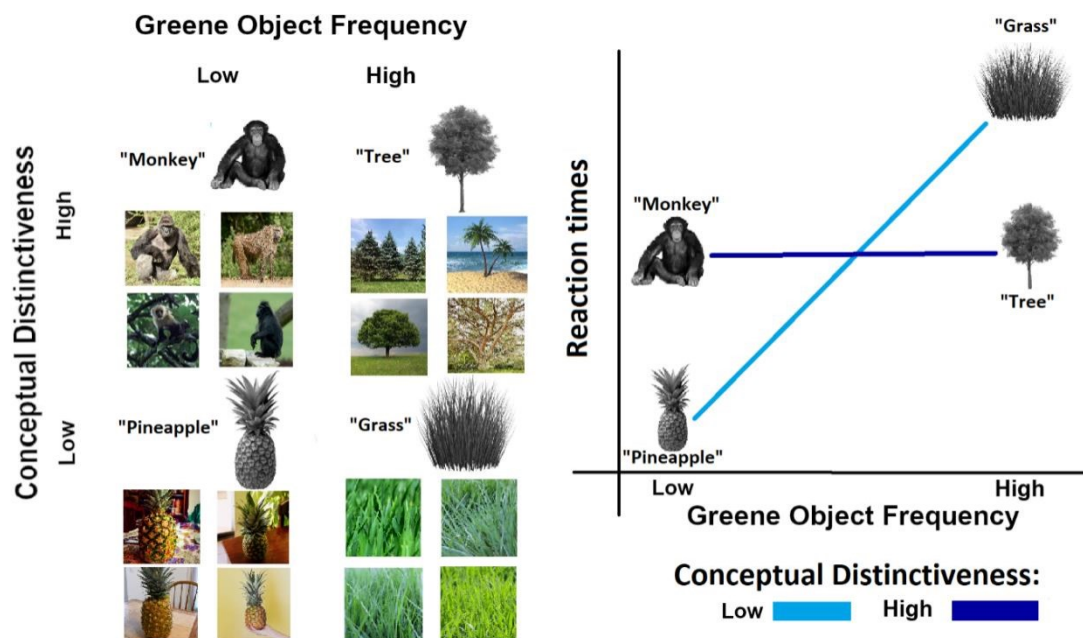


Figure 4 – Study 1: Example of interaction between object frequency and conceptual distinctiveness – On the left, examples of object concepts that have low and high object frequency (OF) in combination with high and low conceptual distinctiveness (CD); the black-and-white pictures and the written words (translated) are the stimuli used in the study, the coloured pictures (taken from Hebart et al., 2019) offer an example of the diversity of kinds within each category; on the right, the expected pattern of interaction between the two dimensions (OF and CD) on reaction times during priming task according to our interpretation based on Konkle et al. (2010).

Indeed, when re-running the model on data of Experiment 2 including an Object frequency (OF) x Conceptual Distinctiveness (CD) interaction, this interaction was found to be significant and stronger in Cross-modal matching trials, where OF effect was previously found. The interaction revealed that when CD is low (low variability of kinds of exemplars), the OF effect remains strong, with the interference found in Experiment 2; however, when CD is high (high variability), the OF effect is counterbalanced and no interference is found (see Figure 5).

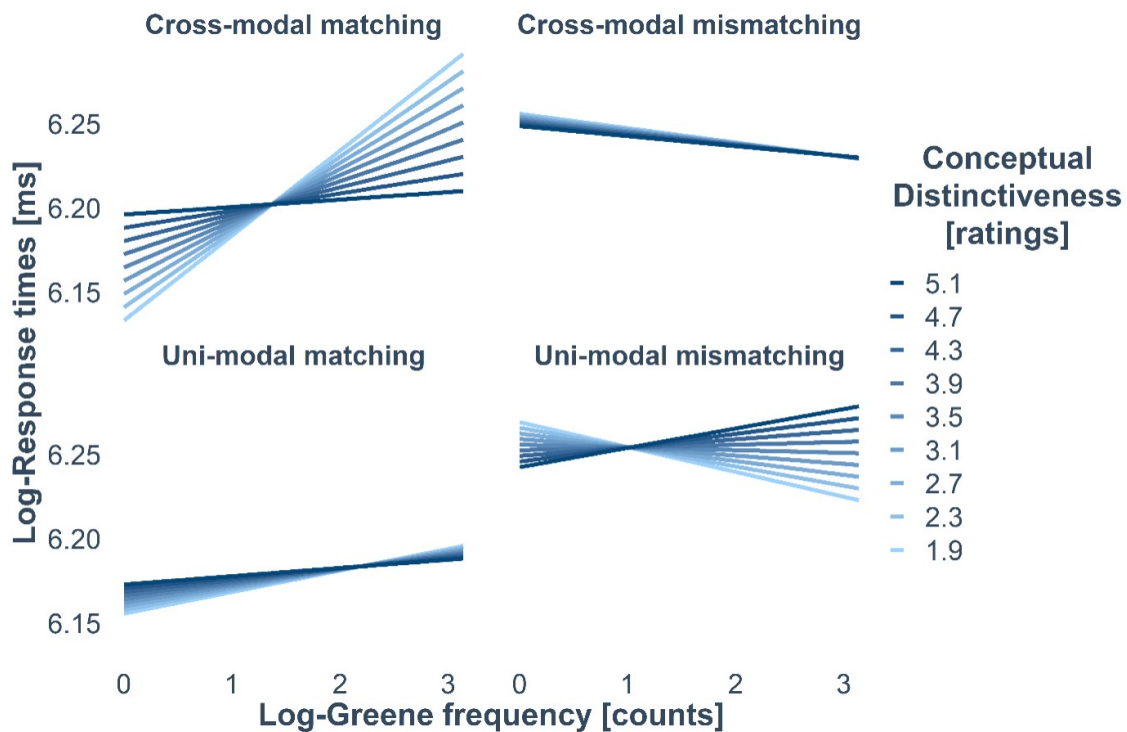


Figure 5 – Study 1: Results interaction of object frequency and conceptual distinctiveness – Participants’ response times (y-axis), estimated from the model including a OFxCD interaction, as a function of Greene OF (x-axis) and Conceptual Distinctiveness (CD, different shades of colour), in the different experimental conditions of the priming task (top-left, top-right, bottom-left, bottom-right). Lines represent linear fitting of individual concepts averaged across participants.

The same model of Experiment 2, considering SUBTLEX WF and Greene OF, was run on data from Experiment 3. We replicated the results of Experiment 2 showing that both Greene OF and SUBTLEX WF have stronger effects in Cross-modal matching trials than in Uni-modal matching trials, and again going in opposite directions (see Figure 6). When considering Cross-modal matching trials alone though, only SUBTLEX WF showed a significant facilitatory effect for more frequent concepts, while Greene OF effect did not reach significance, despite being qualitatively in the opposite direction than SUBTLEX WF effect, as in Experiment 2.

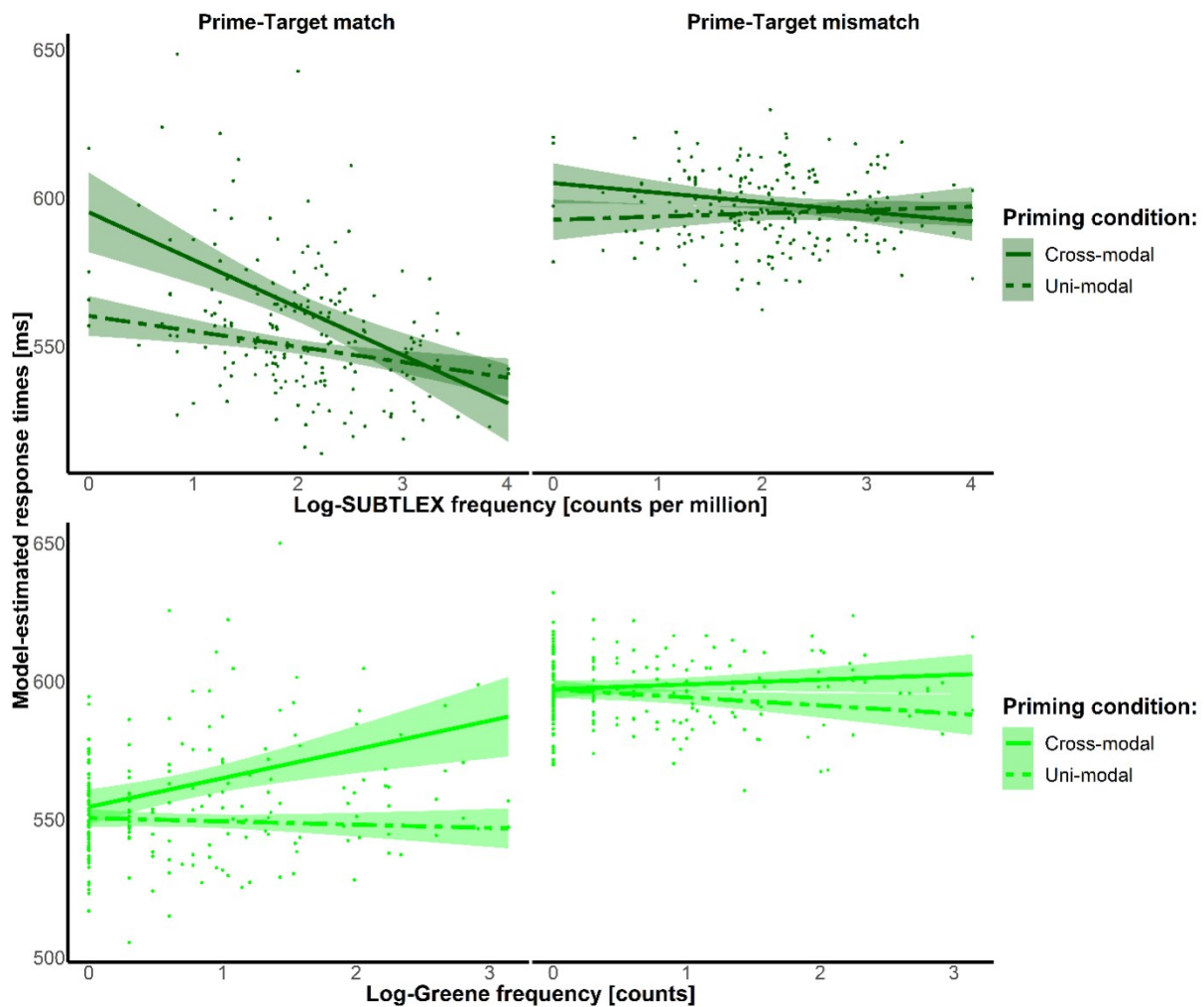


Figure 6 – Study 1: Results Experiment 3 – Participants’ response times (y-axis) estimated from the post-hoc model, as a function of SUBTLEX WF (x-axis up, darker green) and Greene OF (x-axis down, lighter green) in the different conditions of priming (Cross-modal = solid lines; Uni-modal = dashed-dotted lines), and in the different conditions of prime-target matching (Matching = left; Mismatching = right). Points represent response times for individual

To conclude, in this study we have shown how strength of linguistic experience shapes semantic processing (WF for words AND objects, WF in cross-modal priming); at the same time, and for the first time, we have shown that when an object picture or word needs to be predicted using a semantic representation (cross-modal priming), the frequency of occurrence of that object in our visual world also influence recognition (OF effect), but in a less straightforward way, that depends on the structure of object categories (interaction of OF with CD).

Study 2: “Hierarchical organization of objects in scenes is reflected in mental representations of objects”

Introduction

The arrangement of objects in visual scenes is not random, but follows a structure that has been called “Scene Grammar”, that is learnt and used by our cognitive system to efficiently perceive and interact with the environment (Vö, 2021). It has been proposed that this structure might be organized according to a hierarchy (Vö et al., 2019): the first level is the level of the “scene” where objects are arranged; within a scene, there are multiple spatial and functional clusters of objects, that represent the “phrase” level; within every phrase, objects hold different status, with one of them (“anchor object”) offering strong prediction of the identity and the position of other objects (“local objects”; see Figure 7A).

The aim of this study was to test if mental representation of objects are organized according to this hierarchy, that we measured in two ways: 1) using common-sense and intuition (“a priori hierarchy”); 2) using statistics computed from annotated image datasets (“data-driven hierarchy”). Additionally, we tested whether this hierarchical organization depends on stimulus modality (object pictures vs written words) or not.

Methods

We considered 45 everyday object concepts, that we organized in the a priori hierarchy assigning each one of them to one of 5 indoor *scenes*. Within every scene, each object was assigned to one of 3 *phrases*, where one object was assigned the status of *anchor object* and two objects were assigned the status of *local objects*. Instead, for the data-driven hierarchy we considered a dataset of images with annotated and segmented objects (Greene, 2013). In each image (“scene level”), we identified clusters of objects (“phrase level”), in which the largest object was assigned the status of anchor and to the other objects the status of local objects (see Figure 7B).

We then collected behavioural similarity judgements for the 45 objects from two groups of participants: in Experiment 1 (N=43) participants judged object pictures, in Experiment 2 (N=43) participants judged written words of the same objects. Similarity judgments were

collected using an odd-one-out triplet task (Hebart et al., 2020): participants saw triplets of stimuli (words or pictures) and were asked to select the object they considered the least similar to the other two. No explicit definition of similarity was given. From each response, three pairwise similarity judgements are computed: given the triplet A, B and C, if C is selected as the least similar (“odd-one-out”), similarity between A and B will be maximal, while similarity between A and C, and between B and C will be considered minimal (see Figure 7C).

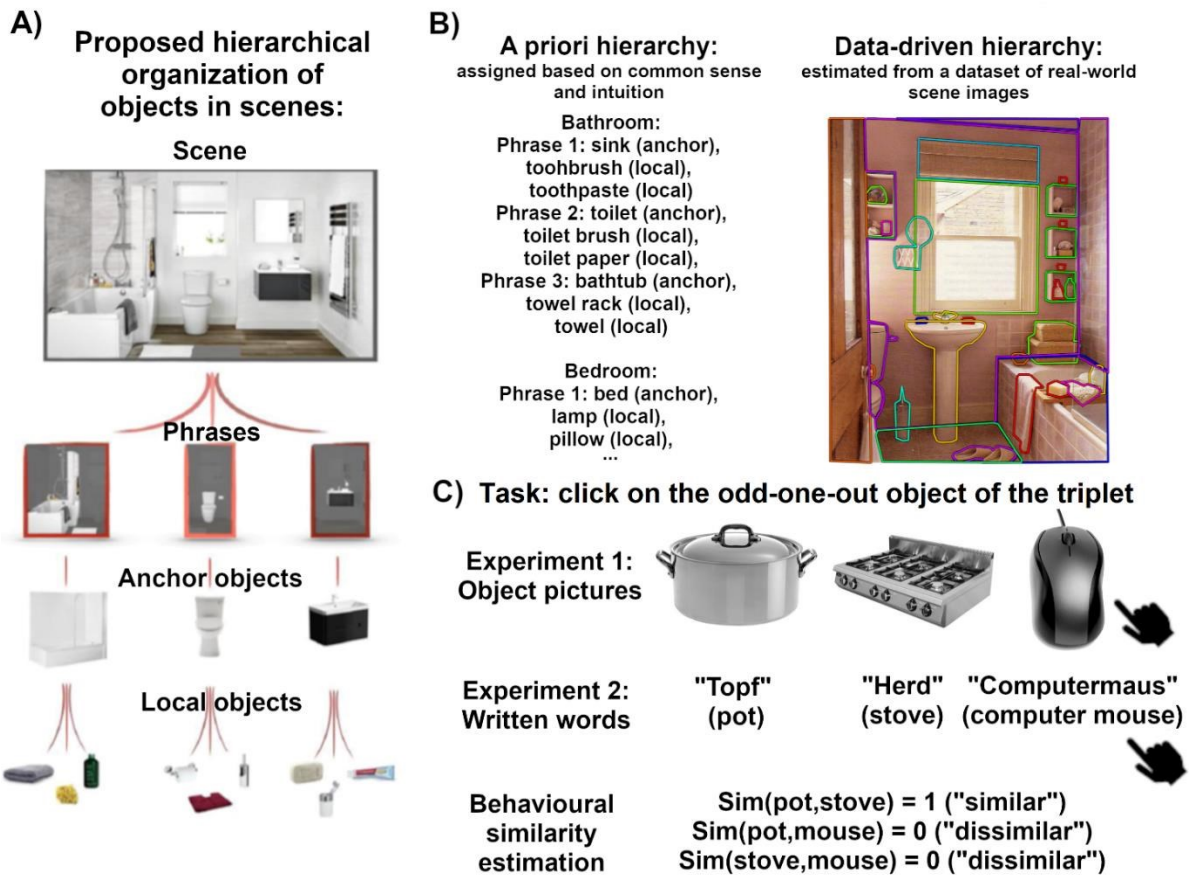


Figure 7 – Study 2: Presentation of the study – A) the proposed hierarchical organization of objects in scenes (adapted from Vö et al., 2019); B) the same hierarchical organization measured using an “a priori” approach (left) and a “data-driven” approach (right; image taken from Greene, 2013 and visualized using LabelMe, Russel et al., 2008); C) example trials for the odd-one-out triplet tasks with pictures (Exp 1, top) and written words (Exp 2, bottom) used in the study, as well as the way we computed behavioural similarity.

To investigate if measures of hierarchy model behavioural similarity judgements, we combined Representational Similarity Analysis (RSA; Kriegeskorte et al., 2008) with Generalized Linear Mixed-effect Models (GLMMs; McCulloch & Neuhaus, 2005). RSA allows to compare data that have different dimensionalities (e.g., behavioural responses, computational models, stimulus features) by comparing their representational spaces instead. Representational space are structured creating Representational (Dis)similarity Matrices (RDMs), which are symmetric matrices where row and column entries are the object stimuli we used, and each cell represents similarity for a pair of objects. These RDMs were created for each of the hierarchical measures we have computed (see Figure 8), as well as for the behavioural similarity judgements collected with the triplet tasks (see Figure 9). Additionally, we have computed RDMs for visual and linguistic covariates that we used to control for potential confounds.

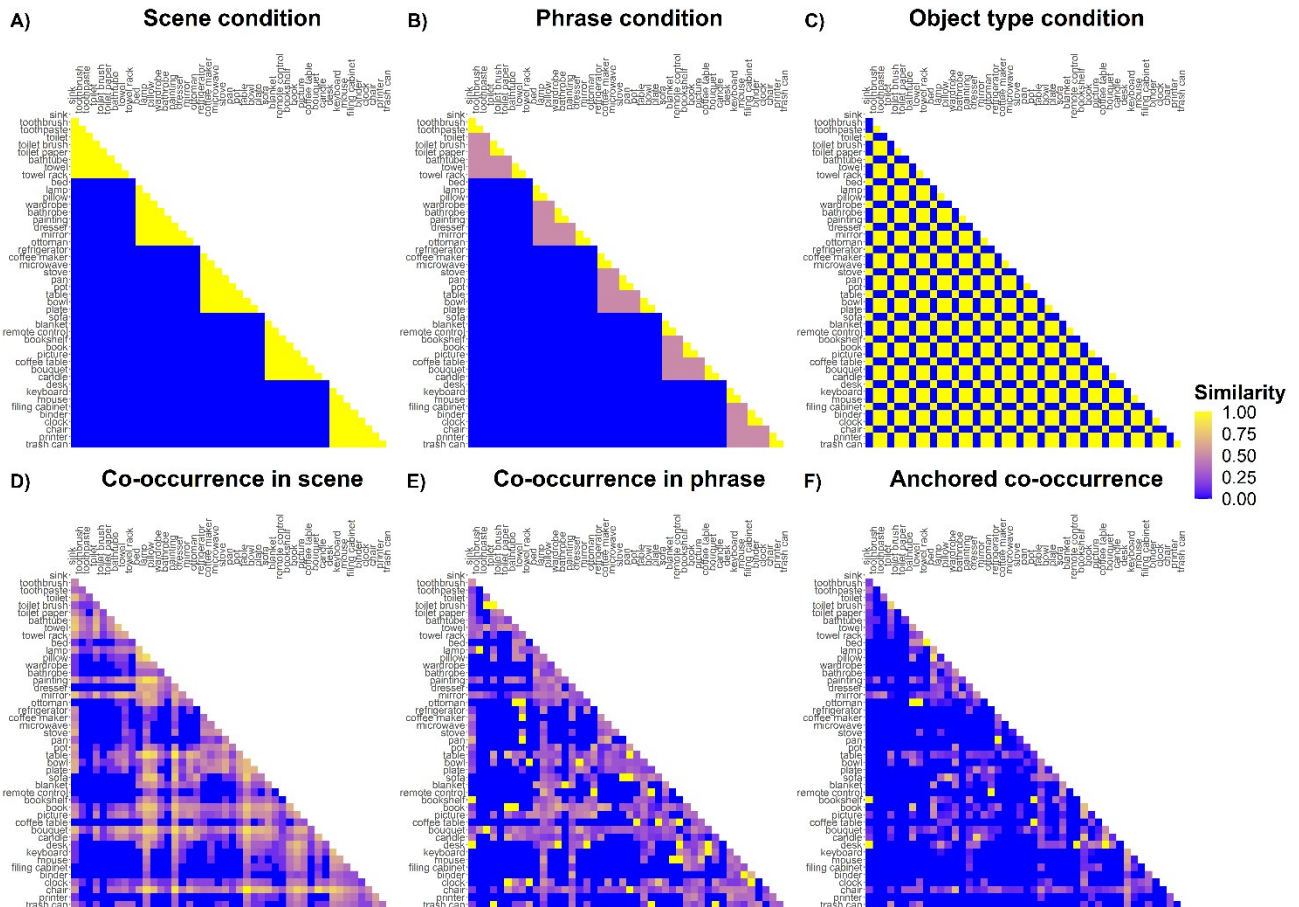


Figure 8 – Study 2: Representation (Dis)similarity Matrices (RDMs) for the hierarchical measures – A), B) and C) are the RDMs for the “a priori” hierarchical measures, where yellow (maximal similarity) indicates pairs that belong to the same scene (A), phrase (B) or are from the same object type (C), while blue (minimal similarity) indicates pairs that belong to different scenes (A, B) or are from different object types (C), and gray (medium similarity) indicates pairs that belong to different phrases within the same scene; D), E) and F) are the RDMs for the “data-driven” hierarchical measures, where yellow (maximal similarity) indicates pairs with large level of co-occurrence in scene (D), in phrase (E) or in an anchor-local relationship (F). Only half of each RDM is shown since the matrices are symmetrical and repeated along the top-bottom diagonal.

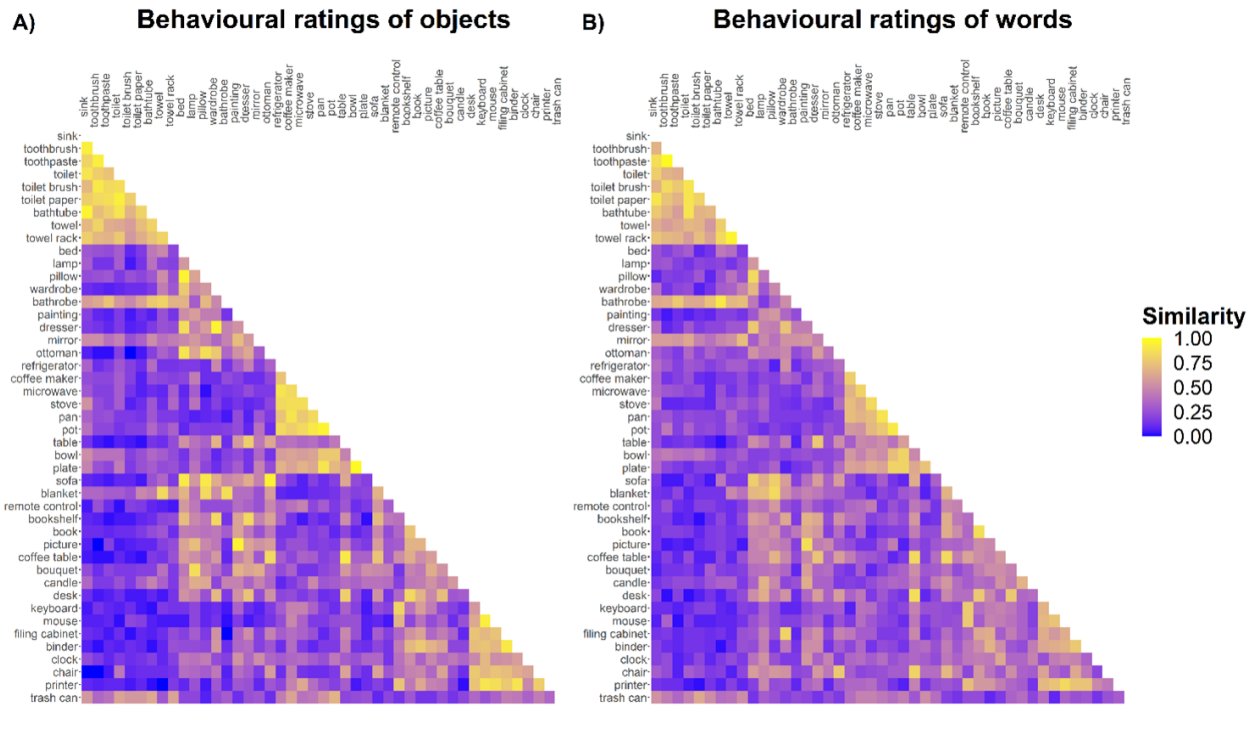


Figure 9 – Study 2: Behavioural Representational (Dis)similarity Matrices (RDMs) – RDMs for the behavioural similarity ratings from Experiment 1 (object pictures, A) and Experiment 2 (written words, B), where yellow represents pairs with maximal similarity and blue pairs with minimal similarity. Pairwise similarity is averaged across all the triplets that contain the pair, and therefore estimated in the context of each of the other objects.

The a priori hierarchy resulted in 3 categorical predictors: *scene condition* (pairs from the same scene vs pairs from different scenes), *phrase condition* (pairs from same phrase vs pairs from different phrases of the same scene), and *object type condition* (pairs of the same object type vs pairs of different object types), where object type means anchor or local objects. The data-driven hierarchy resulted in 3 continuous predictors instead: *co-occurrence in scene* (number of occurrences in the same image for a given pair), *co-occurrence in phrase* (proportion of co-occurrence where pairs are also in the same cluster), and *anchored co-occurrence* (proportion of co-occurrence where the pairs are in an anchor-local relationship).

Results and Discussion

We found significant main effects for all the three a priori hierarchy predictors, so that object pairs that were assigned to the same scene, to the same phrase or to the same object type, were estimated to be more similar than object pairs from different scenes, different phrases or different object types. We found also significant main effects for the co-occurrence in scene and co-occurrence in phrase predictors, with pairs that co-occurred more often being judged more similar. No significant main effect of anchored co-occurrence was found. When comparing these effects between modalities (objects vs words), we only found stronger effects of scene condition and co-occurrence in scene for objects than for words (see Figure 10).

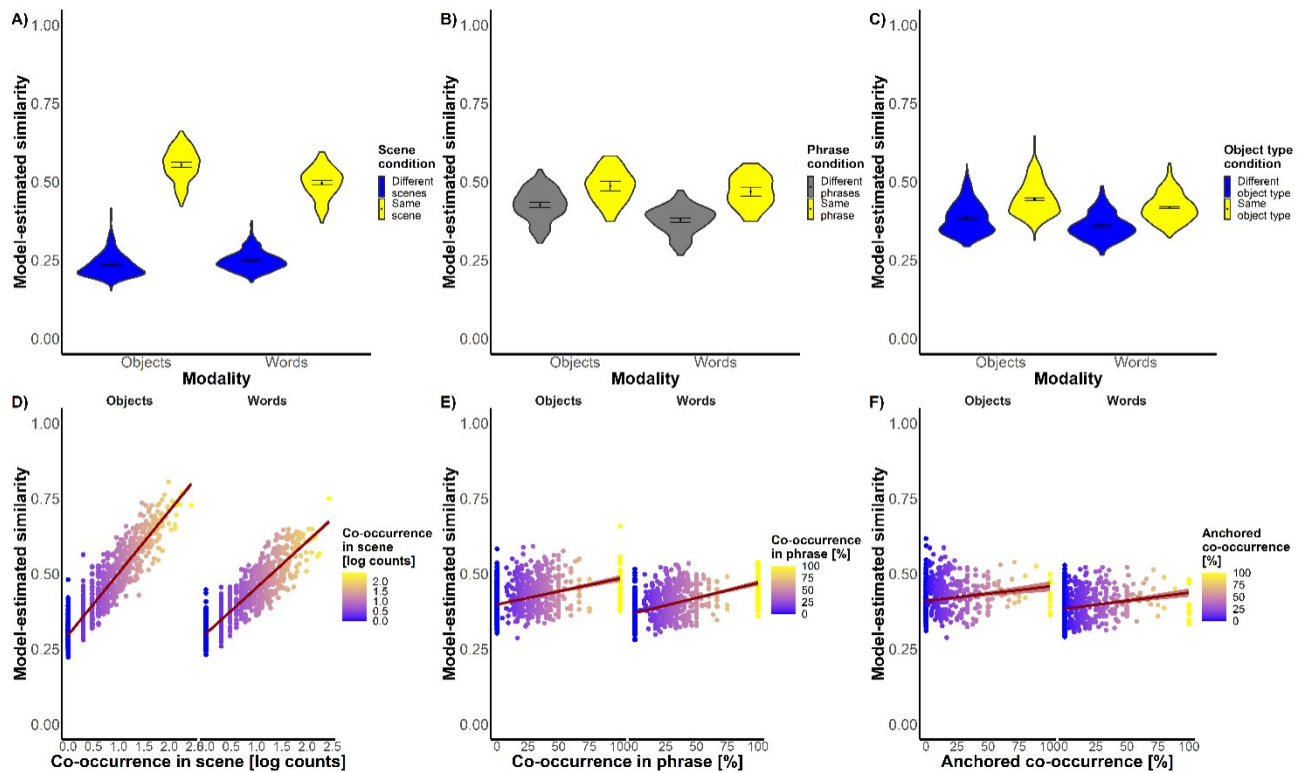


Figure 10 – Study 2: Results – Behavioural similarity judgments (y-axis), estimated from the model, as a function of the “a priori” hierarchical measures (A, B and C) and of the “data-driven” hierarchical measures (D, E and F). Violins and points represent pairs of objects averaged across the different object context (i.e., the third object of the triplet). Different stimulus modalities are represented in different position on the x-axis (in A, B and C) or different facet (in D, E and F). For A, B and C, the black points represent the mean, while for D, E and F the lines represent linear fitting of points; 95 % confidence interval is indicated by error bars (A, B and C) or shaded area (D, E and F).

We managed to show, for the first time, that mental representation of objects are organized according to a scene hierarchy that encompasses three different levels. Crucially, we used both typical a priori assignment and a data-driven approach that estimated hierarchical measures from a set of real-world scene images. Besides, this hierarchical organization seems quite stable across modalities and independent from other visual and linguistic covariates, indicating that this structure is learnt and incorporated in a more abstract representation of objects.

Study 3: “Scene hierarchy structures mental object representations while flexibly adapting to varying task demands”

Introduction

In the previous study we have shown how the scene hierarchy proposed by the “Scene grammar” framework (Võ et al., 2019; Võ, 2021) has a match in the way mental representation of objects are organized. This indicates that the complex structure of the visual environment is learnt and used to guide efficient behaviour.

What remains unanswered is: which purpose does this mental organization primarily serve? Is this organization stable across task or it is enhanced and reduced flexibly adapting to different task demands? It has been shown that scenes are categorized by the set of actions one can perform in them, and that this is the primary and stronger organizational factor (Greene et al., 2016). Therefore, it seems reasonable that the hierarchy of objects in scene might also be primarily relevant in offering efficient representation of objects as prediction for potential action-directed interactions with them.

In this study, we aimed at investigating this problem by comparing the effect of scene hierarchy on behavioural responses during different tasks, one focusing on the actions we can perform with objects, one focusing on the visual appearance of the objects and one unconstrained control task.

Methods

Materials and methods are mostly identical to the one used in the previous study (“Study 2”). We used the same object pictures and the same a priori and data-driven measures of hierarchy; we also used the same visual covariates, computed as unit activations in different layers of a Deep Neural Network (AlexNet, Krizhevsky et al., 2017). This is a computer vision algorithm trained to perform human-like object categorization. From the DNN, we computed low-level visual features (“early layer”), mid-level visual features (“mid layer”) and high-level visual features (“late layer”) for the 45 object images used in the study.

occurrence in scene and co-occurrence in phrase, but not of anchored co-occurrence, with pairs that co-occurred more often being judged as more similar.

Crucial for the study were the interaction effects, comparing the effect of predictors across tasks: scene condition has similar effect between the “action task” and the “no task”, but in the “visual task” this effect is significantly reduced; co-occurrence in scene is significantly stronger in the “no task” than in the other two tasks, while co-occurrence in phrase is significantly stronger in the “action task” than the “no task”, and the anchored co-occurrence was similarly strong in the “action task” and “no task”, but significantly reduced in the “visual task”. Finally, for the visual covariates, we found that similarity in the early layer (low-level visual features) and similarity in the mid layer (mid-level visual features) had both stronger effect in the visual task than the others (see Figure 12).

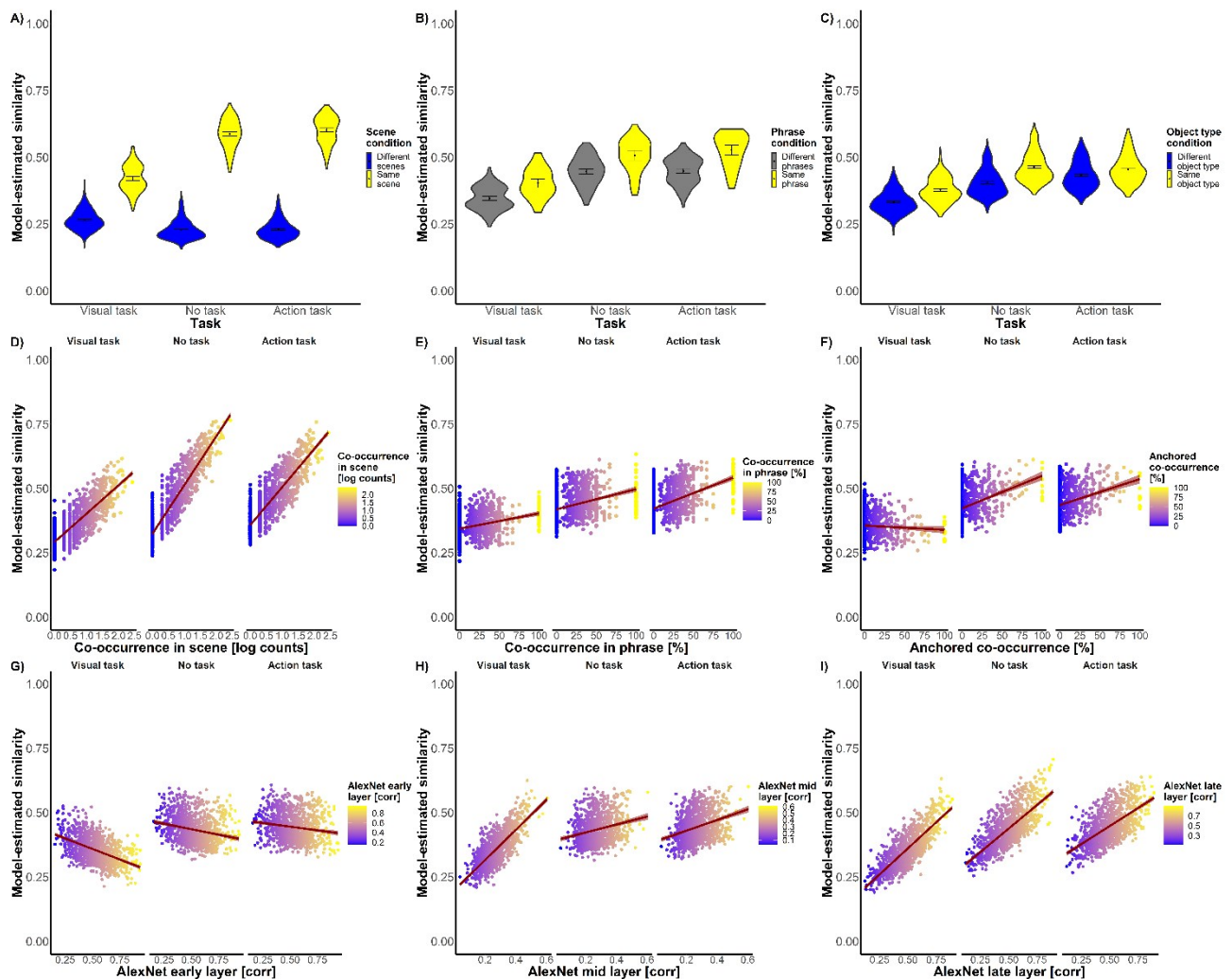


Figure 12 - Study 3: Results – Behavioural similarity judgments (y-axis), estimated from the model, as a function of the “a priori” hierarchical measures (A, B and C), of the “data-driven” hierarchical measures (D, E and F) and of the visual covariates (G, H and I). Violins and points represent pairs of objects averaged across the different object context (i.e., the third object of the triplet). Different tasks are represented in different position on the x-axis (in A, B and C) or different facet (in D, E, F, G, H and I). For A, B and C, the black points represent the mean, while for D, E, F, G, H and I the lines represent linear fitting of points; 95 % confidence interval is indicated by error bars (A, B and C) or shaded area (D, E, F, G, H and I).

The results indicates that hierarchical structure is overall stable (main effects) in shaping mental representations of objects, but at the same time different task demands can enhance task-related aspects of these representations, while reducing task-unrelated aspects of the representations. In the action task, finer-grained information of the hierarchy (phrasal organization) is enhanced, putatively because the more action-oriented task requires to prioritize information that might make interaction more efficient. On the contrary, during visual task, hierarchical predictors become less important, and what is enhanced is the information about low-level and mid-level visual features, like brightness, contours, shape, that all help in defining similarity based on visual appearance.

Importantly, the results also suggest that in absence of explicit similarity criteria (“no task”), the effect of hierarchical predictors and of the visual predictors are more similar to what is found in the “action task” than what is found in the “visual task”. This would point to the idea that the primary role of learning and using this hierarchical organization might be to have a structured understanding of objects in the environment in order to efficiently find them, reach them and use them to perform actions.

General discussion

Advantages and problems in using image datasets to extract object statistics

We have shown how measures of object statistics extracted from segmented and annotated image datasets can explain behavioural response times and similarity judgements during processing of objects, interpreting this fact as an indication that our cognitive system is sensitive to the structure of the world and able to use this information to make efficient predictions. However, it is important to consider whether the datasets from which these measures were computed are a good representation of the structure of the world.

Every dataset (of images, text or other materials) presents us with three main challenges: first, a dataset represents an approximation of the world, even the largest ones available nowadays; second, being an approximation, a dataset contains biases from the process of selecting the materials and from the nature of the sources from which the materials are taken; third, the assumption that the structure of the world captured by a dataset is reflected in the mind is only valid if we assume that experience with that world structure happens in the same way from one person to the other. We will discuss these challenges, focusing in particular on the datasets that have been used in our empirical investigations (Greene, 2013; Zhou et al., 2019).

Once again, we can consider what we have learned from the usage of text corpora in language research to understand the role of dataset size on its reliability. Brysbaert et al. (2011) proposed that a text corpus needs at least 20-30 million words to offer reliable estimates of word frequency, but even larger corpora do not seem to explain more variance (Brysbaert & New, 2009). This consideration of size came from empirical comparisons of different sizes with large dataset of reaction times for lexical-decision tasks. Therefore, for our image datasets the considerations of size should be, where possible, based on empirical comparisons.

This was possible for the results of Study 1 presented here (Gregorová et al., 2021): we compared the effect of object frequency (OF) measures on reaction times for semantic categorization of object pictures and words, as well as for cross-modal and uni-modal priming. The two datasets have very different sizes: the Greene dataset (Greene, 2013), on one hand, is based on 3499 individual scene images, grouped into 16 scene categories (8 indoor, 4 outdoor natural, 4 outdoor man-made). In these images, there are in total about 48,000 individual object

annotations, grouped into 617 object categories; ADE20K dataset (Zhou et al., 2019), on the other hand, has a much larger size, with more than 20,000 individual scene images from about 900 scene categories, and more than 400,000 individual object annotations grouped into more than 2,500 object categories. The two datasets' sizes are still quite far away from what has been found empirically for language corpora, but the current panorama of annotated image datasets hardly offers much more (e.g., COCO dataset for example have many more instances of objects, but also very few object categories, Lin et al., 2014), with datasets of ADE20K size and variability representing state-of-the-art in computer vision semantic segmentation (Yu et al., 2018).

The direct comparison of the Greene dataset and ADE20K dataset in Study 1 showed that in terms of model fitting, Greene OF showed better predictive power on response times data, although ADE20K OF also showed significant improvement in model fitting, and a similar pattern of results than Greene OF (Gregorová et al., 2021). Additionally, and importantly, the two OF measures had a strong correlation ($r=0.81$; Gregorová et al., 2021). A further but more indirect and qualitative comparison between the two datasets could be done by contrasting the results of Study 2-3 (Turini & Vö, 2022a; 2022b), which use information on object pairs co-occurrence and clustering from Greene dataset, with the results obtain by a complex model of distributional semantics of objects computed from ADE20K (i.e., the object2vec of Bonner & Epstein, 2021). Both these models, despite the differences, have shown to match either behavioural or neural representations of objects. To conclude on this, we could state that the different sizes of the Greene dataset and ADE20K dataset do not seem to have produced differences in the empirical efficacy that measures computed from them have on accounting for behavioural (or neural data). Clearly the aim for further developments of this kind of datasets is to expand their size, but quantity does not seem to be the main issue, in our opinion, but rather the quality of the images and annotations. This leads us to the second issue in dealing with datasets, namely the biases they contain (Torralba & Efros, 2011).

When considering biases in scene image datasets, one source of potential problems comes from the fact that images might not cover the wide spectrum of scene categories in real-world, and subsequently include a limited and skewed distribution of objects appearing in them. Consider the case of a dataset of images of only bathrooms, or a dataset of images of only indoor

scenes. Even if these two datasets were based on 20 million unique images, they would be unreliable in estimating the occurrence of certain objects because of the limited variety of the images in them (e.g., for the “bathroom only” dataset we would not be able to understand how frequently stove tops appear in the world; for the “indoor only” dataset we would not be able to understand how frequently trees appear in the world). Both the datasets used in our studies addressed this problem by having images from scene categories that were both indoors and outdoors, and within the outdoors they collected images from both natural and man-made scenes. However, ADE20K dataset offers a hugely bigger variety of scene categories (about 900; Zhou et al, 2019) compared to Greene dataset (16 categories; Greene, 2013), although in the latter we have categories that intuitively we could consider the most common ones (and therefore most commonly used in cognitive science research): kitchen, bedroom, living room, bathroom, dining room, office and conference room (for the indoor scenes); mountain, forest, country-side field, coast beach (for the outdoor natural scenes); street, city center, sky-line, highway (for the outdoor man-made scenes).

Another source of biases to consider is the configuration of scenes as it is depicted in the images. Often times, downloading scene images from copyright free websites biases the datasets because most of the images might come in an artificially cleaned, stylized and idealized view because photographers wanted to capture the whole scene, they wanted it to be tidy and harmonious, which results in a less naturalistic description of what a bedroom or a street look like (e.g., consider the case of images used by design / architecture studios to present potential furnishing solutions). On the one hand, with ADE20K that includes images taken from the SUN dataset (Xiao et al., 2010) and the Place dataset (Zhou et al., 2014), which are both taken from the internet, this potential bias might have been addressed by looking in search engines using the scene name + different adjectives (“messy, spare, sunny, desolate, etc”; Zhou et al., 2014) in order to avoid collection of scenes with too uniform point of view and configuration. On the other hand, with Greene dataset, this issue was addressed by cross-validating the image dataset with an auxiliary dataset compiled using images that were uploaded by people on LabelMe (Russel et al., 2008) and therefore lacking the “artificiality” of typically collected images (Greene, 2013).

The final biasing aspect that we are going to discuss is also the most relevant for this type of datasets: labelling. The individuation of objects in these image datasets is strongly dependent on the label they received from the annotators and their labelling decisions. Assigning a label to an object can be easy sometimes, but other times it requires that a criterion is used. For examples, a tree can be labelled as “tree” but also as “oak”; a car can be labelled “car”, but also more generally “vehicle” or more specifically “Ferrari”. This is particularly relevant if object categories are very heterogeneous, as it emerged from the interaction of this dimension with object frequencies as computed in Study 1 (Gregorová et al., 2021). A related problem is the usage of synonyms or repetition error (e.g., using singular and plural labels, or capitalizing the first letter of the label on some occasion and not in others). Both ADE20K and Greene datasets have addressed this problem by training annotators in their labelling procedure: assigning entry-level label when possible (e.g., “dog”, neither “pet” nor “golden retriever”), using singular nouns, using all lowercase characters, etc. Besides, they both used additional annotators on a subset of the images to test reliability of label assignment across different individuals (Greene, 2013; Zhou et al., 2019). One thing that differentiates the two datasets in terms of quality of annotation is the fact that Greene dataset, despite being smaller (or maybe because of this), has more complete and clean annotations, where synonyms are checked and repetition are avoided. Applying this same care to the annotations of ADE20K would help to overcome the limitations posed by the smaller number of scenes in the Greene dataset.

Finally, the third main challenge in using image datasets (but also other types of datasets) is the assumption that a dataset, even if large and unbiased, is capturing aspects of the world that are experienced in similar ways by everyone. Even for the knowledge acquired primarily with perception, it seems likely the physical and cultural context in which the perceptual experience takes place play a role, as well as the individual expertise with the content of the experience. Consider for example, the outdoor environments experienced by a person in North America in opposition to a person in Europe or in Africa or in East Asia. Landscapes are different, streets and roads are different, urban structures are different, even objects within the environments are different, like for example cars. This is also true if we consider design of indoor scenes: kitchens, bathrooms, living rooms, bedrooms have differences reflecting the influence of cultural habits,

with presence and arrangement of objects that differ sometimes in subtle ways, sometimes in more evident ways. Although in a globalized and digital world we can have a visual experience of outdoor and indoor environments and objects that are far away from our own culture, it would be important to acknowledge cultural differences in objects' presence and arrangement, and to find a way to incorporate these differences when building image datasets.

Similarly, we know how important the effect of expertise is to visual recognition (for a review, Harel, 2016): some objects and some environments are more familiar to us because we spend a lot of time performing actions with and in them; objects that are more familiar to us can be differentiated better, while other people who are not familiar may consider them to be the same, influencing the reliability of a label in an image dataset. Going back to what is known from using text corpora, recently it became more relevant when estimating measures like word frequency to acknowledge not simply the frequency per se, but in how many contexts a word appears, if those contexts and their linguistic registers are familiar to the participants from which behaviour is measured, and more generally, the role of subjective familiarity of individuals with certain words (Brysbaert et al., 2011; 2018; Kuperman & Dyke, 2013). We have tried to take care of these aspects when using word frequency measures in Study 1 (Gregorová et al., 2021), but our (and of other people's) few attempts to use object measures from image datasets still have to properly incorporate these suggestions.

To conclude on this topic, the first positive results of using annotated and segmented image datasets are promising for the future of visual cognitive science, with some of the challenges of large datasets being addressed, but with a lot of further work that needs to be done to improve their quality and reliability.

The future of Scene Grammar

If on one side one of the goal of our studies was to consolidate the findings of Scene Grammar research (i.e., measuring object-to-scene and object-to-object consistency in a more data-driven and continuous way), on the other side it is important to discuss how our findings might suggest future directions and development of questions related to Scene Grammar.

The results of Study 3 leaves us with the suggestion that aspects of the hierarchy, in which objects seem to be organized within a scene (and in the mind), are enhanced and functional to efficiently perform actions using those objects within the scene (Turini & Vö, 2022b). This idea is not new in scene perception research, having roots, as we have seen, in the ecological approach of Gibson (Gibson, 1986), for which visual input contains already information about efficient motor pattern elicited by certain objects (“affordances”). This intuition were developed showing evidence that neural processing (Greene & Hansen, 2020) and categorization behaviour (Greene et al., 2016; Groen et al., 2018) of scenes are strongly influenced by the set of actions that can be performed in a certain scene.

This evidence makes intuitively sense. If on one side the activation of Scene Grammar knowledge starts with the recognition of the environment, that we have seen it can be achieved using many different global and local scene properties (Wiesmann & Vö, 2022; Lauer & Vö, 2022); and then, this Scene Grammar knowledge is used to efficiently recognize (Lauer & Vö, 2022), search (Helbing et al., 2022) and interact (Draschkow & Vö, 2017) with objects within scenes; on the other side it seems that the end goal of all these processes is to allow the usage of the objects that are recognized, found and reached (Malcom et al., 2016). This action-centered processing of Scene Grammar knowledge seems particularly important for indoor and man-made environments, that are designed by humans exactly to serve some purposes (e.g., in the bathroom one showers, in the kitchen one cooks, in the bakery one shops). Given these considerations, on our opinion, future investigations of scene grammar should incorporate the role of action goals in more typically investigated object recognition and search task. For example, what is the influence of object-to-object spatial relationship (e.g., distance, consistent vs inconsistent relative positive) when one needs to search for multiple objects that serve a specific task (e.g., a pan and a spatula when cooking eggs)? What is the influence of consistently and inconsistently placed anchor objects (e.g., stove top and refrigerator) in this sequential and goal-directed search? How actions are differentiated between the ones involving multiple phrases of a scene and the ones linked to a specific cluster of objects (see studies on the so called “reachspaces”, Josephs & Konkle, 2019; 2020)?

Adding such complexity to typically used paradigms might come with the risk of losing experimental control. We believe that this problem could be addressed continuing in two directions that have been already approached when investigating Scene Grammar. The first one relates to this current investigation, and recommends the employment of large set of images from which to extract more and more objects-to-scene information. The second one is the employment of Virtual Reality (VR) to test behaviour in a more ecological settings.

Investigating object recognition and search with the idea of reaching and using the objects, especially with VR, also opens to incorporate in the study of Scene Grammar another fundamental task we perform in scene, which is navigation (Bonner & Epstein, 2017; Malcolm et al., 2016). When searching in the real world (and in VR) we do not just use our eyes, but we move our head and also our body. Going back to the example of cooking an egg, even within a small-scale environment such as a kitchen, we need to move from one anchor to the other (e.g., cabinet, stove top, refrigerator) in order to find and grab the objects we need.

Considering navigation also offers the occasion to reflect on the differences between outdoor and indoor environments, and how these differences impact Scene Grammar knowledge. Although we can identify object-to-scene relationship in outdoor scenes as we do for indoor scenes (e.g., a car in the street is in a consistent relationship), it becomes harder to think about object-to-object hierarchical organization and the distinction between anchor and local objects when dealing with outdoor environments. This has again to do likely with the fact that indoor environments are design to perform specific actions, while outdoor environments, also because of their larger scale, they seem to be predominantly designed for navigation, or for actions related to navigation (e.g., driving, cycling). One direction we suggest would be to investigate how to include outdoor scenes into a hierarchical structure, where the crucial dimension is spatial scale: we navigate space through large scale outdoor environments (e.g., a street) until we reach smaller scale environments, often times indoor ones (e.g., a bakery); within these scenes, in order to perform an action (e.g., buying something), we search for our target using anchor objects, which point to a specific reachspace/phrase (e.g., the space around the counter or the space around a shelf); finally, once the reachspace/phrase is found, we use our scene grammar knowledge about relative spatial arrangement between anchor and local objects

to find and reach our target (e.g., a loaf of bread). Although this is only a relatively simple example of behaviour, it poses already a challenge of complexity that is rarely addressed in cognitive science research.

A further issue, that goes beyond this action-centered proposal of scene grammar knowledge but that is in some sense related, is the investigation of scenes and objects that do not have a clear specificity in terms of arrangement and functions. To clarify, consider the case of a living room compared to a kitchen. The design of a kitchen is functional to a specific sets of actions, mainly related to cooking and consuming food. This is quite standard and independent from room size, quality of furniture and even on some extent from culture to culture. Living rooms, instead, are designed to be more general in their purposes, although with some typical ones (e.g., sitting on a couch / armchair to relax, or welcoming guests). Many people do not have a living room, though; many people in their living rooms do not have a TV, while many others do; many people read and / or listen to music in their living rooms, but many others do not; some people have a piano in their living rooms, most of the people do not. Keeping into account the differences in scales addressed before, this difference in specificity of actions becomes more evident if we compare indoor (e.g., bathroom) and outdoor scenes (e.g., a forest). Similar considerations can be done for certain objects: a book is mostly found in a living room, in a bedroom or in a home office, but it would not come as inconsistent to find it in a kitchen, or even in a bathroom; the same applies to chairs that, with some differences in design and materials, they can be found with no surprise in a lot of different context, indoor and outdoor. It seems that for some objects and some scenes is difficult to find a clear semantic relationship (here to be intended as “using object X in scene Y to perform action Z”). Thus, in this case our suggestion, following some empirical findings in literature, would be to depart from scene semantics and focus on an aspect that is pre-semantic or to some extent independent from semantics, which is space (Castelhano & Kryzś, 2020).

The idea that spatial arrangement is an independent dimension from semantic relationship of objects has been investigated for some years now. For example, it has been shown, using the typical paradigm of consistent vs inconsistent arrangements, that an object that is semantically inconsistent but is placed in a consistent spatial configuration is processed better

than a semantically consistent object placed in an inconsistent position (Castelhano & Heaven, 2011). Besides, an object's function can be informative of its spatial arrangement within a scene, even in absence of clear semantic information from the scene (Castelhano & Whitterspoon, 2016). A scene spatial information that seems to be used in advance to scene semantics is the distribution of surfaces (Pereira & Castelhano, 2014; 2019). According to this "surface guidance" model, during search, attention is deployed to surfaces within a scene, and the surfaces that most likely would "host" the target object are further selected, simplifying the search task (Castelhano & Kryzś, 2020).

All this evidence are not new to the Scene Grammar framework, but highlights how spatial dimension could support object-to-scene and object-to-object knowledge with reduced or absent semantic information. Future investigations should try to integrate these different suggestions in a unique and coherent research plan.

Scene Grammar and applications to Artificial Intelligence

The incredible developments in computer vision of the last decade, started with the introduction of deep Convolutional Neural Network for image classification (dCNNs, or CNNs or DNNs, Krizhevsky et al., 2017) and continued with the "Deep Learning revolution" (Sejnowski, 2018), has brought a lot of enthusiasm in the field of Artificial Intelligence (AI) and in the idea of developing algorithm that are able to extract complex information from images in order to achieve complex visual task.

Many of the datasets we have discussed above (ADE20K, Zhou et al., 2019; COCO, Lin et al., 2014; Places, Zhou et al., 2014) were developed exactly within this field and specifically to train algorithm to perform efficient semantic segmentation of scenes (Yu et al., 2018). Semantic segmentation of scenes is a complex task that requires segmenting the image / space into meaningful entities, such as objects and more extended "stuff" (e.g., sky, ground, sand, sea, etc.). This task is crucial, as we have seen from scene understanding research in human (Malcolm et al., 2016), to interact with space and objects in a more efficient way, for example in searching for target or in navigating through obstacles. Implementing these trained algorithms in machine

could have, and in some cases already has, important application in automated driving vehicles or for rescue robots, just to name two examples.

And if these datasets developed for AI became a resource for research in cognitive science in humans, as the current investigation has tried to present, it became also clear how the understanding of human perception and cognition can guide further developments in AI. Already Oliva and Torralba (2007) pointed to the need for computer vision algorithms to integrate contextual information of scenes, as found in human vision, to improve object recognition (Oliva & Torralba, 2007). For examples, it has been showed that combining two CNNs, one processing object information and one processing background scene information, the scene processing CNN is able to extract the “gist” of the scene and improve the classification performance of the object processing CNN (Wu et al., 2018; Zhang et al., 2020). With a similar goal, it has been shown how embeddings built from object co-occurrence distributional measures (as in Bonner & Epstein, 2021) can improve scene recognition and object recognition CNNs (Treder et al., 2020).

Relating these applications to the insights on scene structure discussed in the current investigation, some researchers in AI have started using notion from Graph Theory to train DNNs in learning the hierarchical structure of objects in scenes and the complex web of object-to-object relationships. Graph theory is the branch of mathematics that study “graphs”, which are mathematical structures that model a set of entities and their complex network of pairwise relationships. Entities are represented as “vertices” (the points), while pairwise relationships are represented as “edges” (the lines) in this network (Zhang & Chartrand, 2006). This graph-like representations can be applied to scene images, where vertices are objects (a table, a chair) and the edges represent pairwise relationship between objects. “Scene graphs” can further be expanded defining objects’ attributes (“chair is made of wood”, “table is red”) and the nature of the relationship (“chair” -> “on the left of” -> “table”; Johnson et al., 2015). This kind of representation of real-world images have been used to train DNNs to perform image retrieval matching the scene graph description with success, outperforming models that used purely text-based description (Johnson et al., 2015). The limitation of this work was in the fact that it required human-aided scene graphs as an input (i.e., graph descriptions made by human workers) to train the networks in performing image retrieval. A system that is able to understand object-to-object

relationship would need to be able to create scene graphs given any kind of scene image. This is what a following work has focused on (Xu et al., 2017): a network takes an image as input and output a scene graph, with objects, objects' attributes and objects' relationship represented as bounding boxes with connections among them. Similarly, other attempts have tried to improve scene graphs generation by integrating both local information (connections centered in one object) and global information (all the connections between all the objects; Woo et al., 2018). This type of networks have been also shown to capture in their scene graph descriptions relational regularities at different scales within the images, suggesting that they are able to extract a hierarchical organization (Zellers et al., 2018).

The intuition and development of these so-called “relational modules” in DNNs is opening many interesting challenges in AI, that go beyond visual domain (Santoro et al., 2017). Being able to detect and understand relationships between items is crucial not just for interacting with the visual environment but also for complex reasoning, action planning and communication. These initial applications open to a promising future of research in implementing structured and complex task in AI algorithms.

Our mind and our world

In this dissertation we have started discussing the relationship between the mind and the world, their complex dynamics and the necessary balance between “what’s outside” and “what’s inside” that is reached in typical cognitive functioning.

We have focused on perception as a paradigmatic set of mental functions where this exchange between internal and external contributes to produce a sense of being here and now, with our own body immersed in a complex and changing but also stable environment, despite the mediation of important mental structures: representations. These representations are fundamental to understand the world and, crucially, to predict its functioning.

From this, we have zoomed on some of these predictive processes, called Implicit Statistical Learning, where the mind learns the structure of the world by tuning and adapting to its structure in ways that seem almost effortless for us. Then, we have further narrowed our

interest considering implicit statistical learning for complex spatial configurations, in particular the visual environment in which we move and do stuff, i.e., visual scenes.

We have discussed how scenes have a complex structure in some sense similar to the structure of language, where things (objects or words) are arrangement in specific positions and in a meaningful way (i.e., what has been called “Scene Grammar”). What became relevant, and we tried to address empirically with our work, is the possibility of using large set of data, in this case annotated real-world images, to extract measures of objects in scenes, measures of our visual world, and to compare these measures with the “measures” that our mind computes from the world, as a results of the adaptation processes of implicit statistical learning.

Our experiments offered positive results, showing that frequency of occurrence of an object influences semantic representations in memory, interacting with the categorical structure of an object concept (“Study 1”); besides, we were able to show that these semantic memory representations of objects are organized according to a complex hierarchical structure present In our environment, that depends on variables such as co-occurrence, distance, size and positional stability of objects (“Study 2”); additionally, this hierarchical organization seems to be enhanced when objects are processed considering their action goals than when objects are processed in terms of visual appearance, suggesting that learning of these visual contextual regularities might serve the purpose of efficiently interacting with and acting on the objects in the environment.

In light of the results, we have discussed the advantages and limitations of using above mentioned image datasets and proposed some future directions of investigating Scene Grammar in relation to its putative role of supporting efficient action. This leads us to a final consideration. The need to fill the gap of individual differences in the experience of the world that is missing from the image datasets in use, as well as the individual differences in terms of actions that are performed in real-world scenes. This is shading light to a broader issue in cognitive science and psychology. Although experimental psychology focuses on understanding the mind in its more general and shared aspects, that belongs to all human beings (and sometimes to other animals as well), when we investigate some complex behaviours that involve the influence of many socio-cultural factors (as in our case, consider e.g., the influence of culture on interior design), it seems

crucial to find a way to integrate the more “universal” dimension of experience to the more personal one.

On how this can be achieved, we are afraid we cannot offer specific and definitive answers, but we can suggest to focus on combining the extremely valuable resources coming from big data, as we have tried to do here, together with a more tailored set of measures that are aiming at capturing the individual differences we have previously discussed. In this way, the impact of the objective measures can be weighted and adjusted taking into account more subjective measures. The aim is to find a way to comprehend mental processes like implicit statistical learning in a way that offers an understanding of the mind and the world in general, but also an understanding of our minds and our experience of the world.

Original manuscripts

Access to meaning from visual input: Object and word frequency effects in categorization behavior.

Klara Gregorová^{1, *}, Jacopo Turini^{1, *}, Benjamin Gagl^{1, 2, ^} & Melissa Le-Hoa Võ^{1, ^}

1. Department of Psychology and Sports Sciences, Goethe University,

Frankfurt am Main, Germany

2. Department of Special Education and Rehabilitation, University of Cologne,

Cologne, Germany

* Joint first authors, ^ Joint senior authors

Corresponding author info: Jacopo Turini, Scene Grammar Lab, Institut für Psychologie, PEG Room 5.G105, Theodor-W.-Adorno Platz 6, 60323 Frankfurt/Main, Germany, turini@psych.uni-frankfurt.de, +49 (0)69 798-35310

Data and materials: <https://osf.io/d3j9h/files/>; Word count: 13,191;

Previous dissemination

Data, results, and interpretation from the current study have been previously shared in the form of preprint on PsyArXiv and ResearchGate; besides, they have been presented in the form of poster during the virtual Vision Science Society meeting of 2021.

Acknowledgments

We want to thank Michelle Greene for generously making her dataset available to us (Greene, 2013) as well as three anonymous reviewers for their valuable and constructive comments on this work. This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) project number 222641018 SFB/TRR 135, subproject C7 to MLV and Hessisches Ministerium für Wissenschaft und Kunst (HMWK; project “The Adaptive Mind”).

Abstract

Object and word recognition are both cognitive processes that transform visual input into meaning. When reading words, the frequency of their occurrence ("word frequency", WF) strongly modulates access to their meaning, as seen in recognition performance. Does the frequency of objects in our world also affect access to their meaning? With object labels available in real-world image datasets, one can now estimate the frequency of occurrence of objects in scenes ("object frequency", OF). We explored frequency effects in word and object recognition behavior by employing a natural vs. man-made categorization task (Experiment 1) and a matching-mismatching priming task (Experiment 2-3). In Experiment 1, we found a WF effect for both words and objects but no OF effect. In Experiment 2, we replicated the WF effect for both stimulus types during Cross-modal Priming but not during Uni-modal Priming. Moreover, in Cross-modal Priming, we also found an OF effect for both objects and words, but with faster responses when objects occur less frequently in image datasets. We replicated this counterintuitive OF effect in Experiment 3 and suggest that better recognition of rare objects might interact with the structure of object categories: While access to the meaning of objects and words is faster when their meaning often occurs in our language, the homogeneity of object categories seems to also impact object recognition, particularly when semantically processing contextual information. These findings have major implications for studies wanting to include frequency measures into their investigations of access to meaning from visual inputs.

Keywords: word recognition, object recognition, frequency, distinctiveness, priming.

Introduction

Visual recognition is the cognitive process that maps sensory input from the retina onto meaningful representations stored in semantic memory (Clarke et al., 2013; Grill-Spector & Weiner, 2014); this process supports many tasks like action planning, navigation, reading, social interaction, etc. The types of visual input for these tasks, e.g., objects, scenes, written words, or faces, already pose a high level of complexity, so that research in cognitive science has often investigated different types of visual input separately, focusing on the specificities of each domain (Capitani et al., 2003; Downing et al., 2006). Notably, investigations of the ventral visual stream, i.e., the core neural substrate of high-level vision, compared the brain activation in response to these different types of stimuli (for a review, Grill-Spector & Weiner, 2014). Their main finding was that different stimulus types activated distinct but neighboring regions (e.g., fusiform face area; Kanwisher, et al., 1997; the visual word form area, Dehaene & Cohen, 2011). At the same time, other researchers focused on comparing different visual inputs, e.g., objects and words, to understand the process of accessing the same semantic representation, i.e., the identical meaning (Shelton & Caramazza 1999; Shinkareva et al., 2011; Devereux et al., 2013; Fairhall & Caramazza, 2013). We followed this approach and investigated how different types of visual input can access identical meanings. We were particularly interested in the frequency of occurrence in the world, operationalized by both word and object frequency measures.

In the fields of visual word recognition (Balota et al., 2004) and reading (Kliegl et al., 2006; Rayner, 2009), the so-called “word frequency effect” is a well-established finding. The word frequency (WF) effect shows that words that occur more often in our language (e.g., the article "the") are processed faster than words that are rare (e.g., “platypus”). Common naming and lexical-semantic categorization tasks, e.g., lexical decision tasks (Balota et al., 2004; for a

review, Brysbaert et al., 2011; Brysbaert et al., 2018), consistently show WF effects, i.e., longer response times and more errors for low frequency words. Even though there have been various attempts to identify more reliable estimates of WF and its nature, it is generally agreed upon that the WF effect emerges as an effect of learning and exposure to a language (for a review, Brysbaert et al., 2018). Thus, despite different assumptions and implementations, most models of visual word recognition and reading took WF into account as a crucial parameter representing the difficulty in accessing lexical representation in the so-called “mental lexicon” (Forster & Chambers, 1973; Morton, 1979; McClelland & Rumelhart, 1981; Coltheart et al., 2001; Engbert et al., 2005).

Object recognition models (Riesenhuber & Poggio, 2000), on the other hand, are primarily concerned with assigning images to different categories, irrespective of their frequency of occurrence (Morrison et al., 1992; Criss & Malmberg, 2008; Taikh et al. 2015). In the rare cases when studies compared recognition performance of written words and matched object images, typically frequency effects were investigated based on word measures. For example, Taikh and colleagues (2015) found faster object than word recognition performance in a semantic categorization task, but WF only affected word recognition performance (Taikh et al., 2015). When behavioral investigations used naming aloud tasks, object recognition performance also showed frequency effects based on word-based estimates (Bates et al., 2001; Almeida et al. 2007; Taikh et al. 2015). However, the WF effects found in object naming studies are likely related to the process of accessing the verbal output representation (i.e., the spoken word; Almeida et al., 2007). Thus, tasks that involve linguistic representations, e.g., as part of the output modality, might be more sensitive for word frequency effects on object recognition performance.

A potential limitation of previous investigations comparing the two domains (words and objects) is that these studies included only frequency estimations that rely on linguistic input:

i.e., large text corpora (e.g., books and newspapers, like dlexDB, > 20 million words; Heister et al., 2011) or spoken language corpora (e.g., from tv-movie subtitles, like SUBTLEXDE, about 25,4 million words; Brysbaert et al., 2011). Typically, WF estimates represent the number of occurrences per million words. Across languages, WF estimates have a better explanatory power for reaction time data from word recognition tasks when extracted from TV and movie subtitles than from book and newspaper texts (e.g., for German, see Brysbaert et al., 2011; for English, see Brysbaert & New, 2009). This finding likely reflects that participants in psycholinguistics experiments (often young students) are more exposed to popular TV shows and movies than the content of classic text corpora, which often include highly specialized texts. Thus, subtitle-based WF measures are, to date, the best representation of the number of occurrences of words in everyday life (Brysbaert et al., 2011; 2018). However, it is still unclear how these more precise measures estimated from subtitles might also explain recognition performance in the object domain. Furthermore, it is essential to explore if newly developed frequency measures, based on the occurrence of objects in images of real-world scenes, could also be valid estimates of access to meaning or not, and could also shed more light on the phenomenon underlying the WF effect. Thus far, the lack of such object frequency (OF) measures has likely been due to a lack of easy access to fully labeled image databases.

Recent advances in computer vision have made annotated image datasets with segmentations and labels of all objects within a scene readily available. Usually, these labels come from human annotators (Russel et al., 2008). For example, the ADE20K dataset contains over 20,000 real-world images from 900 different scene categories, with hundreds of thousands of object annotations categorized into more than 2,500 object categories (Zhou et al., 2019). Despite having been developed for computer vision research, these datasets allow us to extract quantitative measures about contextual regularities of objects in the environment (e.g., objects that appear more often in a specific scene category). These newly available object-in-scene

statistics have inspired new investigations regarding which aspects of a scene our cognitive system exploits to efficiently process objects and scenes (Greene, 2013; Vö et al., 2019). Notably, we can now efficiently compute an object-based frequency measure based on these image datasets. This OF measure uses the same logic as word-based frequency: counting the number of occurrences of a labeled object in a given image dataset.

It is important to note that current research on WF measures suggests that corpora should include at least 20 million words (Brysbaert et al., 2011) in order to yield a reliable frequency estimate. We cannot expect such a high number of objects for the currently available annotated image datasets, and we should consider that - as is the case even with well-established text corpora - every measure computed from a dataset represents only an approximation of real-world properties. In the specific case of real-world image datasets, biases could arise not just from the limited number, but also from limited variety of scene categories, limited points of view of photographs, artificiality of image composition, lack of clutter, etc. Nevertheless, there have been some successful attempts to use measures from existing image datasets to model neural response to object recognition (e.g., from ADE20K; Bonner & Epstein, 2021; Bracci et al., 2021). Thus, in this study, we explore the potential of these newly computed object-based frequency measures on capturing aspects of visual recognition behavior and compare them to well-established word-based frequency measures. To do so, and to limit biases from specific datasets, we employed not only one, but two measures of OF computed from two datasets that differ in size and quality of annotations (Greene, 2013; Zhou et al., 2019), as well as two measures of WF from datasets that differ in the source of the linguistic input (Brysbaert et al., 2011; Heister et al., 2011). The effect of these measures on accessing meaning during visual recognition was assessed in three experiments.

The first experiment used a semantic categorization task in which participants had to decide whether a concept, presented via an object image or via a written word, was natural or

artificial (i.e., man-made). During the procedure, we recorded response times and error rates from participants. The response time data allowed us to investigate whether word-based or object-based frequency measures modulated the speed of semantic access. We expected to replicate the WF effect for words. Besides, we wanted to test whether a WF effect on object recognition would emerge without an explicit linguistic response. Importantly, for the first time we explored possible effects of newly developed OF measures on both object and word recognition behavior.

Observing an OF effect only in object recognition and a WF effect only for words would indicate that recognizing and learning visual stimuli (words vs. objects) occurs separately within each modality (e.g., by means of a verbal vs pictorial representation). Alternatively, if one frequency measure would affect both modalities alike (e.g., WF affecting word and object recognition), this finding would indicate that a frequency measure is not just a proxy for the repeated experience with a modality-specific stimulus (e.g., a word) but for the repeated experience with the semantic representation connected to that stimulus (i.e., its meaning). Therefore, the strength of the semantic representation given by the repeated experience would also be present when that semantic representation is accessed from a different modality (e.g., a picture). This scenario is in line with the idea that semantic representations are shaped by different kinds of experiences: perceptual, motor, affective, but also linguistic. In this view, for example, language is not just a means of representing and communicating conceptual knowledge but has a transformative power on this knowledge as well (Lupyan & Lewis, 2019). These transformations derived from modality-specific experience then generalize to other modalities.

In the second experiment, the same participants completed a priming task in which they had to decide whether the meaning of the prime and target stimuli matched. By implementing either Uni-modal or Cross-modal Priming, we were able to modulate the degree of semantic

processing in the task and examine how frequency effects change as a function of the varying semantic demands. Uni-modal Priming (Scarborough et al., 1977) occurs solely on the perceptual level, as matching prime and target pairs not only have the same meaning but are also identical in their visual appearance (i.e., word primes word or object primes object). Thus, we expected a lower involvement of semantic processing. In contrast, Cross-modal Priming (Tversky, 1969) necessarily requires semantic processing because participants must relate two visually distinct stimuli to one meaning (i.e., object priming word or vice versa) to solve the task. If the effects were most substantial in Cross-modal Priming, this would provide further evidence that the frequency effects reflect an aspect of semantic rather than merely perceptual processing. The same participants of Experiment 1 and 2 also performed a rating study from which we have extracted stimulus-specific measures that we have used as covariates in the analysis.

To avoid potential carry-over effects from Experiment 1 to 2 when testing the same participants, we conducted a third experiment which included two new sets of participants — one performing only the Cross-modal and another performing only the Uni-modal Priming trials. This additionally reduced the number of concept repetitions per person. We again hypothesized that if frequency effects reflect processing of semantic representation rather than only perceptual representation, stronger frequency effects should emerge in the group exposed to Cross-modal Priming rather than Uni-modal Priming. Finally, further sets of ratings were collected from a new group of participants different from the ones of Experiments 1,2, and 3, again with the idea of extracting covariate measures to use during the analysis.

Materials and Methods

Participants

We required all participants taking part in our study to have normal or corrected-to-normal vision, be German native speakers, and have no history of linguistic, psychiatric, or neurological disorders. Additionally, we only included participants who did not report having technical problems during the online procedures and who completed both sessions. Participants were recruited by sharing the link to the studies on through platforms and mailing lists of students at the Goethe University of Frankfurt.

To prevent an overestimation of underpowered correlations, which may be expected when N is below 30 participants (e.g., see Yarkoni, 2009), we tested 60 participants (of whom 42 fit the above-mentioned criteria) in Experiments 1 and 2, as well as the rating study judging typicality and familiarity of the used stimuli (age: $M = 23.55$, $SD = 8.88$, Range: 15-59 y.; Gender: 34 F, 7 M, one person did not report; 5 bi/multilingual with German as one of the native languages).

For the replication in Experiment 3, we recruited 53 additional participants for the Cross-modal Priming task (age: $M = 22.87$, $SD = 6.83$, Range: 18-50 y.; Gender: 43 F, 10 M; 10 bi/multilingual with German as one of the native languages), and yet another 53 participants took part in the Uni-modal Priming task (age: $M = 22.66$, $SD = 4.81$, Range: 18-39 y.; Gender: 35 F, 16 M, 2 NB; 13 bi/multilingual with German as one of the native languages). The sample size for the replication ($N = 53 + 53 = 106$) was obtained by taking the sample size in Experiment 2 ($N=42$), which had a within-participant design, and adapting it to a betweenparticipants design in the replication, following this formula: $N_{between} = (N_{within} * 2) / (1 - \rho)$, where 2 represents the number of groups / conditions (in our case: Cross-modal and Uni-modal Priming) and ρ represents the correlation between the two groups / conditions

(in our case, from Exp 2, $\rho = 0.208$). The formula was then solved for $N_{between} = (42 * 2) / (1 - 0.208) = 106.061$ (Maxwell et al., 2017).

Additionally, two distinct groups of participants were recruited to collect further ratings regarding the stimuli used: (1) One group of 20 participants (age: $M = 21.65$, $SD = 2.66$, Range: 19-29 y.; Gender: 10 F, 10 M; 4 bilinguals/multilinguals with German as one of the native languages) performed a rating study judging typicality and familiarity of the stimuli. This further set of ratings for typicality and familiarity were collected anew as part of the replication. (2) A second group of 16 participants performed a rating study judging the “Conceptual distinctiveness” (as in Konkle et al., 2010) of the concept used in the studies (age: $M = 23$, $SD = 4.75$, Range: 18-33 y.; Gender: 9 F, 7 M).

All participants gave their informed consent and received course credits or monetary compensation for their participation. The Ethics Committee of the Goethe University Frankfurt approved all experimental procedures (approval # 2014-106).

Stimuli

For this study, we selected 100 noun concepts that can be depicted by a single word and an image of an object in isolation. We use the phrase “object concept” here and below, to refer to the semantic representation common to a word denoting an object (e.g., “apple”) and the object itself (e.g., a physical apple, an image of it). Half of the concepts could be categorized as natural (e.g., apple) and the other half as man-made (e.g., bicycle). We restricted our search to objects with word labels in the ADE20K dataset, a set of real-world images of scenes with segmented and annotated objects (Zhou et al., 2019). After selection, we translated the English word labels to German. For presentation, we displayed German nouns with an uppercase initial-letter (i.e., correct spelling in German) and in white Arial font on a grey background (hexadecimal color

#424242; jsPsych, de Leeuw, 2015). We downloaded the object images from internet databases (e.g., <https://pnghunter.com/>, <http://pngimg.com/>, <https://www.cleanpng.com/>). They were pasted on a white background, grey-scaled, and resized to 392 x 392 pixels.

Object and word characteristics

For all concepts, we computed four selected frequency measures (two word-based and two object-based). In addition, we computed several stimulus characteristics identified to influence recognition behavior (i.e., to consider as covariates in the statistical analysis).

Object-based frequency measures (object frequency - OF). OF measures represent the logtransformed (base 10) number of occurrences of an object in a dataset of segmented and labeled scene images (e.g., cars on the street). Implementing the log-transformation for frequency measures reduces the skewness of the frequency distribution as typically only few objects have high frequency, while majority of objects have a low frequency (Zipf's-law-like distribution; Greene, 2013). We determined the OF based on two datasets. One used more than 20,000 scene images (from 900 categories), and objects (more than 400,000 instances grouped in more than 2,500 categories) were segmented and labeled by a single expert worker and used to train an image recognition algorithm to identify objects in scenes (*ADE20K OF*; Zhou et al., 2019). Since we based our stimulus selection on objects present in the ADE20K dataset, we tried to represent all the different levels of frequency we could find there (i.e., from few appearances to tens of thousands of appearances). The second dataset used 3,499 scene images (from 16 categories; indoors, outdoors, natural, artificial), labeled by four different workers and carefully cleaned of misspellings, synonyms, and other errors, to measure statistical regularities of objects in a scene (more than 48,000 instances grouped in more than 800 object categories; *Greene OF*; Greene, 2013). Only 78 of our 100 object labels selected from ADE20K

were present in the Greene dataset. When an object was missing in the Greene dataset, we assigned an OF value of 1 count (i.e., 0 log₁₀-counts). Density distribution of ADE20K OF and Greene OF for the set of stimuli can be found in *Supplementary Materials 1*.

Word-based frequency measures (word frequency - WF). WF measures are based on the number of occurrences of a word in a corpus of linguistic materials. Specifically, as for object frequency, the numeric parameter was computed as the logarithm (base 10) of the number of occurrences per million words in a dataset (to turn the Zipf's-law-like distribution into a normal distribution, Li, 1992). When a word was not included in a corpus, which was the case for one concept, the WF was set to 1 count per million (i.e., 0 log₁₀-counts per million). The WF was determined based on two corpora, one using German subtitles from films and tv-shows, *SUBTLEX-DE WF* (Brysbaert et al., 2011) and the other including a large set of German written material, such as books and newspapers, *dlexDB WF* (Heister et al., 2011). The density distributions of SUBTLEX WF and dlexDB WF for the set of stimuli can be found in *Supplementary Materials 1*.

Covariates. In order to estimate and control for the contribution of other variables, we collected subjective ratings from participants, as well as we computed object- and word-specific visual predictors.

Ratings: As part of the replication, we obtained two sets of concept familiarity and image typicality ratings: one from the participants who have taken part in the original study (i.e., Experiments 1 and 2) and one from a different group of participants who had not previously taken part in any of the experiments. We measured *concept familiarity* as the subjective familiarity with an object concept to serve as a subjective counterpart of the objective frequency measures of words and objects computed from a text or image dataset (see Kuperman & van

Dyke, 2013). *Typicality*, on the other hand, represents how an object exemplar is typical of its category. In the original study, individual ratings for each concept and each participant were used to model each participant's performance on each concept in the main tasks. In contrast, in the replication experiment, ratings were averaged across participants and used to model performance on each concept since participants of the rating study differed from those of the original study.

To substantiate the interpretation of some of our results, it became important to further investigate the relationship between OF and *Conceptual Distinctiveness* (Konkle et al., 2010). For this purpose, we set up yet another rating study in which we collected ratings regarding our stimuli's Conceptual Distinctiveness from participants who had not taken part in any of the previous experiments. The rating study followed the methodology described in Konkle et al. (2010). They defined a concept as having a high Conceptual Distinctiveness if it is relatively easy to make subdivisions among the category members it denotes and where these subdivisions are not simply based on perceptual features (e.g., color or shape). Conceptual Distinctiveness ratings were obtained for every concept by averaging ratings across participants.

Visual and visuo-orthographic predictors: In addition to the subjective ratings, we computed and included various object- and word-specific measures from which we extracted visual and visuo-orthographic predictors using a Principal Component Analysis (PCA).

To assess the *visual characteristics of object images*, we computed several measures based on pixel-level input: *Entropy* (Shannon, 1948), which measures the level of “disorder” and visual variance of an image (entropy equals zero means no variance); *Signal-to-noise ratio (SNR) of pixel values* (computed as mean of all pixel values divided by the standard deviation of all pixel values), which we used as a proxy of how the content of the image differs from the background (larger negative values indicate that the content is closer to the background, values

close to zero indicates that the content is more different than the background); *graphic-based visual saliency* (Harel et al. 2007), which measures saliency of the image based on bottom-up features (every pixel has a value between 0 and 1, where zero indicates not salient and a value of one indicates high saliency); *GIST descriptor* (Oliva & Torralba, 2001), which gives us the orientation and spatial frequency in different parts of the image; finally, *Deep Convolutional Neural Network activation from convolutional layer 1, 4 and fully-connected layer 7 of the AlexNet model* (Krizhevsky et al., 2012); they represent low-level (layer 1), mid-level (layer 4) and high-level (layer 7) visual features of our images, as processed by a deep learning algorithm trained to perform human-like object categorization. From a PCA on these visual predictors, we extracted 3 orthogonal principal components (PCs) that we named *Image visual PC1*, *Image visual PC2* and *Image visual PC3* (for more info on their impact and interpretations, see *Supplementary Materials 1*)

To assess the *visual and orthographic characteristics of words*, we performed another PCA. For this, we considered two visual properties, *entropy* and *SNR*, computed as described above for object images but now applied to the images of written words. In addition, we computed two orthographic measures, *word length* (i.e., the number of letters) and *distance from orthographic neighbors* (i.e., Orthographic Levenshtein Distance, Yarkoni, et al., 2008). One PC was selected from this process and was labeled *Visuo-orthographic PC*. Correlations between all predictors and between PCs and original measures, as well as PCA loadings, can be found in *Supplementary Materials 1*.

Apparatus

Participants performed the experiments online, hosted on a web server at the Goethe University Frankfurt. We used jsPsych (de Leeuw, 2015) for stimulus presentation and response recording. Participants were instructed to ensure that they started the experiments only when seated in a

quiet environment without potential interruptions and when they had enough time to dedicate to it. Besides, they were instructed to perform the experiment only on laptops or desktop computers. To account for differences in screen size and resolution, we implemented an adaptation mechanism based on the measurement of a credit card (<https://www.jspsych.org/plugins/jspsych-resize/>).

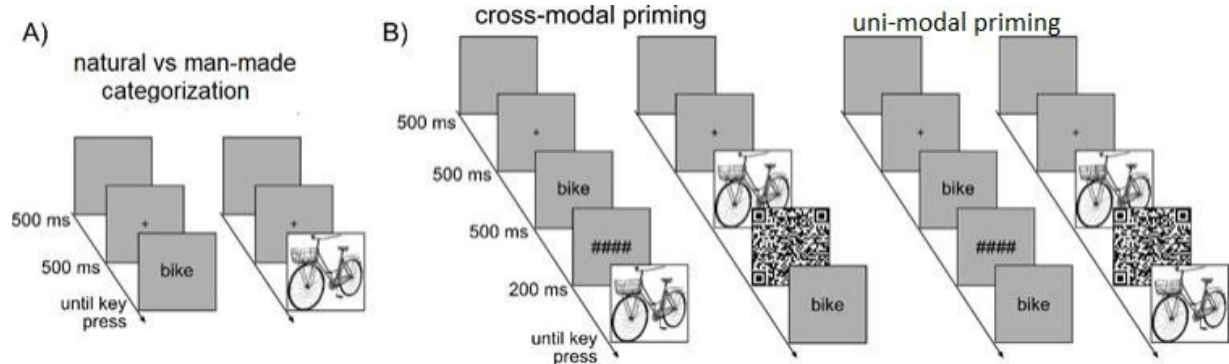
Before the experiments started, participants had to adapt a rectangle presented in the center of the screen to the size of a credit card. This information was used to ensure that the size of stimuli on screen was the same for every participant (object images: 6.7 x 6.7 cm; words, uppercase letter: circa 0.7 cm). In all parts of the experiment, the screen background was grey (hexadecimal color #424242). The Conceptual Distinctiveness rating experiment was programmed in Python using PsychoPy (version 2020.2, Builder GUI; Peirce et al., 2019) and administered online through the hosting platform Pavlovia (<https://pavlovia.org/>). Stimulus words were presented in black Arial text of 1.5 cm vertical size on white background.

Procedure

Experiment 1. *Figure 1A* shows an example of the natural (e.g., apple) vs. man-made (e.g., bicycle) categorization task of Experiment 1. The two stimulus modalities were presented in two separated blocks (100 stimuli each). Block order was randomized across participants, and within each block, the stimulus order was randomized for each participant. The stimulus presentation sequence started with a fixation cross at the screen center (500 ms) followed by the presentation of an object image/word. After the participants responded, the presentation was terminated. We asked participants to press a key as quickly and as accurately as possible: *j* when a “natural” stimulus was presented and *f* when a “man-made” stimulus was presented. A blank screen was presented for 500 ms between two trials (*Figure 1A*). After each block, we asked the participants to take a break.

Figure 1. Experimental design

A) Experiment 1. Categorization of natural vs. man-made object images and words. B) Experiment 2. Categorization of prime-target matches vs. mismatches. Cross-modal Priming: words are primed with objects, and objects are primed with words. Uni-modal Priming: words are primed with words and objects with objects.



Experiment 2. In the second experiment, we implemented a priming task that included Uni-modal and Cross-modal prime-target pairs, consisting of object images and words. Participants evaluated if both the prime and the target had the same meaning or not. They started with two Cross-modal Priming blocks (i.e., word-priming-object, object-priming-word; see *Figure 1B*). After that, participants completed two Uni-modal Priming blocks (word-priming-word, object priming-object). Within Cross-modal and Uni-modal blocks, we randomized block order across participants. We presented all 100 object concepts twice as a target within each block (200 trials) in a randomized order. Every target was once paired with a matching and once with a mismatching prime stimulus. Mismatching pairs were randomly generated and kept constant for all blocks of each participant. We instructed the participants to evaluate whether the target and prime concepts matched or mismatched. Again, they should indicate this by pressing a key (*j* for match and *f* for mismatch) as quickly and as accurately as possible. Like in Experiment 1, trials started with a fixation cross presented in the screen center for 500 ms. After that, the prime was presented for 500 ms followed by a backward mask for 200 ms (“#####”) for words or QRcode-like for objects; see *Figure 1B*). The presentation of the target was

terminated by the response of the participant. Again, we asked participants to take a break in between blocks and one break halfway through every block.

Typicality and familiarity ratings. Finally, we asked participants to perform an additional session the following day to collect demographic data and stimulus ratings. This procedure was again performed online. Participants rated all stimuli on a one to six Likert scale. We assessed *concept familiarity* by presenting the concept as a written word in the screen center. In addition, we presented the question “*How familiar are you with the object that the word represents, in your everyday life?*” plus the Likert scale. *Image typicality* was assessed, presenting the object picture in the center of the screen, and the object word on top. In addition, we presented the question, “*How typical is this image in relation to the category designated by the word?*” with the Likert scale.

In total, data collection lasted for about 75 minutes on day 1 (Experiment 1 and Experiment 2) and about 30 minutes on day 2 (ratings).

Experiment 3. Experiment 3 was run to replicate the findings of Experiment 2. It therefore has the same structure of Experiment 2, except that two separate groups of new participants either performed only the two Cross-modal Priming blocks or only the two Uni-modal Priming blocks. Data collection lasted about 30 mins each.

Replication typicality and familiarity ratings. This procedure resembled that of the original typicality and familiarity rating task, with the exception of having two blocks for *concept familiarity*, one with words (“*How familiar are you with the object that the word represents, in your everyday life?*”) and one with pictures (“*How familiar are you with the object that the picture represents, in your everyday life?*”), presented in counterbalanced order across participants. For the analysis, we aggregated familiarity ratings for words and objects on

concept level within each participant before averaging across participants. *Image typicality* ratings were aggregated for each concept averaging across participants. Data collection lasted about 30 minutes.

Conceptual Distinctiveness ratings. Finally, we performed a new rating study that was aimed at measuring *Conceptual Distinctiveness (CD)* as it was defined in Konkle et al. (2010). We first carefully instructed participants on the definition of CD as it was done in Konkle et al. (2010), and by presenting a set of example objects rated either as being low on Conceptual Distinctiveness or high in the original investigation. By definition, concepts with high Conceptual Distinctiveness are those whose category members can be easily divided into subgroups of different kinds, regardless of visual appearance. After this introduction, each trial presented a word from our stimulus set in the center of the screen. In addition, the question “*How distinctive are the members of the category denoted by this word?*” was presented along with a six-point scale spanning from one (very similar) to six (very distinctive). Participants responded by clicking with the mouse on a circle corresponding to the number representing their rating. Once they clicked, they saw a black fixation cross in the screen center for about 500 ms before the next word was presented. In total, participants rated all 100 object concepts. We presented the words in randomized order, and participants could take as long as they wanted to make their judgment. Data collection lasted about 15 minutes.

Analysis

Data analysis was performed using R (version 3.6.3, R Core Team, 2020). First, we excluded response times smaller than 200 ms and larger than 1500 ms from further analysis. We set a lower cut-off for excluding response times at 200 ms as typically faster response times are highly likely so-called “fast guesses” (Luce, 1986; Whelan, 2008). Since we had instructed

participants to perform the task as quickly and accurately as possible, we assumed that a cutoff at 1500 ms would prevent the inclusion of response times that did not fit this criterion. Our exclusion criteria led to the removal of only 2.7 % of collected RTs in Experiment 1 and of 1 % of collected RTs in Experiment 2 (1.4 % of the total considering the two experiments together); in Experiment 3, 2.0 % of RTs collected were removed. We implemented a logtransformation to obtain a normal distribution to account for the ex-Gauss distribution of reaction time measures. No further pre-processing was administered.

We used linear mixed-effects models (LMMs; Bates et al., 2014) for statistical analyses of log-transformed response times. Independent variables considered in the models were the four frequency measures described above (object frequencies based on the ADE20K and Greene datasets, word frequencies based on SUBTLEX and dlexDB corpora), several continuous covariates and categorical predictors for the experimental conditions (see *Supplementary Materials 1*). The main advantage of LMMs is that one can consider each trial from each participant simultaneously (i.e., estimating crossed random effects of items and participants; Baayen et al., 2008). In all our LMMs, we included intercept-only random effects for participants and object/word meanings. Note that by including random slope estimates the models did not converge, so we followed the recommendations of Bates et al. (2015).

Our analysis was divided into three steps (more details in *Supplementary Materials 1*):

- 1) First, we implemented a model comparison based on the Akaike Information Criterion (AIC, Akaike, 1981). This step allowed us to compare our four frequency measures and select the frequency measures with the best fit in both modalities. To implement this, we first fit one model per frequency measure (i.e., SUBTLEX, dlexDB, ADE20K, and Greene frequency) separately for the word and the object recognition trials, and then compared the four models of each modality to a “baseline” model that did not include the frequency measure, but that was estimated on the same subset of data. We selected the frequency measures following

these criteria: in the best case, we would have selected two measures, i.e., the best fitting OF and the best fitting WF measure. In the worst-case, none of the frequency measures would have explained variance in both object and word trials. While, in between, we would have selected either only an OF or a WF measure.

2) After selecting the best frequency measures, we ran a LMM estimating the effects of those selected frequencies on the entire dataset (word trials + object trials), and including all categorical factors and continuous covariates, as well as random factors for participants and concepts.

3) When we detected significant interactions between frequency measures and categorical predictors, we also ran post-hoc LMMs to understand the different effects of frequency *between different conditions* (e.g., SUBTLEX in Cross-modal trials vs. SUBTLEX in Uni-modal trials) and *within each condition* (e.g., the simple effect of SUBTLEX in Crossmodal trials and simple effect of SUBTLEX in Uni-modal trials). Note that the estimation of frequency effects, given the structure of linear models, was independent (i.e., controlled for) from the effect of the other predictors/covariates included in the models.

Data, analysis scripts and stimulus materials are all available at the following link: <https://osf.io/d3j9h/files/>; for more details, see *Supplementary Materials 1*.

Results

Results Experiment 1

The initial model comparison showed that, in the man-made vs natural categorization task, for word recognition trials, only the SUBTLEX and dlexDB measures produced a significantly better fit when included in the models (SUBTLEX WF: $\chi^2=29.153$, $p<0.001$; dlexDB WF:

$\chi^2=15.447, p=0.001$), while considering OF measures did not produce a better fit (ADE20K OF: $\chi^2=3.228, p=0.072$; Greene OF: $\chi^2=0.867, p=0.352$). For object recognition trials, only SUBTLEX WF resulted in a significant improvement of the model fit ($\chi^2=6.163, p=0.013$; dlexDB WF: $\chi^2=1.646, p=0.200$; ADE20K OF: $\chi^2=0.051, p=0.821$; Greene OF: $\chi^2=0.310, p=0.578$; for more details, see *Supplementary Materials 2*; no multicollinearity was detected: variance inflation factors < 5). The result of this initial model comparison showed that the SUBTLEX measure was the best fitting parameter in both word and object trials, with no significant increase in explained variance for any of the two object-based predictors. Thus, we implemented a detailed investigation of the SUBTLEX WF effect with both word and object datasets merged.

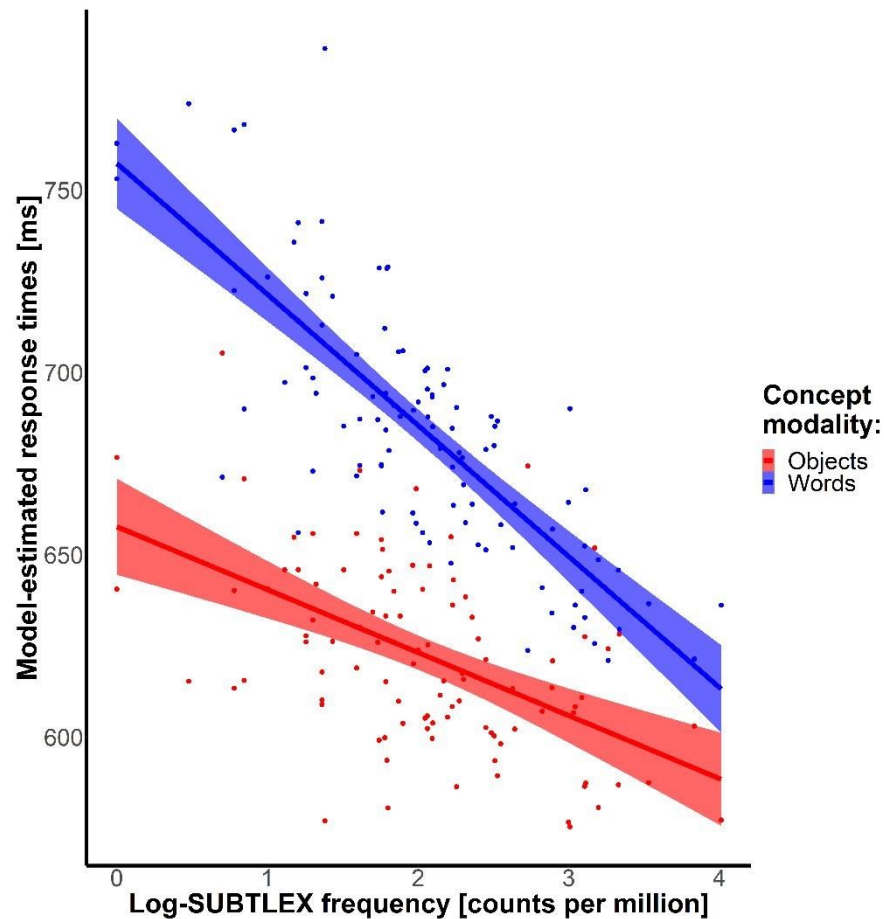
The LMM describing all response times together included a SUBTLEX WF by Concept modality (i.e., words vs. objects) interaction and nine further covariates (see *Supplementary Materials 3* for R-based formula; no multicollinearity detected: variance inflation factors < 5). We found a significant SUBTLEX WF by Concept modality interaction ($\beta=-0.019, SE=0.005, t=-4.160, p<0.001$), showing a more substantial facilitatory SUBTLEX WF effect (i.e., faster RTs for high frequency items) for words compared to objects (see Figure 2; for details see *Supplementary Materials 3*).

Two post-hoc models, for objects and words separately, showed significant SUBTLEX WF effects for both words ($\beta=-0.041, SE=0.007, t=-5.794, p<0.001$) and objects ($\beta=-0.022, SE=0.009, t=-2.524, p=0.012$), but the effect size for words was almost double (1.86 times higher; for more details, see *Supplementary Materials 4 and 5*).

Figure 2. Main results of Experiment 1.

Semantic categorization response times as a function of logarithmic SUBTLEX frequency, separated for objects and words; RTs were estimated based on the SUBTLEX WF x Concept modality interaction term from the selected model. Points present participant-based mean reaction times separated for stimulus type (red: object stimuli; blue:

word stimuli) in the different frequency levels. Lines represent linear fitting of points, and shaded areas represent 95 % confidence interval.



Discussion Experiment 1

The first experiment replicated the well-established SUBTLEX WF effect in word recognition (Brysbaert et al., 2011; Gagl et al., 2020). In contrast to previous literature (Taikh et al. 2015), we also found a SUBTLEX frequency effect for object recognition performance, although the effect for object recognition was weaker than for word recognition. However, all together, findings from this experiment suggest that - given that WF has an effect on both object and word recognition - this effect might reflect processing of what word and object recognition have in common, i.e., the same semantic representation being accessed from two different visual inputs. The phenomenon producing the WF effect during language experience may not just be based on the strengthening of modality-specific representations (WF effect for words),

but also the strengthening of domain-general semantic representation (WF effect also for objects). Interestingly, neither OF measure improved the fit. Thus, OF seems to be less relevant in this simple categorization task.

In Experiment 2, we implemented a priming task to investigate the effect of the novel object-based frequency measures in a paradigm where context is given by a prime allowing prediction of an upcoming visual stimulus. Additionally, we wanted to test the role of WF effects during semantic processing of visual stimuli. The critical manipulation therefore contrasted Cross-modal and Uni-modal Priming (Tversky, 1969; Scarborough et al., 1977; Eisenhauer et al., 2019; 2021). As described earlier, Cross-modal Priming does not involve perceptual processing but rather conceptual/semantic information transfer from prime to target processing. Thus, frequency effects in Cross-modal Priming would signify an involvement of these effects with semantic rather than perceptual processing.

Results Experiment 2

First, we again implemented a model comparison procedure to determine which frequency measure should be part of a detailed analysis. Here, we found that all four frequency measures improved model fit in both stimulus modalities (words and objects; ADE20K OF, objects: $\chi^2=10.105$, $p=0.039$, words: $\chi^2=27.302$, $p<0.001$; Greene OF, objects: $\chi^2=27.547$, $p<0.001$, words: $\chi^2=43.409$, $p<0.001$; SUBTLEX WF, objects: $\chi^2=52.695$, $p<0.001$, words: $\chi^2=43.409$, $p<0.001$; dlexDB WF, objects: $\chi^2=33.014$, $p<0.001$, words: $\chi^2=19.105$, $p<0.001$; for detailed information, see *Supplementary Materials 6*). We found that both Greene and SUBTLEX frequencies had stronger fit improvements than their alternatives in both stimulus modalities. Thus, we selected the Greene and SUBTLEX frequency measures for further investigation.

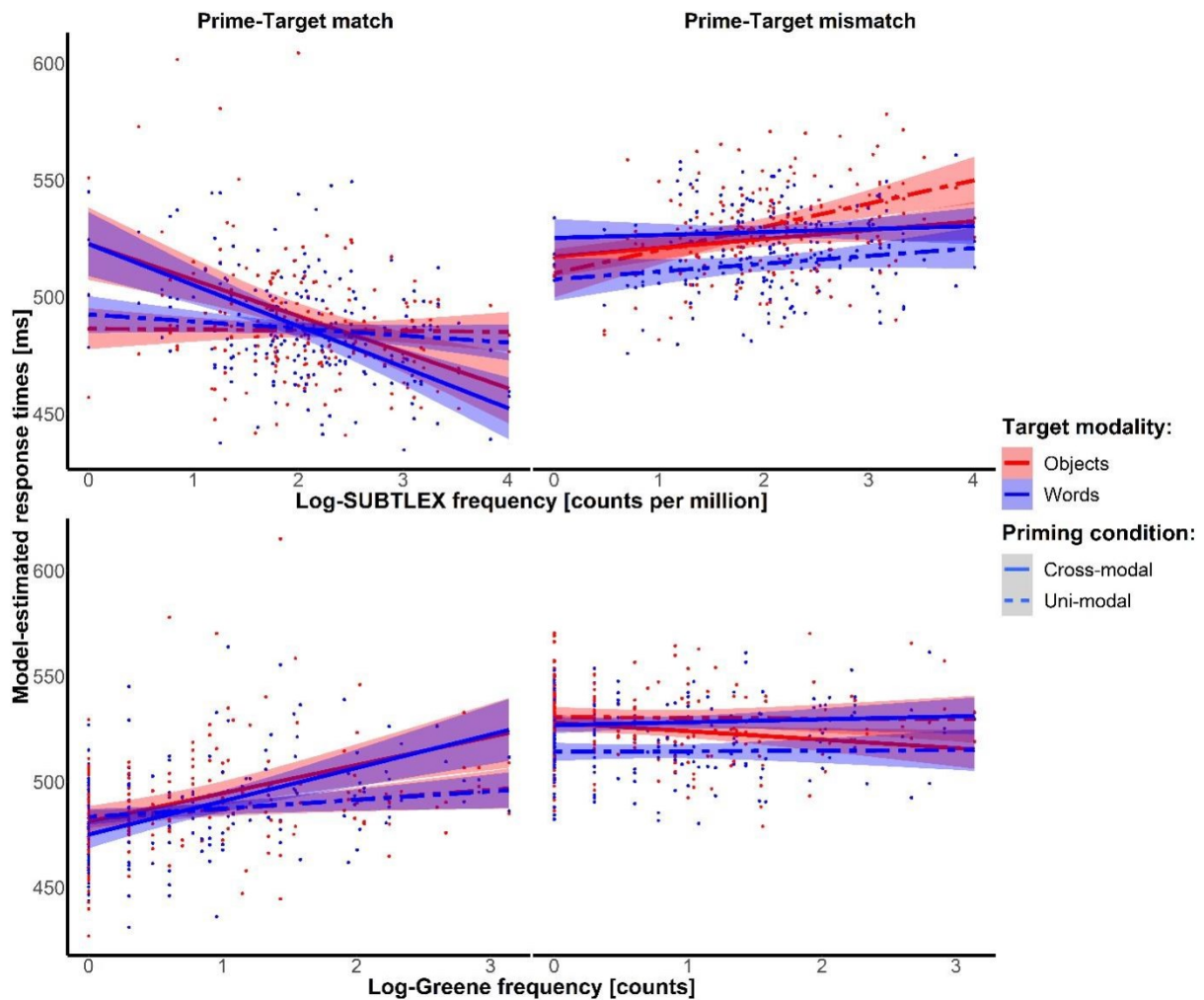
We entered the two measures into a single model, including covariates, categorical predictors and random effects, to describe the response times from the entire dataset of

Experiment 2. Further model comparisons indicated that the interaction between SUBTLEX WF and Greene OF did not improve the model fit beyond the simpler model without the interaction ($\chi^2=5.455$, $p=0.708$). So, the selected model included each of the two frequency measures in interaction with the experimental conditions (Priming condition: Cross-modal vs Uni-modal; Matching condition: Mismatching vs. Matching; Target modality: Words vs Objects) separately, but not in interaction with each other (for the model formula and other details, see *Supplementary Materials 7*).

When participants had to judge whether prime and target had the same meaning, we found a significant 3-way interaction between frequency, Matching condition, and Priming condition, for both SUBTLEX WF ($\beta=0.017$, $SE=0.005$, $t=3.687$, $p<0.001$; *Figure 3 top*) and Greene OF measures ($\beta=-0.020$, $SE=0.005$, $t=-4.256$, $p<0.001$; *Figure 3 bottom*, for more detailed information see *Supplementary Materials 7*). Importantly, we found that these interactions had opposite effects for Greene OF and for SUBTLEX WF. However, we found no evidence for Target modality effects, i.e., WF and OF effects in Matching and Priming conditions were similar for words and objects (SUBTLEX WF: $\beta=0.006$, $SE=0.009$, $t=0.612$, $p=0.541$; Greene OF: $\beta=0.002$, $SE=0.009$, $t=0.227$, $p=0.821$).

Figure 3. Main results of Experiment 2.

Response times as a function of logarithmic SUBTLEX frequency (top plots) and Greene frequency (bottom plots) in the different conditions of Experiment 2; RTs were estimated based on the selected model. Points present participant-based mean response times separated for stimulus type (red: object stimuli; blue: word stimuli) in the different frequency levels. Lines represent linear fitting of points (solid: Cross-modal; dashed: Uni-modal), and shaded areas represent 95 % confidence interval. Top-left and bottom-left plots represent the effects in primetarget matching condition, while top-right and bottom-right plots represent the effects in prime-target mismatching condition.



Post-hoc models showed that the frequency effects were stronger in *Crossmodal Matching* trials than in *Uni-modal Matching* trials (SUBTLEX WF: $\beta=-0.023$, $SE=0.003$, $t=-7.094$, $p<0.001$; Greene OF: $\beta=0.018$, $SE=0.003$, $t=5.379$, $p<0.001$), while, no differential effects were found between *Cross-modal Mismatching* and *Uni-modal Mismatching* trials (SUBTLEX WF: $\beta=-0.006$, $SE=0.003$, $t=-1.870$, $p=0.062$; Greene OF: $\beta=0.002$, $SE=0.003$, $t=-0.644$, $p=0.520$). Besides, we only found strongly significant effects of SUBTLEX frequency ($\beta=-0.019$, $SE=0.006$, $t=-3.230$, $p=0.001$) and Greene frequency ($\beta=0.023$, $SE=0.005$, $t=4.710$, $p<0.001$) in *Cross-modal Matching* trials. The WF and OF effects went in opposite directions: while we observed faster responses for more frequent concepts when investigating the SUBTLEX WF, the Greene OF effect was characterized by faster response for more rare

concepts (*see Supplementary Materials 8 and 9*). In a further control analysis, we showed a substantial stability of the effects for the individual participants and individual concepts across the two modalities (for details, *see Supplementary Materials 10*) which again suggests non-different (i.e., statistically equivalent) processes across modalities.

Discussion Experiment 2

In Experiment 2, we replicated the facilitatory effect of the SUBTLEX WF found in Experiment 1 for both words and objects. It is important to note that we found the SUBTLEX WF effect only when participants categorized objects or words after seeing a semantically matched prime from the other stimulus modality (e.g., a bike image primed by the word “bike” and vice versa), a condition that requires the integration of semantic information from the prime in preparation for the target. The Cross-modal condition specifically includes a prediction process from one modality to the other: it requires processing both object exemplars and their verbal labels within one trial.

A novel aspect that became evident in Experiment 2 was that we also found an effect of the Greene OF in the Cross-modal Matching trials. However, the effect went in the opposite direction, i.e., better performance for low-frequency object concepts than for high-frequency concepts. Both frequency effects were stable across modalities when investigated within each participant and each concept, as shown by our exploratory analysis. Regarding the presence of these two opposite frequency effects, it is worth noting that the model that included an interaction between Greene and SUBTLEX frequency measures did not increase the model fit, implying that the two effects might represent distinct, independent processes. Also, note that the match/mismatch task of Experiment 2 did not result in a global processing advantage for objects compared to words. This was only found in Experiment 1 and replicated previous studies showing the same effect (e.g., Taikh et al., 2015). We believe that since our task in

Experiment 2 was only concerned with the prime-target matching, it resulted in this task being equally difficult for object and word target stimuli.

At this point, one might wonder why the OF improved the fit (and showed significant effect) only in the priming task of Experiment 2 (to be precise, only in Cross-modal Matching trials), and not in the semantic categorization of Experiment 1 (man-made vs. natural). Unfortunately, it is difficult to offer an easy explanation for this unexpected result. It seems that the Greene OF has an effect only when a semantic representation (i.e., concept) is part of a process to predict upcoming input. This process is not part of Uni-modal Priming and unprimed categorization (Experiment 1), where the task does not demand semantic processing (Uni-modal Priming) and it does not use semantic representations to make predictions (Experiment 1). We will now try to explain the WF and OF frequency effects and why both effects occur specifically in the Cross-modal Matching condition, which is important given the high involvement of semantic processing and the predictability of the upcoming stimulus.

The SUBTLEX WF effect is in line with results of Experiment 1 and with typically reported WF effects, reflecting how often a concept has been processed during receptive language processing. One could interpret this effect to reflect the *strength of linguistic experience* with a concept (Brysbaert et al., 2011), based on repeated experiences with that concept during regular language use. It is important to note that this frequency measure is only mildly correlated with the subjective familiarity we additionally collected via ratings, which did not show any relevant impact on reaction times either here or in Experiment 1. This would suggest that there might be a dissociation between what people experience, and therefore rate as being familiar, and how often objects truly occur in the world (Greene, 2016).

In contrast, the Greene OF effect emerged in the opposite direction, i.e., showing facilitation for concepts encountered less often in our visual world. It seems counterintuitive that fewer occurrences could strengthen mental representations, but we can speculate on two

interpretations to explain this effect that has been found for both words and object targets in our study. One possible explanation is that one can remember a concept better when presented with fewer exemplars of that category because more frequent encounters with variable exemplars create interference that weakens the memory trace (Konkle et al., 2010). Based on these findings, we could infer that the facilitation found for low Greene OF concepts (e.g., pineapple) could be due to reduced interference from fewer encounters with exemplars of that object concept during the visual perceptual experience. In contrast, more frequently encountered object categories (e.g., tree) might produce a weaker representation due to exposure to more exemplars creating the abovementioned interference.

Alternatively, the OF effect, which is only detected in congruent Cross-modal Priming, could be explained based on the predictability of the stimulus features from conceptual representations. Objects that are less frequent in the databases might be the expression of more narrow categories (less exemplars and more homogeneous), and their features would be well predictable in contrast to concepts from more broad and thus frequent categories (more exemplars and more heterogeneous). This explanation also relates to theories more deeply concerned with the neuronal preparation for highly predicted incoming stimuli, like predictive coding theories (Rao & Ballard, 1999) or sharpening (Kok et al. 2012; 2017). Evidence from similar experiments using words (Eisenhauer et al., 2019, 2021; Gagl et al., 2020), objects (Summerfield et al., 2008; Richter et al., 2018), faces (Olkkonen et al., 2017) or Cross-modal Priming paradigms (Kok et al., 2012; 2017) have provided findings that indicate feature-based prediction effects. To reevaluate this finding, we performed additional analyses on the data from Experiment 2 and collected a replication dataset in Experiment 3 to shed more light on the explanation of this initially counterintuitive effect and how it could relate to the interference process presented by Konkle and colleagues (2010), as well as to the categorical structure of the investigated concepts (see next section).

To sum up, these results suggest that when participants perform a task where contextual information (i.e., the prime) is semantically processed, different types of information in semantic memory (supposedly derived from linguistic and visual experience) are being preactivated to facilitate the processing of an upcoming input (i.e., the target). The present findings suggest that these processes seem to be at least partially domain-general and thus might depend less on the modality of the stimuli.

Results Conceptual Distinctiveness ratings

In Experiment 2, we unexpectedly found opposite effect of Greene OF on visual recognition, with less frequent concepts being recognized faster than more frequent ones. Having fewer encounters with an object may constitute an advantage in recognizing these compared to concepts for which we have experienced more exemplars, as higher frequency of occurrence has been shown to produce interference in long-term memory (LTM; Konkle et al., 2010).

To further explore this idea, we have collected ratings of *Conceptual Distinctiveness* adapting a procedure from Konkle et al., (2010; for more details, see the *Materials and Methods* section), which has been used to demonstrate how memory interference for objects presented in many exemplars (i.e., comparable to our high Greene frequency concepts) is reduced for objects whose category can easily be separated into many different subcategories (i.e., categories with a high Conceptual Distinctiveness; Konkle et al., 2010). We would expect that including CD in our model will reduce the Greene frequency effect for concepts with high CD, while the effect of Greene frequency would remain the same for concepts with low CD. To illustrate how this relates to the concepts we used in our experiment, see the examples provided in *Figure 4*.

First, we found that CD and Greene OF had a moderate correlation ($r=0.43$), where concepts with low Greene OF tended to also be less easily dividable in subcategories, while

concepts with high Greene OF tended to more easily dividable. Then we compared the original main LMM of Experiment 2, fitted on the data of Experiment 2, with an identical model including CD in interaction with Greene and the experimental conditions (for details, see *Supplementary Materials 11*). Despite this new model being more complex in terms of number of parameters, it showed a significantly better fit than the original model ($\chi^2=36.691$, $p=0.004$; no multicollinearity detected: variance inflation factors < 5). In the new model including CD, results showed that the *interaction between Greene OF and CD* was stronger in *Cross-modal Matching* than in *Uni-modal Matching trials* ($\beta=0.010$, $SE=0.003$, $t=-3.139$, $p=0.002$), while no difference of the *Greene OF by CD interaction* was found between *Cross-modal Mismatching* and *Uni-modal Mismatching trials* ($\beta=0.002$, $SE=0.003$, $t=0.495$, $p=0.621$). Additionally, the *Greene OF by CD interaction* was found to be stronger in *Cross-modal Matching* than in *Cross-modal Mismatching trials* ($\beta=-0.012$, $SE=0.003$, $t=-3.774$, $p<0.001$; for more details, see *Supplementary Materials 11*). As shown in *Figure 5*, in *Cross-modal Matching* trials, higher Conceptual Distinctiveness was associated with weaker Greene frequency effects (the slope reduced towards zero). *Cross-modal Matching* trials, the condition with strongest semantic processing and predictable semantic context, was also the only condition that had previously shown strong Greene OF effects and, as hypothesized based on findings by Konkle et al. (2010), the condition with the strongest modulation of Greene OF by CD.

Figure 4. Example of interaction between Greene frequency and Conceptual Distinctiveness.

An example of the hypothesized interaction, using object concepts from our stimulus set. Concepts are shown as the black and white pictures used in the experiments and the associated written words (in the English translation). Exemplar pictures to show different levels of conceptual distinctiveness were taken from the THINGS dataset (Hebart et al., 2019). These were not part of the actual experiment.

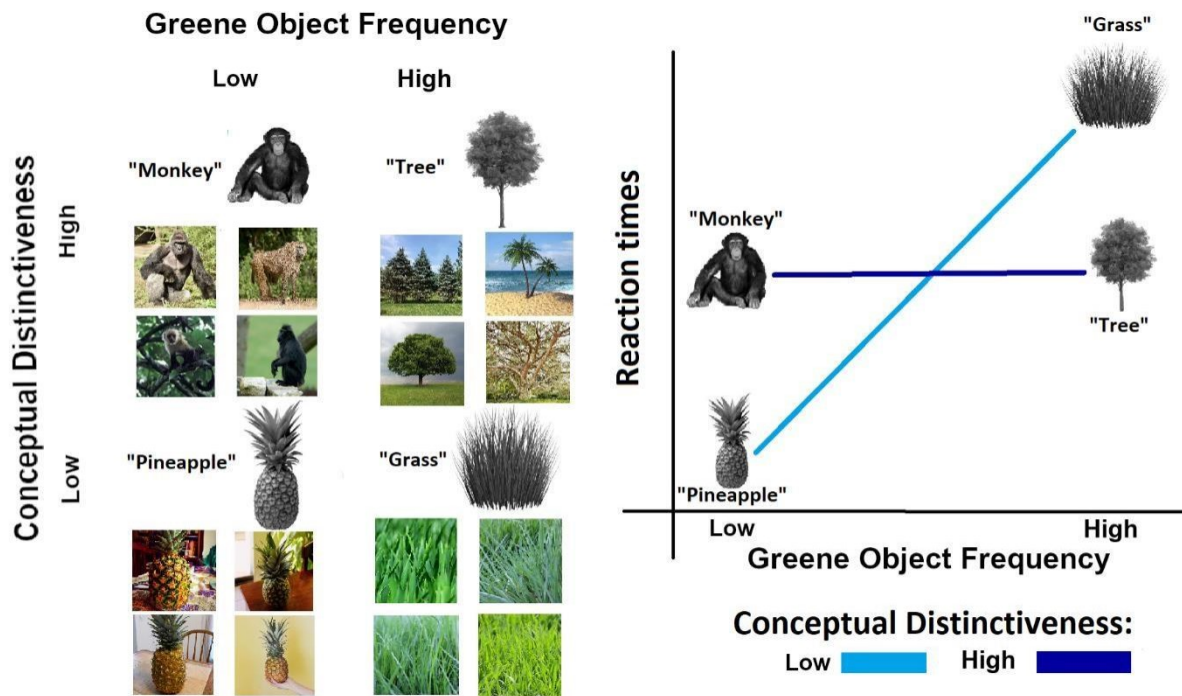
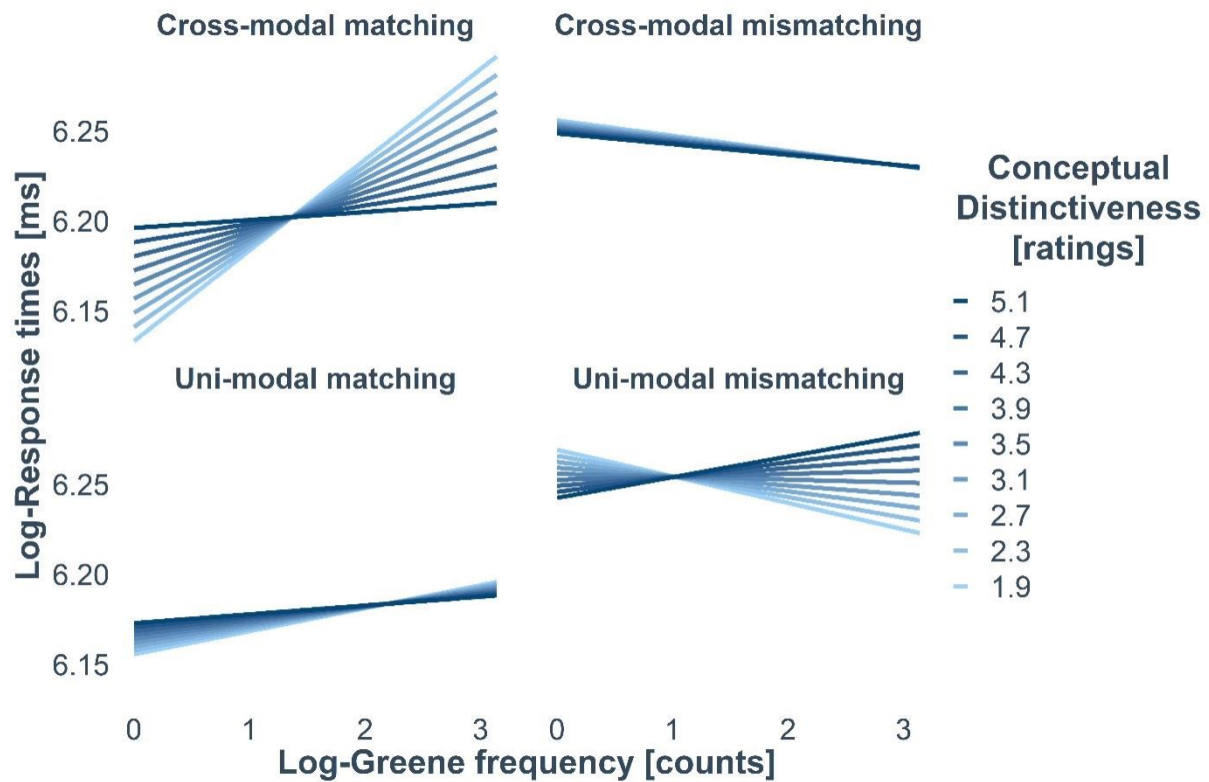


Figure 5. Results interaction between Greene frequency and Conceptual Distinctiveness.

Response times as a function of logarithmic Greene OF in interaction with Conceptual Distinctiveness across Matching conditions and Priming conditions (Cross-modal Matching vs Uni-modal Matching; Cross-modal Mismatching vs Uni-modal Mismatching; Cross-modal Matching vs Cross-modal Mismatching). RTs were estimated based on the selected model. Lines represent linear fitting of log Response times (y axis) by Greene frequency (x axis) for different values of Conceptual Distinctiveness (line colours: lighter = low CD, darker = high CD), in different experimental conditions (top-bottom-left-right panes).



Discussion Conceptual Distinctiveness ratings

When the unexpected processing facilitation for more rare concepts found in the Greene dataset (i.e., a frequency effect with opposite direction) first emerged in Experiment 2, we speculated that the Greene OF measure may reflect memory interference linked to perceptual experience with exemplars of an object concept (Konkle et al., 2010). Konkle et al. (2010) showed that memorability of an object depends on how many exemplars of that category were previously encountered, with more encounters creating a stronger interference that weakened the memory trace. Crucially, this interference for higher number of exemplars of an object was reduced for object categories with higher Conceptual Distinctiveness (i.e., whose members were more easily distinguishable into subgroups of different kinds; Konkle et al., 2010). Therefore, when object categories have low Conceptual Distinctiveness (i.e., it is difficult to divide their

exemplars in subcategories), the number of occurrences of an object has a strong impact on the mental representation (many occurrences = strong interference, few occurrences = weak interference); however, when object categories have high Conceptual Distinctiveness (i.e., it is easy to divide their exemplars in subcategories), the number of occurrences of an object does not influence mental representation to the same degree (few occurrences and many occurrences = similar weak interference). As stated in the Discussion of Experiment 2, this interpretation of Greene OF reflecting an interference process is only one possible explanation. One may also argue that less frequent objects reflect more narrow categories, which would offer more precise predictions of upcoming sensory input in Cross-modal Matching trials. We believe that this analysis of Greene OF in relation to Conceptual Distinctiveness of object categories might offer new, valuable insights on both these interpretations (for more detailed discussions please see *Supplementary Materials 11*).

Similar to Konkle et al. (2010) we found impaired performance for object categories that are encountered in more exemplars (higher object frequency) compared to object categories that are encountered in less exemplars (lower object frequency). And like Konkle et al. (2010), when Conceptual Distinctiveness (CD) was considered, the facilitation for more rare objects (low Greene frequency) was strongly reduced for those objects concepts that have more distinctive subgroups (high CD).

The example in *Figure 4* illustrates the influence of Conceptual Distinctiveness on the effect of the frequency of objects occurrence. *High Conceptual Distinctiveness* identifies the various visual experiences from a diverse set of exemplars (e.g., pine tree or palm tree; gorilla or macaque) that are connected to both frequently encountered (e.g., tree) and rarely encountered (e.g., monkey) objects. Instead, *Low Conceptual Distinctiveness* identifies the similar visual experiences from a homogeneous set of exemplars that are connected to both frequently encountered (e.g., grass) and rarely encountered (e.g., pineapple) objects. Following

the interference explanation of Konkle et al. (2010), the concepts that are encountered in many exemplars (high Greene OF) but have a diverse set of exemplars (high CD) are somehow privileged as the interference from other exemplars or different visual encounters is limited and counteracted for. For concepts with low CD, where it is less easy to distinguish between exemplars, an interference effect can be expected if many exemplars are encountered (high Greene OF).

These considerations also allow us to discuss the alternative explanation according to which the Greene OF effect is due to less frequent objects having more narrow categories allowing more precise predictions. This interpretation is especially interesting as we, again, found the interaction most strongly in the Cross-Modal Priming condition. CD is a way to measure if a category is narrow or wide in terms of the kinds of exemplars. We have shown that low OF concepts can have low CD (in line with this alternative explanation) but also high CD (opposing this alternative explanation). Indeed, the analysis of interactions between CD and the Greene OF measure could be used to show how the narrowness/width of a category impacts the frequency of occurrence: for narrower categories the frequency of occurrence has a strong impact on behavior, while for wider categories the frequency is less relevant. That is, when predicting an upcoming word or object from a low CD category it seems to be particularly beneficial for performance when the OF is low. However, clearly, the two dimensions (Greene OF and CD) do not overlap.

To conclude our discussion on Conceptual Distinctiveness, our analyses have shown how the effect of frequency of objects occurrence in real-world scenes is related to and dependent on the subcategorical structure of object concepts. In the next section, we present Experiment 3, a large-scale replication of the priming experiment, with the goal to reduce potential cross-experiment carry-over effects and object concept repetitions. In the original study (Experiments 1 and 2), participants performed the tasks in every condition (repeated

measures / within-participants design), which exposed them to many repetitions (18 times) of each concept (as either object picture or written word, as either prime or target). Despite the statistical advantages of within-participants designs, e.g., the reduction of variance from individual differences, potential carry-over effects could have created artificial frequency effects (especially since the Greene OF effect was unexpectedly going in the opposite direction of the WF effect). In Experiment 3, we therefore reduced the number of repetitions from 18 times to 8 by including two separate groups of new participants each of which performed either the Cross-modal or Uni-modal Priming tasks (between-participants design).

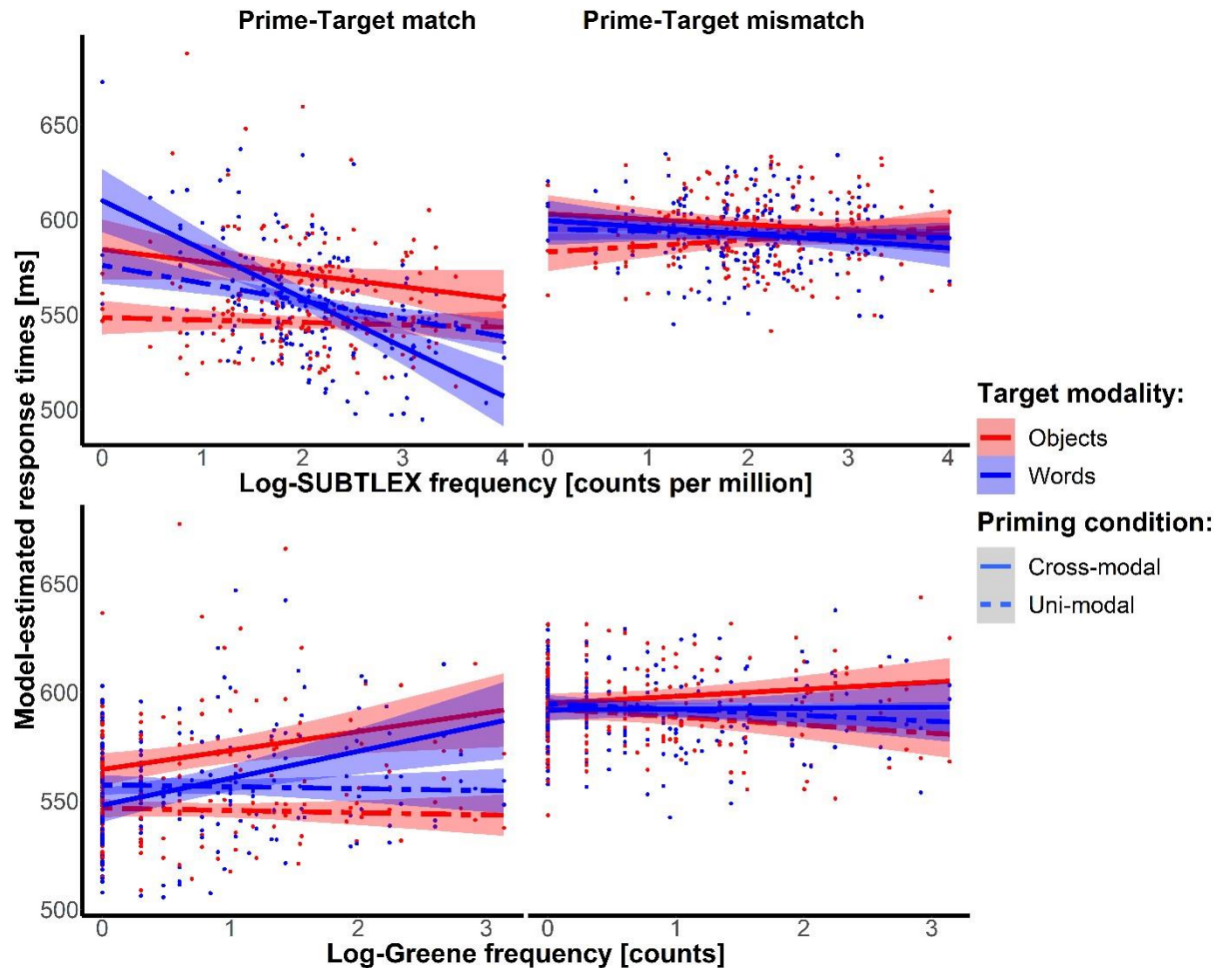
Results Experiment 3

Given that our aim was to replicate Experiment 2, we followed the same analysis, starting with the main model (i.e., no AIC-based frequency selection implemented). The only difference in the model structure was that in the current experiment (Experiment 3), we included the newly collected ratings of concept familiarity and image typicality from an independent participant sample, whereas in Experiments 1 and 2 we used ratings from the same participants who performed the task (no multicollinearity detected: variance inflation factors < 5; see *Supplementary Materials 12*).

Again, participants had to judge whether the meaning of the prime and target matched. We replicated the significant interactions between Greene OF, Matching condition, and Priming condition ($\beta=-0.010$, $SE=0.005$, $t=-2.165$, $p=0.030$); however, the interaction of SUBTLEX, Matching condition, and Priming condition was not significant ($\beta=0.009$, $SE=0.005$, $t=1.880$, $p=0.060$), but qualitatively in the same direction as in Experiment 2. Replicating Experiment 2, the interactions again revealed an effect in the opposite direction for SUBTLEX WF and Greene OF, while the two interaction effects were reduced in their effect size (i.e., about half of the effect size compared to Experiment 2).

Figure 6. Main results of Experiment 3.

Response times as a function of logarithmic SUBTLEX frequency (top plots) and Greene frequency (bottom plots) in the different conditions of Experiment 3; RTs were estimated based on the selected model. Points present participant-based mean response times separated for stimulus type (red: object stimuli; blue: word stimuli) in the different frequency levels. Lines represent linear fitting of points (solid: Cross-modal; dashed: Uni-modal), and shaded areas represent 95 % confidence interval. Bottom-left and top-left plots represent the effects in prime-target matching condition, while bottom-right and top-right plots represent the effects in prime-target mismatching condition.



In a post-hoc analysis that disentangled the interaction effects, we replicated the finding that the frequency effects were stronger in *Cross-modal Matching* trials than in *Uni-modal Matching* trials (SUBTLEX WF: $\beta=-0.014$, $SE=0.003$, $t=-4.324$, $p<0.001$; Greene OF: $\beta=0.017$, $SE=0.003$, $t=5.165$, $p<0.001$). Again, no difference was found for the frequency effects in *Mismatching* trials between *Cross-modal* and *Uni-modal* Priming (SUBTLEX WF: $\beta=-0.006$, $SE=0.003$, $t=-1.664$, $p=0.097$; Greene OF: $\beta=0.006$, $SE=0.003$, $t=1.959$, $p=0.050$;

see *Supplementary Materials 13*). Compared to Experiment 2, the effect size of difference of the SUBTLEX WF effect between *Cross-modal Matching* and *Uni-modal Matching* trials was reduced by more than 1/3 (Beta in Exp. 2: -0.023; Beta in Exp. 3: -0.014), while the difference of effects of the Greene OF between the two conditions was similar to Experiment 2 (Beta in Exp. 2: 0.018; Beta in Exp. 3: 0.017; for more details, see *Supplementary Materials 13*).

Again, the strongest frequency effects were found in *Cross-modal Matching* trials. With less trials per person, only the SUBTLEX frequency effect was significant ($\beta=-0.013$, $SE=0.006$, $t=-2.253$, $p=0.024$), while the Greene OF effect was not ($\beta=0.010$, $SE=0.005$, $t=1.873$, $p=0.061$). Qualitatively, the two effects again went in opposite directions. That is, we again found facilitatory effects for more frequent concepts for the SUBTLEX WF (faster RTs for high frequency items), while facilitatory effects emerged for more rare concepts for the Greene OF (faster RTs for low frequency items).

Contrary to Experiment 2, we found a significant interaction involving SUBTLEX, Matching condition, Priming condition, and Target modality ($\beta=0.021$, $SE=0.009$, $t=2.269$, $p=0.023$; see the Prime-Target match pane for SUBTLEX WF in *Figure 6* and for more details *Supplementary Materials 12*), which indicates a different modulation of words and objects as a function of SUBTLEX WF. Post-hoc investigations found that the difference of SUBTLEX WF effects between *Cross-modal* and *Uni-modal Matching* trials was stronger for words than for objects ($\beta=-0.016$, $SE=0.007$, $t=-2.401$, $p=0.016$).

Discussion Experiment 3

Experiment 3 investigated whether the WF and OF effects would still emerge when potential carry-over effects from previous exposure to the same concepts were minimized. One of the main motivations was that multiple presentations of the same concept may alter the perceived frequency of individual concepts, causing spurious effects. To reduce the number of

presentations, we exposed one group of participants only to the Uni-modal and another group to only the Cross-modal Priming condition of Experiment 2.

In general, we largely replicated the main interaction effect found in Experiment 2: That is, SUBTLEX WF and Greene OF had opposing effects and these effects differed as a function of matching and priming condition. More specifically, we replicated the findings that suggested that frequency effects are stronger when deeper semantic processing is required (i.e., frequency effect in Cross-modal Matching trials vs. Uni-modal matching trials), and that these effects seem to reflect a pre-activation from a semantically matched stimulus. Moreover, as in Experiment 2, the WF effect qualitatively indicated faster responses to frequently occurring concepts, while the OF effect was characterized by faster responses to rare concepts.

Of note, our post-hoc analyses revealed some differences. Specifically, we found reduced effect sizes for WF and OF, which led to the Greene OF effect not reaching significance (reduction of $1/3$ for the WF and $> 1/2$ for the OF effect). One explanation would be that fewer repetitions could reduce effect sizes. However, to account for this issue, we controlled for the number of concept repetitions using a covariate in both Experiments 2 and 3, ensuring that the confound of this variable on the frequency effects was minimal. Adding the parameter to the model increased model fit but did not affect the effect size estimates or the t-statistics. Potentially, the reduced number of occurrences of concepts in Experiment 3 compared to 2 might have resulted in less strong semantic associations of the words and object images, explicitly influencing the effects in Cross-modal Priming. Alternatively, or additionally, the between-participant design of Experiment 3 could be an explanation for this difference considering that this experiment showed a higher variance from individual differences compared to Experiment 2, which used a within-participant design. Importantly, we estimated the random effects for participants in both analyses, which should reduce the influence of the differences in design. Based on these considerations, we can summarize that the word

frequency effect, as expected, reliably occurs across experiments, while the object frequency effect seems to be more volatile.

It is also worth mentioning that two other effects emerged in the replication: A) a SUBTLEX WF effect was found for Uni-modal Matching trials with similar size and direction of the one in Cross-modal Matching trials. However, the post-hoc analysis between conditions showed that the effect in Cross-modal Matching trials remained stronger than the one in Unimodal Matching trials, supporting our hypothesis that frequency effects are strengthened by deeper semantic processing reflecting aspects of conceptual representation; B) a four-way interaction between SUBTLEX x Matching condition x Priming condition x Target modality was found, which, when explored, revealed that the effect was mainly driven by a significant SUBTLEX WF facilitation in *Cross-modal Matching trials with words* as target, while it was less pronounced for *Cross-modal Matching trials with objects* as target. Despite this difference to the original Experiment 2, the weaker influence of SUBTLEX WF on object processing resembles the one found in the semantic categorization task of Experiment 1. This stronger frequency-mediated priming effect for words might reflect the fact that words are visually more homogenous than objects, resulting in a more precise prediction of the visual aspects of the upcoming target (Gagl et al., 2020).

General Discussion

Investigating how semantic representations are accessed via different input modalities is a critical step in better understanding how humans store and organize knowledge about the world. The three experiments described in this manuscript provide evidence that high linguistic exposure to a semantic concept (i.e., how often it occurs or is used in our language) increases

recognition performance of both written words and object images (as measured by SUBTLEX WF effect). Furthermore, we present findings suggesting that semantic access might be facilitated not only when concepts are used frequently in language but also when they occur rarely in our visual world (as measured by the Greene OF effect). This phenomenon is possibly modulated by the specific categorical structure of each concept (i.e., the interaction of Greene OF effect with Conceptual Distinctiveness). Finally, we provide insights suggesting that these two effects reflect independent factors affecting visual word and object perception. All frequency effects seem to be substantially strengthened by a greater depth of semantic processing, as seen in the dependence of frequency effects on the type of task. In the following section, we will discuss the various findings in more depth.

SUBTLEX word frequency effect and strength of linguistic experience

The observed effect of subtitle-based frequency measure (SUBTLEX WF) replicated previous findings on word recognition (e.g., Brysbaert et al. 2011; Eisenhauer et al. 2021) and again showed that subtitle-based frequency estimates predict performance better than frequency estimates based on written text corpora (e.g., dlexDB, Heister et al., 2011; see *Supplementary Materials 14*). The novel aspect here is that contrary to Taikh and colleagues (2015), a wordbased frequency measure was also found to influence object recognition. Crucially, previous studies included multiple predictors of semantic richness in their regression models, which were not available for the stimulus material used here. These semantic richness measures could be of interest as they previously showed moderate correlations with the SUBTLEX WF measure (Taikh et al., 2015). However, a reanalysis of the WF effect in Experiment 1 that included Conceptual Distinctiveness (a measure that likely correlates with semantic richness) as a covariate did not change the pattern of effects described above. Thus, it is unlikely that the

observed WF effect in object recognition would have emerged as a confound (see *Supplementary Materials 15*). Nevertheless, future studies should include a larger set of semantic richness measures in order to determine the unique contributions of semantic richness on the one hand and WF on the other.

The finding that subtitle-based (i.e., SUBTLEX) but not text-based (i.e., dlexDB) WF effects were present in both stimulus modalities (i.e., words and objects) confirmed that subtitles are a more reliable source of estimation, and this measure is interpreted not just as reflecting the strength of experience with a word (effect in word recognition), but the *strength of experience with a concept* (effect in both object and word recognition). Indeed, as we control for many perceptual and linguistic variables, we suggest that this effect was modulated by the access to semantic representation required by the tasks. We could speculate that this strength is built through linguistic experience and, after that, transfers to other non-linguistic modalities. Such an interpretation would be in line with the idea that language would be not merely a means of communicating semantic information but also shaping semantic representations (Lupyan & Lewis, 2019).

Greene object frequency effect and its relationship with structure of object categories

In contrast to the SUBTLEX-based word frequency effect for objects and words, the Greene OF measure showed an opposite frequency effect: recognition performance in response to less frequent concepts was faster when compared to frequently encountered concepts. This inverted OF effect was surprising as we had computed the two measures based on a similar logic, i.e., counting occurrences in a dataset and capturing properties of the word. Furthermore, the OF effect did not emerge when we presented objects or words in isolation, but only in the matching trials of the Cross-modal Priming task, i.e., in context of a predictable prime stimulus, irrespective of modality. Note that in the same condition, we observed a substantial WF effect.

In these trials, the primed concept is retrieved from semantic memory to prepare participants for the upcoming stimulus, which is visually different but semantically matched.

This semantic memory involvement led us to reevaluate our findings based on the results reported in Konkle et al. (2010), who investigated memory interference processes when the number of exemplars belonging to a category was manipulated. They showed that we have the worse memory for the specific instance of frequently encountered objects (e.g., cars) because the increased number of exemplars creates interference. Conversely, we remember objects that we rarely encounter (e.g., pineapple) better because they suffer less from the interference of different exemplars (Konkle et al., 2010). Crucially, we found that the OF effect was only found when the objects came from a category that is not easily dividable into subgroups of different kinds, as measured by Conceptual Distinctiveness (CD). This seems to be due to the fact that when concepts can be easily divided into subgroups, this more complex division counterbalanced the interference effect produced by repeated encounters with exemplars of that category.

We want to stress that although CD and Greene OF are moderately correlated ($r= 0.43$), our finding of an interaction of the two measures showed that they explain different parts of variance. Thus, one should interpret the Greene OF effect beyond the effect of homogeneity/heterogeneity of object categories on the prediction of upcoming input. However, this explanation has highlighted the relevant issue of how categorical structure (more or less homogeneity) interacts with object occurrence and how this can impact the predictability of upcoming input.

Frequency effects and semantic processing

The results from the priming tasks (Experiments 2 and 3) are crucial to supporting the notion that frequency effects are also semantic. We found that they are more robust in Cross-modal

(i.e., integration of information across modalities) than Uni-modal Priming tasks (i.e., integration of information within modalities). Besides, these effects seem to reflect the processing of a corresponding prime-target combination rather than just recognizing the target, as the frequency effects were much more substantial in Cross-modal Matching trials than in Cross-modal Mismatching trials. Nevertheless, in Experiment 3, we only found a WF effect when an object picture primed a matching word but not when a word primed a matching object. It could be that the priming effect mediated by frequency is more substantial when words are the target stimulus. A potential explanation could be that words are more visually homogeneous stimuli than objects, making the upcoming word target easier to predict down to the individual pixel level (Zhao et al., 2019; Gagl et al., 2020; Wang & Maurer, 2020).

In sum, the present findings point to the semantic nature of the measured frequency effects. Moreover, these frequency effects might reflect processing common to both word and object recognition. Since they show similar patterns for word and object trials and given that what our word and object stimuli have in common is their meaning, one could speculate that this typical processing relates to accessing abstract conceptual representations.

Possible mechanisms underlying frequency effects

Regarding the mechanisms underlying the observed frequency effects in Cross-modal Matching trials, one could hypothesize that they resemble the neural processes described in Kok and colleagues (2017), where an auditory prime pre-activated a representation of a previously matched visual stimulus before its presentation (Kok et al., 2017). Furthermore, in line with our results (i.e., pre-activation facilitation based on frequency), they found that the pre-activation strength could predict behavioral responses. These findings suggest a mechanism of *sharpening* visual representation compatible with expected upcoming input, modulated by some aspects of previous experience. In our case, these aspects might be the strength built

through linguistic experience and the encounters during visual experience that are incorporated into the conceptual representations evoked by the prime.

Analogously, one could speculate that similar processes are occurring during Unimodal Priming too. The crucial difference is that what modulates sharpening is not a semantic representation but a more perceptual representation (e.g., orthographic for words, visual for objects), therefore producing hardly any frequency effects. This finding is in line with the behavioral and MEG evidence reported by Eisenhauer and colleagues (2021) who found frequency effects for words presented in isolation (i.e., as in Experiment 1 described above), but not in a Uni-modal Priming context (i.e., a word primed by the same word as in our Experiments 2 and 3). Notably, they found a modulation of neural activity by orthographic information following the prime and preceding the target word, similarly indicating a sharpening process on the neuronal level (Eisenhauer et al., 2021).

However, given the study's design and methods, we cannot yet draw firm conclusions about the nature of the mechanisms underlying our frequency effects. For example, we cannot rule out that the involved predictive processing (Rao & Ballard, 1999) functions by inhibiting the most common features of upcoming input instead of sharpening it (Gagl et al., 2020). Further investigations are needed to specify the neuronal mechanisms on representations in perceptual and or semantic processes in Cross-modal Priming. Here, electrophysiological measures (M/EEG) would allow for a more fine-grained and better temporally resolved investigation of how optimization of recognition behavior in Cross-modal Priming is implemented on the neuronal level.

Choosing the right dataset for frequency estimations

In general, any decision to use one dataset over another in order to compute frequency measures needs to be approached with great care. One problem lies in the assumption that a chosen

dataset is a good representation of the state of the world, but every dataset, even the largest available, remains an approximation. Besides, the composition of the datasets often reflects biases in the way they were composed and the sources that were used to create them. Moreover, the assumption that a dataset captures universally shared concept representations might not be valid. Factors like expertise and physical or cultural context have a different impact on the individual experience of the world (Kuperman & Van Dyke, 2013).

Of course, the quantity and variety of scene images of the datasets are lower than the corpora usually used for computing WF measures (more than 20 million words of the SUBTLEX database vs the 400,000 object annotations in the ADE20K and 48,000 object annotations in the Greene database). Concerns about the representability of selected image datasets are therefore always valid and must be considered carefully. To account for this concern - and to start somewhere - we decided to include both image datasets (the ADE20K dataset, Zhou et al., 2019, and the Greene dataset, Greene, 2013). Both datasets are widely used by computer vision scientists and cognitive psychologists working on visual cognition (Bonner & Epstein, 2021; Bracci et al., 2021). While the Greene dataset includes fewer annotations than ADE20K, it has the advantage of thoroughly cleaning up spelling mistakes, synonyms, and other issues affecting any frequency analyses based on labelled image databases. So, which frequency measure should be used?

Even though our results confirm that as a word frequency measure, the SUBTLEXbased frequency is the better predictor for categorization behavior, the situation seems less clear for object frequencies, especially given that Greene and ADE20K produce similar result patterns. In our primary analysis, Greene was preferred to ADE20K, given its more robust improvement of model fit in both words and object trials (see details of the AIC-based selection method in the *Analysis* section of *Materials and Methods*). However, also ADE20K showed a significant improvement in model fit in both modalities in Experiment 2.

The two datasets have both pros and cons, as pointed out previously: ADE20K is clearly superior when it comes to dataset size and variety of images, while the Greene dataset would be the preferred choice when looking for high quality annotations. Ideally, revising ADE20K annotations with the same approach offered by Greene (2013) would likely create the best of both worlds. However, more practical ways to decide which measure to employ would be to consider aspects like the number and types of object stimuli and the scenes they are typically found. For larger and more diverse sets (e.g., natural vs man-made, public vs private), it is more likely to find good estimates in the ADE20K dataset. For smaller and more homogeneous sets (e.g., objects found in a house), the quality of Greene’s annotations could beat the quantity of ADE20K’s ones. In general, one goal for the future would be a database with a high number of quality annotations that, similar to word databases, contains a sufficient number of examples for a more appropriate estimation of object frequency (i.e., at least 20 million; Brysbaert et al., 2011).

Pros and cons of using labeled image databases for cognitive studies

As previously discussed, estimating any type of frequency from databases can create unwanted biases in the frequency measures being extracted. In addition to these database dependent biases, calculating object frequency measures includes further hurdles. For instance, linguistic image databases make the evaluation of the visual domain dependent of the linguistic domain. In addition, labeling decisions must be made for each object. At times labeling decisions can be easy (e.g., pineapple), but sometimes there are very explicit decisions to make (e.g., are all types of cars simply labelled as “cars” or by their brand name/type, e.g., “Porsche” vs. “Jeep”). The problem might be more severe for highly general concepts (i.e., trees, cars, animals, and others).

Crucially, these decisions can and will have an impact on the computed frequency of occurrence, and could create differences between datasets (although ADE20K and Greene OF show strong correlation $r=0.81$ and led to similar results, see *Supplementary Materials 16*). The annotators of the images in the Greene database were instructed to use entry-level labels (e.g., “car”, not “vehicle” or “Mercedes”), and labels were inspected and corrected for synonyms and similar confounds. We believe that the issue of biases from labelling has been addressed in our study in three ways: (i) the OF effect was always estimated independently of the WF effect, since both were included in the same model. This would allow to rule out differences arising from common vs uncommon labels; (ii) we have shown that the Greene OF effect is present only for concepts with low Conceptual Distinctiveness, which have a more homogeneous set of exemplars and thus should be less prone to be biased by possible variabilities in labelling; (iii) the OF effect was estimated independently of image typicality (included as covariate), which measures how the employed image stimuli are a typical exemplars of the categories denoted by the employed word stimuli (i.e., the labels). In this sense, it represents how strongly image-word pairs are associated and therefore predictable. Still, the labelling procedure of image datasets remains an important issue that needs to be considered in the future.

We want to stress that - to our knowledge - this is the first time that metrics such as object frequencies were computed and used to predict response times in a way traditionally done with WF measures. Even the other OF measure used in this study, ADE20K OF (Zhou et al., 2019), was found to produce similar patterns of effects in terms of size, direction, and probability when repeating the analysis of Experiment 2, substituting Greene OF with it (for more details, see *Supplementary Materials 16*). Similar fit and result patterns indicate that datasets of annotated and segmented objects capture aspects of the world and our experience with it, which are relevant for our cognitive system in general. Therefore, we hope that this first attempt at studying object-based frequency measures gives rise to broader investigations, as it

was done by some studies in cognitive neuroscience that already started with investigations in this direction (Bonner & Epstein, 2021; Bracci et al., 2021).

Conclusion

To conclude, this study aimed to expand and innovate previous investigations of semantic access from words and objects by employing new measures of object frequencies and comparing them to established word frequency measures. In a first attempt, we identified language-based and image-based frequency measures and demonstrated how they differentially influence recognition processes which might reflect two organizational principles for conceptual knowledge. Moreover, we showed that very different visual information (words vs. objects) could lead to relatively similar processing when accessing conceptual knowledge, providing further evidence for the strong interrelation between language and vision. We hope that this study will lead to further investigations of both word- and object-based frequency measures to increase our understanding of accessing meaning from visual input.

Context

The word frequency (WF) effect in visual word recognition is a well-established empirical finding, while there is little evidence about object frequency (OF)'s role in object recognition. Word and object recognition have the common goal of accessing meaning based on visual input. This similarity raises questions about whether similar parameters modulate object and word recognition. Since more frequent words are recognized more efficiently, we investigate whether the frequency of occurrence also similarly affects object recognition. This team of researchers - with expertise in visual word and object recognition - joined forces to investigate the process of accessing the meaning of objects and words using object and word frequency

measures. We, therefore, applied new metrics of object frequency based on state-of-the-art datasets of annotated images and evaluated them in comparison to widely used metrics of word frequency. Beyond this, we aimed to determine common aspects of object and word processing that would give further evidence for the strong interrelation between language and vision while providing a starting point for future investigations.

References

- Akaike, H. (1981). Likelihood of a model and information criteria. *Journal of Econometrics*, 16(1), 3–14. [https://doi.org/10.1016/0304-4076\(81\)90071-3](https://doi.org/10.1016/0304-4076(81)90071-3)
- Almeida, J., Knobel, M., Finkbeiner, M., & Caramazza, A. (2007). The locus of the frequency effect in picture naming: When recognizing is not enough. *Psychonomic Bulletin & Review*, 14(6), 1177–1182. <https://doi.org/10.3758/BF03193109>
- Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4), 390–412. <https://doi.org/10.1016/j.jml.2007.12.005>
- Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. J. (2004). Visual Word Recognition of Single-Syllable Words. *Journal of Experimental Psychology: General*, 133(2), 283–316. <https://doi.org/10.1037/0096-3445.133.2.283>
- Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious Mixed Models. *ArXiv:1506.04967 [Stat]*. <http://arxiv.org/abs/1506.04967>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting Linear Mixed-Effects Models using lme4. *ArXiv:1406.5823 [Stat]*. <http://arxiv.org/abs/1406.5823>
- Bates, E., Burani, C., D’Amico, S., & Barca, L. (2001). Word reading and picture naming in Italian. *Memory & Cognition*, 29(7), 986–999. <https://doi.org/10.3758/BF03195761>
- Bonner, M. F., & Epstein, R. A. (2021). Object representations in the human brain reflect the co-occurrence statistics of vision and language. *Nature Communications*, 12(1), 1-16.
- Bracci, S., Mraz, J., Zeman, A., Leys, G., & de Beeck, H. O. (2021). Object-scene conceptual regularities reveal fundamental differences between biological and artificial object vision. *bioRxiv*.
- Brehm, L., & Alday, P. M. (2020). *A decade of mixed models: It’s past time to set your contrasts*. Talk presented at the 26th Architectures and Mechanisms for Language Processing Conference (AMLap 2020). Potsdam, Germany. 2020-09-03 - 2020-09-05.
- Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The Word Frequency Effect: A Review of Recent Developments and Implications for the Choice of Frequency Estimates in German. *Experimental Psychology*, 58(5), 412–424. <https://doi.org/10.1027/1618-3169/a000123>
- Brysbaert, M., Mander, P., & Keuleers, E. (2018). The Word Frequency Effect in Word Processing: An Updated Review. *Current Directions in Psychological Science*, 27(1), 45–50. <https://doi.org/10.1177/0963721417727521>
- Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior research methods*, 41(4), 977-990.
- Capitani, E., Laiacona, M., Mahon, B., & Caramazza, A. (2003). What are the facts of Semantic Category-specific deficits? A Critical review of the clinical evidence. *Cognitive Neuropsychology*, 20(3–6), 213–261. <https://doi.org/10.1080/02643290244000266>
- Clarke, A., Taylor, K. I., Devereux, B., Randall, B., & Tyler, L. K. (2013). From Perception to Conception: How Meaningful Objects Are Processed over Time. *Cerebral Cortex*, 23(1), 187–197. <https://doi.org/10.1093/cercor/bhs002>
- Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, 108(1), 204–256. <https://doi.org/10.1037/0033-295X.108.1.204>

- Criss, A. H., & Malmberg, K. J. (2008). Evidence in favor of the early-phase elevated-attention hypothesis: The effects of letter frequency and object frequency. *Journal of Memory and Language*, 59(3), 331–345. <https://doi.org/10.1016/j.jml.2008.05.002>
- de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, 47(1), 1–12. <https://doi.org/10.3758/s13428-014-0458-y>
- Dehaene, S., & Cohen, L. (2011). The unique role of the visual word form area in reading. *Trends in Cognitive Sciences*, 15(6), 254–262. <https://doi.org/10.1016/j.tics.2011.04.003>
- Devereux, B. J., Clarke, A., Marouchos, A., & Tyler, L. K. (2013). Representational Similarity Analysis Reveals Commonalities and Differences in the Semantic Processing of Words and Objects. *Journal of Neuroscience*, 33(48), 18906–18916. <https://doi.org/10.1523/JNEUROSCI.3809-13.2013>
- Downing, P. E., Chan, A. W.-Y., Peelen, M. V., Dodds, C. M., & Kanwisher, N. (2006). Domain Specificity in Visual Cortex. *Cerebral Cortex*, 16(10), 1453–1461. <https://doi.org/10.1093/cercor/bhj086>
- Eisenhauer, S., Fiebach, C. J., & Gagl, B. (2019). Context-Based Facilitation in Visual Word Recognition: Evidence for Visual and Lexical But Not Pre-Lexical Contributions <sup>/>. *Eneuro*, 6(2), <https://doi.org/10.1523/ENEURO.0321-18.2019>
- Eisenhauer, S., Gagl, B., & Fiebach, C. J. (2021). Predictive pre-activation of orthographic and lexical-semantic representations facilitates visual word recognition. *Psychophysiology*, e13970.
- Engbert, R., Nuthmann, A., Richter, E. M., & Kliegl, R. (2005). SWIFT: A Dynamical Model of Saccade Generation During Reading. *Psychological Review*, 112(4), 777–813. <https://doi.org/10.1037/0033-295X.112.4.777>
- Fairhall, S. L., & Caramazza, A. (2013). Brain Regions That Represent Amodal Conceptual Knowledge. *Journal of Neuroscience*, 33(25), 10552–10558. <https://doi.org/10.1523/JNEUROSCI.0051-13.2013>
- Forster, K. I., & Chambers, S. M. (1973). Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior*, 12(6), 627–635. [https://doi.org/10.1016/S0022-5371\(73\)80042-8](https://doi.org/10.1016/S0022-5371(73)80042-8)
- Gagl, B., Sassenhagen, J., Haan, S., Gregorova, K., Richlan, F., & Fiebach, C. J. (2020). An orthographic prediction error as the basis for efficient visual word recognition. *NeuroImage*, 214, 116727. <https://doi.org/10.1016/j.neuroimage.2020.116727>
- Greene, M. R. (2013). Statistics of high-level scene context. *Frontiers in Psychology*, 4. <https://doi.org/10.3389/fpsyg.2013.00777>
- Greene, M. R. (2016). Estimations of object frequency are frequently overestimated. *Cognition*, 149, 6–10. <https://doi.org/10.1016/j.cognition.2015.12.011>
- Grill-Spector, K., & Weiner, K. S. (2014). The functional architecture of the ventral temporal cortex and its role in categorization. *Nature Reviews Neuroscience*, 15(8), 536–548. <https://doi.org/10.1038/nrn3747>
- Harel, J., Koch, C., & Perona, P. (2007). Graph-Based Visual Saliency. In B. Schölkopf, J. C. Platt, & T. Hoffman (Eds.), *Advances in Neural Information Processing Systems 19* (pp. 545–552). MIT Press. <http://papers.nips.cc/paper/3095-graph-based-visual-saliency.pdf>
- Hebart, M. N., Dickter, A. H., Kidder, A., Kwok, W. Y., Corriveau, A., Van Wicklin, C., & Baker, C. I. (2019). THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PloS one*, 14(10), e0223792.
- Heister, J., Würzner, K.-M., Bubenzer, J., Pohl, E., Hanneforth, T., Geyken, A., & Kliegl, R. (2011). DlexDB – eine lexikalische Datenbank für die psychologische und linguistische Forschung. *Psychologische Rundschau*, 62(1), 10–20. <https://doi.org/10.1026/0033-3042/a000029>

- Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of neuroscience*, 17(11), 4302-4311.
- Kliegl, R., Nuthmann, A., & Engbert, R. (2006). Tracking the mind during reading: The influence of past, present, and future words on fixation durations. *Journal of Experimental Psychology: General*, 135(1), 12–35. <https://doi.org/10.1037/0096-3445.135.1.12>
- Kok, P., Jehee, J. F., & De Lange, F. P. (2012). Less is more: expectation sharpens representations in the primary visual cortex. *Neuron*, 75(2), 265-270.
- Kok, P., Mostert, P., & de Lange, F. P. (2017). Prior expectations induce prestimulus sensory templates. *Proceedings of the National Academy of Sciences*, 114(39), 10473–10478. <https://doi.org/10.1073/pnas.1705652114>
- Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of Experimental Psychology: General*, 139(3), 558–578. <https://doi.org/10.1037/a0019165>
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6), 84–90. <https://doi.org/10.1145/3065386>
- Kuperman, V., & Van Dyke, J. A. (2013). Reassessing word frequency as a determinant of word recognition for skilled and unskilled readers. *Journal of Experimental Psychology: Human Perception and Performance*, 39(3), 802.
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, 82(13). <https://doi.org/10.18637/jss.v082.i13>
- Li, W. (1992). Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE Transactions on information theory*, 38(6), 1842-1845.
- Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization* (No. 8). Oxford University Press on Demand.
- Lüdtke, D., Ben-Shachar, M. S., Patil, I., Waggoner, P., & Makowski, D. (2021). performance: An R package for assessment, comparison and testing of statistical models. *Journal of Open Source Software*, 6(60).
- Lupyan, G., & Lewis, M. (2019). From words-as-mappings to words-as-cues: The role of language in semantic knowledge. *Language, Cognition and Neuroscience*, 34(10), 1319–1337. <https://doi.org/10.1080/23273798.2017.1404114>
- Maxwell, S. E., Delaney, H. D., & Kelley, K. (2017). *Designing experiments and analyzing data: A model comparison perspective*. Routledge.
- McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological review*, 88(5), 375.
- Morrison, C. M., Ellis, A. W., & Quinlan, P. T. (1992). Age of acquisition, not word frequency, affects object naming, not object recognition. *Memory & Cognition*, 20(6), 705–714. <https://doi.org/10.3758/BF03202720>
- Morton, J. (1979). Facilitation in word recognition: Experiments causing change in the logogen model. In *Processing of visible language* (pp. 259-268). Springer, Boston, MA.
- Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3), 145-175.
- Olkkonen, M., Aguirre, G. K., & Epstein, R. A. (2017). Expectation modulates repetition priming under high stimulus variability. *Journal of vision*, 17(6), 10-10.
- <https://osf.io/d3j9h/files/>
- Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., ... & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior research methods*, 51(1), 195-203.

- R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1), 79–87. <https://doi.org/10.1038/4580>
- Rayner, K. (2009). Eye movements in reading: Models and data. *Journal of eye movement research*, 2(5), 1.
- Richter, D., Ekman, M., & de Lange, F. P. (2018). Suppressed sensory response to predictable object stimuli throughout the ventral visual stream. *Journal of Neuroscience*, 38(34), 7452–7461.
- Riesenhuber, M., & Poggio, T. (2000). Models of object recognition. *Nature neuroscience*, 3(11), 1199–1204.
- Robinson, A. P., & Froese, R. E. (2004). Model validation using equivalence tests. *Ecological Modelling*, 176(3–4), 349–358. <https://doi.org/10.1016/j.ecolmodel.2004.01.013>
- Russell, B. C., Torralba, A., Murphy, K. P., & Freeman, W. T. (2008). LabelMe: a database and web-based tool for image annotation. *International journal of computer vision*, 77(1), 157–173.
- Scarborough, D. L., Cortese, C., & Scarborough, H. S. (1977). Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology: Human perception and performance*, 3(1), 1.
- Schad, D. J., Vasishth, S., Hohenstein, S., & Kliegl, R. (2020). How to capitalize on a priori contrasts in linear (mixed) models: A tutorial. *Journal of Memory and Language*, 110, 104038. <https://doi.org/10.1016/j.jml.2019.104038>
- Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3), 379–423. <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>
- Shelton, J. R., & Caramazza, A. (1999). Deficits in lexical and semantic processing: Implications for models of normal language. *Psychonomic Bulletin & Review*, 6(1), 5–27. <https://doi.org/10.3758/BF03210809>
- Shinkareva, S. V., Malave, V. L., Mason, R. A., Mitchell, T. M., & Just, M. A. (2011). Commonality of neural representations of words and pictures. *NeuroImage*, 54(3), 2418–2425. <https://doi.org/10.1016/j.neuroimage.2010.10.042>
- Summerfield, C., Trittschuh, E. H., Monti, J. M., Mesulam, M., & Egner, T. (2008). Neural repetition suppression reflects fulfilled perceptual expectations. *Nature neuroscience*, 11(9), 1004–1006.
- Taikh, A., Hargreaves, I. S., Yap, M. J., & Pexman, P. M. (2015). Semantic classification of pictures and words. *Quarterly Journal of Experimental Psychology*, 68(8), 1502–1518. <https://doi.org/10.1080/17470218.2014.975728>
- Tversky, B. (1969). Pictorial and verbal encoding in a short-term memory task. *Perception & Psychophysics*, 6(4), 225–233. <https://doi.org/10.3758/BF03207022>
- Vö, M. L.-H., Boettcher, S. E., & Draschkow, D. (2019). Reading scenes: How scene grammar guides attention and aids perception in real-world environments. *Current Opinion in Psychology*, 29, 205–210. <https://doi.org/10.1016/j.copsyc.2019.03.009>
- Wang, F., & Maurer, U. (2020). Interaction of top-down category-level expectation and bottom-up sensory input in early stages of visual-orthographic processing. *Neuropsychologia*, 137, 107299.
- Whelan, R. (2008). Effective analysis of reaction time data. *The Psychological Record*, 58(3), 475–482.
- Yarkoni, T. (2009). Big correlations in little studies: Inflated fMRI correlations reflect low statistical power—Commentary on Vul et al. (2009). *Perspectives on psychological science*, 4(3), 294–298.
- Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart’s N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, 15(5), 971–979. <https://doi.org/10.3758/PBR.15.5.971>
- Zhao, J., Maurer, U., He, S., & Weng, X. (2019). Development of neural specialization for print: Evidence for predictive coding in visual word recognition. *PLoS biology*, 17(10), e3000474.

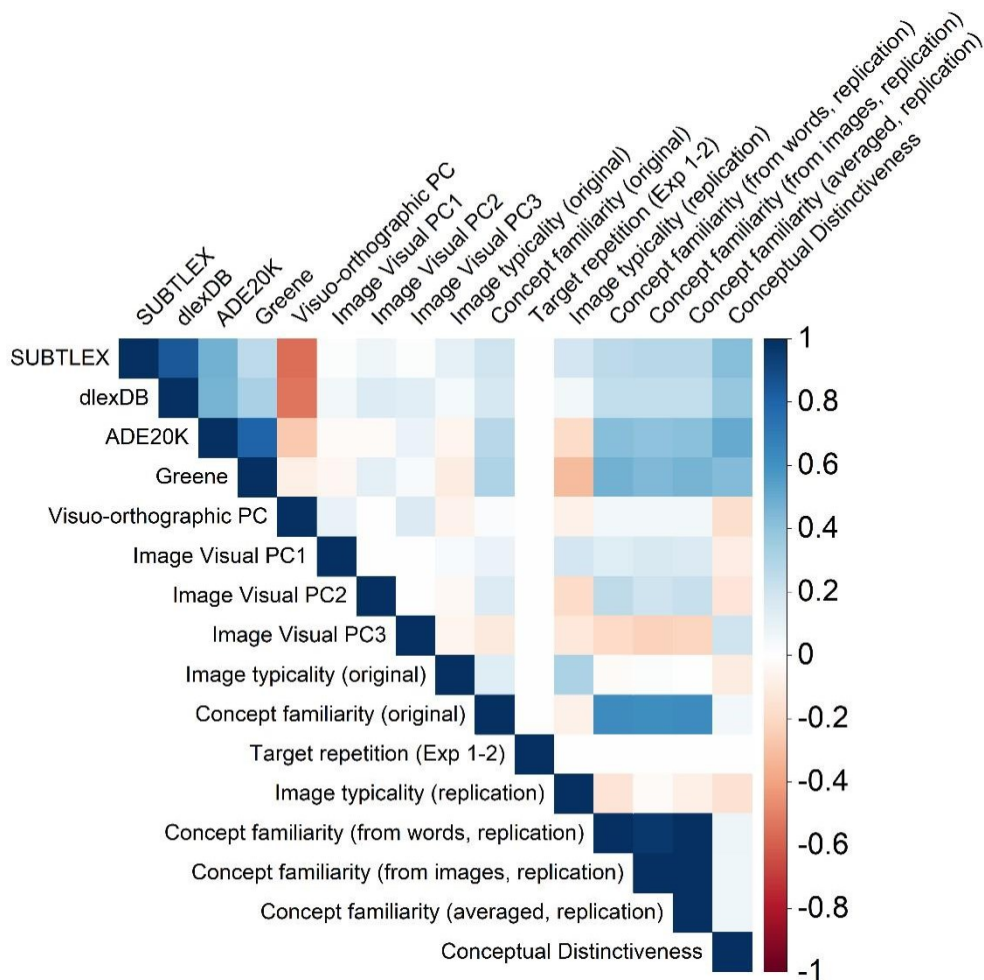
- Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., & Torralba, A. (2019). Semantic Understanding of Scenes Through the ADE20K Dataset. *International Journal of Computer Vision*, 127(3), 302–321. <https://doi.org/10.1007/s11263-018-1140-0>

Supplementary Materials

Supplementary materials 1 – Factor correlations, distributions, analysis details

Supplementary figure 1 – Correlations between factors measured for the study

Product-Moment Correlation Coefficients for each pair of predictors used in the experiment.

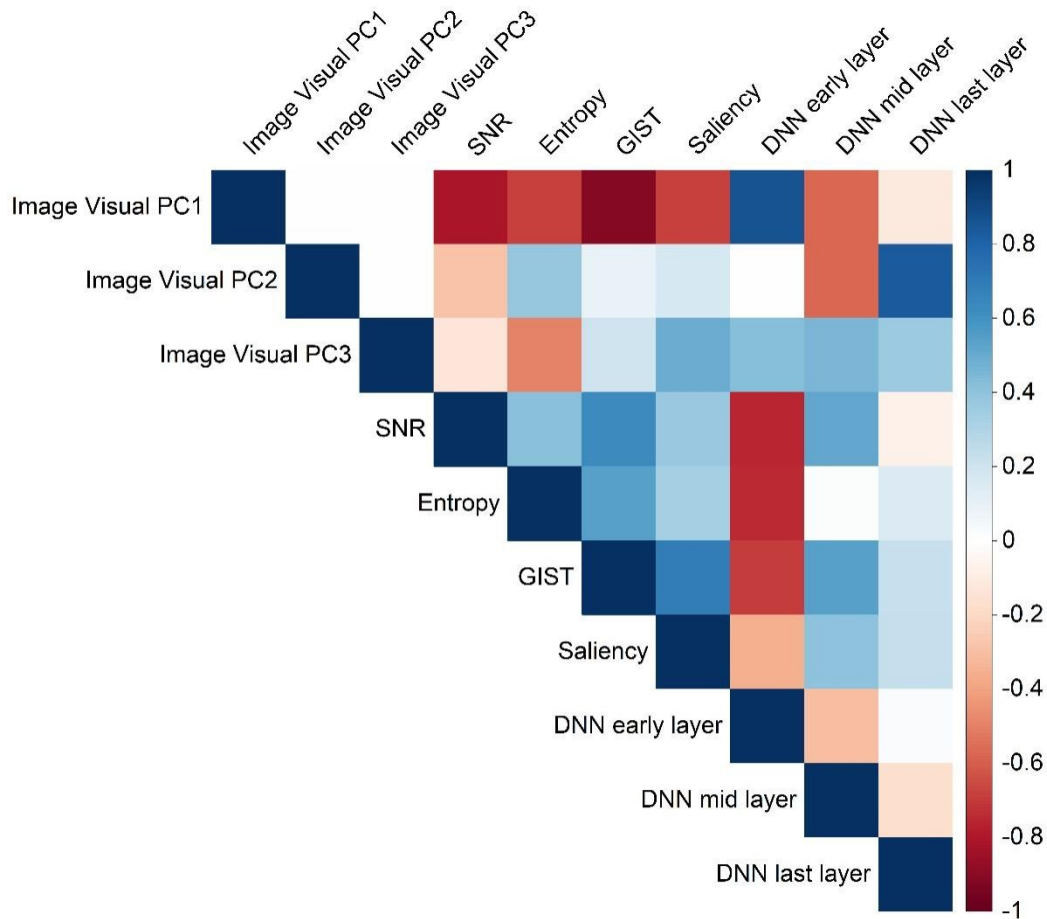


PCA procedure for visual and visuo-orthographic predictors: measures with multidimensional output were averaged to obtain a single value for every image. As a rule of thumb, we selected the Principal Components (PCs) that, alone, explained more variance than what a variable would explain if they all explained the same amount of variance.

Image visual PCs: 7 variables \rightarrow threshold: $100 / 7 = 14.29\%$. We extracted three orthogonal PCs, explaining more than 84 % of the variance. We labeled the first PC *Image visual PC1* (about 50 % of variance explained, strong positive correlation with convolutional layer 1 of AlexNet, and strong negative correlation with SNR, GIST, Entropy, Saliency and AlexNet layer 4). The second PC was named *Image visual PC2* (about 18 % of variance explained, strong positive correlation with AlexNet layer 7, strong negative correlation with AlexNet layer 4). The third PC was named *Image visual PC3* (about 15 % of variance explained, medium positive correlation with AlexNet layers and Saliency, medium negative correlation with Entropy). We interpret the PC1 as an estimate of low-to-mid-level visual features of the images (stronger weights from AlexNet early layer, SNR, saliency, but also from AlexNet mid layer and GIST), while the PC2 seems to capture more complex mid-to-high-level visual features (stronger weights from AlexNet mid layer and entropy, but also from AlexNet late layer). PC3, however, has a less clear interpretation, capturing part of variance from both low-level and high-level visual features estimates (higher weights for all the three AlexNet layers).

Supplementary figure 2 – Correlations of visual predictors and extracted PCs

Product-Moment Correlation Coefficients for each pair of visual predictors of objects and the Principal Components (PCs) extracted from the PCA on those predictors.



Supplementary table 1. Object image PCA loadings for every variable in every extracted principal component (PCs). They represent the weights of every variable on the extracted PCs.

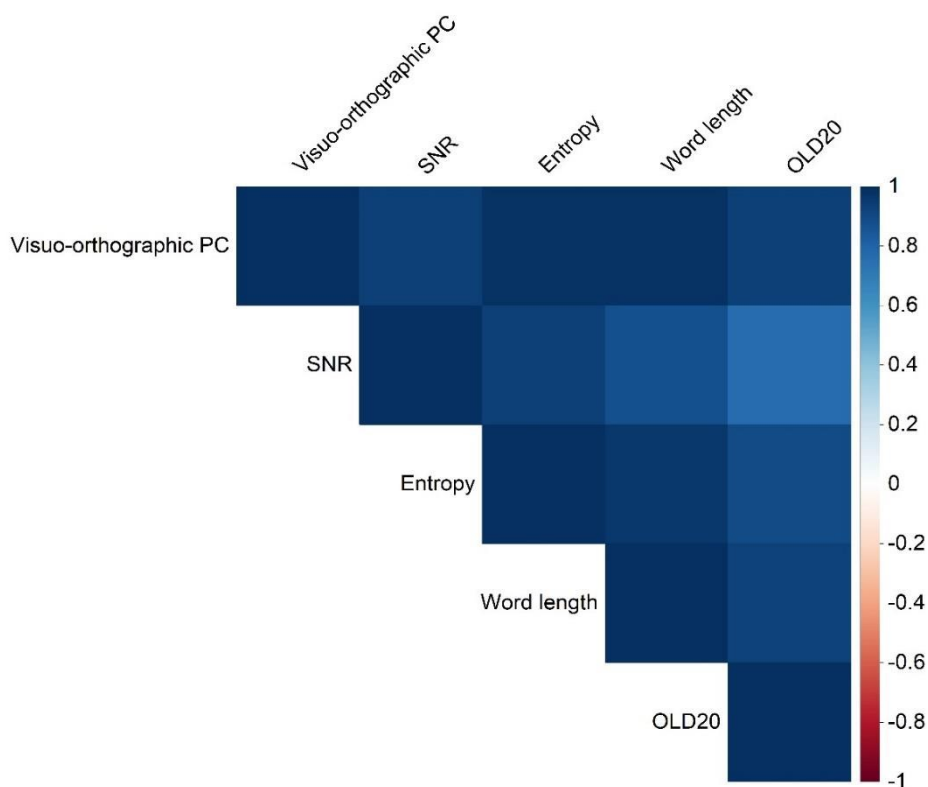
Variables	PC1 loadings	PC2 loadings	PC3 loadings
AlexNet conv1	0.459	0.005	0.413
AlexNet conv4	-0.305	-0.505	0.436
AlexNet fc7	-0.062	0.735	0.352
Saliency	-0.366	0.154	0.477
GIST	-0.486	0.080	0.194

Entropy	-0.366	0.337	-0.482
SNR	-0.434	-0.249	-0.134

Visuo-orthographic PC: 4 variables -> threshold: $100 / 4 = 25 \%$; one principal component (PC) was extracted and was labeled *Visuo-orthographic PC* (variance explained circa 92 %; strong positive correlation with all the original variables). Being all the variables highly correlated between them and with the PC, interpretation seems straightforward and difficult at the same time. The rationale for including many variables that were expected to be highly correlated was to acknowledge the different levels (visual and orthographical) from which we wanted to extract a covariate able to control for perceptual aspects of a word.

Supplementary figure 3 – Correlations between visuo-orthographic predictors and the extracted PC

Product-Moment Correlation Coefficients for each pair of visuo-orthographic predictors of words and the Principal Component (PC) extracted from the PCA on those predictors.

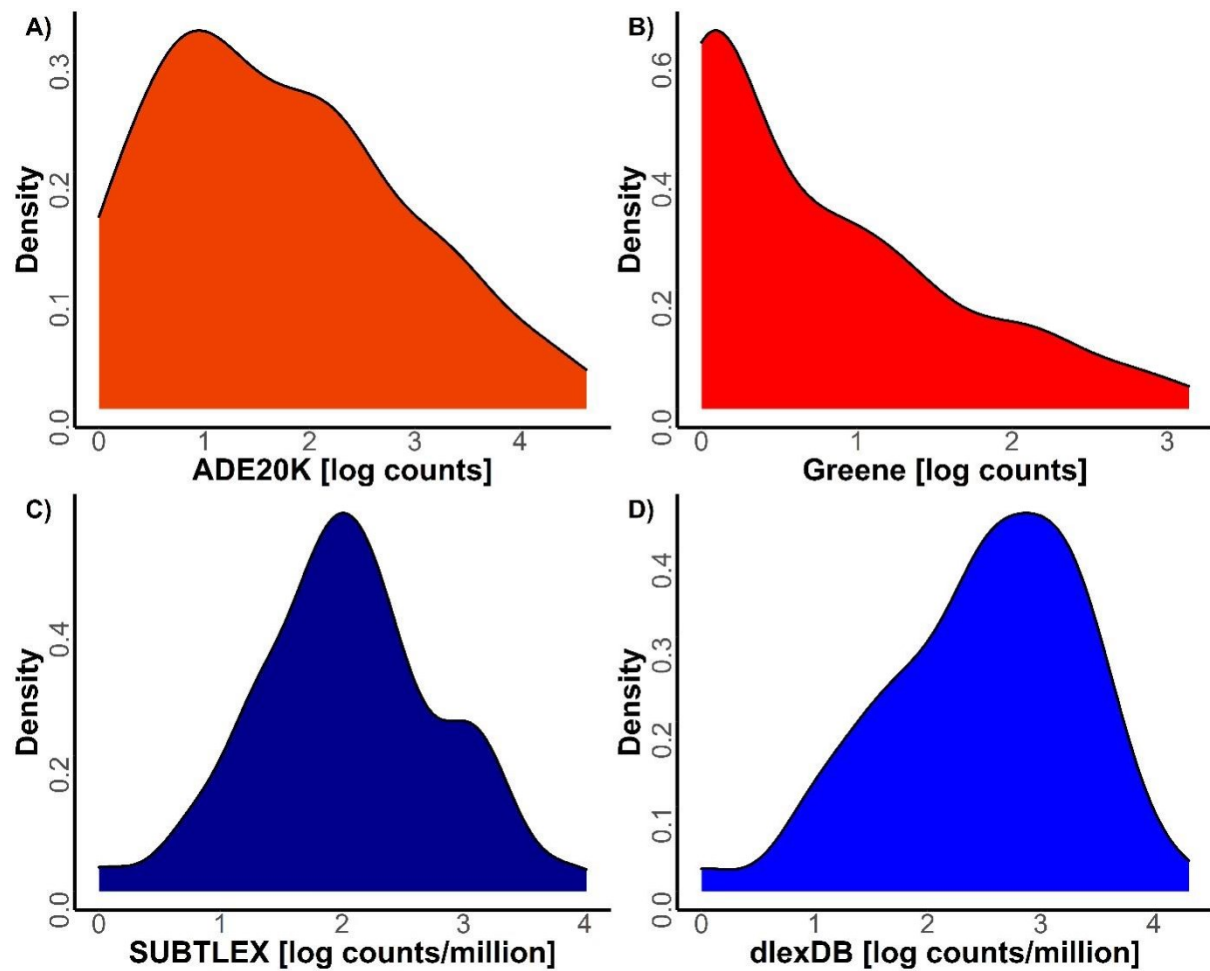


Supplementary table 2. Word image PCA loadings for every variable in the extracted principal component (PC). They represent the weights that every variable has on the extract PC.

Variables	PC1 loadings
OLD20	0.487
Word length	0.512
Entropy	0.515
SNR	0.485

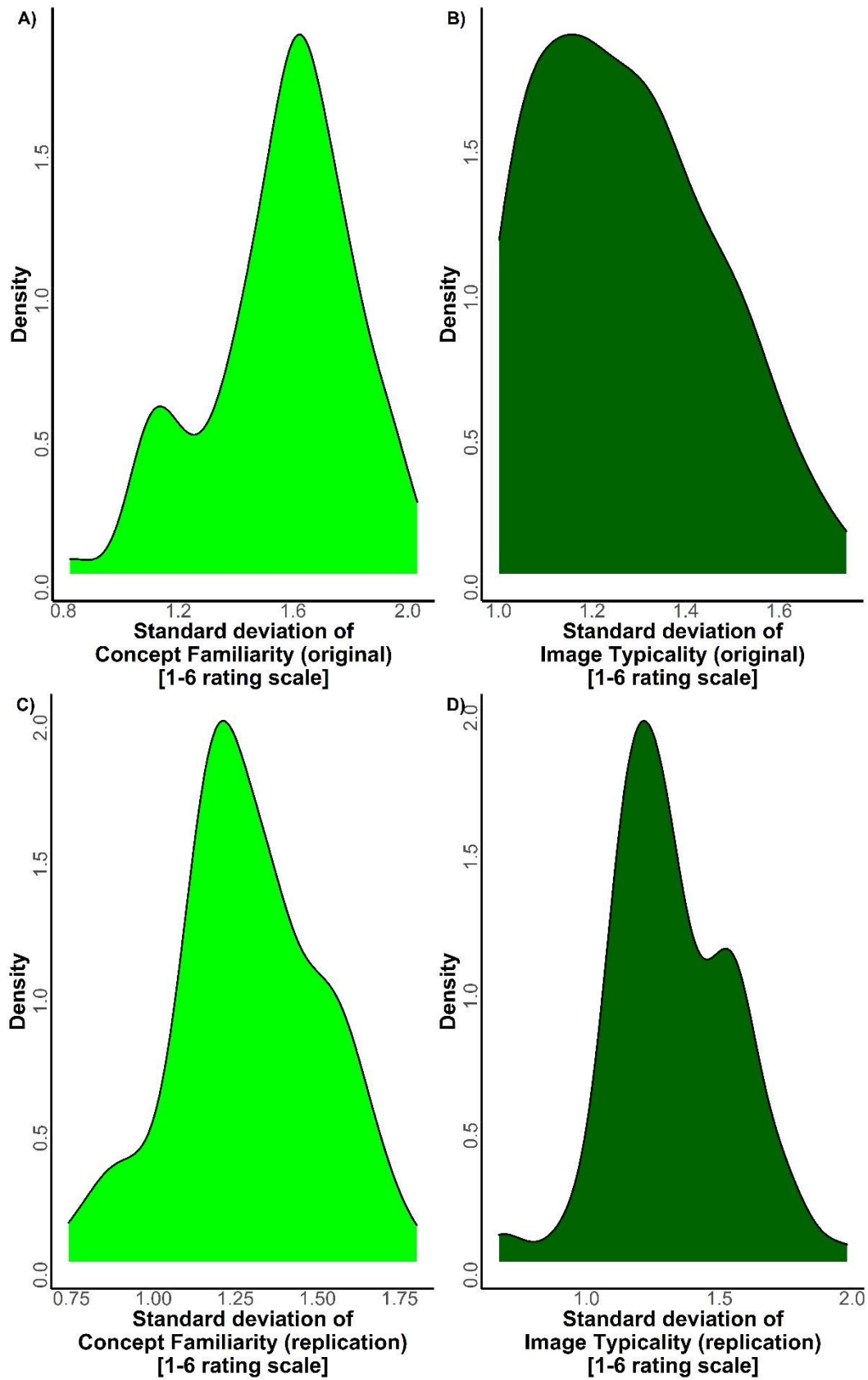
Supplementary figure 4 – Distribution of the frequency measures

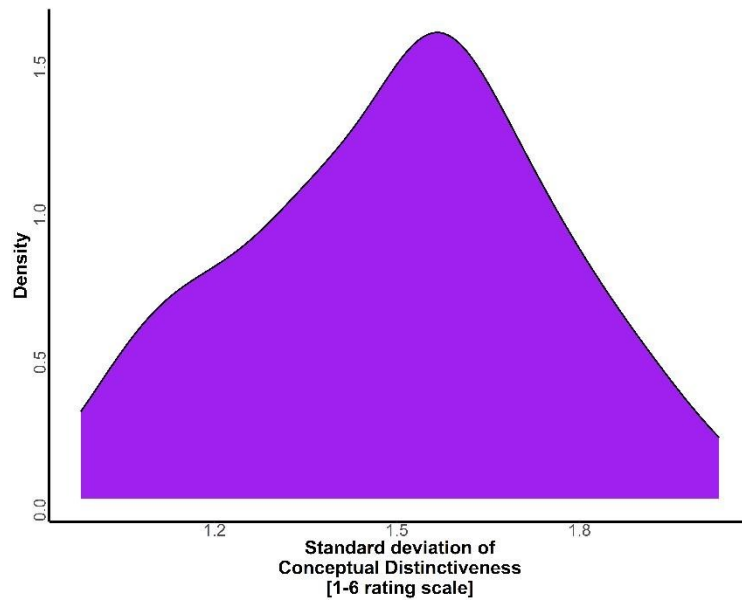
Density distribution of the frequency values for our set of stimuli.



Supplementary figure 5 – Distribution of concept variability across participants for the rating measures

Density distribution of concepts' standard deviation across participants for the collected ratings (scale 1-6)





Analysis details: We fitted Linear Mixed-effects Models (LMMs) via maximum likelihood estimation, and Satterthwaite's method was used to obtain p-values (package lmerTest, Kuznetsova et al., 2017). Using the *scale()* function in R, we transformed each continuous predictor variable onto a common scale which improves model fitting procedures. These continuous predictors are the four frequency measures (SUBTLEX, dlexDB, Greene, ADE20K) and the covariates (Concept familiarity [different in Exp 1-2 and Exp 3], Image typicality [different in Exp 1-2 and Exp 3], Image visual PC1, Image visual PC2, Image visual PC3, Visual-orthographic PC, Target repetition [different in Exp 1-2 and Exp 3]).

For the coding of contrasts in categorical predictors (*Exp 1*: Concept modality: Words – Objects; Concept category: Natural – Man-made; Trial accuracy: Correct – Incorrect; *Exp 2 and Exp 3*: Target modality: Words - Objects; Priming condition: Cross-modal – Uni-modal; Matching condition: Mismatching – Matching; Trial accuracy: Correct - Incorrect), we used sum contrast coding, which in our case gave us an estimate of the difference between the two levels of each of our categorical variables, like main effects in a multi-way repeated measures ANOVA (Schad et al., 2020; Brehm & Alday, 2020).

Including trial response accuracy as a categorical covariate in the LMMs allows us to consider the variance explained by the output of the task (i.e., correct or incorrect trial), but at the same time to estimate the impact of the other variables independently from the output of the task itself. Besides, this way, we did not have to exclude further trials from the analysis, and we could exploit the flexibility offered by LMMs.

To account for the multiple repetitions of the same object concepts within participants and a potential carry-over effect that could confound frequency effects, we included in the models a numeric covariate Target repetition that represents the number of times that the

current target concept has been presented (as either a word or an object image, as either target or prime).

To prevent misinterpretation of the effects and confounds due to high correlation of the predictors, we assessed potential multicollinearity of the models by computing the variance inflation factors (VIFs) for each term in each model, using the *check_collinearity()* function in R (package performance; Lüdtke et al., 2021). When variance inflation factors are below 5, there are low correlations between predictors and therefore no predictors need to be excluded to avoid confounds in the interpretation of the results. When the variance inflation factors are higher than 5, those predictors should be excluded from the model and the analysis should be repeated.

1) We implement a model comparison based on the Akaike Information Criterion (AIC, Akaike, 1981). This step allowed us to compare our four frequency measures and select the frequency measures with the best fit. To implement this, we first fit one model per frequency measure (i.e., SUBTLEX, dlexDB, ADE20K, and Greene frequency) separately for the word and the object recognition trials (four frequency measures times two modalities: eight models in total). All models implemented the same covariates and random-effects structure. Then we compared the four models of each modality to a “baseline” model that did not include the frequency measure, but that was estimated on the same subset of data and implemented the same structure of covariates and random effects (2 baseline models in total, one for words and one for objects data). With this procedure, we could estimate the singular fit of each frequency measure in each stimulus modality. From that, we selected the frequency measures that explained a considerable amount of variance in both modalities for further analysis. A better fit was determined by a significant decrease in the AIC, which was tested by implementing the *anova()* function in R. Given the different sources from which word and object frequencies are estimated, they might provide a distinct contribution in representing the occurrence of

objects/words in the world. Therefore, we operated the AIC-based selection following these criteria: in the best case, we would have selected two measures, i.e., the best fitting OF and the best fitting WF measure. In the worst-case, none of the frequency measures would have explained variance in both object and word trials. While, in between, we would have selected either only an OF or a WF measure.

2) After selecting the best frequency measures, we ran a LMM estimating the effects of those selected frequencies on the entire dataset (word trials + object trials), and including all categorical factors and continuous covariates, as well as random factors for participants and concepts.

3) When we detected significant interactions between frequency measures and categorical predictors, we also ran post-hoc LMMs in order to understand the different effects of frequency between different conditions (e.g., SUBTLEX in Cross-modal trials vs. SUBTLEX in Uni-modal trials) and within each condition (e.g., the simple effect of SUBTLEX in Cross-modal trials and simple effect of SUBTLEX in Uni-modal trials). Note that the estimation of frequency effects, given the structure of linear models, was independent (i.e., controlled for) from the effect of the several continuous covariates included in the models.

Supplementary materials 2 – Model selection in Experiment 1

Formula of the models computed in the selection process (1 model x 4 frequency measures x 2 modalities + baseline model without frequency measures x 2 modalities = 10 models):

Exp1_logRT ~ FREQUENCY MEASURE +
Concept category + Concept familiarity + Image typicality +
Image visual PC1 + Image visual PC2 + Image visual PC3 +
Visuo-orthographic PC + Target repetition + Trial accuracy +
(1|Participants) + (1|Concepts)

Supplementary table 3. Summary table of the models included in the selection process. “Frequency” indicates the frequency measure included in the model, where ‘Baseline’ means no measures included. “AIC” is the criterion used to evaluate the fit of the model. “Modality” indicates which subset of data was considered. “AIC difference” is the difference in AIC between every model and the baseline model of the same modality. More negative differences indicate a better fit of the model including the frequency measure; significant improvements of fit are highlighted in bold.

Frequency	AIC	Modality	AIC difference
Baseline	-1214.816	Objects	0
SUBTLEX WF	-1218.978	Objects	-4.163
ADE20K OF	-1212.867	Objects	1.949
Greene OF	-1213.126	Objects	1.690
dlexDB WF	-1214.462	Objects	0.354
Baseline	-1442.289	Words	0
SUBTLEX WF	-1469.443	Words	-27.153
ADE20K OF	-1443.517	Words	-1.228
dlexDB WF	-1455.736	Words	-13.447
Greene OF	-1441.156	Words	1.133

Supplementary materials 3 – Results of the selected model in Experiment 1

*Exp1_logRT ~ SUBTLEX WF * Concept modality +
 Concept category + Concept familiarity + Image typicality +
 Image visual PC1 + Image visual PC2 + Image visual PC3 +
 Visuo-orthographic PC + Target repetition +
 Trial accuracy + (1|Participants) + (1/Concepts)*

Supplementary table 4. Results from the selected model for semantic categorization

<i>Predictors</i>	<i>β</i>	<i>SE</i>	<i>t</i>	<i>p</i>
(Intercept)	6.479	0.021	302.552	< 0.001
Concept modality (Words – Objects)	0.094	0.005	20.529	< 0.001
SUBTLEX WF	-0.031	0.007	-4.417	< 0.001
Visuo-orthographic PC	-0.006	0.007	-0.818	0.413
Concept familiarity	-0.003	0.003	-0.983	0.326
Image typicality	-0.004	0.003	-1.302	0.193
Image visual PC1	-0.002	0.006	-0.316	0.752
Image visual PC2	0.019	0.006	3.253	0.001
Image visual PC3	0.008	0.006	1.410	0.159
Target repetition	-0.011	0.002	-4.933	< 0.001
Trial accuracy (Correct – Incorrect)	-0.017	0.009	-1.936	0.053
Concept category (Natural – Man-made)	0.001	0.012	0.116	0.908
SUBTLEX x (Words – Objects)	-0.019	0.005	-4.160	< 0.001

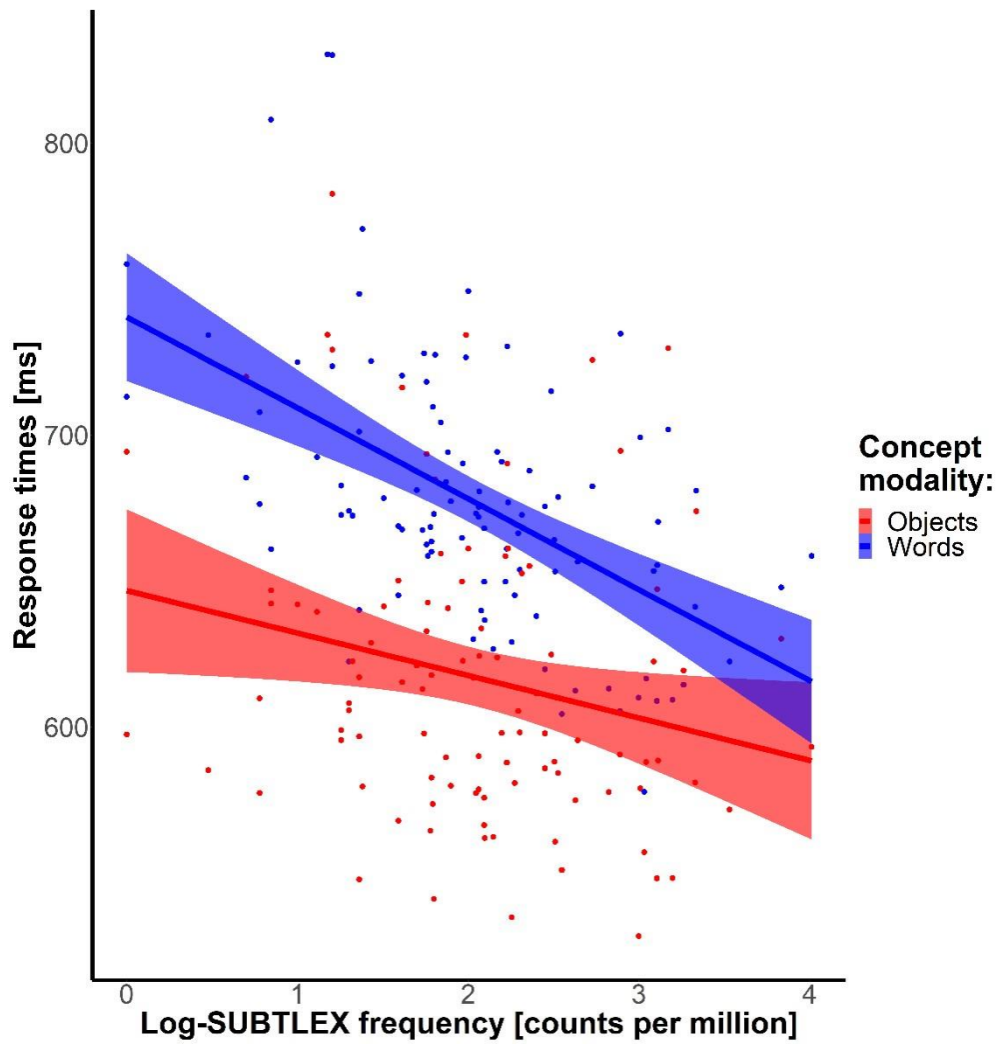
Supplementary table 5. Variance Inflation Factors for the estimated effects of the main model of Experiment 1

Term	VIF
Concept modality (Words – Objects)	1.012
SUBTLEX WF	1.535
Visuo-orthographic PC	1.698
Concept familiarity	1.043
Image typicality	1.015
Image visual PC1	1.032
Image visual PC2	1.027
Image visual PC3	1.063
Trial accuracy (Correct – Incorrect)	1.017
Concept category (Natural – Man-made)	1.180
Concept modality x SUBTLEX	1.000
Target repetition	1.010

The measured SUBTLEX WF effect was independent of visual and visuo-orthographic information of the stimuli, as well as of image typicality, subjective familiarity, concept repetition, concept category and accuracy of categorization.

Supplementary figure 6 – Raw RTs from Experiment 1

Raw response times for object (red) and word (blue) trials as a function of SUBTLEX frequency in Experiment 1. Points show concepts with different level of frequency, averaged across participants; lines represent linear fitting of points and shaded areas represent 95 % confidence interval



Supplementary materials 4 - Post-hoc of interaction in Experiment 1

2 post-hoc models are estimated, with the same formula, but on 2 different subsets of the data (Object trials and Word trials):

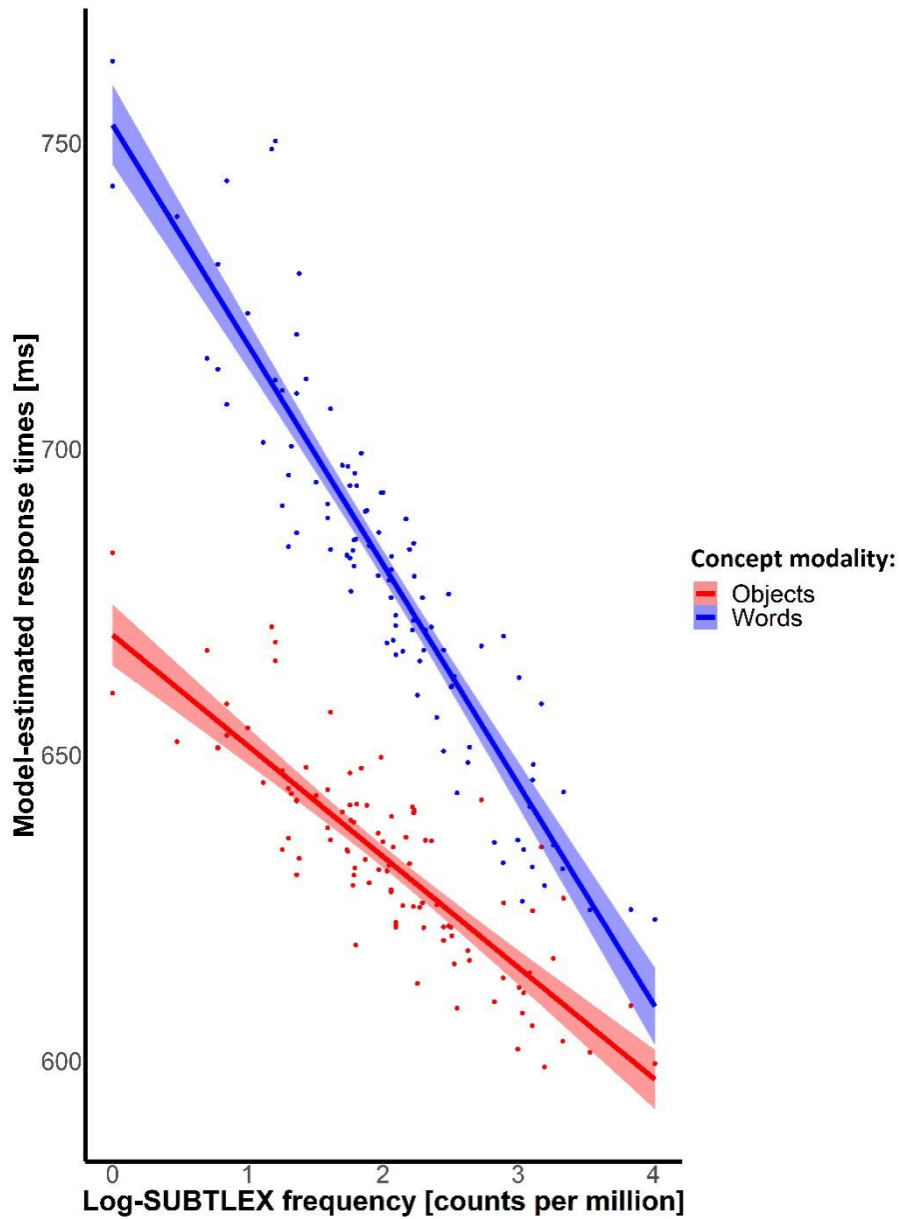
$$\begin{aligned}
 \text{Exp1_logRT} \sim & \text{SUBTLEX WF} + \\
 & \text{Concept category} + \text{Concept familiarity} + \text{Image typicality} + \\
 & \text{Image visual PC1} + \text{Image visual PC2} + \text{Image visual PC3} + \\
 & \text{Visuo-orthographic PC} + \text{Target repetition} + \\
 & \text{Trial accuracy} + (1|\text{Participants}) + (1|\text{Concepts})
 \end{aligned}$$

Supplementary table 6. Results from the post-hoc models for semantic categorization

<i>Predictors</i>	Objects				Words			
	β	<i>SE</i>	<i>t</i>	<i>p</i>	β	<i>SE</i>	<i>t</i>	<i>p</i>
(Intercept)	6.449	0.022	289.694	< 0.001	6.520	0.024	266.828	< 0.001
SUBTLEX WF	-0.022	0.009	-2.524	0.012	-0.041	0.007	-5.794	< 0.001
Concept category	0.009	0.015	0.610	0.542	-0.008	0.012	-0.626	0.531
Visuo-orthographic PC	-0.006	0.009	-0.671	0.502	-0.006	0.007	-0.823	0.411
Concept familiarity	-0.000	0.005	-0.050	0.960	-0.005	0.004	-1.151	0.250
Image typicality	-0.009	0.004	-1.977	0.048	-0.001	0.004	-0.258	0.797
Image visual PC1	-0.004	0.007	-0.598	0.550	0.000	0.006	0.069	0.945
Image visual PC2	0.026	0.007	3.581	< 0.001	0.011	0.006	1.949	0.051
Image visual PC3	0.005	0.007	0.692	0.489	0.012	0.006	2.034	0.042
Target repetition	0.009	0.021	0.430	0.667	-0.030	0.024	-1.293	0.196
Trial accuracy	-0.059	0.013	-4.379	< 0.001	0.005	0.011	0.483	0.629

Supplementary figure 7 – RTs estimated from post-hoc models of Experiment 1

Estimated response times from individual post-hoc models for object (red) and word (blue) trials as a function of SUBTLEX frequency in Experiment 1. Points show concepts with different level of frequency, averaged across participants; lines represent linear fitting of points and shaded areas represent 95 % confidence interval



Supplementary materials 5 - Results of new ratings on original data of Exp 1

$Exp1_logRT \sim SUBTLEX\ WF * Concept\ modality +$
 $Concept\ category + Concept\ familiarity\ (replication) +$
 $Image\ typicality\ (replication) +$
 $Image\ visual\ PC1 + Image\ visual\ PC2 + Image\ visual\ PC3 +$
 $Visuo-orthographic\ PC + Target\ repetition +$
 $Trial\ accuracy + (1|Participants) + (1/Concepts)$

Supplementary table 12. Results from main model of Exp 1 including ratings from replication study

Predictors	β	<i>SE</i>	<i>t</i>	<i>p</i>
(Intercept)	6.479	0.021	303.428	< 0.001
Concept modality (Words – Objects)	0.094	0.005	20.527	< 0.001
SUBTLEX WF	-0.035	0.008	-4.690	< 0.001
Visuo-orthographic PC	-0.009	0.007	-1.148	0.251
Concept familiarity (replication)	0.012	0.007	1.857	0.063
Image typicality (replication)	-0.009	0.006	-1.513	0.130
Image visual PC1	-0.001	0.006	-0.233	0.816
Image visual PC2	0.013	0.006	2.316	0.021
Image visual PC3	0.010	0.006	1.741	0.082
Target repetition	-0.011	0.002	-4.933	< 0.001
Trial accuracy (Correct – Incorrect)	-0.017	0.009	-1.905	0.057
Concept category (Natural – Man-made)	0.012	0.013	0.925	0.355
SUBTLEX x (Words – Objects)	-0.019	0.005	-4.157	< 0.001

Supplementary Materials 6 - Model selection in Experiment 2

Formula of the models computed in the selection process (1 model x 4 frequency measures x 2 modalities + baseline model without frequency measures x 2 modalities = 10 models):

$$\begin{aligned}
 &Exp2_logRT \sim FREQUENCY MEASURE * Priming condition * Matching condition + \\
 &Concept familiarity + Image typicality + \\
 &Image visual PC1 + Image visual PC2 + Image visual PC3 + \\
 &Visuo-orthographic PC + Target repetition + Trial accuracy + 379 \\
 &(1/Participants) + (1/Concepts)
 \end{aligned}$$

Supplementary table 7. Summary table of the models included in the selection process. “Frequency” indicates the frequency measure included in the model, where ‘Baseline’ means no measures included. “AIC” is the criterion used to evaluate the fit of the model. “Modality” indicates which subset of data was considered. “AIC difference” is the difference in AIC between every model and the baseline model of the same modality. More negative differences indicate a better fit of the model including the frequency measure; significant improvements of fit are highlighted in bold.

Frequency	AIC	Modality	AIC difference
Baseline	-5150.153	Objects	0
SUBTLEX WF	-5194.848	Objects	-44.695
ADE20K OF	-5152.258	Objects	-2.105
DlexDB WF	-5175.167	Objects	-25.013
Greene OF	-5169.700	Objects	-19.547
Baseline	-6366.279	Words	0
SUBTLEX WF	-6401.687	Words	-35.409
ADE20K OF	-6385.581	Words	-19.302
DlexDB WF	-6377.383	Words	-11.105
Greene OF	-6401.694	Words	-35.415

Supplementary materials 7 Results selected model in Experiment 2

Exp2_logRT ~ *SUBTLEX WF* * *Priming condition* * *Matching condition* * *Target modality* +
Greene OF * *Priming condition* * *Matching condition* * *Target modality* +
Concept familiarity + *Image typicality* +
Image visual PC1 + *Image visual PC2* + *Image visual PC3* +
Visuo-orthographic PC + *Target repetition* + *Trial accuracy* +
(1|*Participants*) + (1|*Concepts*)

Supplementary table 8. Results from the selected model for priming task

Predictors	β	SE	t	p
(Intercept)	6.225	0.020	315.728	<0.001
Greene OF	0.008	0.002	4.474	<0.001
Matching condition (Mismatch – Match)	0.073	0.002	32.684	<0.001
Target modality (Words – Objects)	-0.008	0.002	-3.618	<0.001
Priming condition (Cross-modal – Uni-modal)	0.006	0.005	1.173	0.241
SUBTLEX WF	-0.004	0.002	-1.812	0.070
Visuo-orthographic PC	0.010	0.002	4.877	<0.001
Concept familiarity	-0.001	0.002	-0.845	0.398
Image typicality	-0.004	0.002	-2.562	0.010
Image visual PC1	0.002	0.002	1.217	0.224
Image visual PC2	0.004	0.002	2.788	0.005
Image visual PC3	-0.001	0.002	-0.832	0.405
Target repetition	-0.036	0.003	-13.357	<0.001
Trial accuracy (Correct – Incorrect)	0.030	0.005	5.444	<0.001
Greene x Matching condition	-0.017	0.002	-7.454	<0.001
Greene x Target modality	0.003	0.002	1.396	0.163
Matching condition x Target modality	-0.009	0.004	-1.939	0.053
Greene x Priming condition	0.008	0.002	3.347	0.001
Matching condition x Priming condition	0.003	0.004	0.737	0.461
Priming condition x Target modality	0.013	0.004	2.872	0.004

SUBTLEX x Matching condition	0.021	0.002	9.074	<0.001
SUBTLEX x Target modality	-0.005	0.002	-2.378	0.017
SUBTLEX x Priming condition	-0.015	0.002	-6.335	<0.001
Greene x Matching condition x Target modality	0.003	0.005	0.668	0.504
Greene x Matching condition x Priming condition	-0.020	0.005	-4.256	<0.001
Greene x Priming condition x Target modality	0.007	0.005	1.416	0.157
Matching condition x Priming condition x Target modality	0.046	0.009	5.220	<0.001
SUBTLEX x Matching condition x Target modality	-0.002	0.005	-0.472	0.637
SUBTLEX x Matching condition x Priming condition	0.017	0.005	3.687	<0.001
SUBTLEX x Priming condition x Target modality	0.003	0.005	0.569	0.570
Greene x Matching condition x Priming condition x Target modality	0.002	0.009	0.227	0.821
SUBTLEX x Matching condition x Priming condition x Target modality	0.006	0.009	0.612	0.541

Supplementary table 9. Variance Inflation Factors for the estimated effects of the main model of Experiment 2

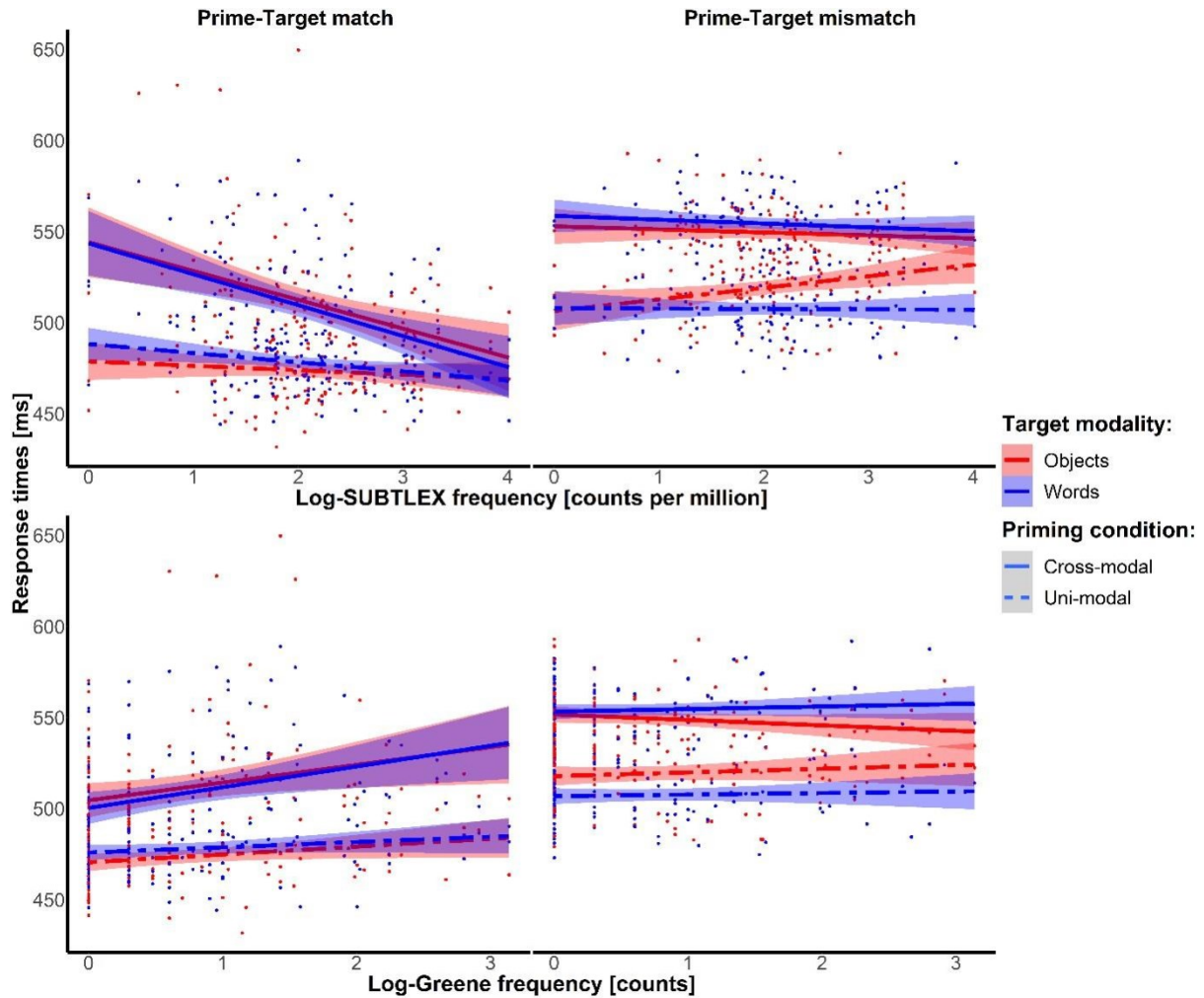
Term	VIF
Greene OF	1.190
Matching condition (Mismatch – Match)	1.022
Target modality (Words – Objects)	1.032
Priming condition (Cross-modal – Uni-modal)	5.799
SUBTLEX WF	1.673
Visuo-orthographic PC	1.583
Concept familiarity	1.168
Image typicality	1.037
Image visual PC1	1.026
Image visual PC2	1.026
Image visual PC3	1.069
Trial accuracy (Correct – Incorrect)	1.007
Greene x Matching condition	1.078

Greene x Target modality	1.077
Matching condition x Target modality	1.000
Greene x Priming condition	1.077
Matching condition x Priming condition	1.000
Priming condition x Target modality	1.004
SUBTLEX x Matching condition	1.078
SUBTLEX x Target modality	1.077
SUBTLEX x Priming condition	1.077
Greene x Matching condition x Target modality	1.077
Greene x Matching condition x Priming condition	1.078
Greene x Target modality x Priming condition	1.077
Matching condition x Priming condition x Target modality	1.000
SUBTLEX x Matching condition x Target modality	1.077
SUBTLEX x Matching condition x Priming condition	1.077
SUBTLEX x Target modality x Priming condition	1.077
Greene x Matching condition x Priming condition x Target modality	1.077
SUBTLEX x Matching condition x Priming condition x Target modality	1.077
Target repetition	5.857

The model showed moderate collinearity (VIFs = 5.8 and 5.9) between the Priming condition and Target repetition. This was expected because, despite counterbalancing block order for modalities (word-object or object-word) across participants, all participants performed the cross-modal blocks before the uni-modal blocks (and after Experiment 1). We kept the term for further analysis since collinearity was only just above the threshold for these terms, and because we deemed it important to account for potential carry-over effect. The measured SUBTLEX WF and Greene OF effects were independent of visual and visuo-orthographic information of the stimuli, as well as of image typicality, subjective familiarity, target repetition, and accuracy of categorization.

Supplementary figure 8 - Raw RTs from Experiment 2

Raw response times for object (red) and word (blue) trials in the priming conditions (Cross-modal solid lines, Uni-modal: dashed-dotted) and matching condition (Matching on the left, Mismatching on the right), as a function of SUBTLEX frequency (top) and Greene frequency (bottom) in Experiment 2. Points show concepts with different level of frequency, averaged across participants; lines represent linear fitting of points and shaded areas represent 95 % confidence interval



Supplementary Materials 8 – Post-hoc of interactions in Experiment 2

Recorded factor is a factor we obtained merging *Priming condition* and *Matching condition* to explore the interaction between frequency x Priming condition x Matching condition. This new factor has 4 levels (*Cross-modal Matching*, *Uni-modal Matching*, *Cross-modal Mismatching*, *Uni-modal*

Mismatching) and 3 contrasts of interest are computed (*Cross-modal Matching – Uni-modal Matching*, *Cross-modal Mismatching – Uni-modal Mismatching*, *Cross-modal Matching – Uni-modal Mismatching*)

$$\begin{aligned} \text{Exp2_logRT} \sim & \text{SUBTLEX WF} * \text{Recoded factor} * \text{Target modality} + \\ & \text{Greene OF} * \text{Recoded factor} * \text{Target modality} + \\ & \text{Concept familiarity} + \text{Image typicality} + \\ & \text{Image visual PC1} + \text{Image visual PC2} + \text{Image visual PC3} + \\ & \text{Visuo-orthographic PC} + \text{Target repetition} + \text{Trial accuracy} + \\ & (1|\text{Participants}) + (1|\text{Concepts}) \end{aligned}$$

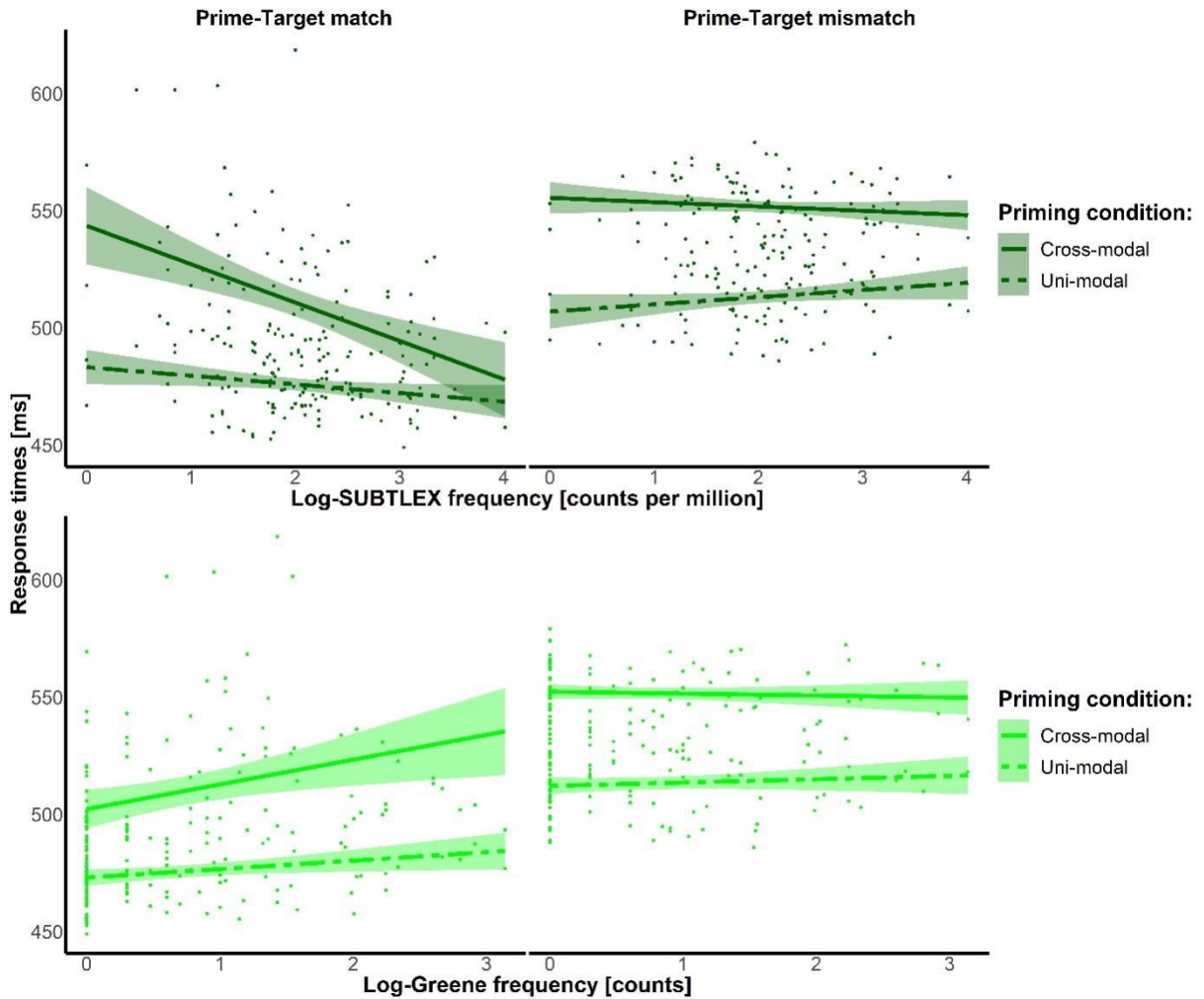
Supplementary table 10. Results from the post-hoc model with re-coded contrasts

Predictors	β	SE	t	p
(Intercept)	6.225	0.020	315.730	<0.001
SUBTLEX WF	-0.004	0.002	-1.812	0.070
Cross-modal matching – Uni-modal matching	0.005	0.006	0.805	0.421
Cross-modal mismatching – Uni-modal mismatching	0.008	0.006	1.360	0.174
Cross-modal matching – Cross-modal mismatching	-0.075	0.003	-23.733	<0.001
Target modality (Words – Objects)	-0.008	0.002	-3.618	<0.001
Greene OF	0.008	0.002	4.474	<0.001
Visuo-orthographic PC	0.010	0.002	4.877	<0.001
Concept familiarity	-0.001	0.002	-0.845	0.398
Image typicality	-0.004	0.002	-2.562	0.010
Image visual PC1	0.002	0.002	1.217	0.224
Image visual PC2	0.004	0.002	2.788	0.005
Image visual PC3	-0.001	0.002	-0.832	0.405
Target repetition	-0.036	0.003	-13.357	<0.001
Trial accuracy (Correct – Incorrect)	0.030	0.005	5.444	<0.001

SUBTLEX x (Cross-modal matching – Uni-modal matching)	-0.023	0.003	-7.094	<0.001
SUBTLEX x (Cross-modal mismatching – Uni-modal mismatching)	-0.006	0.003	-1.870	0.062
SUBTLEX x (Cross-modal matching – Cross-modal mismatching)	-0.029	0.003	-9.027	<0.001
SUBTLEX x Target modality	-0.005	0.002	-2.378	0.017
(Cross-modal matching – Uni-modal matching) x Target modality	-0.010	0.006	-1.655	0.098
(Cross-modal mismatching – Uni-modal mismatching) x Target modality	0.036	0.006	5.717	<0.001
(Cross-modal matching – Cross-modal mismatching) x Target modality	-0.015	0.006	-2.320	0.020
Greene x (Cross-modal matching – Uni-modal matching)	0.018	0.003	5.379	<0.001
Greene x (Cross-modal mismatching – Uni-modal mismatching)	-0.002	0.003	-0.644	0.520
Greene x (Cross-modal matching – Cross-modal mismatching)	0.027	0.003	8.276	<0.001
Greene x Target modality	0.003	0.002	1.396	0.163
SUBTLEX x (Cross-modal matching – Uni-modal matching) x Target modality	-0.000	0.007	-0.031	0.975
SUBTLEX x (Cross-modal mismatching – Uni-modal mismatching) x Target modality	0.005	0.007	0.834	0.404
SUBTLEX x (Cross-modal matching – Cross-modal mismatching) x Target modality	-0.001	0.007	-0.099	0.921
Greene x (Cross-modal matching – Uni-modal matching) x Target modality	0.005	0.007	0.842	0.400
Greene x (Cross-modal mismatching – Uni-modal mismatching) x Target modality	0.008	0.007	1.161	0.246
Greene x (Cross-modal matching – Cross-modal mismatching) x Target modality	-0.004	0.007	-0.632	0.527

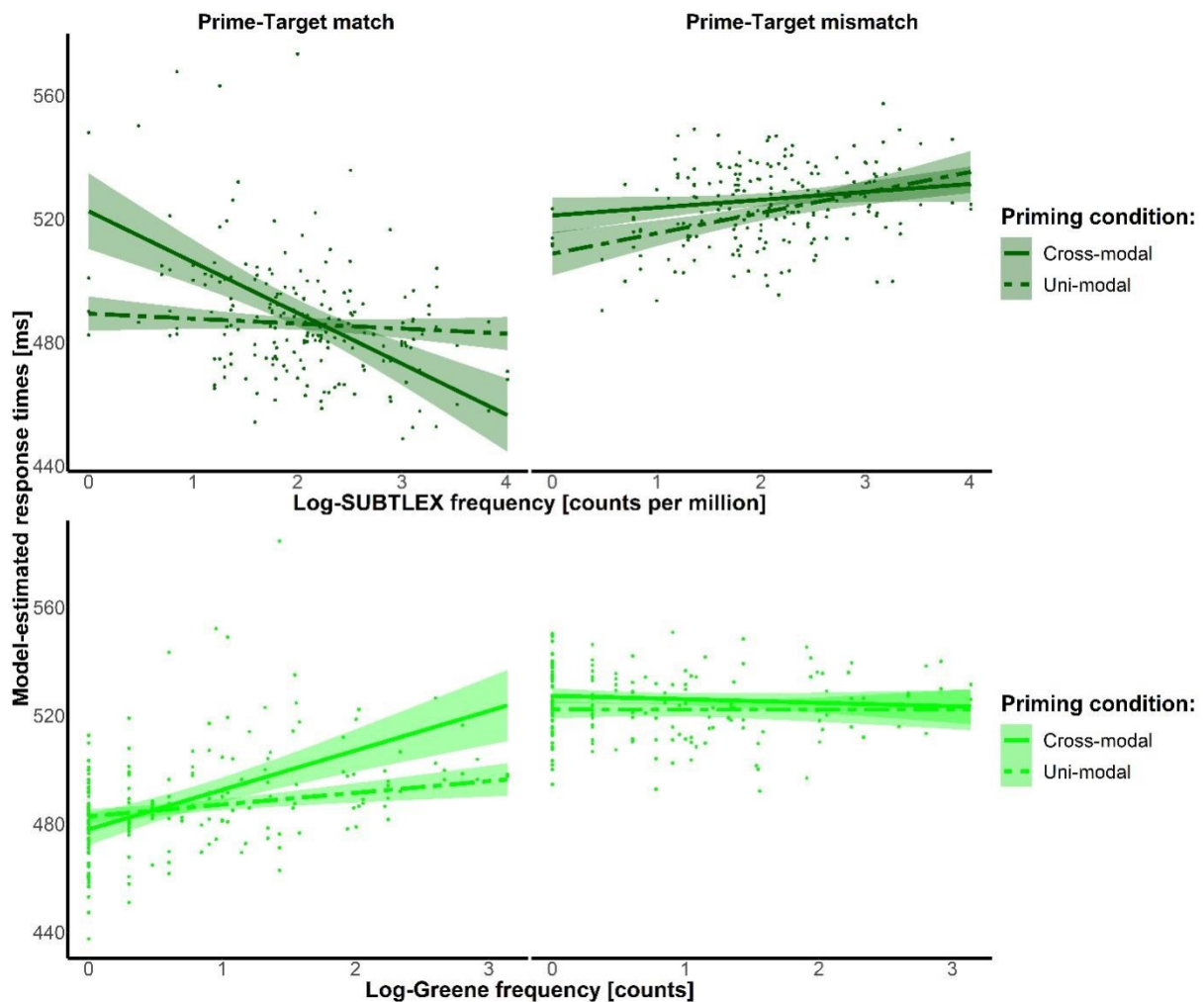
Supplementary figure 9. Raw RTs from post-hoc conditions of Experiment 2

Raw response times in the priming conditions (Cross-modal: solid lines, Uni-modal: dashed-dotted) and matching condition (Matching on the left, Mismatching on the right), as a function of SUBTLEX frequency (top, dark green) and Greene frequency (bottom, light green) in Experiment 2. Points show concepts with different level of frequency, averaged across participants; lines represent linear fitting of points and shaded areas represent 95 % confidence interval



Supplementary figure 10 - RTs estimated from post-hoc of interaction effects of Experiment 2 (between condition)

Response times as a function of logarithmic SUBTLEX frequency (top plots, dark green) and Greene frequency (bottom plots, light green) in the 3-way significant interaction with Matching condition and Priming condition (Cross-modal matching vs. Uni-modal matching; Cross-modal mismatching vs. Uni-modal mismatching; Cross-modal matching vs. Cross-modal mismatching). RTs were estimated based on the selected model. Points present participant-based mean response times for concepts in the different frequency levels. Lines represent linear fitting of points (solid: cross-modal; dashed: uni-modal), and shaded areas represent 95 % confidence interval. Bottom left and top-left plots represent the effects in prime-target matching condition, while bottom-right and top-right plots represent the effects in prime-target mismatching condition



4 post-hoc models are additionally computed, one for every level of the re-coded factor (Cross-modal Matching, Uni-modal Matching, Cross-modal Mismatching, Uni-modal Mismatching)

Exp2_logRT ~ SUBTLEX WF * Target modality + Greene OF * Target modality +
 Concept familiarity + Image typicality +
 Image visual PC1 + Image visual PC2 + Image visual PC3 +
 Visuo-orthographic PC + Target repetition + Trial accuracy +
 (1|Participants) + (1|Concepts)

Supplementary table 11. Results from the post-hoc individual models for conditions of interest

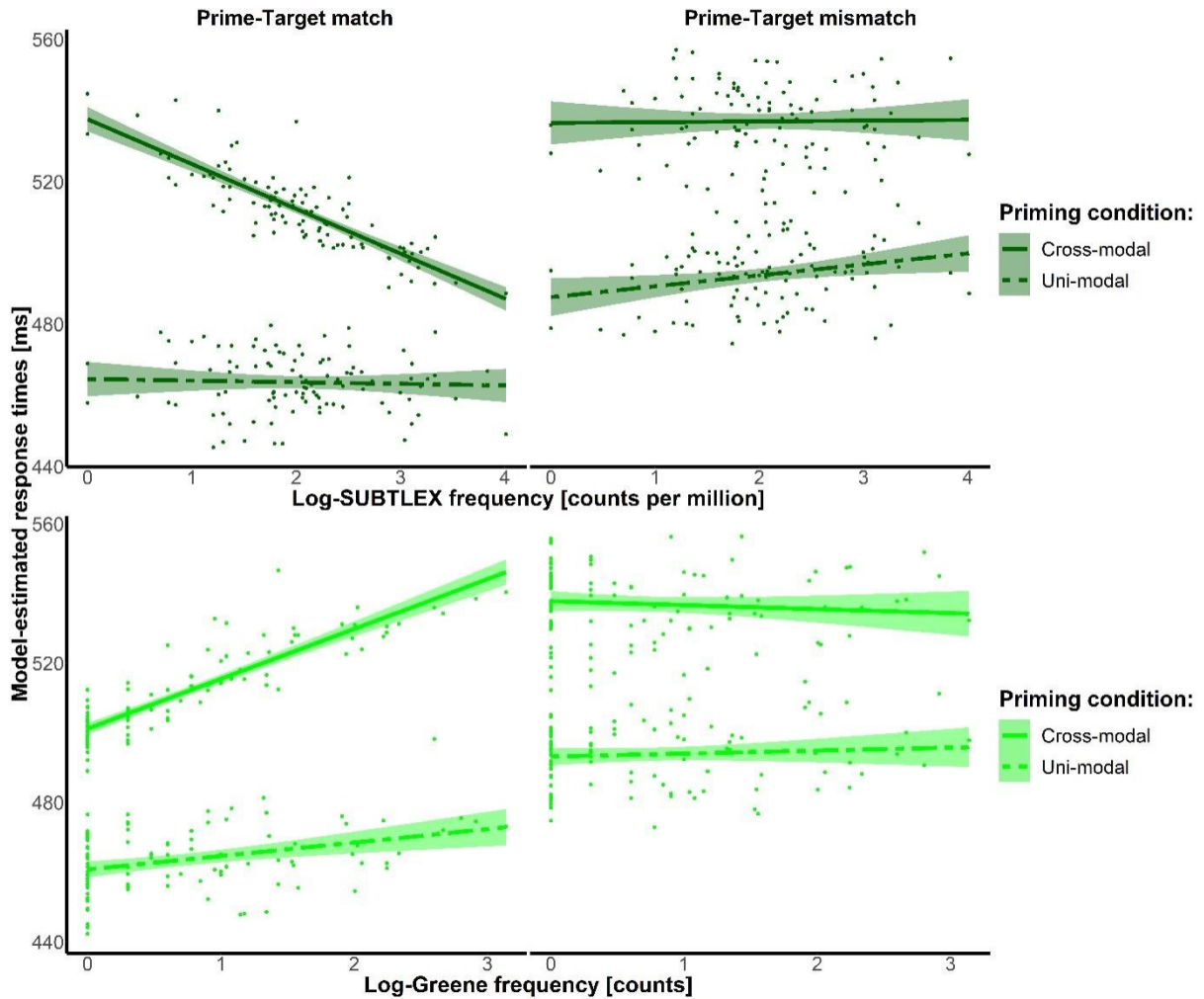
<i>Predictors</i>	Uni-modal Matching				Cross-modal Matching			
	β	<i>SE</i>	<i>t</i>	<i>p</i>	β	<i>SE</i>	<i>t</i>	<i>p</i>
(Intercept)	6.139	0.020	309.972	<0.001	6.238	0.021	290.842	<0.001
SUBTLEX WF	-0.001	0.003	-0.245	0.807	-0.019	0.006	-3.230	0.001
Target modality (Words – Objects)	0.008	0.004	1.790	0.073	-0.012	0.005	-2.620	0.009
Greene OF	0.007	0.003	2.725	0.006	0.023	0.005	4.710	<0.001
Visuo-orthographic PC	0.013	0.003	4.418	<0.001	0.018	0.006	3.077	0.002
Concept familiarity	-0.000	0.003	-0.150	0.881	-0.003	0.003	-0.858	0.391
Image typicality	-0.005	0.003	-1.835	0.067	-0.013	0.003	-3.825	<0.001
Image visual PC1	0.003	0.002	1.289	0.197	0.003	0.005	0.593	0.553
Image visual PC2	0.004	0.002	1.758	0.079	0.001	0.005	0.254	0.799
Image visual PC3	-0.003	0.002	-1.173	0.241	0.001	0.005	0.178	0.859
Target repetition	-0.002	0.002	-0.792	0.428	-0.028	0.002	-12.095	<0.001
Trial accuracy (Correct – Incorrect)	0.058	0.010	6.001	<0.001	-0.008	0.010	-0.760	0.448
SUBTLEX x (Words – Objects)	-0.004	0.004	-0.982	0.326	-0.005	0.005	-0.948	0.343
Greene x (Words – Objects)	-0.001	0.004	-0.304	0.761	0.005	0.005	1.012	0.311

<i>Predictors</i>	Uni-modal Mismatching				Cross-modal Mismatching			
	β	<i>SE</i>	<i>t</i>	<i>p</i>	β	<i>SE</i>	<i>t</i>	<i>p</i>
(Intercept)	6.202	0.020	312.978	<0.001	6.286	0.022	279.822	<0.001
SUBTLEX WF	0.005	0.003	1.535	0.125	0.000	0.003	0.108	0.914
Target modality (Words – Objects)	-0.026	0.004	-5.943	<0.001	0.004	0.004	0.863	0.388
Greene OF	0.001	0.003	0.543	0.587	-0.002	0.002	-0.733	0.463
Visuo-orthographic PC	0.001	0.003	0.407	0.684	0.005	0.003	1.978	0.048
Concept familiarity	-0.003	0.003	-1.225	0.221	0.002	0.003	0.586	0.558
Image typicality	0.001	0.003	0.221	0.825	0.003	0.003	1.109	0.268
Image visual PC1	0.000	0.002	0.024	0.981	0.002	0.002	0.751	0.452
Image visual PC2	0.008	0.002	3.082	0.002	0.005	0.002	2.194	0.028
Image visual PC3	-0.001	0.003	-0.322	0.748	-0.003	0.002	-1.172	0.241
Target repetition	-0.007	0.002	-3.341	0.001	-0.023	0.002	-10.344	<0.001
Trial accuracy (Correct – Incorrect)	0.083	0.012	6.712	<0.001	0.063	0.012	5.361	<0.001

SUBTLEX x (Words – Objects)	-0.009 0.004	-2.113 0.035	-0.004 0.005	-0.951 0.342
Greene x (Words – Objects)	0.001 0.004	0.226 0.821	0.009 0.005	2.027 0.043

Supplementary figure 11 - RTs estimated from post-hoc models of Experiment 2 (within conditions)

Effects of SUBTLEX WF (dark green, top) and Greene OF (light green, bottom) on reaction times estimated from the post-hoc models separately for each Priming condition (continuous and dashed-dotted line types) and Matching condition (left and right plots). Points show concepts with different level of frequency, averaged across participants; lines represent linear fitting of points and shaded areas represent 95 % confidence interval.



Supplementary materials 9 – Results of new ratings on original data of Exp 2

*Exp2_logRT ~ SUBTLEX WF * Priming condition * Matching condition * Target modality +
Greene OF * Priming condition * Matching condition * Target modality +
Concept familiarity (replication) + Image typicality (replication) +
Image visual PC1 + Image visual PC2 + Image visual PC3 +
Visuo-orthographic PC + Target repetition + Trial accuracy +
(1|Participants) + (1|Concepts)*

Supplementary table 13. Results from main model of Exp 2 including ratings from replication study

Predictors	β	SE	t	p
(Intercept)	6.225	0.020	315.766	<0.001
Greene OF	0.005	0.002	2.978	0.003
Matching condition (Mismatch – Match)	0.073	0.002	32.683	<0.001
Target modality (Words – Objects)	-0.008	0.002	-3.613	<0.001
Priming condition (Cross-modal – Uni-modal)	0.006	0.005	1.175	0.240
SUBTLEX WF	-0.001	0.002	-0.576	0.565
Visuo-orthographic PC	0.010	0.002	5.485	<0.001
Concept familiarity (replication)	-0.002	0.002	-0.795	0.427
Image typicality (replication)	-0.008	0.002	-4.829	<0.001
Image visual PC1	0.003	0.002	2.077	0.038
Image visual PC2	0.003	0.002	2.173	0.030
Image visual PC3	-0.002	0.002	-1.518	0.129
Target repetition	-0.036	0.003	-13.355	<0.001
Trial accuracy (Correct – Incorrect)	0.030	0.005	5.453	<0.001
Greene x Matching condition	-0.017	0.002	-7.458	<0.001
Greene x Target modality	0.003	0.002	1.396	0.163
Matching condition x Target modality	-0.009	0.004	-1.938	0.053
Greene x Priming condition	0.008	0.002	3.347	0.001
Matching condition x Priming condition	0.003	0.004	0.736	0.462
Priming condition x Target modality	0.013	0.004	2.874	0.004
SUBTLEX x Matching condition	0.021	0.002	9.075	<0.001
SUBTLEX x Target modality	-0.005	0.002	-2.376	0.018
SUBTLEX x Priming condition	-0.015	0.002	-6.335	<0.001
Greene x Matching condition x Target modality	0.003	0.005	0.666	0.505

Greene x Matching condition x Priming condition	-0.020	0.005	-4.257	< 0.001
Greene x Priming condition x Target modality	0.007	0.005	1.412	0.158
Matching condition x Priming condition x Target modality	0.046	0.009	5.216	< 0.001
SUBTLEX x Matching condition x Target modality	-0.002	0.005	-0.473	0.636
SUBTLEX x Matching condition x Priming condition	0.017	0.005	3.687	< 0.001
SUBTLEX x Priming condition x Target modality	0.003	0.005	0.569	0.569
Greene x Matching condition x Priming condition x Target modality	0.002	0.009	0.228	0.819
SUBTLEX x Matching condition x Priming condition x Target modality	0.006	0.009	0.612	0.540

Supplementary materials 10 – Exploratory analysis in Experiment 2

Since we detected similar frequency effects in both word and object modalities (i.e., no significant interaction including target modality), as well as the presence of these effects only when semantic processing is required (cross-modal trials), we decided to further explore whether the frequency effects found in Experiment 2 represented common semantic processing of objects and words. Thus, we restricted this exploratory analysis to the *Cross-modal matching trials*. Response times from this subset showed substantial word and object frequency effects and, at the same time, included predictive semantic processing to a high degree.

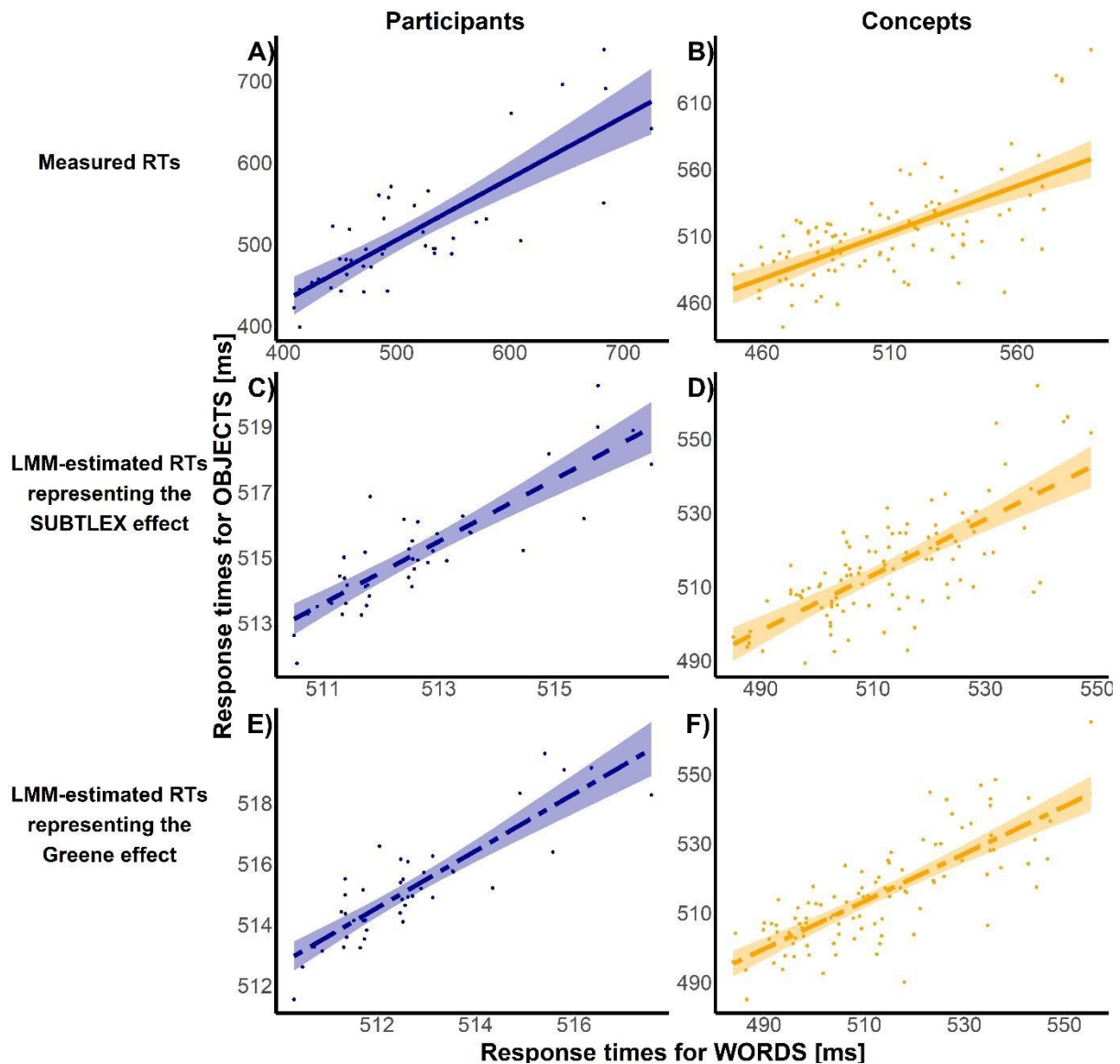
We considered three sources of data: 1) the actual response times from cross-modal matching trials of words and objects; 2) response times estimated from the effect of SUBTLEX WF in cross-modal matching trials of words and objects; 3) response times estimated from the effect of Greene OF in cross-modal matching trials of words and objects. 2) and 3) were estimated using two models (one for word trials and one for object trials) that included SUBTLEX WF, Greene OF, and all the covariates and random effects of the main model of Experiment 2 introduced before. The comparison between word and object processing was made for each of the three datasets considering response times for every participant and for every concept in both modalities. To test the similarity between frequency effects in words and objects, we implemented *paired-samples equivalence tests* and *product-moment correlation tests* between response times from word and object trials.

With the equivalence test, we can check if two samples/conditions come from the same distribution (i.e., they are equivalent). Thus, we computed test statistics for which low probability values allow us to reject the null hypothesis of statistical difference (instead of rejecting the null hypothesis of statistical equivalence of commonly used t-test). For the equivalence test, we needed to set an epsilon parameter, i.e., the maximally allowed difference to consider two conditions non-different; in our case, we used 50 % of the standard deviation of the difference between object and word trials (Robinson & Froese, 2004). The correlation of word-based vs. object-based reaction times could additionally prove whether associated entries show similar behavior.

For actual response time data, we found a significant equivalence (mean of differences = 0.005 log(ms), $\epsilon=0.045$ log(ms), CI = [-0.018 0.028], $p=0.003$) and correlation ($r=0.804$, $t(40)=8.554$, $p<0.001$), between *participants' performance* in object and word trials; also, we found a significant equivalence (mean of differences = 0.006 log(ms), $\epsilon=0.028$ log(ms), CI = [-0.004 0.015], $p<0.001$) and correlation ($r=0.652$, $t(98)=8.503$, $p<0.001$) between *processing of concepts* in the two different modalities (*Supplementary Figure 12A-B*). That implies high interrelation between the processing objects and words which becomes evident when comparing participants and comparing stimuli with the same semantics.

Supplementary figure 12 - Linear relationship between object trials and word trials in Cross-modal matching trials

Correlations among unique participants (dark blue: A, C, and E) and unique concepts (orange: B, D and F). A, B) Actual response times for Cross-modal matching trials (solid lines). C, D) Response times estimated from SUBTLEX WF effect in Cross-modal matching trials (dashed lines); E, F) Response times estimated from Greene OF effect in Cross-modal matching trials (dashed-dotted lines). Points represent performance of individual participants or concepts in the two tasks. Lines represent linear fitting of points, and shaded areas represent 95 % confidence interval.



To get an estimate to which degree the interrelation was driven by the WF effect, we predicted RTs that were influenced by the SUBTLEX WF effect without confounds based on the estimated models (one for cross-modal matching trials of objects, one for cross-modal matching trials for words). For *participants' performance*, we could not reject statistical difference (mean of differences = 0.0049 log(ms), $\epsilon=0.0009$ log(ms), CI = [0.0044 0.0053], $p=1$), but we found a significant correlation ($r=0.860$, $t(40)=10.667$, $p<0.001$) between object trials and word trials; similarly, but *considering single concepts*, we could reject statistical difference (mean of differences = 0.005 log(ms), $\epsilon=0.010$ log(ms), CI = [0.001 0.008],

$p=0.004$), and we found also a significant correlation ($r=0.717$, $t(98)=10.186$, $p<0.001$) (*Supplementary Figure 12C-D*).

We repeated the same procedure for the Greene OF effect and found the same pattern: statistical difference could not be reject for individual participants (mean of differences = $0.0048 \log(\text{ms})$, $\varepsilon=0.0009 \log(\text{ms})$, $\text{CI} = [0.0044 \ 0.0053]$, $p=1$), but it was rejected for individual concepts performance (mean of differences = $0.005 \log(\text{ms})$, $\varepsilon=0.010 \log(\text{ms})$, $\text{CI} = [0.001 \ 0.008]$, $p<0.001$), while both showed a strong correlation between object and word trials (participants: $r=0.868$, $t(40)=11.067$, $p<0.001$; concepts: $r=0.779$, $t(98)=12.29$, $p<0.001$) (*Supplementary Figure 12E-F*). Overall, these exploratory analyses show that even if the WF and OF do not affect object and word processing completely identically, the individual participant's frequency effects for words and object and the frequency effects for single semantic concepts are strongly associated to each other across modalities.

Supplementary Materials 11 – Effect of Greene OF with Conceptual Distinctiveness

We can draw parallels between the experimental visual experience created and tested by Konkle and colleagues (2010) (i.e., manipulating the frequency of visually presented objects in the lab) and what the Greene OF used in our study aims to represent (i.e., the frequency of visually encountered objects in the real world). This comparison might raise some concerns since the two studies seem relatively different at first sight: First, Konkle et al. artificially induced memory interference and second, they specifically measured visual LTM. That said, we believe that Konkle et al. (2010) of course aimed at measuring a phenomenon of memory interference that they think is happening intrinsically when encountering objects in the world. While Konkle's task required retrieving specific exemplars, our task required retrieving a concept (i.e., the prime meaning) from memory. For both tasks, however, interferences from other exemplars are similarly possible. In addition, to correctly perform the Cross-modal priming task in our study, participants in our study had to access representations in semantic long-term memory (LTM), which was also the locus of the memory interferences as highlighted by Konkle et al. (2010). This is in line with the observation that the Greene OF effect in our study only came into play in Cross-modal Matching trials, where semantic processing and LTM involvement was particularly high.

Recoded factor is a factor we obtained merging *Priming condition* and *Matching condition* to explore interaction between frequency x Priming condition x Matching condition. This new factor has 4 levels

(Cross-modal Matching, Uni-modal Matching, Cross-modal Mismatching, Uni-modal Mismatching)
and 3 contrasts of interest are computed (Cross-modal Matching – Uni-modal Matching, Cross-modal
Mismatching – Uni-modal Mismatching, Cross-modal Matching – Uni-modal Mismatching)

Exp2_logRT ~ *SUBTLEX WF* * *Recoded factor* * *Target modality* +
Greene OF * *Conceptual Distinctiveness* * *Recoded factor* * *Target modality* +
Concept familiarity + *Image typicality* +
Image visual PC1 + *Image visual PC2* + *Image visual PC3* +
Visuo-orthographic PC + *Target repetition* + *Trial accuracy* +
(1|*Participants*) + (1|*Concepts*)

Supplementary table 14. Results from model including Conceptual Distinctiveness in interaction with Greene frequency.

Predictors	β	SE	t	p
(Intercept)	6.225	0.020	315.564	<0.001
Conceptual Distinctiveness (CD)	0.002	0.002	0.799	0.424
Greene OF	0.008	0.002	3.709	<0.001
Cross-modal matching – Uni-modal matching	0.009	0.006	1.533	0.125
Cross-modal mismatching – Uni-modal mismatching	0.007	0.006	1.235	0.217
Cross-modal matching – Cross-modal mismatching	-0.070	0.003	-20.276	<0.001
Target modality (Words – Objects)	-0.008	0.002	-3.356	0.001
SUBTLEX	-0.004	0.002	-1.923	0.055
Visuo-orthographic PC	0.009	0.002	4.807	<0.001
Concept familiarity	-0.001	0.002	-0.762	0.446
Image typicality	-0.004	0.002	-2.440	0.015
Image visual PC1	0.002	0.002	1.266	0.206
Image visual PC2	0.005	0.002	2.776	0.006
Image visual PC3	-0.002	0.002	-0.976	0.329
Target repetition	-0.036	0.003	-13.333	<0.001
Trial accuracy (Correct – Incorrect)	0.031	0.005	5.595	<0.001
Conceptual Distinctiveness (CD)x Greene	-0.001	0.002	-0.749	0.454
CD x (Cross-modal matching – Uni-modal matching)	0.004	0.004	1.051	0.293
CD x (Cross-modal mismatching – Uni-modal mismatching)	0.001	0.004	0.246	0.806
CD x (Cross-modal matching – Cross-modal mismatching)	0.007	0.004	1.937	0.053
Greene x (Cross-modal matching – Uni-modal matching)	0.021	0.004	5.364	<0.001
Greene x (Cross-modal mismatching – Uni-modal mismatching)	-0.003	0.004	-0.810	0.418
Greene x (Cross-modal matching – Cross-modal mismatching)	0.030	0.004	7.737	<0.001
Conceptual Distinctiveness (CD)x Target modality	-0.000	0.003	-0.065	0.948
Greene x Target modality	0.003	0.003	1.155	0.248
Target modality x (Cross-modal matching – Uni-modal matching)	-0.009	0.007	-1.252	0.211
Target modality x (Cross-modal mismatching – Uni-modal mismatching)	0.030	0.007	4.432	<0.001

Target modality x (Cross-modal matching – Cross-modal mismatching)	-0.013	0.007	-1.872	0.061
SUBTLEX x (Cross-modal matching – Uni-modal matching)	-0.023	0.004	-6.500	< 0.001
SUBTLEX x (Cross-modal mismatching – Uni-modal mismatching)	-0.007	0.004	-1.873	0.061
SUBTLEX x (Cross-modal matching – Cross-modal mismatching)	-0.030	0.004	-8.512	< 0.001
SUBTLEX x Target modality	-0.005	0.003	-2.181	0.029
CD x Greene x (Cross-modal matching – Uni-modal matching)	-0.010	0.003	-3.139	0.002
CD x Greene x (Cross-modal mismatching – Uni-modal mismatching)	0.002	0.003	0.495	0.621 -
CD x Greene x (Cross-modal matching – Cross-modal mismatching)	0.012	0.003	-3.774	< 0.001
CD x Greene x Target modality	0.000	0.002	0.086	0.932
CD x (Cross-modal matching – Uni-modal matching) x Target modality	-0.001	0.008	-0.072	0.942
CD x (Cross-modal mismatching – Uni-modal mismatching) x Target modality	0.001	0.008	0.129	0.898
CD x (Cross-modal matching – Cross-modal mismatching) x Target modality	-0.002	0.008	-0.275	0.784
Greene x (Cross-modal matching – Uni-modal matching) x Target modality	0.008	0.008	0.984	0.325
Greene x (Cross-modal mismatching – Uni-modal mismatching) x Target modality	0.001	0.008	0.143	0.886
Greene x (Cross-modal matching – Cross-modal mismatching) x Target modality	-0.002	0.008	-0.195	0.845
SUBTLEX x (Cross-modal matching – Uni-modal matching) x Target modality	0.001	0.007	0.078	0.938
SUBTLEX x (Cross-modal mismatching – Uni-modal mismatching) x Target modality	0.003	0.007	0.464	0.643
SUBTLEX x (Cross-modal matching – Cross-modal mismatching) x Target modality	0.001	0.007	0.083	0.934
CD x Greene x (Cross-modal matching – Uni-modal matching) x Target modality	-0.004	0.006	-0.666	0.506
CD x Greene x (Cross-modal mismatching – Uni-modal mismatching) x Target modality	0.013	0.006	2.031	0.042
CD x Greene x (Cross-modal matching – Cross-modal mismatching) x Target modality	-0.004	0.006	-0.631	0.528

Supplementary materials 12 – Results main model in priming task (Exp 3)

*Exp3_logRT ~ SUBTLEX WF * Priming condition * Matching condition * Target modality +
Greene OF * Priming condition * Matching condition * Target modality +
Concept familiarity (replication) + Image typicality (replication) +
Image visual PC1 + Image visual PC2 + Image visual PC3 +
Visuo-orthographic PC + Target repetition + Trial accuracy +
(1|Participants) + (1|Concepts)*

Supplementary table 15. Results from the main model for priming task replication

Predictors	β	SE	<i>t</i>	<i>p</i>
(Intercept)	6.354	0.017	374.088	< 0.001
Greene OF	0.003	0.002	1.661	0.097
Matching condition (Mismatch – Match)	0.062	0.002	27.489	< 0.001
Target modality (Words – Objects)	-0.002	0.002	-0.896	0.370
Priming condition (Cross-modal – Uni-modal)	0.014	0.033	0.413	0.680
SUBTLEX WF	-0.008	0.002	-3.932	< 0.001
Visuo-orthographic PC	0.005	0.002	2.489	0.013
Concept familiarity (replication)	-0.001	0.002	-0.347	0.729
Image typicality (replication)	-0.009	0.002	-5.229	< 0.001
Image visual PC1	0.001	0.002	0.633	0.527
Image visual PC2	0.003	0.002	1.929	0.054
Image visual PC3	0.001	0.002	0.592	0.554
Target repetition	-0.036	0.001	-31.062	< 0.001
Trial accuracy (Correct – Incorrect)	0.013	0.007	1.954	0.051
Greene x Matching condition	-0.008	0.002	-3.529	< 0.001
Greene x Target modality	0.001	0.002	0.371	0.711
Matching condition x Target modality	0.001	0.004	0.195	0.845
Greene x Priming condition	0.012	0.002	5.158	< 0.001
Priming condition x Matching condition	-0.016	0.004	-3.523	< 0.001
Priming condition x Target modality	-0.029	0.005	-6.285	< 0.001
SUBTLEX x Matching condition	0.013	0.002	5.638	< 0.001
SUBTLEX x Target modality	-0.011	0.002	-4.830	< 0.001
SUBTLEX x Priming condition	-0.010	0.002	-4.291	< 0.001
Greene x Matching condition x Target modality	-0.004	0.005	-0.917	0.359
Greene x Matching condition x Priming condition	-0.010	0.005	-2.165	0.030
Greene x Priming condition x Target modality	0.000	0.005	0.019	0.985
Matching condition x Priming condition x Target modality	0.030	0.009	3.406	0.001

SUBTLEX x Matching condition x Target modality	0.016	0.005	3.399	0.001
SUBTLEX x Matching condition x Priming condition	0.009	0.005	1.880	0.060
SUBTLEX x Priming condition x Target modality	-0.006	0.005	-1.230	0.219
Greene x Matching condition x Priming condition x Target modality	-0.011	0.009	-1.169	0.242
SUBTLEX x Matching condition x Priming condition x Target modality	0.021	0.009	2.269	0.023

Supplementary table 16. Variance Inflation Factors for the effects of the main model of Replication experiment.

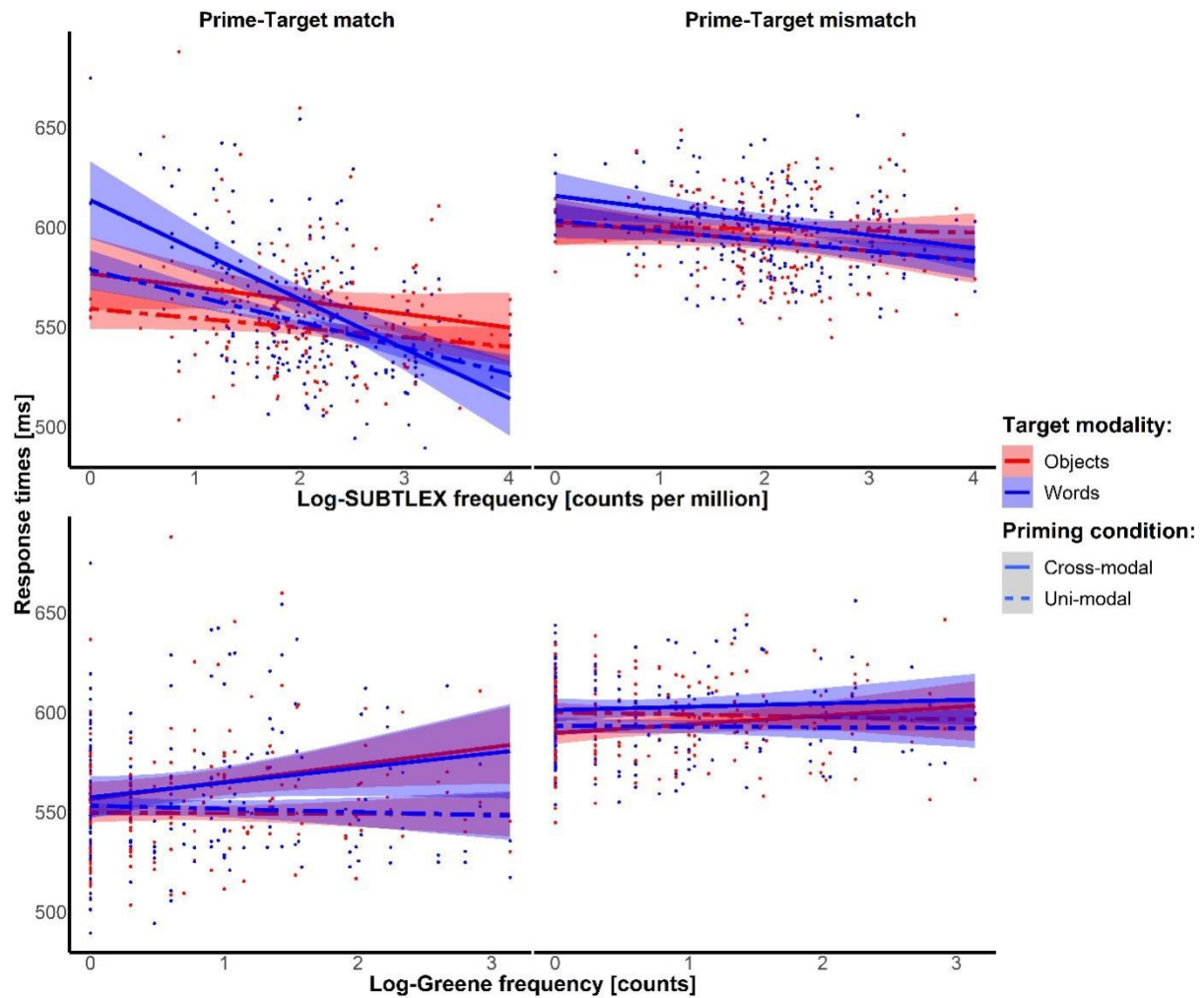
Term	VIF
Greene OF	1.535
Matching condition	1.022
Target modality	1.003
Priming condition	1.000
SUBTLEX WF	1.974
Visuo-orthographic PC	1.725
Concept familiarity (replication)	1.705
Image typicality (replication)	1.337
Image visual PC1	1.084
Image visual PC2	1.093
Image visual PC3	1.206
Target repetition	1.097
Trial accuracy	1.006
Greene x Matching condition	1.076
Greene x Target modality	1.076
Matching condition x Priming condition	1.000
Greene x Priming condition	1.076
Matching condition x Target modality	1.000
Priming condition x Target modality	1.077
SUBTLEX x Matching condition	1.077
SUBTLEX x Target modality	1.076
SUBTLEX x Priming condition	1.076

Greene x Matching condition x Target modality	1.076
Greene x Matching condition x Priming condition	1.076
Greene x Priming condition x Target modality	1.076
Matching condition x Priming condition x Target modality	1.000
SUBTLEX x Matching condition x Target modality	1.076
SUBTLEX x Matching condition x Priming condition	1.076
SUBTLEX x Priming condition x Target modality	1.076
Greene x Matching condition x Priming condition x Target modality	1.076
SUBTLEX x Matching condition x Priming condition x Target modality	1.076

The measured SUBTLEX WF and Greene OF effects were independent of visual and visuo-orthographic information of the stimuli, as well as of image typicality, subjective familiarity, target repetition and accuracy of categorization.

Supplementary figure 13 - Raw RTs from Experiment 3

Raw response times for object (red) and word (blue) trials in the priming conditions (Cross-modal solid lines, Uni-modal: dashed-dotted) and matching condition (Matching on the left, Mismatching on the right), as a function of SUBTLEX frequency (top) and Greene frequency (bottom) in the Replication experiment. Points show concepts with different level of frequency, averaged across participants; lines represent linear fitting of points and shaded areas represent 95 % confidence interval



Supplementary Materials 13 – Post-hoc of interactions in priming task (Exp 3)

Recoded factor is a factor we obtained merging *Priming condition* and *Matching condition* to explore the interaction between frequency x Priming condition x Matching condition. This new factor has 4 levels (*Cross-modal Matching*, *Uni-modal Matching*, *Cross-modal Mismatching*, *Uni-modal Mismatching*) and 3 contrasts of interest are computed (*Cross-modal Matching – Uni-modal Matching*, *Cross-modal Mismatching – Uni-modal Mismatching*, *Cross-modal Matching – Uni-modal Mismatching*)

Exp3_logRT ~ *SUBTLEX WF* * *Recoded factor* * *Target modality* +
Greene OF * *Recoded factor* * *Target modality* +
Concept familiarity (replication) + *Image typicality (replication)* +
Image visual PC1 + *Image visual PC2* + *Image visual PC3* +
Visuo-orthographic PC + *Target repetition* + *Trial accuracy* +
(1|*Participants*) + (1|*Concepts*)

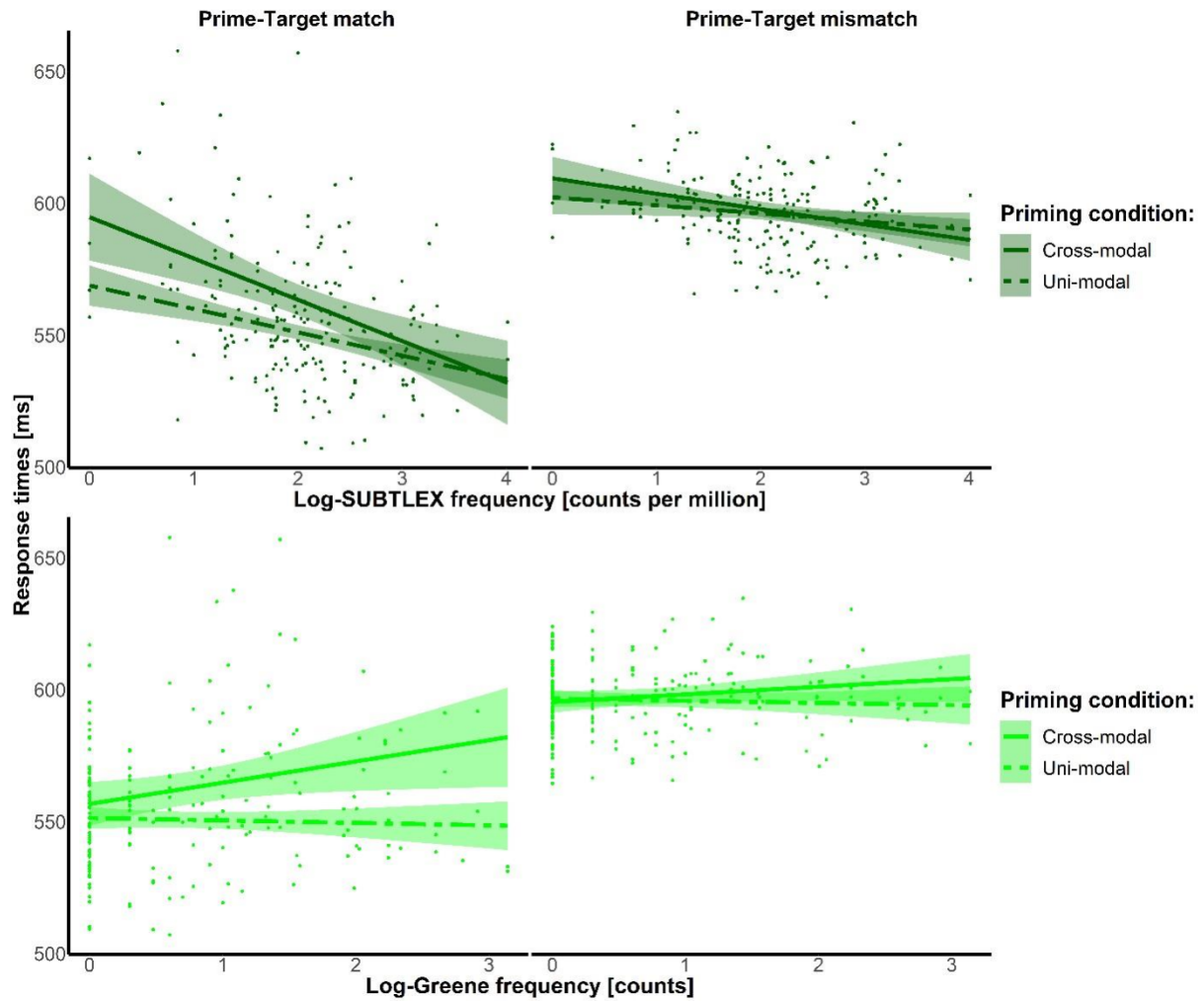
Supplementary table 17. Results from the post-hoc model with re-coded contrasts in the Replication exp **Predictors**

	β	SE	t	p
(Intercept)	6.355	0.017	374.337	<0.001
SUBTLEX WF	-0.008	0.002	-3.865	<0.001
Cross-modal matching – Uni-modal matching	0.022	0.033	0.649	0.517
Cross-modal mismatching – Uni-modal mismatching	0.006	0.033	0.175	0.861
Cross-modal matching – Cross-modal mismatching	-0.063	0.003	-19.787	<0.001
Target modality (Words – Objects)	0.001	0.002	0.614	0.539
Greene OF	0.003	0.002	1.636	0.102
Visuo-orthographic PC	0.005	0.002	2.465	0.014
Concept familiarity (replication)	-0.001	0.002	-0.309	0.758
Image typicality (replication)	-0.009	0.002	-5.220	<0.001
Image visual PC1	0.001	0.002	0.557	0.577
Image visual PC2	0.003	0.002	1.881	0.060
Image visual PC3	0.001	0.002	0.618	0.537
Trial accuracy (Correct – Incorrect)	0.009	0.007	1.317	0.188

SUBTLEX x (Cross-modal matching – Uni-modal matching)	-0.014	0.003	-4.324	<0.001
SUBTLEX x (Cross-modal mismatching – Uni-modal mismatching)	-0.006	0.003	-1.664	0.096
SUBTLEX x (Cross-modal matching – Cross-modal mismatching)	-0.018	0.003	-5.291	<0.001
SUBTLEX x Target modality	-0.011	0.002	-4.786	<0.001
Target modality x (Cross-modal matching – Uni-modal matching)	-0.006	0.006	-0.949	0.343
Target modality x (Cross-modal mismatching – Uni-modal mismatching)	0.025	0.006	3.872	<0.001
Target modality x (Cross-modal matching – Cross-modal mismatching)	-0.017	0.006	-2.591	0.010
Greene x (Cross-modal matching – Uni-modal matching)	0.017	0.003	5.165	<0.001
Greene x (Cross-modal mismatching – Uni-modal mismatching)	0.007	0.003	1.959	0.050
Greene x (Cross-modal matching – Cross-modal mismatching)	0.013	0.003	3.874	<0.001
Greene x Target modality	0.001	0.002	0.344	0.731
SUBTLEX x (Cross-modal matching – Uni-modal matching) x Target modality	-0.016	0.007	-2.401	0.016
SUBTLEX x (Cross-modal mismatching – Uni-modal mismatching) x Target mod	0.005	0.007	0.794	0.427
SUBTLEX x (Cross-modal matching – Cross-modal mismatching) x Target mod	-0.027	0.007	-4.008	<0.001
Greene x (Cross-modal matching – Uni-modal matching) x Target modality	0.005	0.007	0.818	0.414
x (Cross-modal mismatching – Uni-modal mismatching) x Target modality	-0.006	0.007	-0.871	0.384
Greene x (Cross-modal matching – Cross-modal mismatching) x Target modality	0.010	0.007	1.487	0.137

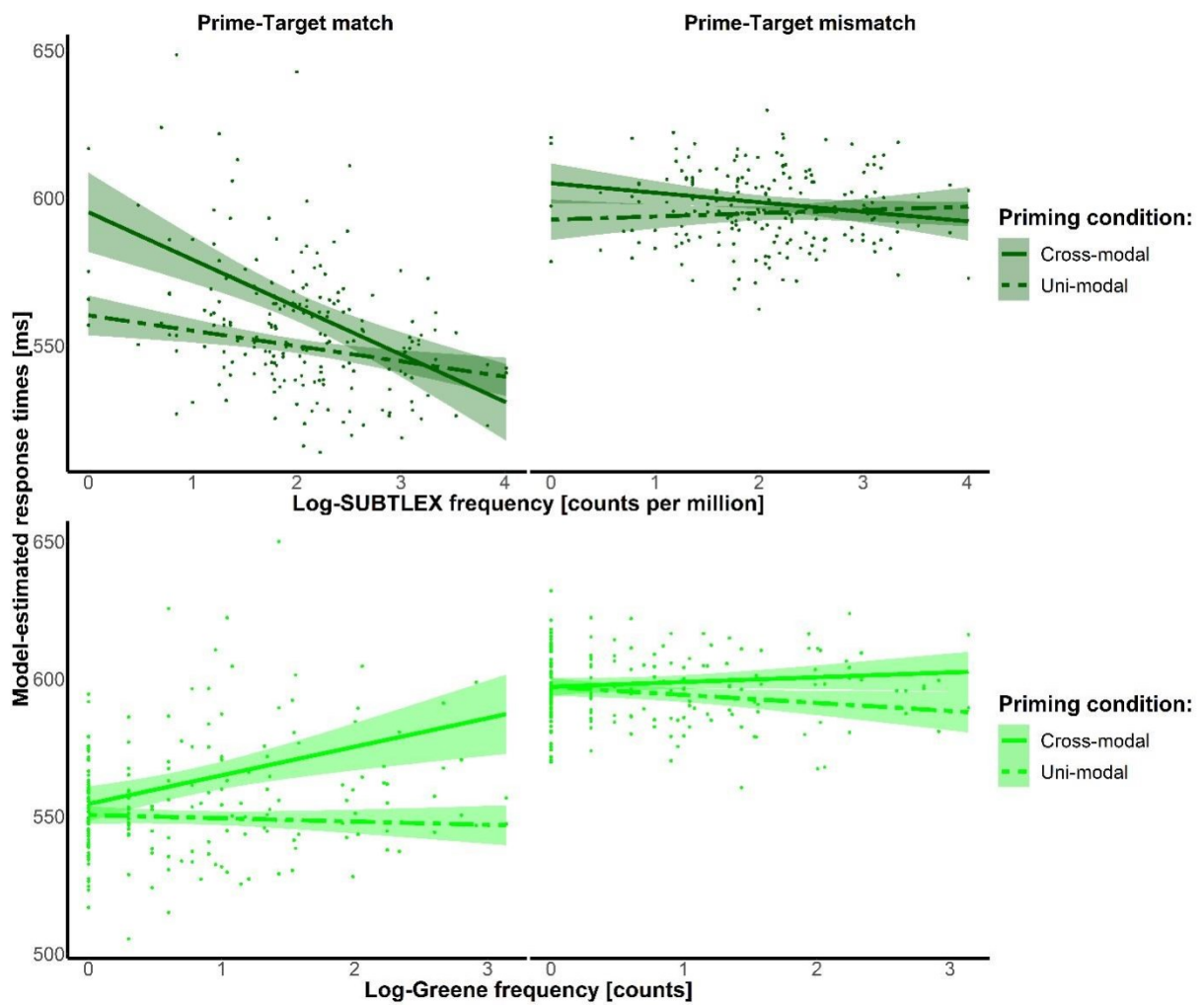
Supplementary figure 14 - Raw RTs from post-hoc conditions of Experiment 3 (between condition)

Raw response times in the priming conditions (Cross-modal solid lines, Uni-modal: dashed-dotted) and matching condition (Matching on the left, Mismatching on the right), as a function of SUBTLEX frequency (top, dark green) and Greene frequency (bottom, light green) in Replication experiment. Points show concepts with different level of frequency, averaged across participants; lines represent linear fitting of points and shaded areas represent 95 % confidence interval



Supplementary figure 15 - RTs estimated from post-hoc of interaction effects of Experiment 3 (between conditions)

Response times as a function of logarithmic SUBTLEX frequency (top plots, dark green) and Greene frequency (bottom plots, light green) in the 3-way significant interaction with Matching condition and Priming condition (Cross-modal matching vs Uni-modal matching; Cross-modal mismatching vs Uni-modal mismatching; Cross-modal matching vs Cross-modal mismatching). RTs were estimated based on the selected model. Points present participant-based mean response times for concepts in the different frequency levels. Lines represent linear fitting of points (solid: cross-modal; dashed: uni-modal), and shaded areas represent 95 % confidence interval. Bottom left and top-left plots represent the effects in prime-target matching condition, while bottom-right and top-right plots represent the effects in prime-target mismatching condition



4 post-hoc models are additionally computed, one for every level of the re-coded factor (Cross-modal Matching, Uni-modal Matching, Cross-modal Mismatching, Uni-modal Mismatching)

$$\begin{aligned} Rep_logRT \sim & \text{SUBTLEX WF} * \text{Target modality} + \text{Greene OF} * \text{Target modality} + \\ & \text{Concept familiarity (replication)} + \text{Image typicality (replication)} + \\ & \text{Image visual PC1} + \text{Image visual PC2} + \text{Image visual PC3} + \\ & \text{Visuo-orthographic PC} + \text{Target repetition} + \text{Trial accuracy} + \\ & (1|\text{Participants}) + (1|\text{Concepts}) \end{aligned}$$

Supplementary table 18. Results from the post-hoc individual models for conditions of interest in the Replication experiment.

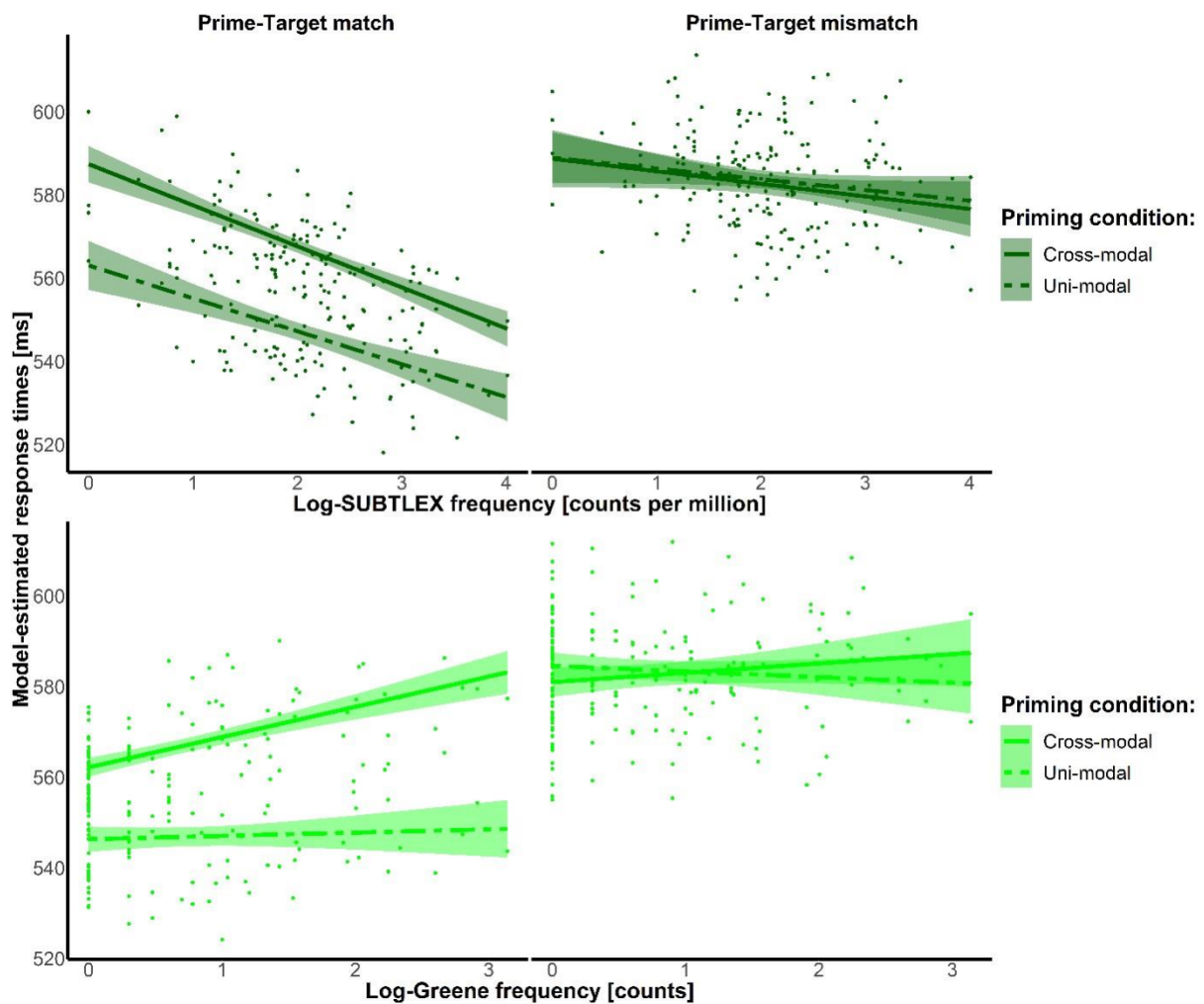
<i>Predictors</i>	Uni-modal Matching				Cross-modal Matching			
	<i>β</i>	<i>SE</i>	<i>t</i>	<i>p</i>	<i>β</i>	<i>SE</i>	<i>t</i>	<i>p</i>
(Intercept)	6.304	0.022	289.192	<0.001	6.340	0.027	237.811	<0.001
SUBTLEX WF	-0.011	0.003	-3.470	0.001	-0.013	0.006	-2.253	0.024
Target modality (Words – Objects)	0.016	0.004	3.748	<0.001	-0.036	0.005	-7.115	<0.001
Greene OF	0.001	0.003	0.398	0.691	0.010	0.005	1.873	0.061
Visuo-orthographic PC	0.002	0.003	0.818	0.413	0.009	0.006	1.613	0.107
Concept familiarity	-0.000	0.003	-0.010	0.992	-0.004	0.006	-0.754	0.451
Image typicality	-0.001	0.003	-0.463	0.643	-0.026	0.005	-5.426	<0.001
Image visual PC1	0.001	0.002	0.422	0.673	0.001	0.004	0.274	0.784
Image visual PC2	0.006	0.002	2.379	0.017	0.002	0.004	0.499	0.618
Image visual PC3	0.003	0.003	1.191	0.234	-0.000	0.005	-0.005	0.996
Target repetition	-0.026	0.002	-12.208	<0.001	-0.051	0.003	-20.200	<0.001
Trial accuracy (Correct – Incorrect)	0.019	0.011	1.781	0.075	-0.012	0.012	-1.051	0.293
SUBTLEX x (Words – Objects)	-0.011	0.004	-2.516	0.012	-0.027	0.005	-5.433	<0.001
Greene x (Words – Objects)	0.000	0.004	0.018	0.985	0.006	0.005	1.141	0.254

<i>Predictors</i>	Uni-modal Mismatching				Cross-modal Mismatching			
	<i>β</i>	<i>SE</i>	<i>t</i>	<i>p</i>	<i>β</i>	<i>SE</i>	<i>t</i>	<i>p</i>
(Intercept)	6.369	0.023	276.110	<0.001	6.367	0.027	233.415	<0.001
SUBTLEX WF	-0.003	0.003	-1.204	0.229	-0.004	0.003	-1.234	0.217
Target modality (Words – Objects)	0.001	0.004	0.337	0.736	-0.010	0.005	-2.018	0.044
Greene OF	-0.002	0.003	-0.699	0.484	0.003	0.003	1.044	0.297
Visuo-orthographic PC	0.002	0.003	0.565	0.572	0.006	0.003	2.018	0.044

Concept familiarity	0.002	0.003	0.692	0.489	-0.000	0.003	-0.094	0.925
Image typicality	-0.000	0.002	-0.205	0.838	-0.007	0.003	-2.695	0.007
Image visual PC1	0.004	0.002	1.896	0.058	-0.003	0.002	-1.071	0.284
Image visual PC2	0.003	0.002	1.489	0.137	0.001	0.002	0.221	0.825
Target repetition	-0.028	0.002	-13.127	<0.001	-0.040	0.002	-16.764	<0.001
Image visual PC3	-0.001	0.002	-0.538	0.591	0.002	0.002	0.733	0.464
Trial accuracy (Correct – Incorrect)	0.048	0.017	2.920	0.004	0.065	0.015	4.251	<0.001
SUBTLEX x (Words – Objects)	-0.006	0.004	-1.340	0.180	-0.000	0.005	-0.072	0.943
Greene x (Words – Objects)	0.001	0.004	0.314	0.754	-0.004	0.005	-0.796	0.426

Supplementary figure 16 – RTs estimated from post-hoc models of Experiment 3 (within conditions)

Effects of SUBTLEX WF (dark green, top) and Greene OF (light green, bottom) on reaction times estimated from the post-hoc models separately for each Priming condition (continuous and dashed-dotted line types) and Matching condition (left and right plots) in the Replication experiment. Points show concepts with different level of frequency, averaged across participants; lines represent linear fitting of points and shaded areas represent 95 % confidence interval.



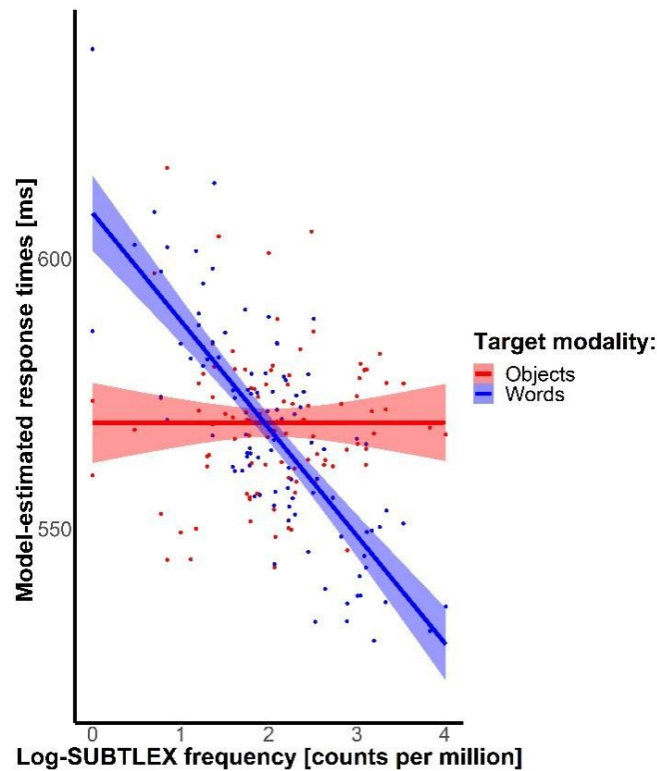
Supplementary table 19. Results from the post-hoc individual models for conditions of interest in Experiment 3.

Cross-modal Matching Words Cross-modal Matching Object

<i>Predictors</i>	β	<i>SE</i>	<i>t</i>	<i>p</i>	β	<i>SE</i>	<i>t</i>	<i>p</i>
(Intercept)	6.341	0.027	239.179	< 0.001	6.345	0.029	215.770	< 0.001
SUBTLEX WF	-0.027	0.007	-3.699	< 0.001	0.000	0.007	0.042	0.966
Greene OF	0.015	0.006	2.307	0.021	0.005	0.006	0.802	0.422
Visuo-orthographic PC	0.008	0.007	1.221	0.222	0.010	0.006	1.528	0.126
Concept familiarity	-0.008	0.007	-1.165	0.244	-0.000	0.006	-0.074	0.941
Image typicality	-0.024	0.006	-4.039	< 0.001	-0.028	0.005	-5.162	< 0.001
Image visual PC1	-0.001	0.005	-0.195	0.846	0.003	0.005	0.669	0.503
Image visual PC2	0.000	0.005	0.082	0.935	0.004	0.005	0.734	0.463
Image visual PC3	0.004	0.006	0.726	0.468	-0.004	0.005	-0.751	0.453
Target repetition	-0.051	0.007	-7.415	< 0.001	-0.055	0.007	-8.106	< 0.001
Trial accuracy	-0.017	0.016	-1.001	0.317	-0.018	0.017	-1.065	0.287

Supplementary figure 17 – RTs estimated from post-hoc models of Experiment 3 (within modalities)

Effects of SUBTLEX WF (left) on reaction times estimated from the post-hoc models for Cross-modal matching trials of words (blue) and Cross-modal matching trials of objects (red) in the Replication experiment. Points show concepts with different level of frequency, averaged across participants; lines represent linear fitting of points and shaded areas represent 95 % confidence interval.



Supplementary Materials 14 – Effect of dlexDB on Experiment 1

*Exp1_logRT ~ dlexDB WF * Concept modality +
 Concept category + Concept familiarity + Image typicality +
 Image visual PC1 + Image visual PC2 + Image visual PC3 +
 Visuo-orthographic PC + Target repetition +
 Trial accuracy + (1|Participants) + (1/Concepts)*

Supplementary table 20. Results from main model of Exp 1 including dlexDB instead of SUBTLEX

Predictors	β	SE	<i>t</i>	<i>p</i>
(Intercept)	6.480	0.021	301.419	< 0.001
Target modality (Words – Objects)	0.094	0.005	20.523	< 0.001
dlexDB WF	-0.021	0.008	-2.801	0.005
Visuo-orthographic PC	-0.000	0.008	-0.050	0.960
Concept familiarity	-0.004	0.003	-1.263	0.206
Image typicality	-0.005	0.003	-1.416	0.157
Image visual PC1	-0.002	0.006	-0.261	0.794
Image visual PC2	0.020	0.006	3.272	0.001
Image visual PC3	0.009	0.006	1.481	0.139
Target repetition	-0.011	0.002	-4.926	< 0.001
Trial accuracy (Correct – Incorrect)	-0.018	0.009	-2.032	0.042
Concept category (Natural – Man-made)	0.001	0.013	0.074	0.941
dlexDB x (Words – Objects)	-0.020	0.005	-4.394	< 0.001

Results from 2 post-hoc models one for each stimulus modality

*Exp1_logRT ~ dlexDB WF +
 Concept category + Concept familiarity + Image typicality +
 Image visual PC1 + Image visual PC2 + Image visual PC3 +
 Visuo-orthographic PC + Target repetition +
 Trial accuracy + (1|Participants) + (1/Concepts)*

Supplementary table 21. Results from post-hoc models of Exp 1 including dlexDB instead of SUBTLEX

Objects trials

Word trials

<i>Predictors</i>	β	<i>SE</i>	<i>t</i>	<i>p</i>	β	<i>SE</i>	<i>t</i>	<i>p</i>
(Intercept)	6.450	0.022	289.121	<0.001	6.520	0.025	265.635	<0.001
dlexDB WF	-0.012	0.009	-1.289	0.197	-0.031	0.008	-4.081	<0.001
Concept category (Natural – Man-made)	0.009	0.016	0.601	0.548	-0.009	0.013	-0.664	0.506
Visuo-orthographic PC1	-0.000	0.010	-0.018	0.986	-0.001	0.008	-0.090	0.928
Concept familiarity	-0.001	0.005	-0.277	0.782	-0.007	0.004	-1.480	0.139
Image typicality	-0.009	0.004	-2.054	0.040	-0.002	0.004	-0.462	0.644
Image visual PC1	-0.004	0.007	-0.597	0.551	0.001	0.006	0.179	0.858
Image visual PC2	0.026	0.007	3.524	<0.001	0.013	0.006	2.172	0.030
Image visual PC3	0.005	0.008	0.665	0.506	0.014	0.006	2.192	0.028
Target repetition	0.009	0.021	0.428	0.668	-0.030	0.024	-1.290	0.197
Trial accuracy (Correct – Incorrect)	-0.059	0.013	-4.403	<0.001	0.004	0.011	0.398	0.690

Supplementary Materials 15 – Conceptual Distinctiveness in Experiment 1

*Exp1_logRT ~ SUBTLEX WF * Concept modality +
 Concept category + Concept familiarity + Image typicality +
 Image visual PC1 + Image visual PC2 + Image visual PC3 +
 Visuo-orthographic PC + Target repetition + Conceptual Distinctiveness +
 Trial accuracy + (1|Participants) + (1|Concepts)*

Supplementary table 22. Results of Experiment 1 including CD as covariate

<i>Predictors</i>	β	<i>SE</i>	<i>t</i>	<i>p</i>
(Intercept)	6.479	0.021	302.570	<0.001
Concept modality (Words – Objects)	0.094	0.005	20.529	<0.001
SUBTLEX WF	-0.032	0.008	-4.150	<0.001
Visuo-orthographic PC	-0.006	0.007	-0.805	0.421
Concept familiarity	-0.003	0.003	-0.980	0.327
Image typicality	-0.004	0.003	-1.279	0.201
Image visual PC1	-0.002	0.006	-0.263	0.792
Image visual PC2	0.019	0.006	3.259	0.001
Image visual PC3	0.008	0.006	1.295	0.195
Target repetition	-0.011	0.002	-4.932	<0.001
Conceptual Distinctiveness (CD)	0.002	0.007	0.295	0.768

Trial accuracy (Correct – Incorrect)	-0.017	0.009	-1.940	0.052
Concept category (Natural – Man-made)	0.003	0.013	0.206	0.837
SUBTLEX WF x (Words – Objects)	-0.019	0.005	-4.160	<0.001

Supplementary Materials 16 – Effect of ADE20K on Experiment 2

$Exp2_logRT \sim SUBTLEX\ WF * Recoded\ factor * Target\ modality +$
 $ADE20K\ OF * Recoded\ factor * Target\ modality +$
 $Concept\ familiarity + Image\ typicality +$
 $Image\ visual\ PC1 + Image\ visual\ PC2 + Image\ visual\ PC3 +$
 $Visuo-orthographic\ PC + Target\ repetition + Trial\ accuracy +$
 $(1|Participants) + (1|Concepts)$

Supplementary table 23. Results from main model of Exp 2 including ADE20K instead of Greene

Predictors	β	SE	t	p
(Intercept)	6.225	0.020	315.721	<0.001
SUBTLEX WF	-0.005	0.002	-2.420	0.016
Cross-modal matching – Uni-modal matching	0.005	0.006	0.829	0.407
Cross-modal mismatching – Uni-modal mismatching	0.008	0.006	1.380	0.168
Cross-modal matching – Cross-modal mismatching	-0.075	0.003	-23.729	<0.001
Target modality (Words – Objects)	-0.008	0.002	-3.612	<0.001
ADE20K OF	0.008	0.002	4.391	<0.001
Visuo-orthographic PC	0.010	0.002	5.256	<0.001
Concept familiarity	-0.001	0.002	-0.739	0.460
Image typicality	-0.004	0.002	-2.638	0.008
Image visual PC1	0.002	0.002	1.085	0.278
Image visual PC2	0.005	0.002	3.463	0.001
Image visual PC3	-0.002	0.002	-1.134	0.257
Target repetition	-0.036	0.003	-13.329	<0.001
Trial accuracy (Correct – Incorrect)	0.029	0.005	5.393	<0.001
SUBTLEX x (Cross-modal matching – Uni-modal matching)	-0.027	0.004	-7.622	<0.001
SUBTLEX x (Cross-modal mismatching – Uni-modal mismatching)	-0.006	0.004	-1.568	0.117
SUBTLEX x (Cross-modal matching – Cross-modal mismatching)	-0.035	0.004	-9.901	<0.001
SUBTLEX x Target modality	-0.007	0.003	-2.651	0.008
Target modality x (Cross-modal matching – Uni-modal matching)	-0.010	0.006	-1.657	0.098
Target modality x (Cross-modal mismatching – Uni-modal mismatching)	0.036	0.006	5.720	<0.001
Target modality x (Cross-modal matching – Cross-modal mismatching)	-0.015	0.006	-2.322	0.020
ADE20K x (Cross-modal matching – Uni-modal matching)	0.018	0.004	5.167	<0.001

ADE20K x (Cross-modal mismatching – Uni-modal mismatching)	-0.002	0.004	-0.617	0.537
ADE20K x (Cross-modal matching – Cross-modal mismatching)	0.028	0.004	7.731	< 0.001
ADE20K x Target modality	0.004	0.003	1.726	0.084
SUBTLEX x (Cross-modal matching – Uni-modal matching) x Target modality	-0.005	0.007	-0.743	0.457
SUBTLEX x (Cross-modal mismatching – Uni-modal mismatching) x Target modality	0.005	0.007	0.762	0.446
SUBTLEX x (Cross-modal matching – Cross-modal mismatching) x Target modality	-0.003	0.007	-0.448	0.654
ADE20K x (Cross-modal matching – Uni-modal matching) x Target modality ADE20K	0.014	0.007	1.937	0.053
x (Cross-modal mismatching – Uni-modal mismatching) x Target modality	0.004	0.007	0.588	0.557
ADE20K x (Cross-modal matching – Cross-modal mismatching) x Target modality	0.003	0.007	0.441	0.660

Additional Bibliography

- Akaike, H. (1981). Likelihood of a model and information criteria. *Journal of Econometrics*, 16(1), 3–14. [https://doi.org/10.1016/0304-4076\(81\)90071-3](https://doi.org/10.1016/0304-4076(81)90071-3)
- Brehm, L., & Alday, P. M. (2020). *A decade of mixed models: It's past time to set your contrasts*. Talk presented at the 26th Architectures and Mechanisms for Language Processing Conference (AMLap 2020). Potsdam, Germany. 2020-09-03 - 2020-09-05.
- Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of Experimental Psychology: General*, 139(3), 558–578. <https://doi.org/10.1037/a0019165>
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, 82(13). <https://doi.org/10.18637/jss.v082.i13>
- Lüdecke, D., Ben-Shachar, M. S., Patil, I., Waggoner, P., & Makowski, D. (2021). performance: An R package for assessment, comparison and testing of statistical models. *Journal of Open Source Software*, 6(60).
- Robinson, A. P., & Froese, R. E. (2004). Model validation using equivalence tests. *Ecological Modelling*, 176(3–4), 349–358. <https://doi.org/10.1016/j.ecolmodel.2004.01.013>
- Schad, D. J., Vasishth, S., Hohenstein, S., & Kliegl, R. (2020). How to capitalize on a priori contrasts in linear (mixed) models: A tutorial. *Journal of Memory and Language*, 110, 104038. <https://doi.org/10.1016/j.jml.2019.104038>

Hierarchical organization of objects in scenes is reflected in mental representations of objects

Jacopo Turini* & Melissa Le-Hoa Võ

Scene Grammar Lab, Department of Psychology and Sports Sciences,
Goethe University, Frankfurt am Main, Germany

*Corresponding author:

Jacopo Turini

Scene Grammar Lab

Institut für Psychologie

PEG, Room 5.G105

Theodor-W.-Adorno Platz 6

60323 Frankfurt/Main

turini@psych.uni-frankfurt.de

+49 (0)69 798-35310

1 **Abstract**

2

3 The arrangement of objects in scenes follows certain rules (“Scene Grammar”), which we exploit to
4 perceive and interact efficiently with our environment. We have proposed that Scene Grammar is
5 hierarchically organized: scenes are divided into clusters of objects (“phrases”, e.g., the sink phrase);
6 within every phrase, one object (“anchor”, e.g., the sink) holds strong predictions about identity and
7 position of other objects (“local objects”, e.g., a toothbrush). To investigate if this hierarchy is
8 reflected in the mental representations of objects, we collected pairwise similarity judgments for
9 everyday object pictures and for the corresponding words. Similarity judgments were stronger not
10 only for object pairs appearing in the same scene, but also object pairs appearing within the same
11 phrase of the same scene as opposed to appearing in different phrases of the same scene. Besides,
12 object pairs with the same status in the scenes (i.e., being both anchors or both local objects) were
13 judged as more similar than pairs of different status. Comparing effects between pictures and
14 words, we found similar, significant impact of scene hierarchy on the organization of mental
15 representation of objects, independent of stimulus modality. We conclude that the hierarchical
16 structure of visual environment is incorporated into abstract, domain general mental
17 representations of the world.

18

19 **Keywords:** scene hierarchy, scene grammar, phrasal structure, object similarity, stimulus modality

20

21

22

23

24

25 **Introduction**

26

27 Objects in our environment are not arranged randomly but usually appear in certain contexts
28 (“semantic rules”) and in certain positions (“syntactic rules”), according to physical laws and typical
29 use [1]. We refer to this set of rules of objects in scenes as “Scene Grammar” (for a recent review
30 see [2]), in analogy with the linguistic grammar that governs words in sentences. It has been shown
31 that Scene Grammar is exploited by our cognitive system to efficiently represent objects during
32 visual perception and to guide allocation of attention during scene perception [3, 4] supporting
33 complex behaviors like object recognition [5], search [6], and object interaction [7].

34 More recently, it has been proposed that Scene Grammar could be structured according to
35 a hierarchy [8]: a scene on the top level is divided into meaningful clusters of spatially related
36 objects, which we refer to as “phrases”; in every phrase, one object holds a special status (“anchor
37 object”), with strong predictions regarding both the identity and position of the other objects within
38 the cluster (“local objects”; *Fig. 1A*). Anchor objects are proposed to be typical (i.e., frequently
39 present) of a scene, bigger in size and rather stationary (e.g., a sink), while local objects tend to be
40 smaller and more moveable (a toothbrush). The proposed role of this hierarchy entails that during
41 complex behavior within a scene, like object search or interaction, we first and foremost process
42 objects based on their phrasal membership within a scene.

43 So far, mostly the top “scene level” as organizing structure of objects has been investigated.
44 It is believed that priors regarding object-to-object and object-to-scene relationships are activated
45 after a quick extraction of a scene’s “gist” [9, 10]. As a result, typically studies have manipulated the
46 consistency between an object and its background scene (e.g., a priest in a church vs. a football
47 court [11]), and have tried to identify which ingredients of a scene are sufficient to retrieve this

48 contextual knowledge (e.g., color and texture [12]; orientation [13]; materials [14]; layout,[15]; for
49 a review [16]).

50 The "phrase level" has hardly received any attention thus far, but there have been attempts
51 to disentangle what the role of pairs and groups of objects is in supporting object identification. For
52 instance, co-occurrence (a pot and a stove) and spatial dependency (a pot on top of a stove)
53 between objects have been also found to be relevant for object processing during visual search [17,
54 18] and object recognition [19, 20], even beyond the effect of background scene information [21].
55 Indeed, the complex network of object-to-object relationships seems to be retrieved even when
56 objects are seen in isolation on a neutral background, as shown by the correlation between fMRI
57 patterns evoked by single object pictures and a computational model that uses distributional
58 statistics of objects in scenes [22]. Besides, typical semantic and spatial arrangements of multiple
59 objects are processed in a more efficient way both at behavioral and neural level [23, 24] supposedly
60 due to a grouping mechanism that allows to reduce the complexity of visual input. This grouping
61 based on meaning and spatial relationship might also be supportive of extraction of action
62 affordances, which seems to play an important role in scene understanding [25] and might be the
63 organizing principle behind the phrasal structure in man-made scenes [2].

64 Finally, for what concerns the "object type level", first empirical results supporting the
65 prominent role of anchor objects in structuring a scene came from a study where participants were
66 asked to arrange objects in a virtual environment according to their scene grammar (creating a
67 typical arrangement of objects in scenes [7]): Anchor objects were preferentially used during initial
68 stages of object arrangements underlining their role as primary building blocks of a scene. The
69 important role of anchor objects in visual search has been further corroborated by a series of eye-
70 tracking experiments where the absence of anchor objects (e.g., the toilet being replaced by a
71 washing machine) resulted in less efficient search performance as seen in faster RTs and reduced

72 gaze coverage of the scene [26]. These results were then replicated in more ecologically valid and
 73 immersive setting provided by virtual reality (VR [27]). Participants had to search for target local
 74 objects within virtual environments that either displayed anchor objects or anchors replaced by gray
 75 cuboids in the same position. The presence of anchors had strong beneficial effects on search
 76 behavior as seen in more efficient gaze and body movements.

77

78 **Fig. 1 – A)** Schema of the hierarchical structure of objects in scenes tested in the study: a scene is divided into clusters 79
 (phrases) and each phrase is formed by one anchor objects and several local objects (figure adapted from [8]); B) 80
 Estimation of hierarchical measures using a priori assignment of objects to a scene, phrase and object type or using a 81
 datasets of annotated and segmented images from which we can extract co-occurrence and clustering information 82
 (image taken from the dataset [28] and visualized through LabelMe [29]); C) Example of a trial from Experiment 1 and 83
 Experiment 2 showing a triplet of objects (pictures or words), as well as the way we measured behavioural similarity 84
 from the response in the trial: pairs including the selected “odd-one” object have minimal similarity while the pair 85
 including the unselected objects have maximal similarity. Object images are taken from [30] and are not the one used 86
 in the real experiment.

A) Proposed hierarchical organization of objects in scenes:



B) A priori hierarchy:
 assigned based on common sense and intuition

- Bathroom:**
 Phrase 1: sink (anchor),
 toothbrush (local),
 toothpaste (local)
 Phrase 2: toilet (anchor),
 toilet brush (local),
 toilet paper (local),
 Phrase 3: bathtub (anchor),
 towel rack (local),
 towel (local)
- Bedroom:**
 Phrase 1: bed (anchor),
 lamp (local),
 pillow (local),
 ...

Data-driven hierarchy:
 estimated from a dataset of real-world scene images



C) Task: click on the odd-one-out object of the triplet

Experiment 1:
 Object pictures



Experiment 2:
 Written words

"Topf" (pot) "Herd" (stove) "Computermaus" (computer mouse)

Behavioural similarity estimation

Sim(pot,stove) = 1 ("similar")
 Sim(pot,mouse) = 0 ("dissimilar")
 Sim(stove,mouse) = 0 ("dissimilar")

87

88 The goal of the current study was to investigate whether the contextual knowledge
89 associated with mental representations of object is organized according to a hierarchy, where the
90 levels of scene, phrase, and object type (anchor vs. local) can be distinguished. Moreover, we
91 wanted to assess whether the organization of object representations is modality-specific or
92 independent of specific modalities (e.g., verbal and non-verbal stimuli [31]).

93 To achieve these goals, we organized a set of everyday objects according to the above-
94 mentioned hierarchical structure in two ways (*Fig. 1B*): one based on common-sense and intuition
95 (a priori hierarchy model), and the other one based on the distribution of objects in a real-world
96 image dataset [28] (data-driven hierarchy model), both organizing objects on three levels: scene,
97 phrases and object types. Then, we collected pairwise similarity ratings for the set of objects,
98 adapting an “odd-one-out” triplet task (*Fig. 1C*) previously used to study perceptual and conceptual
99 dimensions underlying mental representation of objects [32]. Finally, we compared the odd-one-
100 out ratings to the hierarchy models using Representational Similarity Analysis (RSA [33]), which
101 allows to estimate if the representational space underlying behavioural responses is structured
102 according to the levels of our proposed hierarchical organization, representing pairwise similarity of
103 both behaviour and hierarchical models in terms of Representational (Dis)similarity Matrices (RDMs;
104 see *Fig. 2* for the organization of individual objects in the RDMs, and *Fig. 3* for RDMs of each
105 hierarchical predictor). To estimate the simultaneous impact of different levels of the hierarchy and
106 different types of hierarchy, we combined RSA with Generalized Linear Mixed-effects Models
107 (GLMMs [34]).

108

109 **Fig. 2** – One half of a symmetric Representational Dissimilarity Matrix (RDM) showing the organization of individual
110 object pairs based on the a priori hierarchical organization. Gray and black portions of the triangle represent pairs of
111 objects assigned to the same scene category, while black portions represent pairs of objects assigned to the same phrase
112 within the scene. Scene category labels and composition of the phrases are also reported, the letter (A) indicates an anchor
113 object, the letter (L) indicates local objects. The remaining white portion of the triangle represents pairs of

114 objects that are assigned to different scenes. This order of objects is maintained in the RDMs and used to represent
 115 different levels of the hierarchical models (see Fig. 3).

116

117

118

119

120

121

122

123

124

125

126

127

128

129

130

131

132

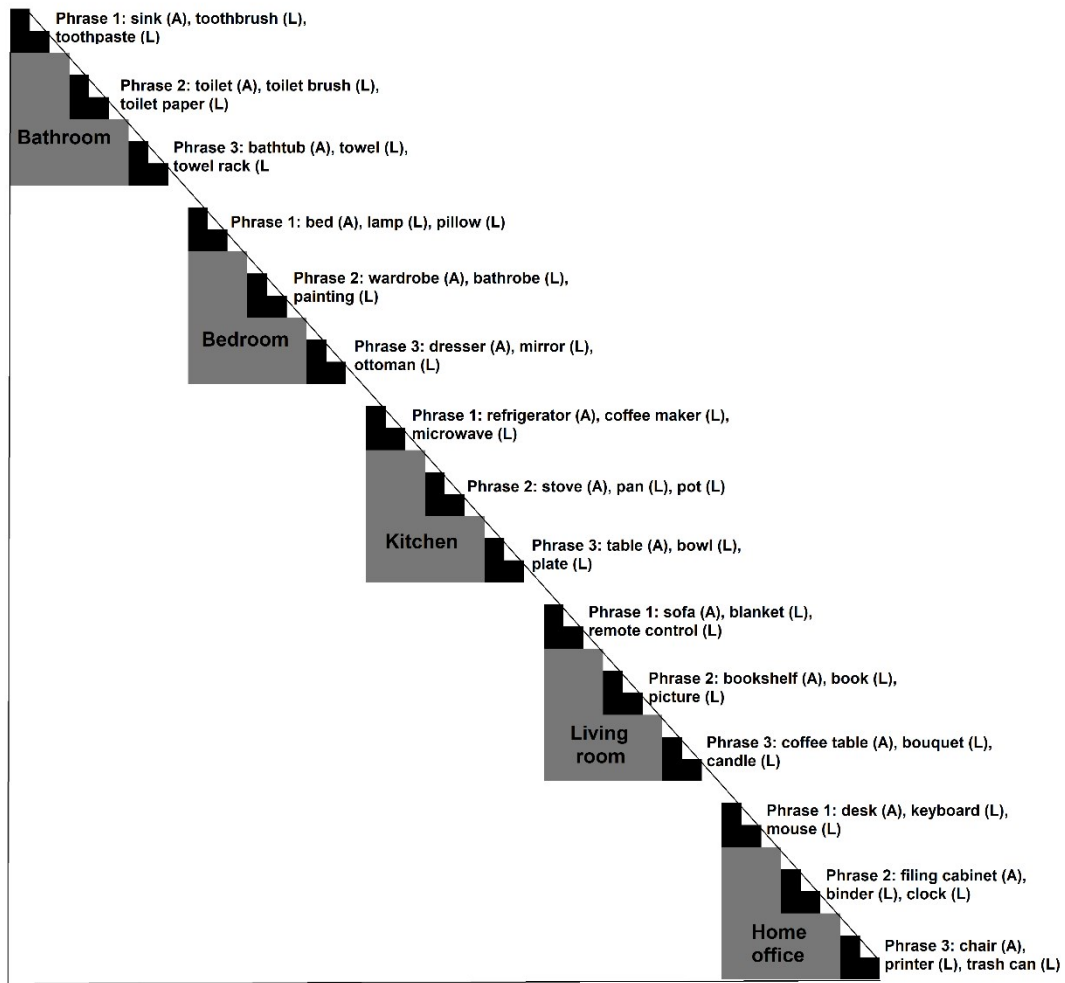
133

134

135

136

137



138

139

140

141

142

143

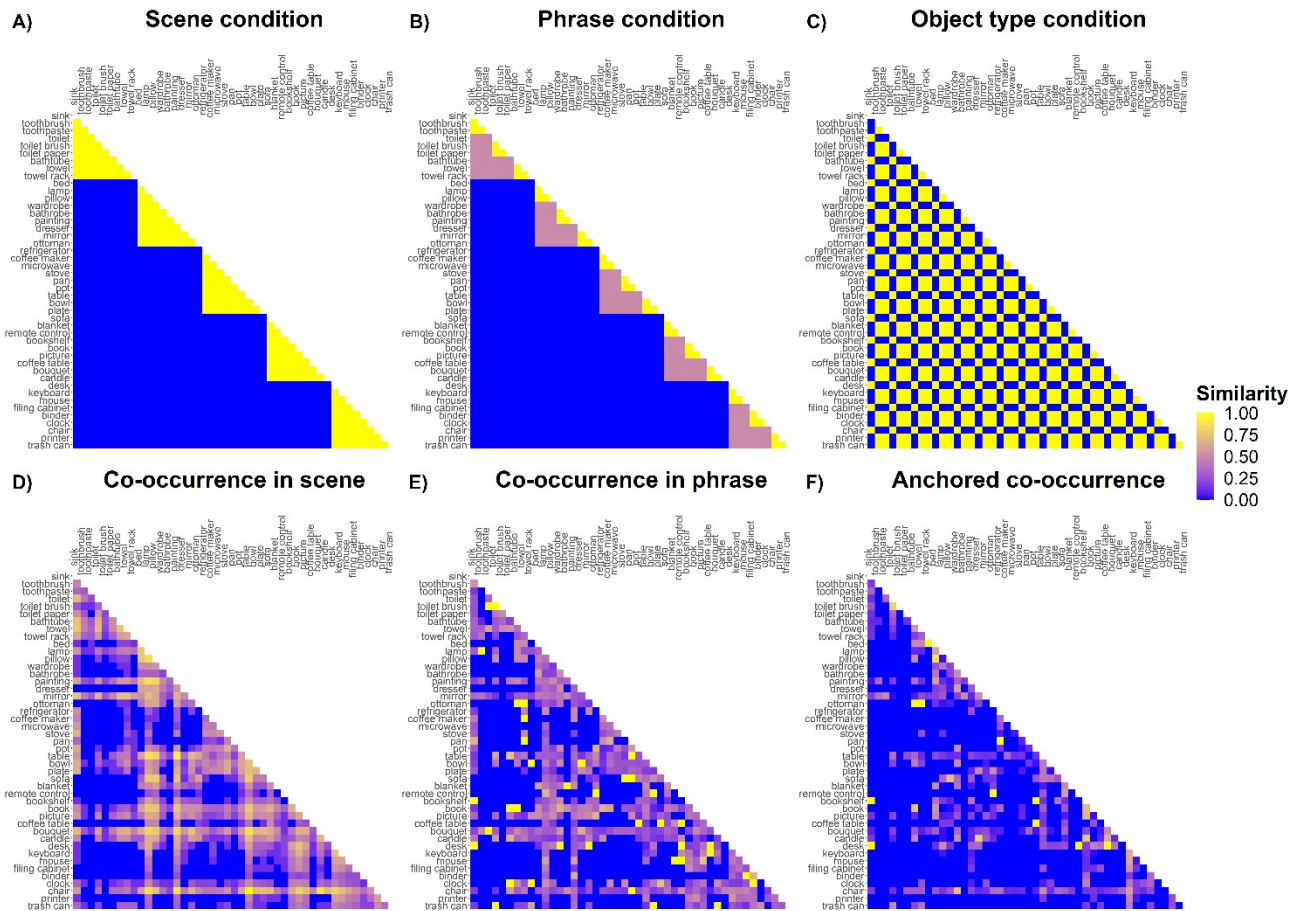
144

145

146

147 **Fig. 3** – Representational (Dis)similarity Matrices (RDMs) for the a priori hierarchical predictors (A, B and C) and for the
 148 data-driven hierarchical predictors (D, E and F). RDMs are symmetric matrices where entries on rows and columns are
 149 the objects stimuli, and cells represent pairwise similarity along a specific dimension. In A, B and C, yellow represents
 150 pairs of objects that are assigned to the same scene, phrase or type (maximal similarity), while blue represents pairs
 151 that are assigned to different scenes, phrases or types (minimal similarity). In D, the $\log_{10}(\text{counts} + 1)$ of co-occurrence

152 in scene is normalized to span between 0 (blue, few counts) to 1 (yellow, many counts). In E and F, the colors represent
 153 proportion of counts to the total co-occurrence counts of each pair.



154

155

156

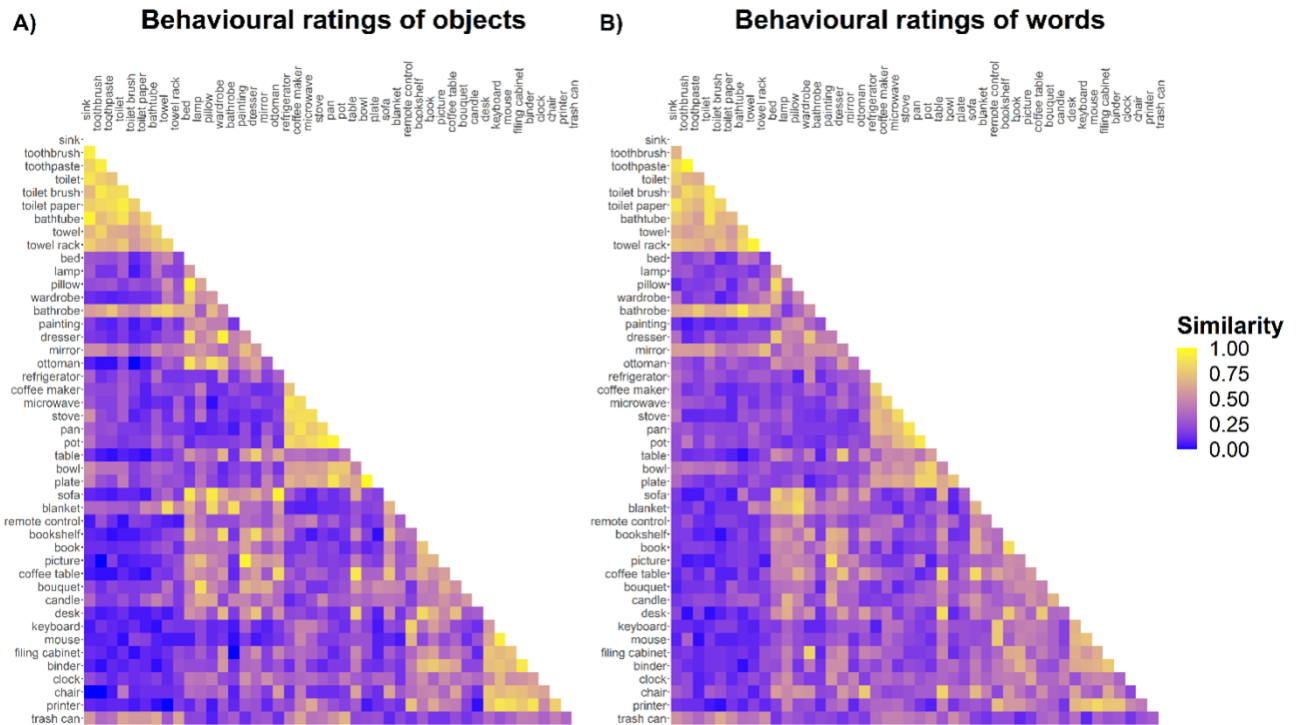
157 Results

158

159 Ratings divided by modality were plotted in the RDM format (Fig. 4), where every cell represents
 160 the pairwise similarity ratings for a given pair averaged across all the triplets where the pair is
 161 present. The GLMM resulted to be singular, due to the random factor term ($1 | participants$)
 162 explaining no variance, since this was already explained by the other two random factors
 163 ($1 | pairs$) and ($1 | context objects$), that identify unique observations.

164

165 **Fig. 4** – Representational (Dis)similarity Matrices (RDMs) for the ratings collected in Exp 1 (object pictures, A) and Exp 2
 166 (words, B). Cells represent pairwise similarity ratings averaged across all the triplets where the pair was present. Every
 167 pair was presented in a triplet with all the other remaining objects (“context object”), and it was judged either as similar
 168 (1) or dissimilar (0), so that In the RDMs pairwise similarity spans from 0 (never judged as similar) to 1 (always judged
 169 as similar).



170
 171
 172
 173

174 To evaluate potential multicollinearity in the model, we computed the variance inflation factors
 175 (VIFs) for each term in the model, using the *check_collinearity* function in R (package “performance”
 176 [35]). Typically, when VIFs are below 5, there is low correlations between predictors and the model
 177 does not need any adjustment, as it was in our case (VIFs and correlations among predictors are
 178 shown in detail in *Supplementary Materials 1*).

179 Results from the GLMM (*Fig. 5*) showed a main effect of stimulus modality ($\beta=-0.107$, $SE=0.031$,
 180 $z=-3.448$, $p=0.001$), with objects pictures estimated to be more similar to each other than object
 181 words. The a priori hierarchical structure was reflected in participants’ similarity ratings, with
 182 significant main effects of scene condition ($\beta=1.078$, $SE=0.075$, $z=14.474$, $p<0.001$), phrase condition

183 ($\beta=0.270$, $SE=0.128$, $z=2.111$, $p=0.035$), and object type condition ($\beta=0.245$, $SE=0.048$, $z=5.106$,
184 $p<0.001$), showing that objects belonging to the same scene / phrase / object type were considered
185 more similar than objects belonging to different scenes / phrase / object types. At the same time,
186 we also found main effects of the data-driven hierarchy predictors measuring co-occurrence in
187 scene ($\beta=0.397$, $SE=0.029$, $z=13.922$, $p<0.001$) and co-occurrence in phrase ($\beta=0.063$, $SE=0.028$, $z=-$
188 2.229 , $p=0.022$), where in both cases the more two objects co-occurred, the more they were judged
189 to be similar. However, the anchored co-occurrence between two objects was not significantly
190 reflected in pairwise similarity ratings ($\beta=0.005$, $SE=0.028$, $z=0.165$, $p=0.869$). Overall, these results
191 already show a hierarchical organization of mental representations not only on the scene level, but
192 also at the phrasal and object type level.

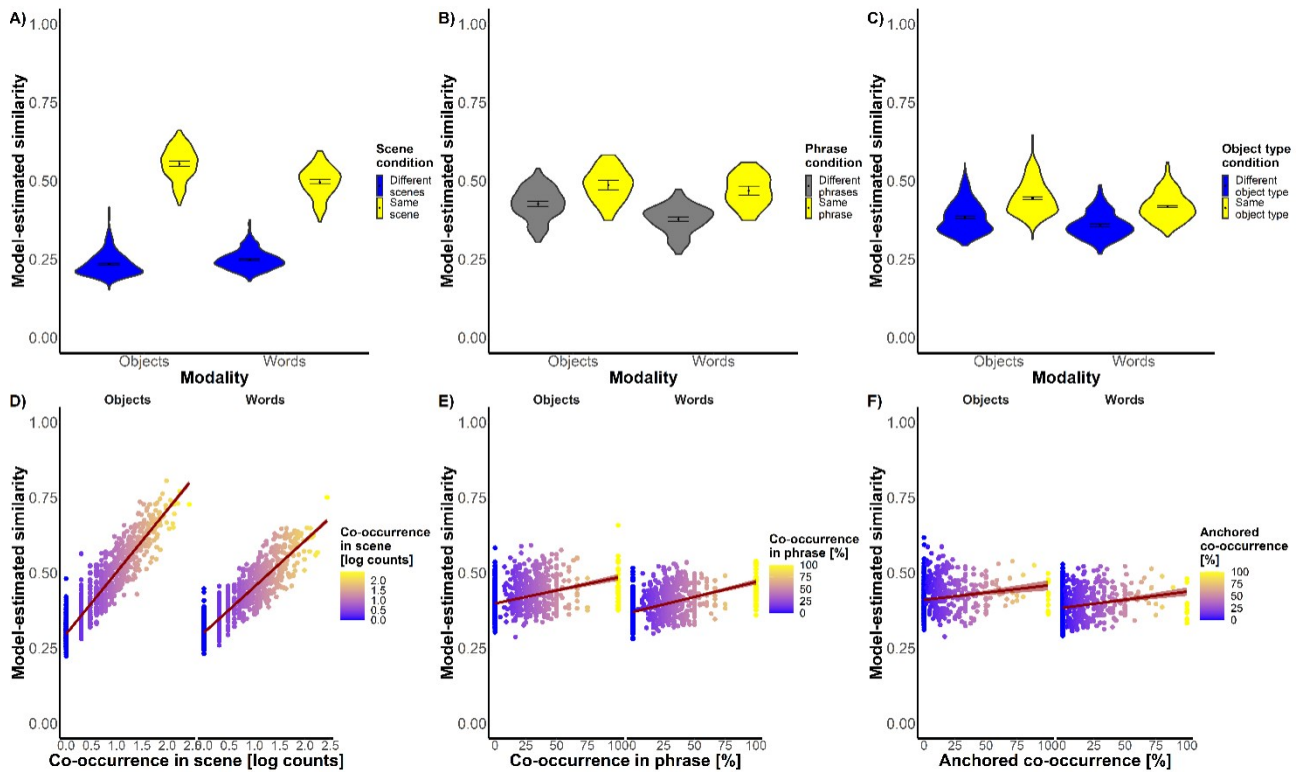
193 Regarding the covariate measures (see *Supplementary Materials 2*), we found main effects of the
194 early layer of AlexNet DNN ($\beta=-0.133$, $SE=0.025$, $z=-5.317$, $p<0.001$), with pairs that looked more
195 similar in terms of low-level visual features being considered less similar at behavioural level, while
196 the main effect of late layer of AlexNet ($\beta=0.126$, $SE=0.031$, $z=4.078$, $p<0.001$) showed that object
197 pairs that looked more similar in terms of high-level visual features were also estimated to be more
198 similar by our participants. Finally, we detected a main effect of word embeddings ($\beta=0.338$,
199 $SE=0.025$, $z=13.363$, $p<0.001$), with object pairs that have stronger similarity in terms of
200 distributional semantics features being considered more similar. These results show that distinction
201 emerging from both complex visual features (AlexNet late layer) and word meaning (Word
202 embeddings) are important factor in determining the mental representation supporting behaviour,
203 while contrary to that, similarity based on low-level visual features (AlexNet early layer) acts as a
204 confound making more similar objects less distinguishable.

205

206 **Fig. 5** – Model-estimated effects of the hierarchy predictors on pairwise similarity ratings for object pictures and words.
207 Colours of violins and points reflect the values of pairs for the given predictor and match the ones in the RDMs showed

208 above. Stimulus modality is indicated by x-axis position (left = objects, right = words). Points and violins reflect estimated
 209 similarity for each pair of objects averaged across all the different contexts (i.e., the third object a triplet) in which they
 210 were presented. 95 % confidence interval are represented by error bars in the violins (point is the mean), and by the 211
 shaded area around lines for continuous predictors.

212



213

214

215 In terms of interaction between stimulus modality and our predictors, the model showed a
 216 significant effect in scene condition (*a priori* predictor, $\beta=-0.280$, $SE=0.050$, $z=-5.601$, $p<0.001$), and
 217 in co-occurrence in scene (data-driven predictor, $\beta=-0.124$, $SE=0.019$, $z=-6.361$, $p<0.001$), where in
 218 both cases the effect of the hierarchical predictor was found to be stronger in ratings of object
 219 pictures than ratings of words. Object ratings had also stronger effect of the late layer of AlexNet
 220 than word ratings ($\beta=-0.112$, $SE=0.022$, $z=-5.157$, $p<0.001$), while word ratings had a stronger effect
 221 of word length than object ratings ($\beta=0.082$, $SE=0.028$, $z=2.977$, $p=0.003$; for more details, see
 222 *Supplementary Materials 2*). This is expected since both predictors are estimated based on their
 223 preferential stimulus modalities (AlexNet activation with object pictures; Word length with words),

224 and signifies that these dimensions are more strongly related to modality specific representations
225 compared to the hierarchical predictors.

226 For more details regarding how object size, manipulability and moveability interact with
227 different object types (anchor and local objects) see *Supplementary Materials 3* and *4*.

228

229 **Discussion**

230

231 Objects in visual scenes are arranged in a structured way. These structural regularities are
232 learnt and stored in long-term memory (“scene grammar”) to make meaningful predictions and
233 efficiently perceive and interact with the environment [2]. In this study, we wanted to explore
234 whether scene grammar is organized in a hierarchical way. We hypothesized that at the top of the
235 hierarchy, objects are grouped together according to whether they appear in the same context
236 (scene level), followed by objects that spatially cluster within that context (phrase level), which
237 again consist of anchor objects that hold strong predictions about identity and position of other
238 local objects within a cluster [8]. Moreover, we wanted to understand if this organization emerges
239 differently in one modality than the other (e.g., object pictures vs. written words). For this purpose,
240 we adopted the odd-one-out task as introduced by Hebart and colleagues [32], a method that has
241 been used to study perceptual and conceptual dimensions underlying mental representation of
242 objects.

243 We have shown that when participants are asked to judge the similarity between pairs of
244 objects, the underlying mental representations seem to be organized according to our proposed
245 hierarchy. That is, pairs of objects that were assigned a priori to the same scene, to the same phrase,
246 or to the same object type, were judged as more similar than pairs of different scenes, phrases and
247 types. This finding largely held up even when the hierarchy was estimated from statistical

248 distributions of objects in real-world images [28]. Besides, we showed that these results were overall
249 consistent and stable across modalities, with only the scene level predictors showing an even
250 stronger effect for object pictures than words. Finally, we highlighted how the a priori division of
251 objects between anchors and local objects is strongly based on object size and moveability, as
252 previously proposed and showed [26].

253 To our knowledge, this is the first attempt to explore whether the hierarchical organization
254 of objects in scenes is incorporated into our mental representations. Previous research either
255 focused on effects of scene context on object processing (e.g., [2]; for a review see [16]) or on the
256 relationship between anchors and related local objects (e.g., [26, 27]). Here, we aimed at bridging
257 the gap between these two levels considering the role of meaningful clusters of objects (“phrase”
258 level) as an intermediate structure within the hierarchy.

259 Employing two different sources of estimation of the hierarchy allowed us to draw some
260 interesting conclusions. The weak correlations between a priori and data-driven hierarchy
261 predictors and the absence of multicollinearity (see *Supplementary Materials 1*) show that, despite
262 the same direction of the effects, the two models of hierarchy are only partly overlapping. We can
263 only speculate about the reasons of these differences, which also might also speak to the limitations
264 of both types of hierarchy estimations: on the one hand, previous research has shown that
265 subjective experience of how frequently objects in the world occur is overestimated [36], which
266 might have resulted in differences between a priori estimations and measures taken from the
267 distribution of objects in labeled image databases; on the other hand, it is important to note that
268 any given dataset of annotated images only represents a rough (and often biased) approximation of
269 the real-world distribution of objects. Compared to word frequency measures based on corpora of
270 at least 20 million words [37], fully annotated image datasets are much smaller in size (in our case,
271 circa 45,000 annotations). The two hierarchical organizations (a priori vs. data-driven) might also

272 reflect object processing in two different ways: for instance, the a priori hierarchy is based on
273 discrete, dichotomic divisions of objects dependent on whether they appear in the same context or
274 not, and therefore might be used when a task requires the processing of rough contextual
275 information; on the other hand, the continuous co-occurrence measures from the data-driven
276 approach might offer a more fine-grained representation of object-to-object contextual information
277 when necessary. Using distributional properties of objects in scenes as calculated from annotated
278 datasets (similar to research on language) is becoming increasingly popular and provides interesting
279 insights on learning statistical regularities in both vision and in language [22, 38], offering an
280 alternative to traditionally employed categorical divisions based on experimenters' intuition or
281 crowd-sourced ratings.

282 The measures that can be extracted from this type of datasets can offer even more fine-
283 grained information than what we highlighted here: for example Boettcher et al. [26] measured that
284 the relationship between anchor and local objects has strong regularities on the vertical axis, that
285 is, it is possible to predict the position of a certain local object from a certain anchor object in terms
286 of "is above" or "is below", but as much on the horizontal axis ("is left of" or "is right of"), similar to
287 linguistic grammar where in most languages the components of a phrase (e.g., subject and object)
288 have predictable positions with respect to each other. This seems to match the intuition that the
289 structure of a room is much more vertically organized: objects typically found on the lower part of
290 a room tend to differ from objects typically found in the top part of the room (e.g., shoes usually
291 are found on the floor, while paintings are hanging up on the wall) , while on the horizontal axis
292 there is much more variability (e.g., the towels can be found either left or right of the shower. This
293 vertical organization of the environment seems to indeed also be reflected in the neural
294 representation of scenes [39].

295 The significant results of both types of hierarchy predictors suggest that, despite some of
296 their limitations, these are capturing aspects of the visual world that seem to be incorporated in our
297 mental representations of objects. This is particularly interesting as these layered representations
298 seem to be triggered by simply viewing isolated objects or words. It is important to point out that –
299 similar to Hebart and colleagues [32] - no explicit definition of similarity or specific instructions on
300 how to judge the (dis)similarity of the three presented objects/words were given to the participants
301 when performing the “odd-one-out” triplet task. The aim was to collect similarity judgements that
302 are not biased towards specific dimensions while allowing different dimensions to emerge in
303 different contexts. For example, “cat” and “elephant” might be similar in a triplet with “table”, based
304 on animacy, but “cat” and “elephant” might be dissimilar in a triplet containing “dog”, where the
305 similarity might be based on whether the animals are pets or not. However, it has been shown that
306 - using the same triplet task with different similarity instructions - it is possible to measure the
307 flexibility of mental representations in highlighting one dimension more than others according to
308 task demands [40]. We believe this could also apply to the hierarchical organization of objects in
309 scenes, whose strength in shaping mental representation might be increased by tasks that require
310 interactions with objects (e.g., judging similarity based on function) and reduced by tasks that rely
311 less on object-to-object contextual relations (e.g., judging similarity based on visual features).
312 Future investigations directly comparing different “odd-one-out” triplet task might shed more light
313 on these aspects.

314 A question that remains open is whether this hierarchical organization is present in every
315 type of scenes. In the present study, we have employed only an organization that relates to indoor
316 man-made environments, because we believe that here the hierarchical structure is optimized to
317 efficiently perform everyday actions like brushing teeth or cooking. Outdoor scenes in general, and
318 natural scenes in particular, might show less of a hierarchical structure. First of all, in the way they

319 are experimentally investigated, they have much bigger scale than indoor environments. This has
320 consequences on navigational and action patterns, which differs from the ones of smaller scale
321 indoor scenes. Second, natural scenes, in which man-made objects are rare or even absent, lack
322 object arrangements that reflect the need for efficient human-object interaction. That said, nature
323 of course has its own “grammar” as well (e.g., the way that rivers flow or rocks fall into place), and
324 it might be worth investigating the hierarchical structure of natural scenes and how these might be
325 mirrored in mental representations.

326 While we did not measure brain responses in this study, it is still worth discussing how such
327 hierarchical organization could be implemented in the brain. For instance, the hierarchical
328 organization of objects in scenes might be represented in the parahippocampal cortex (PHC), in the
329 anterior part of the ventral-temporal cortex. Within the PHC lies the parahippocampal place area
330 (PPA), a scene-selective region which shows stronger activation for scene stimuli rather than single
331 objects [41]. Subsequent investigations have suggested that PPA/PHC might represent spatial and
332 non-spatial context in a more general way [9, 42], and not just based on visual scenes. This is in line
333 with recent findings that viewing single isolated objects evoked a complex representation of objects’
334 co-occurrence in the anterior portion of PPA [22]. Here also lies the perirhinal cortex, which has
335 been proposed to represent semantic information for individual objects [43], and is the medial
336 portion of the Anterior Temporal Lobe (ATL), which has been proposed to be the primary hub of the
337 semantic network [44].

338 Finally, our results - according to which hierarchical predictors show significant main effects
339 and minor differences between modalities - suggest that scene grammar might act on domain-
340 general representations. That is, the hierarchical structure of our visual world might be incorporated
341 into semantic memory representations of objects which are accessed when an object’s meaning is
342 retrieved from processing input from different modalities, here either pictures or words. Some

343 visual and hierarchical features are not completely independent, but we took great care to not have
344 extreme levels of multicollinearity invalidate the interpretation of our results (see Supplementary
345 Materials for correlation plots and VIF estimates). We therefore want to propose that a scene's
346 hierarchical structure is incorporated into the abstract semantic representations of both objects and
347 words that can be used to flexibly form predictions when encountering new visual environments or
348 written text. We believe that with this paper we were able to demonstrate that using several visual
349 and linguistic covariates, as well as measuring effects on both object pictures and words, we can
350 now provide some first evidence that the hierarchical predictors are 1) independent of the visual
351 and linguistic dimensions measured here and 2) are independent of the specific modality of stimulus
352 presentation.

353 To conclude, in the current study we provided first evidence that abstract mental
354 representations of objects in scenes might be hierarchically organized, incorporating not only scene
355 semantic information at the highest level, but also a more fine-grained, mid-level phrasal structure,
356 as well as distinctions of object types. We therefore believe that these phrasal substructures of
357 scenes play an important role in the organization of our mental representations of the world and
358 therefore should be considered when studying visual cognition.

359

360

361 **Materials and Methods**

362

363 **Participants**

364 Eighty-six participants took part in our study. Half of them took part in Experiment 1 (age: M = 24.72
365 yrs, SD = 5.33 yrs, range = 18 – 40 yrs; gender: F = 31, M= 12), the other half took part in Experiment
366 2 (age: M = 22.60 yrs, SD = 5.18 yrs, range = 19 – 50 yrs, 1 person did not report age; gender: F = 28,

367 M= 15). The number of participants in each experiment (N=43) was determined as the optimal ratio
368 between the total number of unique trials and an optimal number of trials to present to a single
369 participant. All participants reported that they had normal or corrected to normal vision and had no
370 history of psychiatric or neurological disorders. Participants of Experiment 2 also reported to be
371 German native speakers. Additionally, a third group of participants (N=20), who did not take part in
372 either Experiment 1 and Experiment 2, participated in a rating experiment to judge some features
373 of objects (age: M = 22.9 yrs, SD = 4.00 yrs, range = 19 – 35 yrs; gender = 12 F, 7 M and 1 NB). These
374 participants matched the same criteria of participants in Experiment 1. No minors participated in
375 the study. All participants gave their informed consent and received course credits or monetary
376 reimbursement for their participation. The Ethics Committee of the Goethe University Frankfurt
377 approved all experimental procedures (approval # 2014-106), that have been performed in
378 accordance with the Declaration of Helsinki.

379

380 **Stimuli**

381 Forty-five everyday indoor object concepts were selected for the study (see section below for more
382 details). For Experiment 1, pictures of the objects in isolation were downloaded from copyright-free
383 internet databases (e.g., <https://pnghunter.com/>, <http://pngimg.com/>,
384 <https://www.cleanpng.com/>), pasted on a white background, grey-scaled to rule out influence of
385 color, and resized to 392 x 392 pixels (jpg format). For Experiment 2, we used the German words
386 associated with the objects, presenting them in bold black Arial font, with the first letter in
387 uppercase and the other letters in lowercase, as by correct German spelling for nouns.

388

389

390

391 **Measures of scene hierarchy**

392 To predict similarity judgments as a function of scene hierarchy, we estimated two sets of scene
393 hierarchy measures.

394 - *A priori hierarchy measures*: these measures were based on intuition of experimenters as
395 well as common sense; therefore, we selected our 45 stimuli as typically belonging to one of
396 5 different indoor *scenes* (bathroom, bedroom, kitchen, living room and home office). For
397 every scene, we divided objects in 3 *phrases*; within every phrase, 1 object was identified as
398 *anchor object*, and the other 2 as *local objects* (Figs. 1B and 2).

399 - *Data-driven hierarchy measures*: these measures were based on a dataset of real-world
400 scene images containing pixel-wise segmentation and annotation of objects [28]. The
401 dataset contained 3499 unique coloured images, grouped into 16 scene categories (both
402 indoor and outdoor, natural and man-made, and including the 5 categories considered in the
403 a priori assignment), with more than 48,000 annotations grouped into 617 different object
404 categories (including the 45 objects selected for the study). Annotations were done by 4
405 different workers using the LabelMe tool [29] and were carefully cleaned of misspelling and
406 synonyms (Fig. 1B).

407 Following the procedure used in Boettcher et al. [26], we first pre-processed the
408 annotation and segmentation data in MATLAB (MathWorks, 2018), extracting identity,
409 coordinates and centroids of each object in the 2D space of pixels of each image. Further
410 analysis were carried on in R (version 3.6.3, R Core Team, 2020). Second, we discarded
411 objects that have a more structural function (e.g., walls, windows, ceiling, doors, pipes)
412 rather than being relevant for the object-to-object relationship we were interested in
413 investigating, leaving us with 567 unique object categories. Given the structure of the data,
414 we could compute how many times *two objects co-occur in the same image*, which is the

415 data-driven counterpart of the *scene level* of the hierarchy. Then, representing the objects
416 in an image through their centroids and the image area as a 2D space, we ran a clustering
417 algorithm to find the optimal spatial grouping of objects in every scene: the algorithm was
418 based on the partitioning around medoids clustering method and estimated the number of
419 clusters using average silhouette width (*pamk* function from R package “fpc” [45]). We
420 identified the resulting *clusters of objects as phrases*, and within every cluster, we identified
421 the *object with the largest area as anchor object*, while the other objects in each cluster were
422 considered *local objects*.

423

424 **Visual and linguistic covariates**

425 Additionally, to ensure that effects of the scene hierarchy did not emerge from a confound of lower-
426 level information, we estimated several measures of visual features (for object pictures in
427 Experiment 1) and linguistic features (for words in Experiment 2):

428 - *Visual measures* (for pictures): we estimated visual features of our object images feeding
429 them to a pre-trained Deep Neural Network (DNN), a state-of-the-art computer vision
430 algorithm that is trained to perform object categorization at human-like level. In our case,
431 we used the popular AlexNet, trained on the ImageNet dataset [46]. AlexNet, like most
432 DNNs, is based on many sequential layers of processing units, which extract and transform
433 features from the previous layer. The first layer extracts features from the input layer, which
434 is formed by the pixel values of an image; then the information is transformed in an
435 increasingly complex way through the many intermediate layers until it reaches the final
436 output layer, which assigns the image to one category (e.g., “cat”). We estimated unit
437 activations for our object images in 3 different layers of AlexNet: convolutional layer 1
438 (*conv1*, “early layer”), which processes low-level visual features (e.g., edges, brightness);

439 convolutional layer 4 (*conv4*, “*mid layer*”), which process mid-level visual features (e.g.,
440 shape); and the fully connected layer 7 (*fc7*, “*late layer*”), which processes high-level visual
441 features (complex configurations, like faces, handles, etc.).

442 - *Orthographic measures* (for words): we estimated orthography of our word stimuli using 2
443 measures: *word length*, as the number of letters in a word; *orthographic distance* from
444 neighboring words (i.e., words that differ for a letter from a target word), computed using
445 the OLD20 measure [47].

446 - *Distributional semantic measures* (for words): distributional semantic is a model of word
447 meaning based on the idea that words that appear in similar linguistic contexts (i.e., they
448 have a similar distribution in text) have similar meaning (for a review [48]). This approach
449 has been widely used in Natural Language Processing (NLP) to create algorithms that use
450 distributional measures from text corpora to build representations of word meaning and
451 perform operations on it. One common way of representing word meaning in NLP is through
452 *Word embeddings* which are multi-dimensional vectors. Words whose embeddings are
453 closer in this vector space have also similar meanings. For our set of word stimuli, we used
454 the embeddings trained on German Wikipedia using fastText and the skip-gram model with
455 default parameters [49].

456

457 **Object features**

458 To better understand what features underlying the division of objects between anchors and local
459 objects, we have collected ratings about three dimensions that have been discussed in connection
460 to the status of anchor and local objects: *real-world size* (how big an object is), *moveability* (how
461 easily an object is moved in space) and *manipulability* (how much the position of an object or of one
462 of its part or its configuration is changed during the interaction with it).

463

464 **Apparatus and Procedure**

465 Apparatus and procedure were mostly identical across Experiments 1 and 2. Where there were
466 differences, those are reported explicitly. For the study, we adapted an “odd-one-out” triplet task
467 introduced by Hebart and colleagues, which elegantly is used to collect pairwise similarity
468 judgments of object pictures [32]. First, we generated all the possible combinations of triplets of
469 stimuli ($45! / (3! * (45 - 3)!) = 14190$ unique triplets). We then divided the triplets randomly into 43
470 groups of 330 triplets, to have a practical number of trials and participants. Every participant,
471 therefore, performed the task on a different subset of triplets.

472 Experiments were programmed in Python using PsychoPy (version 2020.2.4, Builder GUI
473 [50]) and administered online through the hosting platform Pavlovia (<https://pavlovia.org/>).
474 Participants were asked to start the experiment only when they had between 30 min / 1 h of free
475 time and only when they could carry on the procedure with calm and in an undisturbed
476 environment. Instructions told participants they would have seen triplets of stimuli and their task
477 would have been to choose the “odd-one-out” stimulus, i.e., the one they considered the least
478 similar to the other two. No explicit definition of similarity was given to participants, as in the original
479 study. This is in line with the purpose played by the “odd-one-out” triplet task: similarity between a
480 pair of objects is evaluated across multiple trials (i.e., triplets), in which the context keeps varying
481 (i.e., the third object of the triplet). This way, many different dimensions are allowed to emerge and
482 be prioritized to judge the pair similarity, giving back a more complex picture of object
483 representations [32].

484 In our study, triplets were presented on a white background screen, with one stimulus on
485 the left, one stimulus in the center and one stimulus on the right (the position of every stimulus in
486 the triplet was randomized within every triplet before the presentation; *Fig. 1C*). Experiments were

487 programmed so that stimulus size were normalized based on screen size, so that every participant
488 saw stimuli occupying the same proportion of screen: each picture spanned about 1/4 of width and
489 height size, while each word spanned about 1/10 of height size and varying width size according to
490 word length. To choose the odd-one-out stimulus, participants had to press the corresponding
491 arrow (left arrow for the stimulus on the left, down arrow for the stimulus in the center, right arrow
492 for the stimulus on the right). Once they pressed the key, a 500 ms black fixation crossed appeared
493 in the center of the screen and then the next triplet was presented. Trials were divided into 6 blocks,
494 between which participants could take a break. Participants were allowed to take as much time as
495 they wanted to make their “odd-one-out” decision, and if they could not recognize one of the
496 stimuli, they were asked to make their decision based on what they thought the stimuli were.

497 In the object features rating experiment, participants performed the ratings of moveability,
498 manipulability, and real-world size in three different blocks (in this order). Within every block,
499 participants saw the pictures of the object stimuli from Experiment 1 one at the time (in randomized
500 order), together with the rating question (above the picture) and a 6-point likert scale (below the
501 picture). Before the block, they were presented with a definition of the investigated dimension, and
502 were asked to press a number between 1 to 6 corresponding to their judgments.

503

504 **Analysis**

505 To analyze how measures of scene hierarchy predict pairwise similarity judgments, we combined
506 two main analytical approaches: Representational Similarity Analysis (RSA [33]) and Generalized
507 Linear Mixed-effects Models (GLMMs [34]). RSA is a tool that allows comparison of different sources
508 of data that have different dimensionalities (brain data, behavioral data, computational models,
509 stimulus features). To do so, it requires the creation of Representational (Dis)similarity Matrices
510 (RDMs), which are symmetric matrices where column and row entries are typically corresponding

511 to the different stimuli (*Fig.2-3*). Every cell in an RDM contains a measure of (dis)similarity for that
512 pair of stimuli. Once the different sources of data are represented in the same RDM format, it is
513 possible to compare them and estimate how similar two RDMs are, i.e., how the structure of
514 pairwise similarity in one source (e.g., behavior) is predicted by the structure of pairwise similarity
515 in another source (e.g., a computational model).

516 In our study, we followed this approach to compute pairwise similarities from the “odd-one-
517 out” triplet behavioral task, as well as from the measures of hierarchy and covariates introduced
518 above.

519 - *Behavioral similarity*: we estimated behavioral similarity between pairs of stimuli in a
520 dichotomic way: similar (dummy coded as 1) vs dissimilar (dummy coded as 0). This estimate
521 was assigned as a result of the “odd-one-out” choice on every triplet. Given a triplet (e.g., A,
522 B and C), once an “odd-one” stimulus is selected (e.g., C), the similarity between the
523 unselected stimuli results to be maximal ($\text{Sim}(A,B) = 1 \rightarrow$ “similar”), while the similarity
524 between the “odd-one” stimulus and one of the unselected stimuli results to be minimal
525 ($\text{Sim}(C,A) = 0 \rightarrow$ “dissimilar”; $\text{Sim}(C,B) = 0 \rightarrow$ “dissimilar”; *Fig. 1C*).

526 - *A priori hierarchy similarity*: we estimated pairwise similarity based on the hierarchy status
527 assigned *a priori*. This results in 3 categorical predictors. First, we considered *scene condition*,
528 with dichotomic categorization: pairs from the same scene (dummy coded as 1) vs pairs from
529 different scene (dummy coded as 0). Then, we considered *phrase condition*, with three
530 groups: pairs from the same phrase (1) vs pairs from different phrases within the same scene
531 (0.5) vs pairs from different phrases in different scenes (0). Finally, we considered *object type*
532 *condition*, with two categories: pairs of objects of the same type (1) vs pairs of objects of
533 different type (0), where object type refers to the object being either an anchor object or a
534 local object.

- 535 - *Data-driven hierarchy similarity*: we estimated pairwise similarity based on the hierarchical
536 status emerging from the clustering procedure on the labelled image dataset. This results in
537 3 continuous predictors. First, we estimated a measure of *co-occurrence of pairs in a scene*,
538 as the number of times a pair appears in the same image; in the analysis we used \log_{10}
539 (counts + 1), so that we had a more uniform distribution along this dimension and avoid
540 having -Infinite values. Then, we estimated a measure of *co-occurrence of pairs in a phrase*,
541 as the proportion of co-occurrence counts where a pair not only appears in the same image
542 but also in the same cluster. Finally, we estimated a measure of *anchored co-occurrence*, as
543 the proportion of co-occurrence counts where one object of a pair is “anchored” to the
544 other.
- 545 - *Covariates*: for the visual, orthographic, and distributional semantic measures, similarity was
546 estimated in different ways. For multidimensional measures (i.e., the 3 AlexNet layers and
547 the Word embedding), similarity was estimated by computing the product-moment
548 correlation coefficient between pairs of vectors (e.g., the embedding vector for “pan” and
549 the embedding vector for “pot”); for mono-dimensional measures (i.e., word length and
550 orthographic distance), similarity was computed as the absolute value of the difference
551 between the two values of each pair (e.g., the absolute value of the difference between word
552 length for “pot” and word length for “pan”).

553

554 GLMMs are an extension of Linear Mixed-effects Models (LMMs [51]) for responses / dependent
555 variables that have a non-gaussian distribution (in our case, the bimodal dichotomic behavioral
556 similarity). The main advantage of (G)LMMs over simple regression models and ANOVAs is that one
557 can consider each trial from each participant simultaneously, without the need for aggregation or
558 separate estimation of the effects across participants and item (i.e., crossed random effects of items

559 and participants [52]). Therefore, the response is estimated based on several predictors (fixed
560 factors) and considering grouping factors that have common portion of variance (random factors).
561 Using R syntax, our model had this structure:

562

563 *behavioral similarity ~ stimulus modality * (scene condition + phrase condition*
564 *+ object type condition + cooccurrence in scene + cooccurrence in phrase*
565 *+ anchored cooccurrence + covariates)*
566 *+(1 | participants) + (1 | pairs) + (1 | context objects)*

567

568 In the formula, on the left of the tilde (~), we have the response, i.e., the dichotomic behavioral
569 similarity from the triplet task; on the right of the tilde, we have the predictors, i.e., the categorical
570 and continuous pair similarity from the *a priori* and data-driven hierarchical organization, as well as
571 pair similarity for covariate measures; finally, we have the random factors, i.e., participant, pair, and
572 context object (the third object in the triplet). We fitted the statistical models via maximum
573 likelihood estimation, and continuous predictors were scaled, as this typically improves model fit.
574 For categorical predictors, we planned specific contrasts between conditions: for scene condition,
575 the contrast was set to *same scene – different scenes*; for object type condition, the contrast was
576 set to *same object type – different object types*; for phrase condition, one contrast was set to *same*
577 *phrase – different phrases of the same scene*, while the other contrast was set to (*same phrase and*
578 *different phrases of the same scene) – different phrases of different scenes*. Since this last contrast
579 is identical to *same scene – different scenes*, and since the scene similarity and phrase similarity
580 predictors are highly correlated, we removed from the model the *scene condition* predictor and
581 incorporate its contrast in the *phrase similarity* predictor. This way, we removed redundancies and
582 reduced multi-collinearity to an acceptable level. Besides, every measure was put in interaction with
583 the categorical predictor *stimulus modality*, which compares the effect of the measures between

584 words and objects pictures. Finally, for random effects, we included only an intercept term, so that
585 we followed the recommendations of Bates et al. about parsimony in random effect structure [53]

586

587 RSA was previously used in combination with general linear model (e.g., [54, 39]), modeling
588 response RDMs of different participants (from brain or behaviour) as a linear combination of
589 multiple predictors RDMs (from stimulus features or computational models) and going beyond the
590 simple 1-to-1 correlation between response and predictor RDMs originally presented in RSA. In our
591 approach we went one step further: since the similarity of each pair is estimated multiple times in
592 different context (the third object of the triplet), and since each context object appeared multiple
593 times with different pairs, we considered these additional sources of random variance (pairs and
594 context objects) exploiting the flexibility of GLMMs.

595 Analysis was performed using R (version 3.6.3, R Core Team, 2020).

596

597

598

599

600

601

602

603

604

605

606

607 Bibliography

608

- 609 1. Biederman, I., Mezzanotte, R. J. & Rabinowitz, J. C. Scene perception: Detecting and judging objects
610 undergoing relational violations. *Cognitive Psychology* **14**, 143–177 (1982).
- 611 2. Vö, M. L.-H. The meaning and structure of scenes. *Vision Research* **181**, 10–20 (2021).
- 612 3. Vö, M. L. H. & Henderson, J. M. Does gravity matter? Effects of semantic and syntactic inconsistencies
613 on the allocation of attention during scene perception. *Journal of Vision* **9**, 24–24 (2009).
- 614 4. Vö, M. L.-H. & Wolfe, J. M. Differential Electrophysiological Signatures of Semantic and Syntactic
615 Scene Processing. *Psychol Sci* **24**, 1816–1823 (2013).
- 616 5. Cornelissen, T. H. W. & Vö, M. L.-H. Stuck on semantics: Processing of irrelevant object-scene
617 inconsistencies modulates ongoing gaze behavior. *Atten Percept Psychophys* **79**, 154–168 (2017).
- 618 6. Vö, M. L.-H. & Wolfe, J. M. The interplay of episodic and semantic memory in guiding repeated search
619 in scenes. *Cognition* **126**, 198–212 (2013).
- 620 7. Draschkow, D. & Vö, M. L.-H. Scene grammar shapes the way we interact with objects, strengthens
621 memories, and speeds search. *Sci Rep* **7**, 16471 (2017).
- 622 8. Vö, M. L.-H., Boettcher, S. E. & Draschkow, D. Reading scenes: how scene grammar guides attention
623 and aids perception in real-world environments. *Current Opinion in Psychology* **29**, 205–210 (2019).
- 624 9. Bar, M. Visual objects in context. *Nat Rev Neurosci* **5**, 617–629 (2004).
- 625 10. Oliva, A. & Torralba, A. The role of context in object recognition. *Trends in Cognitive Sciences* **11**, 520–
626 527 (2007).
- 627 11. Davenport, J. L. & Potter, M. C. Scene Consistency in Object and Background Perception.
628 *Psychological Science* **15**, 559–564 (2004).
- 629 12. Lauer, T., Cornelissen, T. H. W., Draschkow, D., Willenbockel, V. & Vö, M. L.-H. The role of scene
630 summary statistics in object recognition. *Sci Rep* **8**, 14666 (2018).
- 631 13. Lauer, T., Willenbockel, V., Maffongelli, L. & Vö, M. L.-H. The influence of scene and object orientation
632 on the scene consistency effect. *Behavioural Brain Research* **394**, 112812 (2020).
- 633 14. Lauer, T., Schmidt, F. & Vö, M. L.-H. The role of contextual materials in object recognition. *Sci Rep* **11**,
634 21988 (2021).
- 635 15. Brady, T. F., Shafer-Skelton, A. & Alvarez, G. A. Global ensemble texture representations are critical
636 to rapid scene perception. *Journal of Experimental Psychology: Human Perception and Performance* **43**, 53 (2017).
- 638 16. Lauer, T. & Vö, M. L.-H. The Ingredients of Scenes that Affect Object Search and Perception. in *Human*
639 *Perception of Visual Information: Psychological and Computational Perspectives*. (Springer
International Publishing, 2022). doi:[10.1007/978-3-030-81465-6](https://doi.org/10.1007/978-3-030-81465-6).

- 641 17. Mack, S. C. & Eckstein, M. P. Object co-occurrence serves as a contextual cue to guide and facilitate
642 visual search in a natural viewing environment. *Journal of Vision* **11**, 9–9 (2011).
- 643 18. Hwang, A. D., Wang, H.-C. & Pomplun, M. Semantic guidance of eye movements in real-world scenes.
644 *Vision Research* **51**, 1192–1205 (2011).
- 645 19. Auckland, M. E., Cave, K. R. & Donnelly, N. Nontarget objects can influence perceptual processes
646 during object recognition. *Psychonomic Bulletin & Review* **14**, 332–337 (2007).
- 647 20. Gronau, N. & Shachar, M. Contextual integration of visual objects necessitates attention. *Atten*
648 *Percept Psychophys* **76**, 695–714 (2014).
- 649 21. Wu, C.-C., Wang, H.-C. & Pomplun, M. The roles of scene gist and spatial dependency among objects
650 in the semantic guidance of attention in real-world scenes. *Vision Research* **105**, 10–20 (2014).
- 651 22. Bonner, M. F. & Epstein, R. A. Object representations in the human brain reflect the co-occurrence
652 statistics of vision and language. *Nat Commun* **12**, 4081 (2021).
- 653 23. Kaiser, D., Stein, T. & Peelen, M. V. Object grouping based on real-world regularities facilitates
654 perception by reducing competitive interactions in visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 655
11217–11222 (2014).
- 656 24. Quek, G. L. & Peelen, M. V. Contextual and Spatial Associations Between Objects Interactively
657 Modulate Visual Processing. *Cerebral Cortex* **30**, 6391–6404 (2020).
- 658 25. Greene, M. R., Baldassano, C., Esteva, A., Beck, D. M. & Fei-Fei, L. Visual scenes are categorized by
659 function. *Journal of Experimental Psychology: General* **145**, 82–94 (2016).
- 660 26. Boettcher, S. E. P., Draschkow, D., Dienhart, E. & Vö, M. L.-H. Anchoring visual search in scenes:
661 Assessing the role of anchor objects on eye movements during visual search. *Journal of Vision* **18**, 11
662 (2018).
- 663 27. Helbing, J., Draschkow, D. & Vö, M. L. H. Auxiliary scene context information provided by anchor
664 objects guides attention and locomotion in natural search behavior. *Psychological Science* (2022).
- 665 28. Greene, M. R. Statistics of high-level scene context. *Frontiers in Psychology* **4**, (2013).
- 666 29. Russel, B. C., Torralba, A., Murphy, K. P. & Freeman, W. T. LabelMe: a database and web-based tool
667 for image annotation. *International journal of computer vision* **77**, 157–173 (2008).
- 668 30. Hebart, M.N., Dickter, A.H., Kidder, A., Kwok, W.Y., Corriveau, A., Van Wicklin, C. and Baker, C.I.
669 THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLoS*
one, **14**(10), p.e0223792 (2019).
- 671 31. Shinkareva, S. V., Malave, V. L., Mason, R. A., Mitchell, T. M. & Just, M. A. Commonality of neural
672 representations of words and pictures. *NeuroImage* **54**, 2418–2425 (2011).
- 673 32. Hebart, M. N., Zheng, C., Pereira, F. & Baker, C. I. *Revealing the multidimensional mental*
representations of natural objects underlying human similarity judgments. <https://osf.io/7wrgh>
(2020)

- 676 33. Kriegeskorte, N., Mur, M. & Bandettini, P. Representational similarity analysis – connecting the
677 branches of systems neuroscience. *Frontiers in Systems Neuroscience* (2008)
678 doi:[10.3389/neuro.06.004.2008](https://doi.org/10.3389/neuro.06.004.2008).
- 679 34. McCulloch, C. E. & Neuhaus, J. M. Generalized Linear Mixed Models. *Encyclopedia of Biostatistics*
680 (2005).
- 681 35. Lüdtke, D., Ben-Shachar, M., Patil, I., Waggoner, P. & Makowski, D. performance: An R Package for
682 Assessment, Comparison and Testing of Statistical Models. *JOSS* **6**, 3139 (2021).
- 683 36. Greene, M. R. Estimations of object frequency are frequently overestimated. *Cognition* **149**, 6–10
684 (2016).
- 685 37. Brysbaert, M. *et al.* The Word Frequency Effect: A Review of Recent Developments and Implications
686 for the Choice of Frequency Estimates in German. *Experimental Psychology* **58**, 412–424 (2011).
- 687 38. Gregorova, K., Turini, J., Gagl, B., & Vo, M. L. H. Access to meaning from visual input: Object and word
688 frequency effects in categorization behavior. *PsyArXiv* (preprint).
- 689 39. Kaiser, D., Turini, J. & Cichy, R. M. A neural mechanism for contextualizing fragmented inputs during
690 naturalistic vision. *eLife* **8**, e48182 (2019).
- 691 40. Greene, M. R. & Hansen, B. C. Disentangling the Independent Contributions of Visual and Conceptual
692 Features to the Spatiotemporal Dynamics of Scene Categorization. *J. Neurosci.* **40**, 5283–5299 (2020).
- 693 41. Epstein, R. & Kanwisher, N. A cortical representation of the local visual environment. *Nature* **392**,
694 598–601 (1998).
- 695 42. Aminoff, E. M., Kveraga, K. & Bar, M. The role of the parahippocampal cortex in cognition. *Trends in*
696 *Cognitive Sciences* **17**, 379–390 (2013).
- 697 43. Clarke, A. Dynamic activity patterns in the anterior temporal lobe represents object semantics.
698 *Cognitive Neuroscience* **11**, 111–121 (2020).
- 699 44. Lambon-Ralph, M. A. L., Jefferies, E., Patterson, K. & Rogers, T. T. The neural and computational
700 bases of semantic cognition. *Nat Rev Neurosci* **18**, 42–55 (2017).
- 701 45. Hennig, C. fpc: Flexible procedures for clustering. *R package*. (2020).
- 702 46. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural
703 networks. *Commun. ACM* **60**, 84–90 (2017).
- 704 47. Yarkoni, T., Balota, D. & Yap, M. Moving beyond Coltheart’s N: A new measure of orthographic
705 similarity. *Psychonomic Bulletin & Review* **15**, 971–979 (2008).
- 706 48. Lenci, A. Distributional Models of Word Meaning. *Annu. Rev. Linguist.* **4**, 151–171 (2018).
- 707 49. Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. Enriching Word Vectors with Subword Information.
708 *TACL* **5**, 135–146 (2017).
- 709 50. Peirce, J. *et al.* PsychoPy2: Experiments in behavior made easy. *Behav Res* **51**, 195–203 (2019).

- 710 51. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models using lme4.
711 *arXiv:1406.5823 [stat]* (2014).
- 712 52. Baayen, R. H., Davidson, D. J. & Bates, D. M. Mixed-effects modeling with crossed random effects for
713 subjects and items. *Journal of Memory and Language* **59**, 390–412 (2008).
- 714 53. Bates, D., Kliegl, R., Vasishth, S. & Baayen, H. Parsimonious Mixed Models. *arXiv:1506.04967 [stat]*
715 (2015).
- 716 54. Proklova, D., Kaiser, D. & Peelen, M. V. Disentangling Representations of Object Shape and Object
717 Category in Human Visual Cortex: The Animate–Inanimate Distinction. *Journal of Cognitive* 718
Neuroscience **28**, 680–692 (2016).

719

720 **Acknowledgments**

721 We want to thank Hyojin Kwon for contributing to the project. This work was supported by SFB/TRR
722 26 135 project C7 to Melissa L.-H. Võ and the Hessisches Ministerium für Wissenschaft und Kunst
723 (HMWK; project “The Adaptive Mind”).

724

725 **Authors contributions**

726 JT and MV conceptualized and designed the study together; JT implemented the experiment,
727 collected and analyzed the data; JT and MV interpreted the results and wrote the manuscript
728 together.

729

730 **Data availability statement**

731 Data and scripts are available at the following link: <https://osf.io/tx4m5/>

732

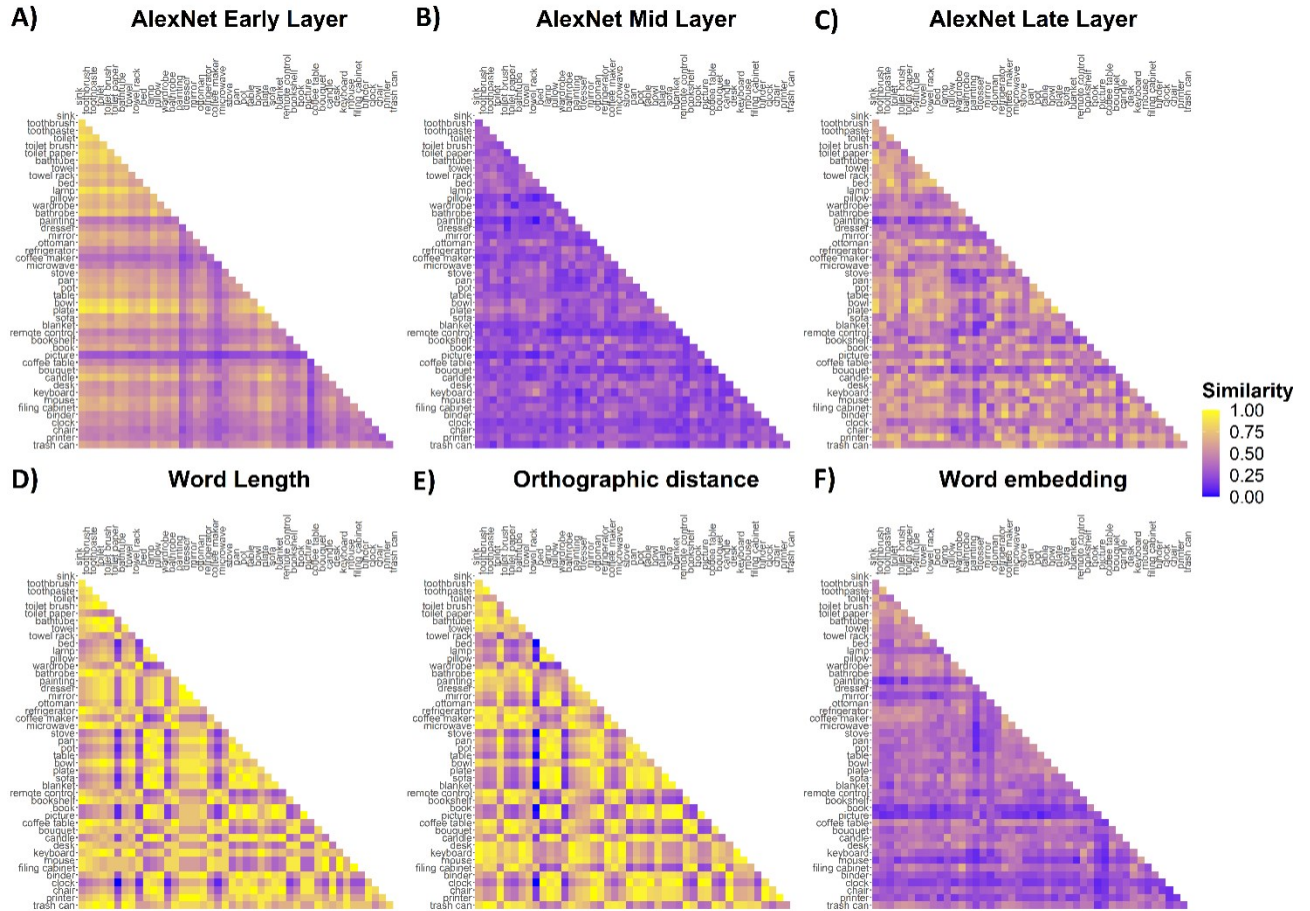
733 **Competing Interests Statement**

734 Authors declare no competing interests

Sup. Table 1 – Variance Inflation Factors (VIFs) for the predictors used in the main model

Predictors	VIF
Modality (Words – Objects)	3.645
Object type condition	1.065
Phrase condition	1.291
Anchored co-occurrence	1.464
Co-occurrence in scene	1.525
Co-occurrence in phrase	1.425
AlexNet early layer	1.184
AlexNet mid layer	1.768
AlexNet late layer	1.771
Word length	2.821
Orthographic distance	2.841
Word embeddings	1.189
Modality x Object type cond	1.084
Modality x Phrase cond	3.909
Modality x Anchored co-oc	1.450
Modality x Co-oc in scene	1.506
Modality x Co-oc in phrase	1.386
Modality x AlexNet early layer	1.190
Modality x AlexNet mid layer	1.771
Modality x AlexNet late layer	1.789
Modality x Word length	2.929
Modality x Orth distance	2.943
Modality x Word embeddings	1.178

Sup. Fig. 2 – Representational (Dis)similarity Matrices (RDMs) for the visual covariates for pictures (A, B and C) and for the orthographic and distributional semantics covariates for words (D, E and F). In A, B, C and D, colours represent the correlation between vectors (blue = 0 no correlation, yellow = 1 maximal correlation). In E and F, absolute value of the difference between word length / old20 of the pair is normalized to span between 0 (blue, bigger difference) to 1 (yellow, smaller difference).

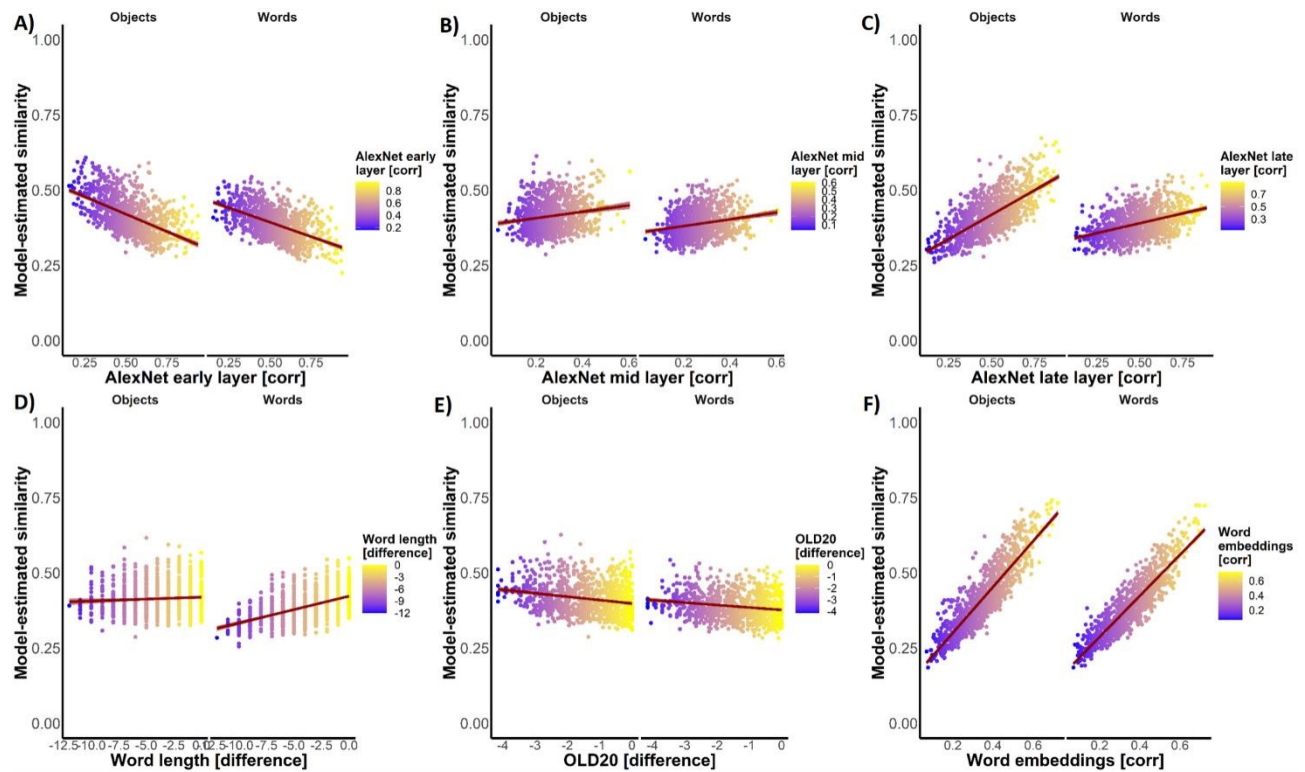


Supplementary Materials 2 – Results of the main model

Sup. Table 2 – Results of the GLMM

Predictors	β	SE	z	p
(Intercept)	-0.321	0.065	-4.809	<0.001
Modality (Words – Objects)	-0.107	0.031	-3.448	0.001
Object type condition (Same – Different)	0.245	0.048	5.106	<0.001
Phrase condition (Same – Different)	0.270	0.128	2.111	0.035
Scene condition (Same – Different)	1.078	0.075	14.474	<0.001
Anchored co-occurrence	0.005	0.028	0.165	0.869
Co-occurrence in scene	0.397	0.029	13.922	<0.001
Co-occurrence in phrase	0.063	0.028	2.292	0.022
AlexNet early layer	-0.133	0.025	-5.317	<0.001
AlexNet mid layer	0.026	0.031	0.846	0.397
AlexNet late layer	0.126	0.031	4.078	<0.001
Word length	0.049	0.039	1.271	0.204
Orthographic distance	-0.050	0.039	-1.270	0.204
Word embeddings	0.338	0.025	13.363	<0.001
Modality x Object type condition	-0.006	0.034	-0.181	0.857
Modality x Phrase condition	0.117	0.087	1.346	0.178
Modality x Scene condition	-0.280	0.050	-5.601	<0.001
Modality x Anchored co-occurrence	0.009	0.019	0.498	0.619
Modality x Co-occurrence in scene	-0.124	0.019	-6.361	<0.001
Modality x Co-occurrence in phrase	0.018	0.019	0.967	0.334
Modality x AlexNet early layer	0.022	0.017	1.242	0.214
Modality x AlexNet mid layer	0.007	0.022	0.333	0.739
Modality x AlexNet late layer	-0.112	0.022	-5.157	<0.001
Modality x Word length	0.082	0.028	2.977	0.003
Modality x Orthographic distance	0.008	0.028	0.302	0.763
Modality x Word embeddings	-0.034	0.018	-1.932	0.053

Sup. Fig. 3 – Model-estimated effects of the covariates on pairwise similarity ratings for object pictures and words. Colours of points reflect the values of pairs for the given predictor and match the ones in the RDMs showed above. Stimulus modality is indicated by x-axis position (left = objects, right = words). Points reflect estimated similarity for each pair of objects averaged across all the different contexts (i.e., the third object a triplet) in which they were presented. 95 % confidence interval are represented by the shaded area around lines for continuous predictors.

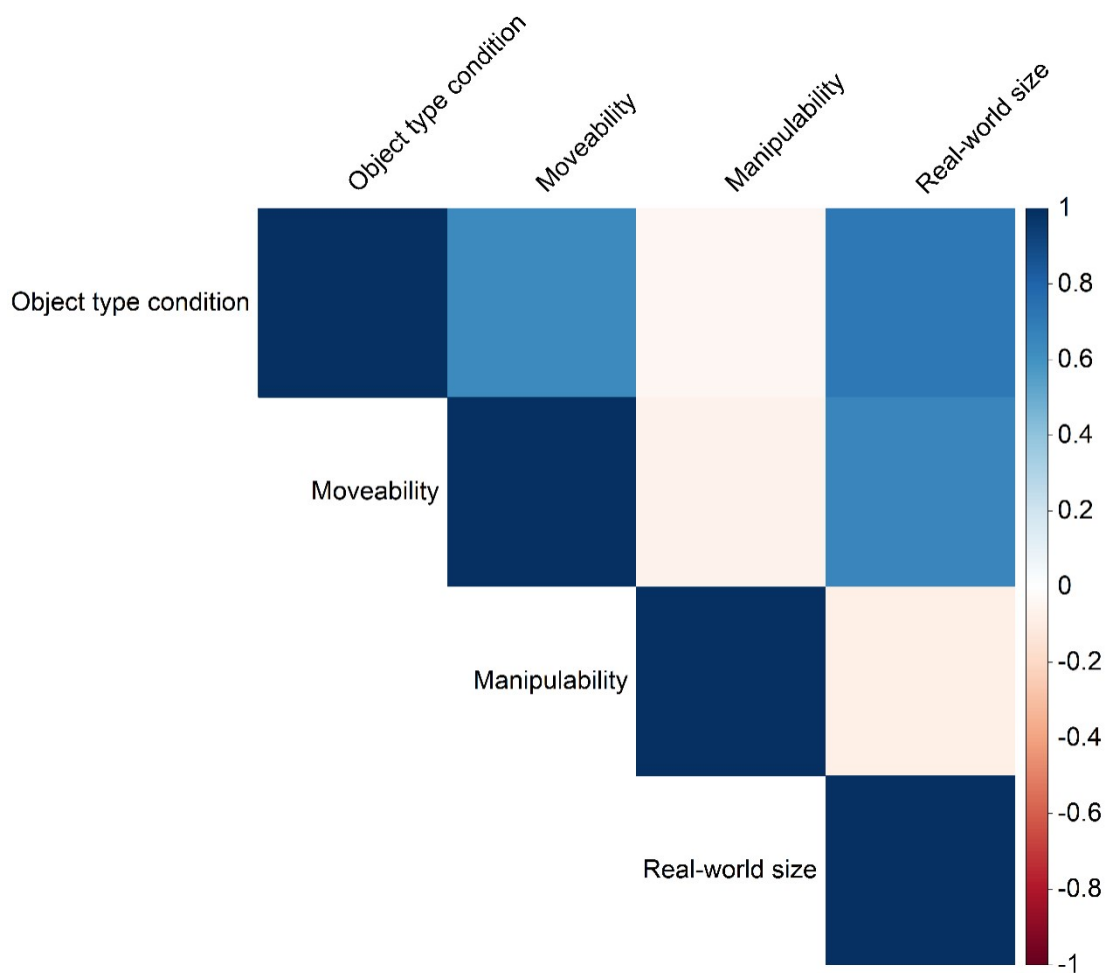


Supplementary Materials 3 – Factor correlations and VIFs in the model with ratings

We explored what makes anchor objects different from local objects (as seen from the effect of the *Object type condition* predictor), comparing this division with the ratings we collected in a separate experiment. First of all, we organized our ratings of *moveability*, *manipulability* and *real-world size* in an RDM format (similarity values were computed as the

absolute value of the difference between the two values of each pair, as done for e.g., word length). We then computed pairwise correlations between each of the ratings RDMs and the object type condition RDM. We found that object type condition had a strong correlation with real-world size ($r = 0.713$) and moveability ($r = 0.639$), with the two measures also being strongly correlated ($r = 0.659$). On the other hand, manipulability did not show to have strong correlation with either object type condition ($r = -0.042$), or moveability ($r = -0.065$) and real-world size ($r = -0.082$).

Sup. Fig. 4 – Matrix of correlations between the ratings and the object type condition factor



Second, we implemented another GLMM modeling the data with the same structure of fixed and random factors, but adding also the three rating predictors:

$$\begin{aligned}
 \text{behavioral similarity} \sim & \text{stimulus modality} * (\text{scene similarity} + \text{phrase similarity} \\
 & + \text{object type similarity} + \text{cooccurrence in scene} + \text{cooccurrence in phrase} \\
 & + \text{anchored cooccurrence} + \textbf{ratings} + \text{covariates}) \\
 & +(1 | \text{pairs}) + (1 | \text{context objects})
 \end{aligned}$$

This new model including the ratings had a significantly better fit compared to the previous one without those measures (AIC difference = 57, $\chi^2 = 58.528$, $p < 0.001$), and despite the new model being more complex in terms of number of parameters. The model also did not show problematic levels of multicollinearity, when inspecting the VIFs of each term.

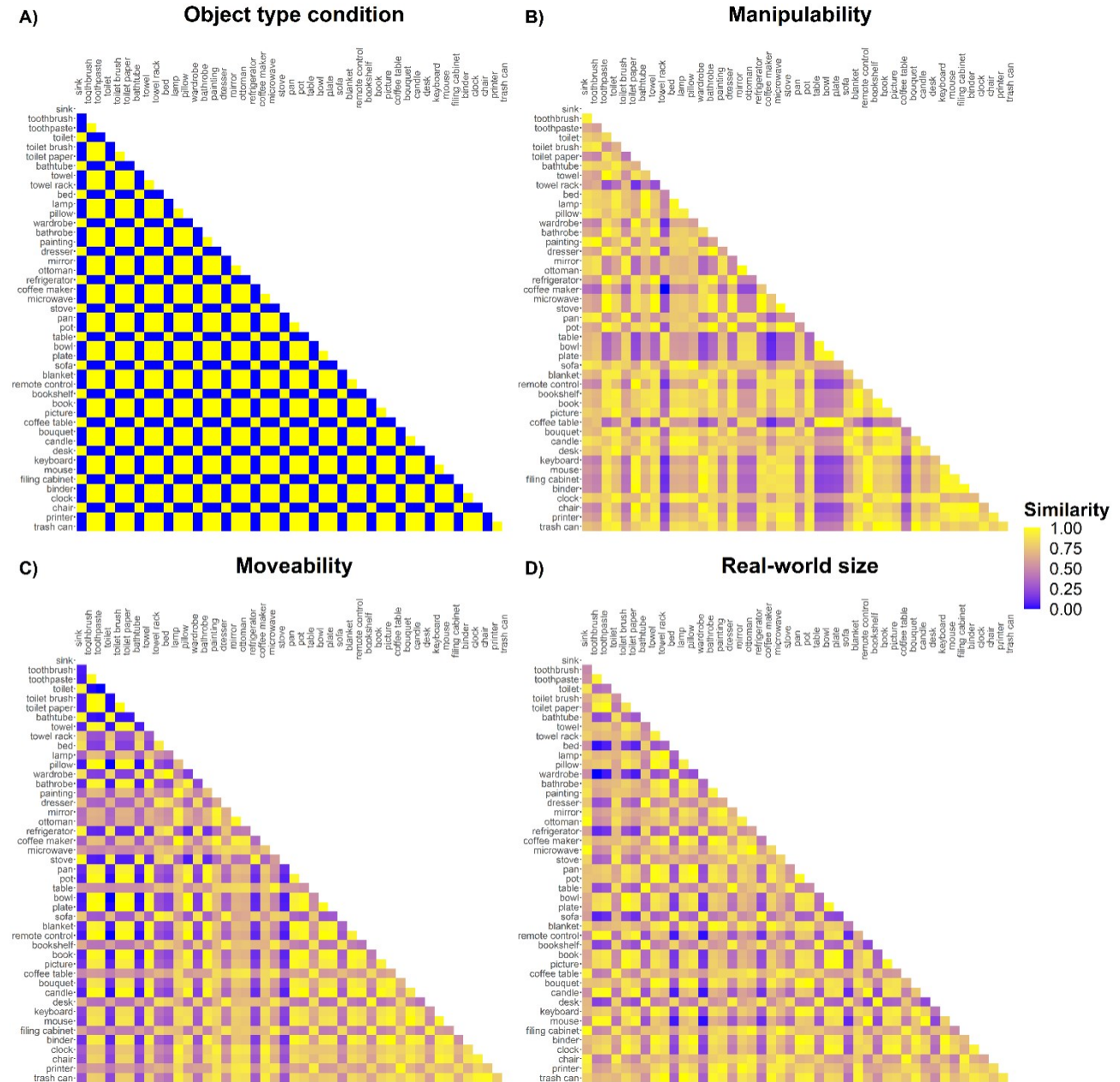
Sup. Table 1 – Variance Inflation Factors (VIFs) for the predictors used in the model including rating measures

Predictors	VIF
Modality (Words – Objects)	3.651
Moveability	2.064
Real-world size	2.518
Manipulability	1.027
Object type condition	2.359
Phrase condition	1.300

Anchored co-occurrence	1.535
Co-occurrence in scene	1.559
Co-occurrence in phrase	1.431
AlexNet early layer	1.234
AlexNet mid layer	1.769
AlexNet late layer	1.776
Word length	2.866
Orthographic distance	2.886
Word embeddings	1.197
Modality x Moveability	2.049
Modality x Real-world size	2.505
Modality x Manipulability	1.029
Modality x Object type condition	2.308
Modality x Phrase condition	3.940
Modality x Anchored co-occur.	1.535
Modality x Co-occur. in scene	1.538
Modality x Co-occur. in phrase	1.392
Modality x AlexNet early layer	1.247
Modality x AlexNet mid layer	1.772
Modality x AlexNet late layer	1.794
Modality x Word length	2.974

Modality x Orthographic distance	2.993
Modality x Word embeddings	1.190

Sup. Fig. 5 – Representational (Dis)similarity Matrices (RDMs) for the a priori object type distinction (A), and for the object features ratings (B, C and D). Every cell represents pairwise similarity for that given dimension. In A yellow represents pairs of objects that belong to the same type (maximal similarity), while blue represents pairs that belong to different types (minimal similarity). In B, C and D, absolute value of the difference between ratings of the pair is normalized to span between 0 (blue, bigger difference) to 1 (yellow, smaller difference).



Supplementary Materials 4 – Model with ratings measures

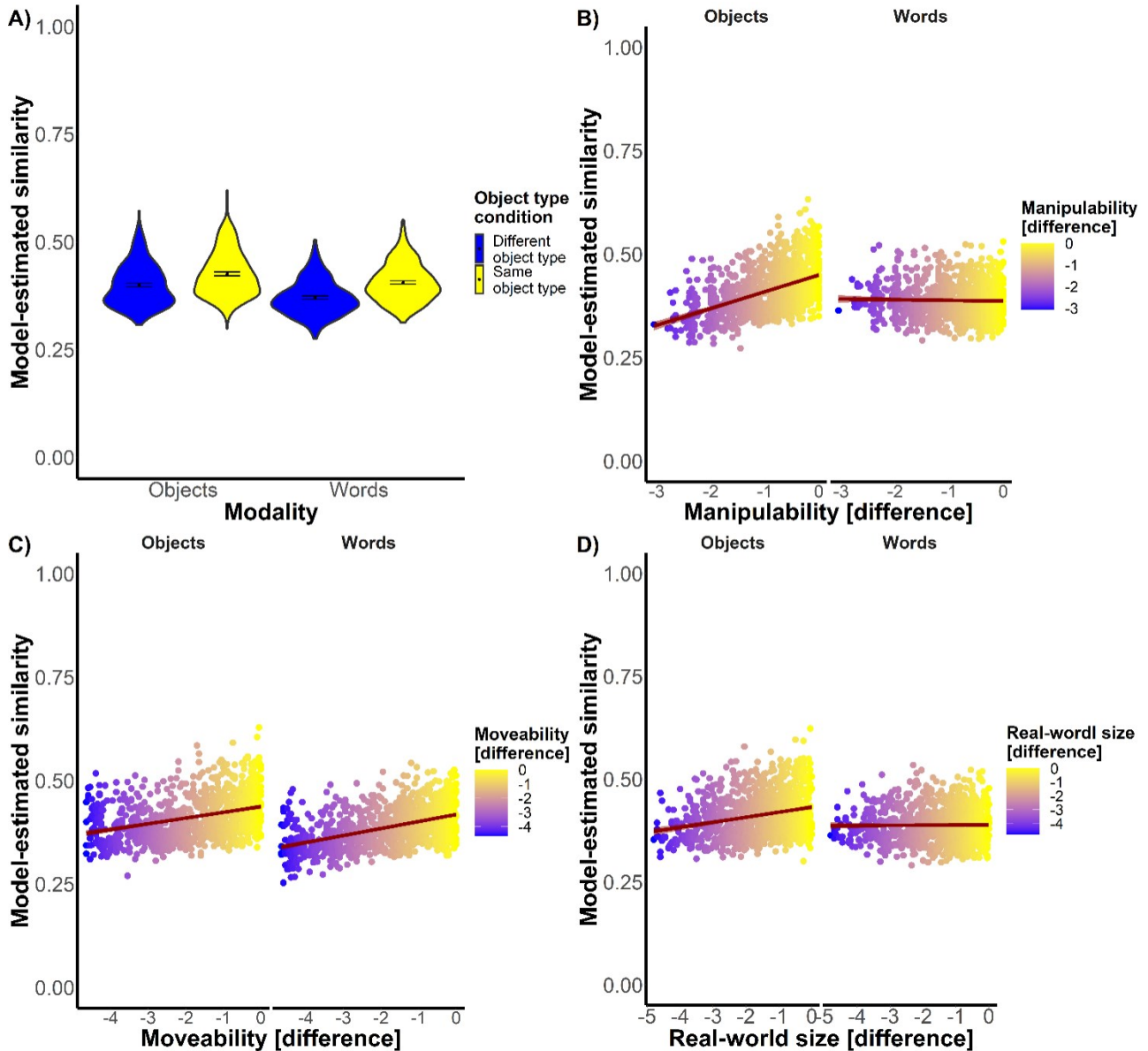
Results overall resembled the one from the previous model, but with some important differences. First, adding the rating measures, the main effect of Object type condition got strongly reduced and was no longer significant ($\beta=0.120$, $SE=0.071$, $z=1.681$, $p=0.093$). On the other hand, we found significant main effects of the newly introduced moveability ($\beta=0.079$, $SE=0.033$, $z=2.386$, $p=0.017$) and manipulability measure ($\beta=0.046$, $SE=0.023$, $z=1.984$, $p=0.047$), both showing that pairs that are similar along those dimensions are also more likely to be judge more similar behaviourally. Real-world size did not show a significant main effect ($\beta=0.022$, $SE=0.037$, $z=0.587$, $p=0.557$), but resulted in having a significant interaction with stimulus modality ($\beta=-0.057$, $SE=0.026$, $z=-2.158$, $p=0.031$), with a stronger effect of this dimension on behavioural similarity for object pictures than for words. Similarly, manipulability had a significant interaction with stimulus modality ($\beta=-0.111$, $SE=0.017$, $z=-6.713$, $p<0.001$), having a stronger effect on perceived similarity for object pictures than for words.

Sup. Table 4 – Results of the GLMM including object features ratings

Predictors	β	SE	z	p
(Intercept)	-0.314	0.065	-4.853	<0.001
Modality (Words – Objects)	-0.101	0.031	-3.252	0.001
Moveability	0.079	0.033	2.386	0.017
Real-world size	0.022	0.037	0.587	0.557
Manipulability	0.046	0.023	1.984	0.047
Object type condition (Same – Different)	0.120	0.071	1.681	0.093
Phrase condition (Same – Different)	0.249	0.128	1.951	0.051
Scene condition (Same – Different)	1.065	0.074	14.339	<0.001

Anchored co-occurrence	0.015	0.028	0.544	0.586
Co-occurrence in scene	0.387	0.029	13.488	<0.001
Co-occurrence in phrase	0.060	0.028	2.184	0.029
AlexNet early layer	-0.119	0.025	-4.658	<0.001
AlexNet mid layer	0.024	0.031	0.793	0.428
AlexNet late layer	0.122	0.031	3.971	<0.001
Word length	0.043	0.039	1.101	0.271
Orthographic distance	-0.048	0.039	-1.227	0.220
Word embeddings	0.343	0.025	13.554	<0.001
Modality x Moveability	0.019	0.024	0.807	0.420
Modality x Real-world size	-0.057	0.026	-2.158	0.031
Modality x Manipulability	-0.111	0.017	-6.713	<0.001
Modality x Object type condition	0.039	0.049	0.791	0.429
Modality x Phrase condition	0.153	0.087	1.750	0.080
Modality x Scene condition	-0.267	0.050	-5.322	<0.001
Modality x Anchored co-occurrence	0.001	0.020	0.057	0.955
Modality x Co-occurrence in scene	-0.116	0.020	-5.870	<0.001
Modality x Co-occurrence in phrase	0.021	0.019	1.107	0.268
Modality x AlexNet early layer	0.021	0.018	1.184	0.236
Modality x AlexNet mid layer	0.010	0.022	0.456	0.648
Modality x AlexNet late layer	-0.114	0.022	-5.267	<0.001
Modality x Word length	0.085	0.028	3.037	0.002
Modality x Orthographic distance	0.016	0.028	0.555	0.579
Modality x Word embeddings	-0.036	0.018	-2.025	0.043

Sup. Fig. 6 – Model-estimated effects of the object type condition predictor as well as for the object features ratings, estimated from the model including the ratings themselves. Colours of violins and points reflect the values of pairs for the given predictor and match the ones in the RDMs showed above. Stimulus modality is indicated by either x-axis position (left = objects, right = words). Points and violins reflect estimated similarity for each pair of objects averaged across all the different contexts (i.e., the third object a triplet) in which they were presented. 95 % confidence interval are represented by error bars in the violins (point is the mean), and by the shaded area around lines for continuous predictors.



Scene hierarchy structures mental object representations while flexibly adapting to varying task demands

Jacopo Turini & Melissa Le-Hoa Võ

Scene Grammar Lab, Department of Psychology and Sports Sciences,

Goethe University, Frankfurt am Main, Germany

Corresponding author:

Jacopo Turini

Scene Grammar Lab

Institut für Psychologie

PEG, Room 5.G105

Theodor-W.-Adorno Platz 6

60323 Frankfurt/Main

turini@psych.uni-frankfurt.de

+49 (0)69 798-35310

1 **Abstract**

2

3 Like words in sentences, objects in our environment do not appear randomly, but follow a
4 hierarchically organized structure. Objects can be part of a scene and, within a scene, are grouped
5 into different spatial clusters, where they serve different functions. Is this organization reflected in
6 mental representations of individual objects, and which behavioural goal does it support? We
7 collected pairwise similarity judgments for objects: 1) based on visual appearance; 2) without
8 explicit instructions; and 3) based on actions. Results showed modulations of mental
9 representations as a function of hierarchy, which were enhanced by the “actions” task, while in the
10 “visual appearance” task visual similarity played a primary role. Crucially, without explicit
11 instructions, judgements were almost identical to the “actions” task. We therefore propose that
12 scene hierarchy seems to organize mental representations of objects and can be flexibly adapted to
13 changing behavioural goals by supporting efficient interactions with objects.

14

15 **Keywords:** object representations, scene hierarchy, scene grammar, task flexibility, action goals,
16 RSA

17

18 **Statement of relevance**

19

20 We constantly look for, recognize, and interact with objects and do so effortlessly despite the many
21 things going on in our mind. One reason for this unmatched ability is that we can use knowledge
22 regarding which objects typically appear where in our environment: a toothbrush is usually found
23 in the bathroom, and therein on the sink. Here, we were interested in whether such external
24 hierarchical structure is also internally mirrored in our minds when seeing individual objects and

25 how action affordances modulate these mental representations. We asked participants to tell us
26 which of three objects is the odd-one-out and by these mere similarity judgements were able to
27 show that a scene's hierarchical structure is indeed reflected in how objects are organized in our
28 mind. Crucially, we showed that this mental organization is driven by possible actions one can
29 perform with the objects and can be flexibly adapted to task demands.

30

31

32 Introduction

33

34 Most of the time we live in and interact with a structured world where objects typically
35 appear in certain contexts and not others (e.g., a toilet appears in the bathroom and not in the
36 kitchen), as well as in certain positions and not others (e.g., a toilet is positioned against a wall, not
37 floating in the middle of the room). We have referred to knowledge regarding “what” objects
38 (“semantics”) appear “where” (“syntax”) within a scene as “Scene Grammar” in analogy to sentence
39 processing (Biederman et al., 1982; for a review, Vö, 2021). Crucially, the structure given by Scene
40 Grammar seems to aid our cognitive system in representing objects during visual perception to
41 efficiently guide attention allocation during perception of complex visual stimuli (i.e., scenes; Vö &
42 Henderson, 2009; Vö & Wolfe, 2013a). This mental organization has been shown to support a wide
43 range of behaviourally crucial processes involving objects in the environment like recognition
44 (Cornelissen & Vö, 2017), search (Vö & Wolfe, 2013b), or interaction (Draschkow & Vö, 2017).

45 More recently, we have suggested that Scene Grammar is organized hierarchically (Vö et al.,
46 2019) with the visual scene at the highest level. Within that scene, there are meaningful clusters of
47 spatially related objects, so-called “phrases” (continuing the analogy with linguistic grammar), and
48 within every phrase one can identify an object with a special status (“anchor object”), which anchors
49 strong predictions regarding both the identity and position of the other objects within the phrase
50 (“local objects”; *Fig. 1A*). According to this framework, anchor objects are supposed to be frequently
51 present in a scene, bigger in size, and mostly stable in their position (e.g., a stove), while local objects
52 are thought to be smaller and more easily moveable during interactions (e.g., a pot). There has been
53 some first evidence that the hierarchical structure of Scene Grammar also organizes mental
54 representations of individual object concepts, independent of their visual features (e.g., shape,
55 brightness) or stimulus modality (i.e., pictorial or verbal; Turini & Vö, 2022).

56 However, it remains unclear whether and how this mental organization of objects changes
57 as a function of different behavioural goals and which purpose it primarily serves. In the realm of
58 high-level vision, it has been shown that different features of complex visual stimuli can be
59 enhanced or inhibited according to different task demands, such as deployment of attention (full
60 vs. reduced, Groen et al., 2016), criteria of categorization (e.g., colour vs. real-world size, Harel et
61 al., 2014) or level of categorization (superordinate vs. basic-level, Clarke et al., 2011). At the same
62 time, it seems reasonable - given the body of previous research presented above - that such a
63 hierarchical structure would also be optimal for efficient action preparation involving the objects
64 present in the environment (Greene et al., 2016). For instance, if you were asked to help a friend
65 cooking in their new house, the hierarchical organization of objects in the scene would quickly allow
66 you to first identify which of the rooms is the kitchen, then efficiently locate the “stove phrase”
67 therein and within that phrase to locate the pot on top of the stove.

68 To investigate the impact of different tasks on the mental representation of objects, we
69 selected a set of everyday objects for which we assume a hierarchical structure (*Fig. 1B*; Turini &
70 Vö, 2022). Then, for each set of objects, pairwise similarity judgements were collected from human
71 participants, using an “odd-one-out” triplet task (*Fig. 1C*; Hebart et al., 2020). Crucially, the task was
72 performed by three different groups of participants, each of which followed a different set of
73 instructions: one group judged similarity based on the visual appearance of the objects (“Visual
74 task”), a second group judged objects’ similarity without an explicit definition of what to base the
75 similarity judgements on (“No task”), while the last group judged the similarity of objects based on
76 the actions that can be performed with those objects (“Action task”). Finally, we compared the odd-
77 one-out ratings to the hierarchy measures across tasks, employing a combination of
78 Representational Similarity Analysis (RSA; Kriegeskorte et al., 2008) and Generalized Linear Mixed-
79 effects Models (GLMMs; McCulloch & Neuhaus, 2005). As the employment of the “odd-one-out”

80 triplet task using different instructions has shown flexible prioritization of task-related features in
81 scene categorization (Greene & Hansen, 2020), we expected that aspects of scene hierarchy would
82 be enhanced by the task based on actions, while visual features would dominate similarity
83 judgements based on visual appearance. Importantly, if the “default setting” of mental
84 representations is based on actions that can be performed with objects, we would expect a strong
85 resemblance between similarity judgments without instructions and similarity judgements based on
86 actions, suggesting that the hierarchical organization of Scene Grammar is supporting efficient
87 interaction with objects in the environment in order to perform goal-oriented behaviour.

88

89

90

91

92

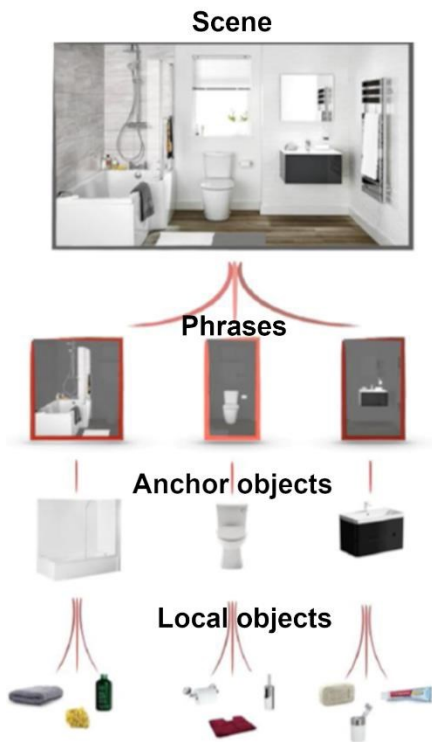
93

94

95

96 **Figure 1** – Schema of Scene hierarchy (A), how we measured it (B) and the tasks to collect similarity judgements (C),
97 adapted from Turini & Vö (2022). A) Presentation of the hierarchical organization of objects in scenes (adapted from Vö
98 et al., 2019): scenes is divided into clusters of objects (phrases), and every phrase contain one anchor objects pointing
99 to several local objects. B) Such hierarchical organization was measured here in two ways: *a priori*, assigning objects to
100 different scenes, to different phrases within a scene and to different types within a phrase (anchor vs local); and *data-*
101 *driven*, using a set of real-world images with annotations and segmentations of objects (image from Greene, 2013,
102 visualized via LabelMe, Russel et al., 2008). C) Example trial (with our stimuli) for the odd-one-out triplet task (Hebart
103 et al., 2020): participants saw a triplet of objects and had to select the one they considered the least similar to the other
104 two. Different groups have different definitions of similarity: one based on visual appearance, one that was not
105 explicated and one based on action goals.

A) Proposed hierarchical organization of objects in scenes:



B) A priori hierarchy:
assigned based on common sense and intuition

Bathroom:
 Phrase 1: sink (anchor),
 toothbrush (local),
 toothpaste (local)
 Phrase 2: toilet (anchor),
 toilet brush (local),
 toilet paper (local),
 Phrase 3: bathtub (anchor),
 towel rack (local),
 towel (local)

Bedroom:
 Phrase 1: bed (anchor),
 lamp (local),
 pillow (local),
 ...

Data-driven hierarchy:
estimated from a dataset of real-world scene images




C) Which object is the least similar to the other two?



Similarity based on visual appearance ("Visual task")



Similarity with no explicit definition ("No task")



Similarity based on action goals ("Action task")

106

107

108 **Materials and Methods**

109

110 Materials and methods used in this study mostly resemble the ones employed in a previous study
 111 by the same authors (Turini & Vö, 2022). We highlighted crucial differences in the text.

112

113 **Participants**

114 One hundred and twenty-nine (129) adult volunteers took part in our study. One third of them
 115 (N=43) participated in Experiment 1 ("Visual task"; age: M = 22.28 yrs, SD = 5.69 yrs, range = 18 – 45
 116 yrs; gender: F = 33, M = 9, NB = 1); another third participated in Experiment 2 ("No task"; age: M =

117 24.72 yrs, SD = 5.33 yrs, range = 18 – 40 yrs; gender: F = 31, M= 12); and the last third participated
118 in Experiment 3 (“Action task”; age: M = 21.95 yrs, SD = 3.64 yrs, range = 18 – 33 yrs; gender: F = 36,
119 M= 6, NB = 1). Data collected in the “No task” (Experiment 2) were already analyzed in a previous
120 study investigating the hierarchical organization of mental representations and the role of different
121 stimulus modalities (Turini & Vö, 2022).

122 The participants’ sample size in each experiment (N=43) was chosen as an optimal number
123 between the total number of unique trials and an optimal number of trials to present to an
124 individual participant. All participants reported to have normal or corrected-to-normal vision and to
125 have no history of neurological or psychiatric disorders. All of them gave their informed consent and
126 received course credits or monetary reimbursement for their participation. The Ethics Committee
127 of the Goethe University Frankfurt approved all experimental procedures that were carried out in
128 line with the Declaration of Helsinki (approval #2014-106).

129

130 **Stimuli**

131 Forty-five objects of familiar usage were considered for the experiment. We collected pictures of
132 the objects in isolation from copyright-free databases on the internet; we pasted them on a white
133 background, grey-scaled them to remove influence of color, and resized them to 392 x 392 pixels.
134 The same stimuli were used in a previous study (Turini & Vö, 2022) and remained the same across
135 the three experiments reported here.

136

137 **Scene hierarchy predictors**

138 To measure if similarity judgments were organized according to a scene hierarchy, two sets of
139 predictors were estimated (for more details see Turini & Vö, 2022).

- 140 - *A priori hierarchy predictors*: we based these predictors on common sense and intuition; we
141 assigned the 45 objects to one of 5 different indoor *scenes* (bathroom, bedroom, kitchen,
142 living room and home office). Within every scene, objects were grouped in 3 different
143 *phrases*; finally, for every phrase, 1 object was assigned the role of *anchor object*, and the
144 other 2 as *local objects*.
- 145 - *Data-driven hierarchy predictors*: we based these predictors on the occurrence in a dataset
146 of real-world scene images with segmented and annotated objects (Greene, 2013; see
147 *Supplementary Materials 1* for more details). From this dataset, we measured how many
148 times *two objects co-occur in the same image*. Then, a clustering algorithm was run to
149 measure how objects grouped together spatially in every image (as in Boettcher et al., 2018,
150 see *Supplementary materials 1*): we identified the obtained *clusters of objects as phrases*,
151 and within each cluster, we considered the *object with the largest area as anchor object*,
152 while the other objects in each cluster were assigned the status of *local objects*.

153

154 **Visual features covariates**

155 As in Turini & Vö (2022), we made sure that effects of the scene hierarchy did not result from a
156 confound of purely perceptual information by considering different measures of visual features of
157 our object pictures: we fed the object images to a pre-trained Deep Neural Network (DNN), a state-
158 of-the-art computer vision algorithm that is trained to perform object categorization with human-
159 like performance. We used AlexNet, trained on the ImageNet dataset (Krizhevsky et al., 2017), which
160 is based on many sequential layers of processing units that extract and transform features from the
161 previous layer. The first layer extracts features from the pixel values of an image; then the
162 information is transformed in the intermediate layers until the final output layer is reached, where
163 the image is finally assigned to a category (e.g., “bed”). Unit activations for our object images were

164 estimated in 3 different layers of the network: convolutional layer 1 (*conv1*, “early layer”) processing
165 low-level visual features (e.g., edges, brightness); convolutional layer 4 (*conv4*, “mid layer”) processing
166 mid-level visual features (e.g., shape); and the fully connected layer 7 (*fc7*, “late layer”) processing
167 high-level visual features (e.g., complex configurations like object parts).

168

169 **Apparatus and Procedure**

170 Apparatus and procedure were almost identical across the three experimental tasks (and mostly
171 replicated the ones of Turini & Vö, 2022). We employed an “odd-one-out” triplet task introduced
172 by Hebart and colleagues (2020), which has been used to collect pairwise similarity judgments of
173 object pictures. As a first step, all the possible combinations of triplets of stimuli were generated for
174 each experiment separately ($45! / (3! * (45 - 3)!) = 14190$ unique triplets). Then, triplets were
175 randomly divided into 43 groups of 330 triplets to ensure a practical number of trials and
176 participants. Therefore, the task was performed on a different subset of triplets for each participant.

177 Stimulus presentation and response recording were programmed in Python using PsychoPy
178 (version 2020.2.4, Builder GUI; Peirce et al., 2019) and the procedure was carried on online hosted
179 by the platform Pavlovia (<https://pavlovia.org/>). We informed participants that the experiment
180 would take between 30-60 min., and asked them to carry it out in a calm and undisturbed
181 environment. Then participants were told that they would see triplets of object pictures and their
182 task was to choose the “odd-one-out” stimulus, i.e., the one they considered the least similar with
183 respect to the other two. In Experiment 1, participants were instructed to base their decision solely
184 on visual similarity. In Experiment 2, no explicit definition of similarity was given to participants (see
185 Turini & Vö, 2022). Finally, in Experiment 3, participants were instructed to base their decision on
186 similarity in terms of action goals.

187 Triplets were presented on a white background screen: one stimulus on the left, one stimulus
188 in the center, and one stimulus on the right (with randomized position of each stimulus within every
189 triplet; see *Fig. 1C*). We programmed the experiment so that the size of object pictures was
190 normalized based on screen size, to have every participant seeing stimuli occupying the same
191 proportion of screen (a picture spanned about 1/4 of screen width and height). We asked
192 participants to press keyboard arrows to choose the odd-one-out stimulus: left arrow for the
193 stimulus on the left, down arrow for the stimulus in the center, right arrow for the stimulus on the
194 right. After the key was pressed, a black fixation cross was presented for 500 ms in the center of the
195 screen, followed by the presentation of the next triplet. Experiments were divided into 6 blocks with
196 breaks in between. Participants could take as much time as they wanted to make their response,
197 and if they were not sure about the answer, they were invited to make the best judgement they
198 could.

199

200 **Analysis**

201 As in Turini & Vö (2022), we combined two approaches for the analysis: Representational Similarity
202 Analysis (RSA, Kriegeskorte et al., 2008) and Generalized Linear Mixed-effects Models (GLMMs,
203 McCulloch & Neuhaus, 2005). RSA is an analytical method that allows to compare the
204 representational organizations of data with different dimensionalities (e.g., brain recordings,
205 behavioural responses, computational models, stimulus features): data are organized into
206 Representational (Dis)similarity Matrices (RDMs), symmetric matrices whose column and row
207 entries typically correspond to the different stimuli. Every cell of an RDM represents a measure of
208 (dis)similarity for that pair of stimuli. Once different data are represented in the same RDM format,
209 it is possible to estimate how similar two RDMs are, i.e., how the organization of pairwise similarity
210 in one source (e.g., behaviour) is predicted by the organization of pairwise similarity in another

211 source (e.g., a computational model). According to this, we have computed pairwise similarity for
212 the behavioural responses in the “odd-one-out” tasks, as well as for the hierarchical predictors and
213 the covariate computational models introduced previously (see RDMS for the hierarchy predictors
214 and the covariate factors in Figure 2).

215

216

217

218

219

220

221

222

223

224

225

226

227

228

229 **Figure 2** – Representational (Dis)similarity Matrices (RDMs) of the a priori hierarchical predictors (A, B and C), for the
230 data-driven hierarchical predictors (D, E and F) and for the visual covariates (G, H and I). In A, B and C, yellow cells 231
232 indicate pairs of objects that appear in the same scene, phrase or belong to the same type (maximal similarity, coded
233 as “1”), while blue cells indicate pairs that appear in different scenes or belong to different types (minimal similarity,
234 coded as “0”). Gray cells in the “phrase condition” indicates pairs that appear in different phrases within the same scene
235 (medium similarity, coded as “0.5”). In D, the $\log_{10}(\text{counts} + 1)$ of co-occurrence is normalized to span between 0 (blue,
236 few counts) to 1 (yellow, many counts). In E and F, the colours represent proportion of counts to the total co-occurrence
237 counts of each pair (blue = 0, no counts; yellow = 1, all counts). In G, H, and I, colours represent the correlation between
activation vectors (blue = 0, no correlation; yellow = 1, maximal correlation).

246 between the “odd-one” stimulus and one of the unselected stimuli was considered minimal
247 ($\text{Sim}(C,A) = 0 \rightarrow$ “dissimilar”; $\text{Sim}(C,B) = 0 \rightarrow$ “dissimilar”).

248 - *A priori hierarchy similarity (Fig. 2, first row)*: we estimated pairwise similarity according to
249 the hierarchical status assigned *a priori*. This resulted in 3 categorical predictors: *scene*
250 *condition*, with a dichotomic categorization in pairs from the same scene (dummy coded as
251 1) vs. pairs from different scene (dummy coded as 0); *phrase condition*, based on three
252 groups, with pairs from the same phrase (1) vs pairs from different phrases within the same
253 scene (0.5) vs pairs from different phrases in different scenes (0); finally, *object type*
254 *condition*, divided between pairs of objects of the same type (1) vs pairs of objects of
255 different type (0; “object type” indicates if objects are either anchor or a local objects).

256 - *Data-driven hierarchy similarity (Fig. 2, second row)*:: we estimated pairwise similarity
257 according to the hierarchical status measured in the labelled image dataset. This resulted in
258 3 continuous predictors: *co-occurrence of pairs in a scene*, which counts how many times a
259 pair of objects appears in the same image (measured as logarithm based 10 of counts + 1, to
260 make distribution more uniform and avoiding -Infinite values); then, *co-occurrence of pairs*
261 *in a phrase*, measuring the proportion of co-occurrence counts in which a pair belongs not
262 just to the same image but to the same cluster; finally, *anchored co-occurrence*, measuring
263 the proportion of co-occurrence counts in which one object is “anchored” to the other in the
264 phrase.

265 - *Covariates (Fig. 2, last row)*:: for the three visual measures (i.e., the activation from the 3
266 layers of AlexNet DNN), similarity was measured via product-moment correlation coefficient
267 of pairs of activation units vectors (e.g., the activation vector for “toilet” in convolutional
268 layer 1 and the activation vector for “sink” in convolutional layer 1).

269

270 GLMMs are a more general extension of Linear Mixed-effects Models (LMMs; Bates et al.,
271 2014) to model responses with a non-gaussian distribution. The main advantage of (G)LMMs over
272 simple regression models and ANOVAs is that it is possible to consider each trial from each
273 participant at the same time, without any need to aggregate data or to estimate separately the
274 effects across participants and stimuli (Baayen et al., 2008). In (G)LMMs, a response variable is
275 measured as a function of many predictors (fixed factors) while considering grouping factors that
276 identify common portions of variance (random factors; see *Supplementary Materials 2* for more
277 details).

278 In our model, the dichotomic behavioural similarity (similar vs. dissimilar) from the triplet
279 task represented the response variable; as fixed factors, we considered the 3 categorical predictors
280 from the a priori hierarchy, the 3 continuous predictors from the data-driven hierarchy as well as
281 the 3 continuous predictors from the visual covariates; additionally, a categorical predictor
282 identifying the tasks (“Visual task”, “No task” and “Action task”) was introduced in interaction with
283 each of the other fixed factors; finally, as random factors, we considered individual participants,
284 individual object pairs, and individual “context” objects (which is the third object present in every
285 triplet; for more details see *Supplementary Materials 2*).

286 Combining RSA with GLMMs (Turini & Vö, 2022), we can measure the impact of many RDMs
287 at the same time, as in a general linear model (Proklova et al., 2016; Kaiser et al., 2019), instead of
288 a single RDM (Kriegeskorte et al., 2008), while additionally taking into account multiple sources of
289 group-based variance (i.e., our random factors: participants, pairs and context objects). Correlations
290 between factors and the test of multicollinearity can be found in *Supplementary Materials 3*.
291 Analyses were performed in R (version 3.6.3, R Core Team, 2020). Data and scripts are available at
292 the following link: <https://osf.io/mxkwz/>.

293

294 **Results**

295 We organized the behavioural ratings divided by task into the RDM format (see *Figure 3*);
296 there, every cell represents the pairwise similarity ratings for a given pair of objects averaged across
297 all the triplets where the pair is present. From mere visual inspection, one can already tell that the
298 structure of the RDM for ratings in the “action task” (A) seems almost identical to the structure of
299 RDM for ratings in the “no task” (B) (no task – action task $r=.92$), while the structure of RDM for
300 ratings in the “visual task” (C) differs more distinctly from the other two (no task – visual task $r=.76$;
301 action task – visual task $r=.70$).

302

303

304

305

306

307

308

309

310

311

312

313

314

315 **Figure 3** – Representational (Dis)similarity Matrices (RDMs) for the behavioural similarity ratings collected in the three
316 experiments (A = Visual task, B = No task, C = Action task). Cells represent pairwise similarity ratings averaged across all
317 the triplets where the pair was present. Every pair was presented in a triplet with all the other remaining objects, and
318 it was judged either as similar (1) or dissimilar (0), so that In the RDMs pairwise similarity spans from 0 (never judged as
319 similar = blue) to 1 (always judged as similar = yellow).

320

A) Behavioural similarity (Visual task)

321

322

323

324

325

326

B) Behavioural similarity (No task)

328

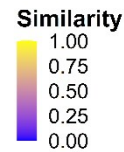
329

330

331

332

333



334

C) Behavioural similarity (Action task)

335

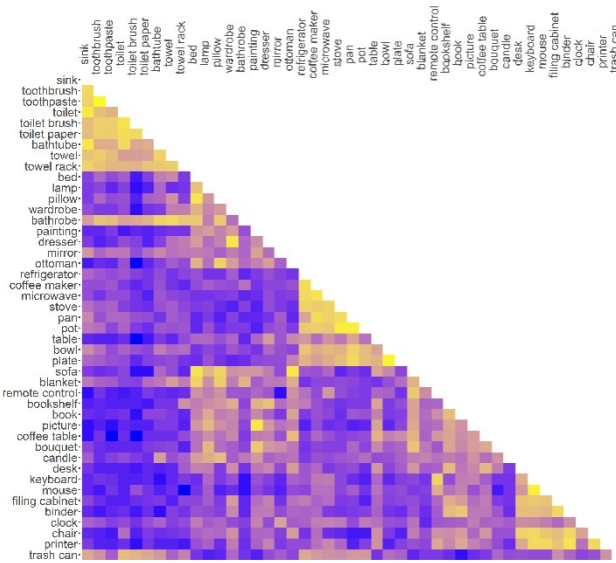
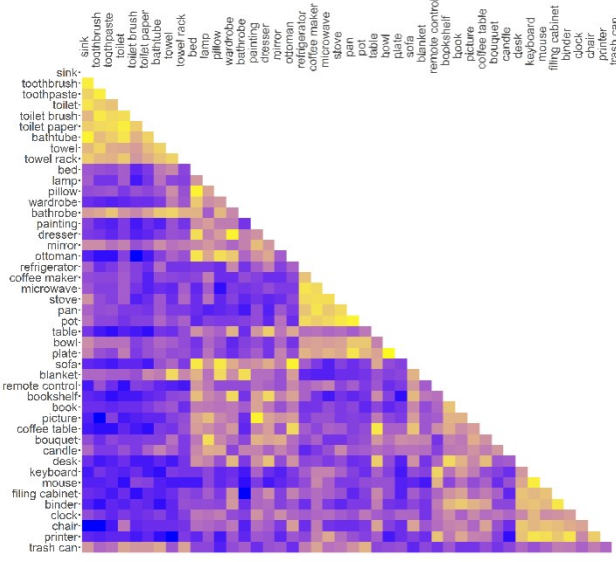
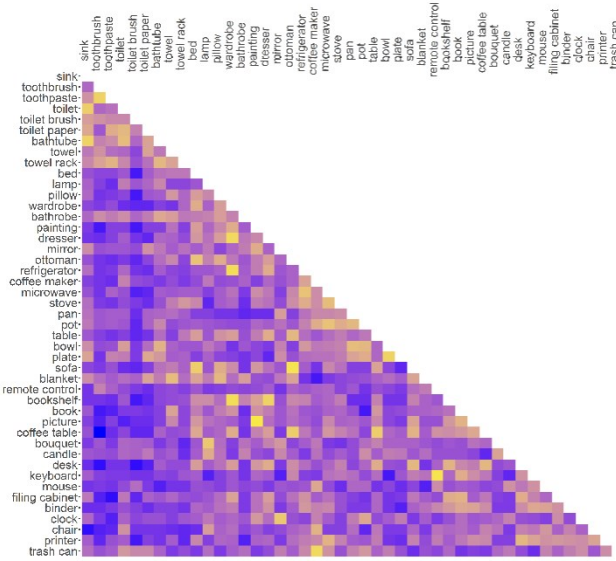
336

337

338

339

340



341

342 Results from the GLMM for the *a priori hierarchy* (see *Figure 4 A-C*) showed a significant main
343 effect of Scene condition ($\beta=1.142$, $SE=0.074$, $z=15.335$, $p<0.001$) and Object type condition
344 ($\beta=0.175$, $SE=0.049$, $z=3.586$, $p<0.001$), and a trend for the Phrase condition ($\beta=0.234$, $SE=0.130$,
345 $z=1.800$, $p=0.072$): Objects that were assigned to the same scene or to the same object type were
346 judged to be more similar than objects assigned to different scenes or types. Numerically, also pairs
347 assigned to the same phrase were judged to be more similar than pairs from different phrases of
348 the same scene. Looking at these effects across the different tasks, we found that the Scene
349 condition effect was similarly pronounced between Action task and No task ($\beta=0.049$, $SE=0.049$,
350 $z=1.005$, $p=0.315$), while in the Visual task condition the effect was largely reduced ($\beta=-0.773$,
351 $SE=0.049$, $z=-15.926$, $p<0.001$).

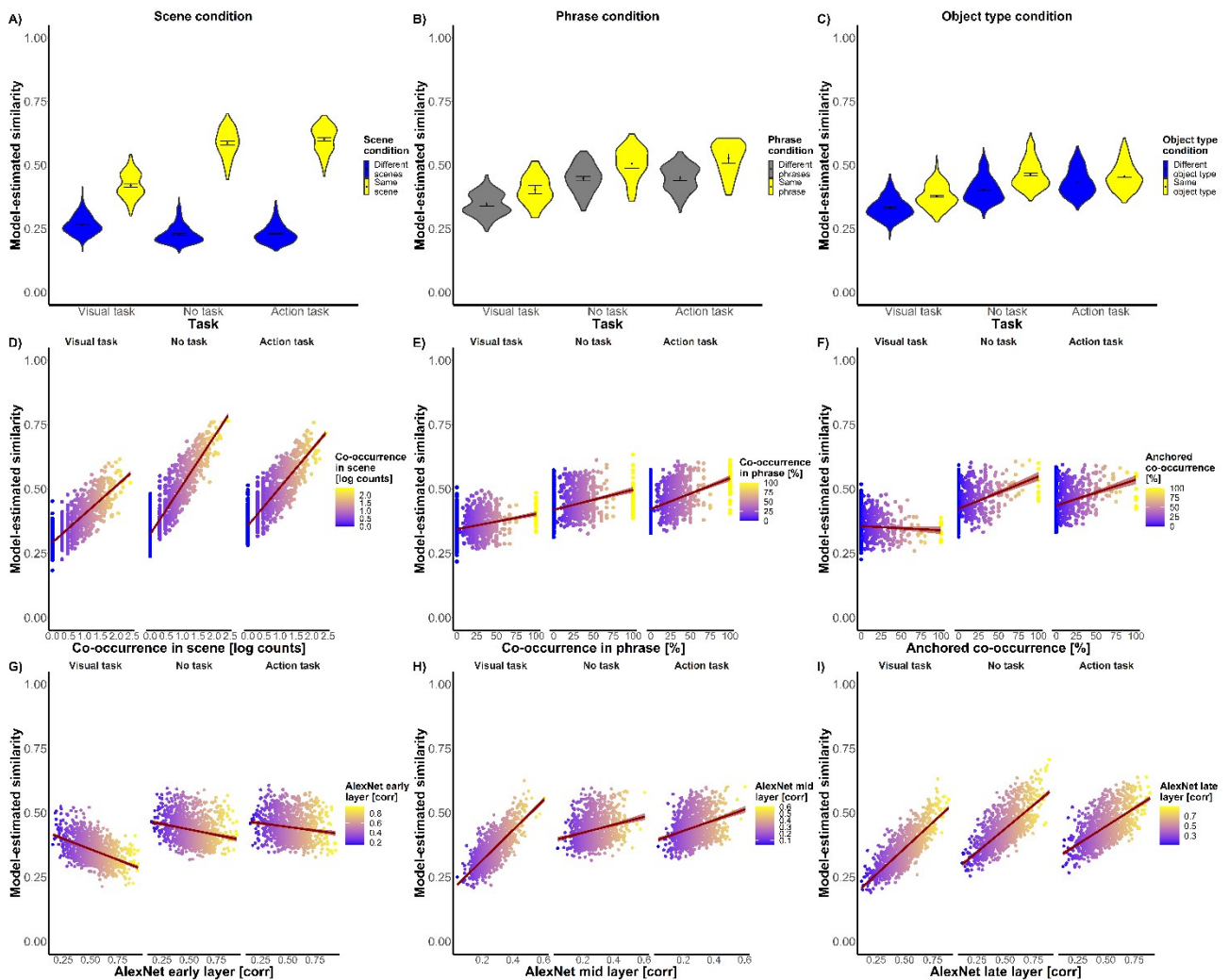
352 For the *data-driven hierarchy* (see *Figure 4 D-F*), we found a significant main effect of Co-
353 occurrence in scene ($\beta=0.325$, $SE=0.029$, $z=11.334$, $p<0.001$) and of Co-occurrence in phrase
354 ($\beta=0.059$, $SE=0.028$, $z=2.111$, $p=0.035$), but not of Anchored co-occurrence ($\beta=0.022$, $SE=0.028$,
355 $z=0.782$, $p=0.434$). That is, objects that co-occurred more often in the same scene or in the same
356 phrase were judged as more similar than objects co-occurring less often in the same scene or
357 phrase. The effect of Co-occurrence in scene was stronger in the No task condition than in the other
358 two (vs. Action task: $\beta=-0.102$, $SE=0.019$, $z=-5.332$, $p<0.001$; vs. Visual task: $\beta=-0.158$, $SE=0.019$, $z=-$
359 8.282 , $p<0.001$), while the effect of Co-occurrence in phrase was enhanced in the Action task
360 compared to No task ($\beta=0.042$, $SE=0.019$, $z=2.204$, $p=0.028$). Finally, the effect of Anchored co-
361 occurrence was similarly strong in the Action task and No task conditions ($\beta=-0.017$, $SE=0.019$, $z=-$
362 0.893 , $p=0.372$), but was reduced in the Visual task ($\beta=-0.083$, $SE=0.019$, $z=-4.447$, $p<0.001$).

363 Finally, for the *visual covariates* (see *Figure 4 G-I*), we found that the effect of Early layer
364 similarity and Mid layer similarity were significant and stronger in the Visual task than in other tasks

365 (Early layer: $\beta=-0.053$, $SE=0.016$, $z=-3.227$, $p=0.001$; Mid layer: $\beta=0.151$, $SE=0.021$, $z=7.119$,
 366 $p<0.001$). In particular for the Early layer, object pairs that were more similar in terms of low-level
 367 visual features (e.g., edges, brightness) were judged as less similar, while for the Mid layer, objects
 368 pairs that showed stronger similarity in terms of mid-level visual features (e.g., shape) were judged
 369 to be more similar. More detailed results can be found in the *Supplementary Materials 4*.

370

371 **Figure 4** – Model-estimated effects of the hierarchy and visual predictors on pairwise similarity ratings for objects in the
 372 three similarity tasks. Colours of violins and points reflect values of pairs for the given predictor and match the ones in
 373 the RDMs showed in Figure 2. Task is indicated by x-axis position for categorical predictors and different facets for
 374 continuous predictors (left = Visual task, center = No task, right = Action task). Points and violins reflect estimated
 375 similarity for each pair of objects averaged across all the different contexts (i.e., the third object in a triplet) in which
 376 they appeared. 95 % confidence interval are represented by error bars in the violins (point is the mean), and by the shaded area around lines for continuous predictors.



378

379

380 **Discussion**

381

382 How do current task demands alter the mental representations of real-world objects? Do
383 they become more or less hierarchical in nature as a function of the degree of action a task affords?
384 By mere use of behavioural similarity judgments, we were able to make visible the otherwise
385 invisible content of our mind, i.e., the hierarchical arrangement of objects in the environment which
386 is also mirrored in the organization of their mental representations (see also Turini & Vö, 2022).
387 More importantly, we were able to show that, despite being stable across different tasks, such
388 organization flexibly adapts to varying task demands. In particular, we were able to show how more
389 detailed aspects of the hierarchical organization, such as the phrasal structure assumed by clusters
390 of objects, become more defined when objects are considered in the context of actions typically
391 performed with them. On the contrary, when objects were considered purely with regard to their
392 visual appearance, not only visual features gained importance in guiding behaviour, but the
393 hierarchical organization of object representations also diminished. Interestingly, when objects
394 were judged without explicit task instructions, their behavioural estimation closely resembled the
395 one based on actions. This suggests that the potential actions one can perform with objects provide
396 a default organizing principle for mental representations.

397 On the one hand, the stability of hierarchical organization of object representations across
398 the different tasks (shown by the significant main effects of hierarchy predictors) is additional
399 evidence of how this structure of the environment has a deep and ubiquitous influence on memory
400 representations for objects. We have shown previously that these object representations are
401 independent of stimulus modality (i.e., pictorial vs verbal) or perceptual aspects of the stimuli (e.g.,
402 visual appearance, orthography; Turini & Vö, 2022). The idea that such rich contextual information

403 (i.e., the hierarchy) could be incorporated in the knowledge associated with single, isolated objects
404 resonates with Bar's proposal of "context frames" (Bar, 2004), according to which an object's
405 memory representation captures, for instance, other associated objects and their relative position
406 to each other. In line with this, it has been shown that single object processing involves the
407 activation of a wide and complex set of information, such as an object's occurrence in relation to
408 other objects both in vision and in language (Bonner & Epstein, 2021).

409 On the other hand, the flexibility shown by the hierarchical organization and the
410 resemblance of object representations between "Action task" and "No task", revealed the
411 importance of action and action preparation in default visual object processing. Already the
412 ecological approach to visual perception proposed by Gibson (Gibson, 1986) emphasized the rapid
413 extraction of motor-related information from visual input ("affordances"), although in our case we
414 instructed participants to evaluate object similarity based on actions performed together or with
415 similar goals (e.g., cooking, cleaning, eating, reading), not simply on the similarity of their motor
416 output (e.g., tilting, lifting, handling, pressing, pushing). In general, these results seem to back the
417 proposal that action, perception, and memory are closely linked processes that support each other
418 in continuously updated cycles by purposefully responding to constantly changing task demands
419 (see Hayhoe, 2017).

420 To conclude, we replicated previous findings showing that the hierarchical organization of
421 objects in the environment is reflected in how objects are represented in the mind. Furthermore,
422 we showed that this mental organization is stable across tasks, yet flexibly prioritizing different task-
423 related information. We therefore propose that incorporating the hierarchical structure of our
424 environments in object representations has the feat of supporting object processing for efficient
425 interactions with them during goal-directed behaviour and is therefore a crucial part of the
426 unmatched human action and perception abilities.

427

428 **Acknowledgments**

429

430 This work was supported by SFB/TRR 26 135 project C7 to Melissa L.-H. Vö and the Hessisches
431 Ministerium fuer Wissenschaft und Kunst (HMWK; project “The Adaptive Mind”).

432

433

434

435

436

437

438

439

440

441

442

443

444

445

446

447

448 **Bibliography**

449

450 Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random

451 effects for subjects and items. *Journal of Memory and Language*, 59(4), 390–412.

452 <https://doi.org/10.1016/j.jml.2007.12.005>

453 Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience*, 5(8), 617-629.

454 Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting Linear Mixed-Effects Models using

455 lme4. *ArXiv:1406.5823 [Stat]*. <http://arxiv.org/abs/1406.5823>

456 Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging
457 objects undergoing relational violations. *Cognitive Psychology*, 14(2), 143–177.

458 [https://doi.org/10.1016/0010-0285\(82\)90007-X](https://doi.org/10.1016/0010-0285(82)90007-X)

459 Boettcher, S. E. P., Draschkow, D., Dienhart, E., & Vö, M. L.-H. (2018). Anchoring visual search in
460 scenes: Assessing the role of anchor objects on eye movements during visual search. *Journal of*

461 *Vision*, 18(13), 11. <https://doi.org/10.1167/18.13.11>

462 Bonner, M. F., & Epstein, R. A. (2021). Object representations in the human brain reflect the co-
463 occurrence statistics of vision and language. *Nature Communications*, 12(1), 4081.

464 <https://doi.org/10.1038/s41467-021-24368-2>

465 Clarke, A., Taylor, K. I., & Tyler, L. K. (2011). The Evolution of Meaning: Spatio-temporal Dynamics of
466 Visual Object Recognition. *Journal of Cognitive Neuroscience*, 23(8), 1887–1899.

467 <https://doi.org/10.1162/jocn.2010.21544>

468 Cornelissen, T. H. W., & Vö, M. L.-H. (2017). Stuck on semantics: Processing of irrelevant object-
469 scene inconsistencies modulates ongoing gaze behavior. *Attention, Perception, &*

470 *Psychophysics*, 79(1), 154–168. <https://doi.org/10.3758/s13414-016-1203-7>

471 Draschkow, D., & Vö, M. L.-H. (2017). Scene grammar shapes the way we interact with objects,
472 strengthens memories, and speeds search. *Scientific Reports*, 7(1), 16471.

473 <https://doi.org/10.1038/s41598-017-16739-x>

474 Gibson, J. J. (1986). *The ecological approach to visual perception*. Hills-dale, NJ: Lawrence.

475 Greene, M. R. (2013). Statistics of high-level scene context. *Frontiers in Psychology*, 4.

476 <https://doi.org/10.3389/fpsyg.2013.00777>

477 Greene, M. R., Baldassano, C., Esteva, A., Beck, D. M., & Fei-Fei, L. (2016). Visual scenes are
478 categorized by function. *Journal of Experimental Psychology: General*, *145*(1), 82–94.
479 <https://doi.org/10.1037/xge0000129>

480 Greene, M. R., & Hansen, B. C. (2020). Disentangling the Independent Contributions of Visual and
481 Conceptual Features to the Spatiotemporal Dynamics of Scene Categorization. *The Journal of*
482 *Neuroscience*, *40*(27), 5283–5299. <https://doi.org/10.1523/JNEUROSCI.2088-19.2020>

483 Groen, I. I. A., Ghebreab, S., Lamme, V. A. F., & Scholte, H. S. (2016). The time course of natural
484 scene perception with reduced attention. *Journal of Neurophysiology*, *115*(2), 931–946.
485 <https://doi.org/10.1152/jn.00896.2015>

486 Harel, A., Kravitz, D. J., & Baker, C. I. (2014). Task context impacts visual object processing
487 differentially across the cortex. *Proceedings of the National Academy of Sciences*, *111*(10).
488 <https://doi.org/10.1073/pnas.1312567111>

489 Hayhoe, M. M. (2017). Vision and Action. *Annual Review of Vision Science*, *3*(1), 389–413.
490 <https://doi.org/10.1146/annurev-vision-102016-061437>

491 Hebart, M. N., Zheng, C., Pereira, F., & Baker, C. I. (2020). *Revealing the multidimensional mental*
492 *representations of natural objects underlying human similarity judgments* [Preprint]. PsyArXiv.
493 <https://doi.org/10.31234/osf.io/7wrgh>

494 Kaiser, D., Turini, J., & Cichy, R. M. (2019). A neural mechanism for contextualizing fragmented
495 inputs during naturalistic vision. *ELife*, *8*, e48182. <https://doi.org/10.7554/eLife.48182>

496 Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis – connecting
497 the branches of systems neuroscience. *Frontiers in Systems Neuroscience*.
498 <https://doi.org/10.3389/neuro.06.004.2008>

499 Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional
500 neural networks. *Communications of the ACM*, *60*(6), 84–90. <https://doi.org/10.1145/3065386>

501 McCulloch, C. E., & Neuhaus, J. M. (2005). Generalized Linear Mixed Models. *Encyclopedia of*
502 *Biostatistics*.

503 Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv,
504 J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1),
505 195–203. <https://doi.org/10.3758/s13428-018-01193-y>

506 Proklova, D., Kaiser, D., & Peelen, M. V. (2016). Disentangling Representations of Object Shape and
507 Object Category in Human Visual Cortex: The Animate–Inanimate Distinction. *Journal of*
508 *Cognitive Neuroscience*, *28*(5), 680–692. https://doi.org/10.1162/jocn_a_00924

- 509 Russel, B. C., Torralba, A., Murphy, K. P., & Freeman, W. T. (2008). LabelMe: A database and web-
510 based tool for image annotation. *International Journal of Computer Vision*, 77(1), 157–173.
- 511 Turini, J., & Võ, M. L. H. (2022). Hierarchical organization of objects in scenes is reflected in mental
512 representations of objects. *PsyArXiv*
- 513 Võ, M. L. H., & Henderson, J. M. (2009). Does gravity matter? Effects of semantic and syntactic
514 inconsistencies on the allocation of attention during scene perception. *Journal of Vision*, 9(3),
515 24–24. <https://doi.org/10.1167/9.3.24>
- 516 Võ, M. L.-H. (2021). The meaning and structure of scenes. *Vision Research*, 181, 10–20.
517 <https://doi.org/10.1016/j.visres.2020.11.003>
- 518 Võ, M. L.-H., Boettcher, S. E., & Draschkow, D. (2019). Reading scenes: How scene grammar guides
519 attention and aids perception in real-world environments. *Current Opinion in Psychology*, 29,
520 205–210. <https://doi.org/10.1016/j.copsyc.2019.03.009>
- 521 Võ, M. L.-H., & Wolfe, J. M. (2013a). Differential Electrophysiological Signatures of Semantic and
522 Syntactic Scene Processing. *Psychological Science*, 24(9), 1816–1823.
523 <https://doi.org/10.1177/0956797613476955>
- 524 Võ, M. L.-H., & Wolfe, J. M. (2013b). The interplay of episodic and semantic memory in guiding
525 repeated search in scenes. *Cognition*, 126(2), 198–212.
526 <https://doi.org/10.1016/j.cognition.2012.09.017>

527

528

529

530

531

1 **Supplementary Materials**

2

3 **Supplementary Materials 1 – Image dataset and data-driven hierarchy**

4

5 **Information on the image dataset**

6 The dataset comprises 3499 unique images from 16 different scene categories (both indoor and
7 outdoor, natural and man-made, and including the 5 categories of the a priori assignment); the
8 images contain more than 48,000 object annotations grouped into 617 different categories
9 (including the 45 objects selected for the study). Annotations were made by 4 different workers
10 using the LabelMe tool (Russel et al., 2008), and were carefully cleaned from typos and synonyms
11

12 **Procedure to extract data-driven hierarchy information**

13 First, annotated and segmented data were pre-processed using MATLAB (MathWorks, 2018),
14 allowing the extraction of identity, coordinates and centroids of each object in the 2D pixel-space
15 of each image. Subsequent analyses were done using R (version 3.6.3, R Core Team, 2020). Second,
16 objects that have mainly a structural function (e.g., walls, windows, ceiling, doors, pipes) rather than
17 being relevant for the object-to-object relationship we were interested in investigating, were
18 discarded, leaving us with 567 unique object categories.

19 We then ran a clustering algorithm to find optimal grouping of objects in the images. The
20 algorithm was based on the partitioning around medoids clustering method and estimated number
21 of clusters using average silhouette width (*pamk* function from R package “fpc”, Hennig, 2020). In
22 this sense, we represented the objects in an image through their centroids and the image area as a
23 2D space.

24 After this procedure, we could measure object-to-object co-occurrence in images, in clusters
25 within an image, and also, considering objects' area, the co-occurrence between anchor and local
26 pairs.

27

28 **Supplementary Materials 2 – Analysis details**

29

30 **Formula of the GLMM and model specifications**

31 Using R syntax, our model had this structure:

32

33 *behavioral similarity* ~ *task * (scene condition + phrase condition*
34 *+ object type condition + cooccurrence in scene + cooccurrence in phrase*
35 *+ anchored cooccurrence + covariates) + (1 | participants) + (1 | pairs)*
36 *+ (1 | context objects)*

37

38 We fitted the statistical models via maximum likelihood estimation, and continuous predictors were
39 scaled, as this typically improves model fit. For random effects, we included only an intercept term,
40 so that we followed the recommendations of Bates et al., (2015) about parsimony in random effect
41 structure (Bates et al., 2015).

42 For categorical predictors, we planned specific contrasts between conditions:

- 43 - for Scene condition, the contrast was set to *same scene – different scenes*;
- 44 - for Object type condition, the contrast was set to *same object type – different object types*;
- 45 - for Phrase condition, one contrast was set to *same phrase – different phrases of the same*
46 *scene*, while the other contrast was set to *(same phrase and different phrases of the same*
47 *scene) – different phrases of different scenes*. Since this last contrast is identical to *same*
48 *scene – different scenes*, and since the Scene condition and Phrase condition predictors are

49 highly correlated, we removed from the model the *scene condition* predictor and
50 incorporate its contrast in the *phrase similarity* predictor. This way, we removed
51 redundancies and reduced multi-collinearity to an acceptable level.

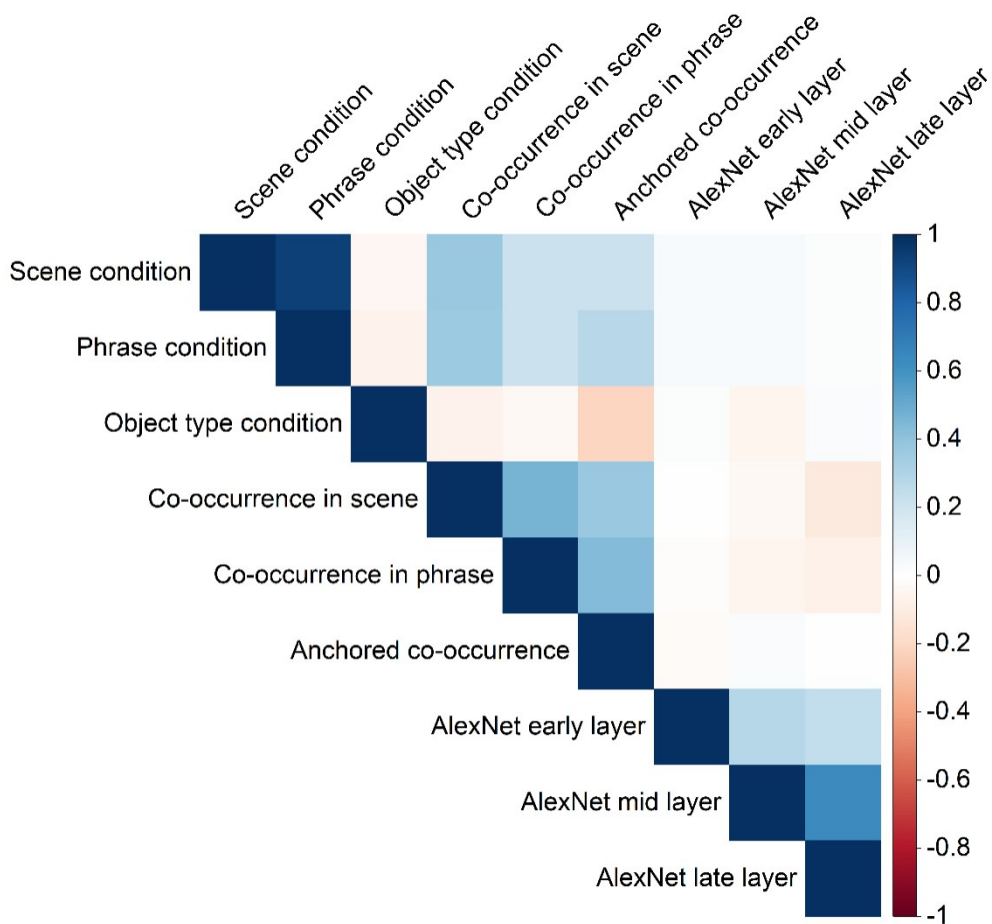
52 - for the Task predictor, we compared the effect of the measures between *action task – no*
53 *task* and between *visual task – no task*.

54

55 **Supplementary Materials 3 – Factor correlations and VIFs in the main model**

56

57 **Sup. Fig. 1** – Matrix of correlations between the predictors used in the model



58

59

60 To evaluate potential multicollinearity in the model, we computed the variance inflation factors
 61 (VIFs) for each term in the model, using the *check_collinearity* function in R (package “performance”;
 62 Lüdecke et al., 2021). Typically, when VIFs are below 5, there is low correlations between predictors
 63 and the model does not need any adjustment, as it was in our case.

64

65

66 **Sup. Table 1** – Variance Inflation Factors (VIFs) for the predictors used in the main model

67

68	Term	VIF
69	Task	1.000
70	Object type condition	1.064
71	Phrase condition	1.238
72	Anchored co-occurrence	1.424
73	Co-occurrence in scene	1.487
74	Co-occurrence in phrase	1.421
75	AlexNet early layer	1.099
	AlexNet mid layer	1.753
	AlexNet late layer	1.737

76

77

78 **Supplementary Materials 4 – Results of the model**

79

80 The GLMM resulted to be singular, due to the random factor term (1 | *participants*) explaining no
 81 variance, since this was already explained by the other two random factors
 82 (1 | *pairs*) and (1 | *context objects*), that identify unique observations. Therefore
 83 (1 | *participants*) random factor term was excluded from the final model.

84

85

86 **Sup. Table 2** – Results of the GLMM

Predictors	β	SE	z	p
(Intercept)	-0.274	0.065	-4.215	<0.001
Action task – No task	0.044	0.031	1.429	0.153
Visual task – No task	-0.332	0.030	-11.038	<0.001

Object type condition (same – different)	0.175	0.049	3.586	<0.001
Phrase condition (same – different)	0.234	0.130	1.800	0.072
Scene condition (same – different)	1.142	0.074	15.335	<0.001
Anchored co-occurrence	0.022	0.028	0.782	0.434
Co-occurrence in scene	0.325	0.029	11.334	<0.001
Co-occurrence in phrase	0.059	0.028	2.111	0.035
AlexNet early layer	-0.067	0.025	-2.706	0.007
AlexNet mid layer	0.095	0.031	3.058	0.002
AlexNet late layer	0.196	0.031	6.329	<0.001
Object type condition x (Action task – No task)	-0.141	0.034	-4.151	<0.001
Object type condition x (Visual task – No task)	-0.052	0.034	-1.562	0.118
Phrase condition x (Action task – No task)	0.064	0.086	0.747	0.455
Phrase condition x (Visual task – No task)	0.012	0.085	0.139	0.889
Scene condition x (Action task – No task)	0.049	0.049	1.005	0.315
Scene condition x (Visual task – No task)	-0.773	0.049	-15.926	<0.001
Anchored co-occurrence x (Action task – No task)	-0.017	0.019	-0.893	0.372
Anchored co-occurrence x (Visual task – No task)	-0.083	0.019	-4.447	<0.001
Co-occurrence in scene x (Action task – No task)	-0.102	0.019	-5.332	<0.001
Co-occurrence in scene x (Visual task – No task)	-0.158	0.019	-8.282	<0.001
Co-occurrence in phrase x (Action task – No task)	0.042	0.019	2.204	0.028
Co-occurrence in phrase x (Visual task – No task)	0.002	0.019	0.106	0.915
AlexNet early layer x (Action task – No task)	0.016	0.017	0.990	0.322
AlexNet early layer x (Visual task – No task)	-0.053	0.016	-3.227	0.001
AlexNet mid layer x (Action task – No task)	0.019	0.022	0.882	0.378
AlexNet mid layer x (Visual task – No task)	0.151	0.021	7.119	<0.001
AlexNet late layer x (Action task – No task)	-0.052	0.021	-2.419	0.016
AlexNet late layer x (Visual task – No task)	0.026	0.021	1.221	0.222

87

88

89 **Additional bibliography**

90

91 Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious Mixed Models.

92 *ArXiv:1506.04967 [Stat]*. <http://arxiv.org/abs/1506.04967>

93 Hennig, C. (2020). fpc: Flexible procedures for clustering. *R Package*. [https://cran.r-](https://cran.r-project.org/web/packages/fpc/index.html)

94 [project.org/web/packages/fpc/index.html](https://cran.r-project.org/web/packages/fpc/index.html)

95 Lüdecke, D., Ben-Shachar, M., Patil, I., Waggoner, P., & Makowski, D. (2021). performance: An R
96 Package for Assessment, Comparison and Testing of Statistical Models. *Journal of Open*
97 *Source Software*, 6(60), 3139. <https://doi.org/10.21105/joss.03139>
98 Russel, B. C., Torralba, A., Murphy, K. P., & Freeman, W. T. (2008). LabelMe: A database and web-
99 based tool for image annotation. *International Journal of Computer Vision*, 77(1), 157–173.
100
101
102

Bibliography

- Aminoff, E. M., Kveraga, K., & Bar, M. (2013). The role of the parahippocampal cortex in cognition. *Trends in Cognitive Sciences*, 17(8), 379–390.
- Auckland, M. E., Cave, K. R., & Donnelly, N. (2007). Nontarget objects can influence perceptual processes during object recognition. *Psychonomic Bulletin & Review*, 14(2), 332–337.
- Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., ... & Treiman, R. (2007). The English lexicon project. *Behavior research methods*, 39(3), 445-459.
- Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience*, 5(8), 617–629.
- Bar, M. (2009). The proactive brain: Memory for predictions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1521), 1235–1243.
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting Linear Mixed-Effects Models using lme4. *ArXiv:1406.5823 [Stat]*.
- Batterink, L. J., Paller, K. A., & Reber, P. J. (2019). Understanding the Neural Bases of Implicit and Statistical Learning. *Topics in Cognitive Science*, 11(3), 482–503.
- Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, 14(2), 143–177.
- Boettcher, S. E. P., Draschkow, D., Dienhart, E., & Vö, M. L.-H. (2018). Anchoring visual search in scenes: Assessing the role of anchor objects on eye movements during visual search. *Journal of Vision*, 18(13), 11.
- Bonner, M. F., & Epstein, R. A. (2017). Coding of navigational affordances in the human visual system. *Proceedings of the National Academy of Sciences*, 114(18), 4793–4798.
- Bonner, M. F., & Epstein, R. A. (2021). Object representations in the human brain reflect the co-occurrence statistics of vision and language. *Nature Communications*, 12(1), 4081.
- Brady, T. F., Shafer-Skelton, A., & Alvarez, G. A. (2017). Global ensemble texture representations are critical to rapid scene perception. *Journal of Experimental Psychology: Human Perception and Performance*, 43(6), 53.
- Brentano, F. (1874). *Psychology from an empirical standpoint*. Routledge.
- Brockmole, J. R., & Vö, M. L.-H. (2010). Semantic memory for contextual regularities within and across scene categories: Evidence from eye movements. *Attention, Perception, & Psychophysics*, 72(7), 1803-1813.
- Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The Word Frequency Effect: A Review of Recent Developments and Implications for the Choice of Frequency Estimates in German. *Experimental Psychology*, 58(5), 412–424.
- Brysbaert, M., Mandera, P., & Keuleers, E. (2018). The Word Frequency Effect in Word Processing: An Updated Review. *Current Directions in Psychological Science*, 27(1), 45–50.
- Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4), 977–990.

- Castelhana, M. S., & Heaven, C. (2011). Scene context influences without scene gist: Eye movements guided by spatial associations in visual search. *Psychonomic Bulletin & Review*, 18(5), 890–896.
- Castelhana, M. S., & Krzyś, K. (2020). Rethinking Space: A Review of Perception, Attention, and Memory in Scene Processing. *Annual Review of Vision Science*, 6(1), 563–586.
- Castelhana, M. S., & Witherspoon, R. L. (2016). How you use it matters: Object function guides attention during visual search in scenes. *Psychological Science*, 27(5), 606–621.
- Chun, M. M., & Jiang, Y. (1998). Contextual Cueing: Implicit Learning and Memory of Visual Context Guides Spatial Attention. *Cognitive Psychology*, 36(1), 28–71.
- Clark, A. (2013). Expecting the World: Perception, Prediction, and the Origins of Human Knowledge: *Journal of Philosophy*, 110(9), 469–496.
- Cornelissen, T. H. W., & Vö, M. L.-H. (2017). Stuck on semantics: Processing of irrelevant object-scene inconsistencies modulates ongoing gaze behavior. *Attention, Perception, & Psychophysics*, 79(1), 154–168.
- Davenport, J. L., & Potter, M. C. (2004). Scene Consistency in Object and Background Perception. *Psychological Science*, 15(8), 559–564.
- Draschkow, D., & Vö, M. L.-H. (2017). Scene grammar shapes the way we interact with objects, strengthens memories, and speeds search. *Scientific Reports*, 7(1), 16471.
- Firestone, C., & Scholl, B. J. (2016). Cognition does not affect perception: Evaluating the evidence for “top-down” effects. *Behavioral and Brain Sciences*, 39.
- Ganis, G., & Kutas, M. (2003). An electrophysiological study of scene effects on object identification. *Cognitive Brain Research*, 16(2), 123–144.
- Garety, P. A., & Freeman, D. (2013). The past and future of delusions research: From the inexplicable to the treatable. *British Journal of Psychiatry*, 203(5), 327–333.
- Gibson, J. J. (1986). *The ecological approach to visual perception*. Hillsdale, NJ: Lawrence.
- Goodale, M. A., & Milner, D. A. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*.
- Gordon, I., & Slater, A. (1998). Nativism and empiricism: The history of. *Perceptual development: Visual, auditory, and speech perception in infancy*, 73.
- Greene, M. R. (2013). Statistics of high-level scene context. *Frontiers in Psychology*, 4.
- Greene, M. R. (2016). Estimations of object frequency are frequently overestimated. *Cognition*, 149, 6–10. •
Greene, M. R., Baldassano, C., Esteva, A., Beck, D. M., & Fei-Fei, L. (2016). Visual scenes are categorized by function. *Journal of Experimental Psychology: General*, 145(1), 82–94.
- Greene, M. R., & Hansen, B. C. (2020). Disentangling the Independent Contributions of Visual and Conceptual Features to the Spatiotemporal Dynamics of Scene Categorization. *The Journal of Neuroscience*, 40(27), 5283–5299.

- Gregorová, K., Turini, J., Gagl, B., & Vö, M. L.-H. (2021). Access to meaning from visual input: Object and word frequency effects in categorization behavior. [Preprint]. *PsyArXiv*.
- Groen, I. I. A., Silson, E. H., & Baker, C. I. (2017). Contributions of low- and high-level properties to neural processing of visual scenes in the human brain. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1714), 20160102.
- Groen, I. I., Greene, M. R., Baldassano, C., Fei-Fei, L., Beck, D. M., & Baker, C. I. (2018). Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. *ELife*, 7, e32962.
- Gronau, N., & Shachar, M. (2014). Contextual integration of visual objects necessitates attention. *Attention, Perception, & Psychophysics*, 76(3), 695–714.
- Harel, A. (2016). What is special about expertise? Visual expertise reveals the interactive nature of real-world object recognition. *Neuropsychologia*, 83, 88–99.
- Hebart, M. N., Dickter, A. H., Kidder, A., Kwok, W. Y., Corriveau, A., Van Wicklin, C., & Baker, C. I. (2019). THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLOS ONE*, 14(10), e0223792.
- Hebart, M. N., Zheng, C., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgments. *Nature Human Behaviour*.
- Heister, J., Würzner, K.-M., Bubbenzer, J., Pohl, E., Hanneforth, T., Geyken, A., & Kliegl, R. (2011). DlexDB – eine lexikalische Datenbank für die psychologische und linguistische Forschung. *Psychologische Rundschau*, 62(1), 10–20.
- Helbing, J., Draschkow, D., & Vö, M. L. H. (2022). Auxiliary scene context information provided by anchor objects guides attention and locomotion in natural search behavior. *Psychological Science*.
- Hwang, A. D., Wang, H.-C., & Pomplun, M. (2011). Semantic guidance of eye movements in real-world scenes. *Vision Research*, 51(10), 1192–1205.
- Johnson, J., Krishna, R., Stark, M., Li, L.-J., Shamma, D. A., Bernstein, M. S., & Fei-Fei, L. (2015). Image retrieval using scene graphs. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3668–3678.
- Josephs, E. L., & Konkle, T. (2019). Perceptual dissociations among views of objects, scenes, and reachable spaces. *Journal of Experimental Psychology: Human Perception and Performance*, 45(6), 715–728.
- Josephs, E. L., & Konkle, T. (2020). Large-scale dissociations between views of objects, scenes, and reachable-scale environments in visual cortex. *Proceedings of the National Academy of Sciences*, 117(47), 29354–29362.
- Kaiser, D., Stein, T., & Peelen, M. V. (2014). Object grouping based on real-world regularities facilitates perception by reducing competitive interactions in visual cortex. *Proceedings of the National Academy of Sciences*, 111(30), 11217–11222.

- Kim, J. (2018). *Philosophy of mind* (3rd ed). Westview Press.
- Koffka, K. (1935). *Principles of gestalt psychology*. Harcourt. New York.
- Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of Experimental Psychology: General*, *139*(3), 558–578.
- Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90.
- Kuiera, H., & Francis, W. (1967). *Computational analysis of present day American English*. Providence, RI: Brown University Press.
- Kuperman, V., & Van Dyke, J. A. (2013). Reassessing word frequency as a determinant of word recognition for skilled and unskilled readers. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(3), 802.
- Kutas, M., & Hillyard, S. A. (1980). Reading Senseless Sentences: Brain Potentials Reflect Semantic Incongruity. *Science*, *207*(4427), 203–205.
- Lauer, T., Boettcher, S. E. P., Kollenda, D., Draschkow, D., & Vö, M. L.-H. (2022). Behavioral and electrophysiological markers of global and local scene context effects. *Under review*
- Lauer, T., Cornelissen, T. H. W., Draschkow, D., Willenbockel, V., & Vö, M. L.-H. (2018). The role of scene summary statistics in object recognition. *Scientific Reports*, *8*(1), 14666.
- Lauer, T., Schmidt, F., & Vö, M. L.-H. (2021). The role of contextual materials in object recognition. *Scientific Reports*, *11*(1), 21988.
- Lauer, T., & Vö, M., L.-H. (2022) The Ingredients of Scenes that Affect Object Search and Perception. In Ionescu, B., Bainbridge, W. A., & Murray, N. (Eds.). (2022). *Human Perception of Visual Information: Psychological and Computational Perspectives*. Springer International Publishing.
- Lauer, T., Willenbockel, V., Maffongelli, L., & Vö, M. L.-H. (2020). The influence of scene and object orientation on the scene consistency effect. *Behavioural Brain Research*, *394*, 112812.
- Lenci, A. (2018). Distributional Models of Word Meaning. *Annual Review of Linguistics*, *4*(1), 151–171.
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common Objects in Context. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *Computer Vision – ECCV 2014* (Vol. 8693, pp. 740–755). Springer International Publishing.
- Lupyan, G., & Clark, A. (2015). Words and the World: Predictive Coding and the Language-Perception-Cognition Interface. *Current Directions in Psychological Science*, *24*(4), 279–284.
- Mack, S. C., & Eckstein, M. P. (2011). Object co-occurrence serves as a contextual cue to guide and facilitate visual search in a natural viewing environment. *Journal of Vision*, *11*(9), 9–9.

- Maffongelli, L., Öhlschläger, S., & Lê-Hoa Võ, M. (2020). The Development of Scene Semantics: First ERP Indications for the Processing of Semantic Object-Scene Inconsistencies in 24-Month-Olds. *Collabra: Psychology*, 6(1), 17707.
- Malcolm, G. L., Groen, I. I. A., & Baker, C. I. (2016). Making Sense of Real-World Scenes. *Trends in Cognitive Sciences*, 20(11), 843–856.
- Matthen, M. (2015). Introduction. In Matthen, M. (Ed. 2015). *The Oxford Handbook of Philosophy of Perception*. OUP Oxford
- McCulloch, C. E., & Neuhaus, J. M. (2005). Generalized Linear Mixed Models. *Encyclopedia of Biostatistics*. • Melcher, D. (2011). Introduction: Visual stability. *Philosophical Transactions: Biological Sciences*, 366(1564), 468–475.
- Nanay, B. (2015). The representationalism versus relationalism debate: Explanatory contextualism about perception. *European Journal of Philosophy*, 23(2), 321-336.
- Nanay, B. (2015). Perceptual representation/Perceptual content. In Matthen, M. (Ed. 2015). *The Oxford Handbook of Philosophy of Perception*. OUP Oxford
- Öhlschläger, S., & Võ, M. L.-H. (2020). Development of scene knowledge: Evidence from explicit and implicit scene knowledge measures. *Journal of Experimental Child Psychology*, 194, 104782.
- Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3), 145-175.
- Oliva, A., & Torralba, A. (2007). The role of context in object recognition. *Trends in Cognitive Sciences*, 11(12), 520–527.
- O'Regan, J. K. (1992). Solving the “real” mysteries of visual perception: The world as an outside memory. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, 46(3), 461–488.
- Osterhout, L., & Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language*, 31(6), 785–806.
- Pereira, E. J., & Castelhana, M. S. (2014). Peripheral guidance in scenes: The interaction of scene context and object content. *Journal of Experimental Psychology: Human Perception and Performance*, 40(5), 2056–2072.
- Pereira, E. J., & Castelhana, M. S. (2019). Attentional capture is contingent on scene region: Using surface guidance framework to explore attentional mechanisms during search. *Psychonomic Bulletin & Review*, 26(4), 1273–1281.
- Perruchet, P., & Pacton, S. (2006). Implicit learning and statistical learning: One phenomenon, two approaches. *Trends in Cognitive Sciences*, 10(5), 233–238.
- Preston, K. A. (1935). The speed of word perception and its relation to reading ability. *The Journal of General Psychology*.

- Quek, G. L., & Peelen, M. V. (2020). Contextual and Spatial Associations Between Objects Interactively Modulate Visual Processing. *Cerebral Cortex*, 30(12), 6391–6404.
- Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1), 79–87.
- Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior*, 6(6), 855–863.
- Russel, B. C., Torralba, A., Murphy, K. P., & Freeman, W. T. (2008). LabelMe: A database and web-based tool for image annotation. *International Journal of Computer Vision*, 77(1), 157–173.
- Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996a). Statistical learning by 8-month-old infants. *Science*, 274(5294), 1926–1928.
- Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996b). Word Segmentation: The Role of Distributional Cues. *Journal of Memory and Language*, 35(4), 606–621.
- Santoro, A., Raposo, D., Barrett, D. G. T., Malinowski, M., Pascanu, R., Battaglia, P., & Lillicrap, T. (2017). A simple neural network module for relational reasoning. *arXiv*.
- Schapiro, A., & Turk-Browne, N. (2015). Statistical Learning. In *Brain Mapping* (pp. 501–506). Elsevier.
- Sejnowski, T. J. (2018). *The Deep Learning Revolution*. MIT Press.
- Sherman, B. E., Graves, K. N., & Turk-Browne, N. B. (2020). The prevalence and importance of statistical learning in human cognition and behavior. *Current Opinion in Behavioral Sciences*, 32, 15–20.
- Siegel, S. (2006). Subject and Object in the Contents of Visual Experience. *The Philosophical Review*, 115(3), 355–388.
- Soteriou, M. (2000). The Particularity of Visual Perception. *European Journal of Philosophy*, 8(2), 173–189. •
- Spence, C. and Driver, J. (2004.). *Crossmodal Space and Crossmodal Attention*, Oxford: Oxford University Press.
- Suddendorf, T., & Whiten, A. (2001). Mental evolution and development: Evidence for secondary representation in children, great apes, and other animals. *Psychological Bulletin*, 127(5), R56–R57.
- Swanson, L. R. (2016). The Predictive Processing Paradigm Has Roots in Kant. *Frontiers in Systems Neuroscience*, 10.
- Telles-Correia, D., Moreira, A. L., & Gonçalves, J. S. (2015). Hallucinations and related concepts—Their conceptual background. *Frontiers in Psychology*, 6.
- Torralba, A., & Efros, A. A. (2011). Unbiased look at dataset bias. *CVPR 2011*, 1521–1528.
- Treder, M. S., Mayor-Torres, J., & Teufel, C. (2020). Deriving Visual Semantics from Spatial Context: An Adaptation of LSA and Word2Vec to generate Object and Scene Embeddings from Images. *arXiv*.
- Treisman, A. (1996). The binding problem. *Current Opinion in Neurobiology*.
- Turini, J., & Vö, M. L.-H. (2022a). Hierarchical organization of objects in scenes is reflected in mental representations of objects [Preprint]. *PsyArXiv*.

- Turini, J., & Vö, M. L.-H. (2022b). Scene hierarchy structures mental object representations while flexibly adapting to varying task demands [Preprint]. *PsyArXiv*.
- Vö, M. L. H., & Henderson, J. M. (2009). Does gravity matter? Effects of semantic and syntactic inconsistencies on the allocation of attention during scene perception. *Journal of Vision*, 9(3), 24–24.
- Vö, M. L.-H. (2021). The meaning and structure of scenes. *Vision Research*, 181, 10–20.
- Vö, M. L.-H., Boettcher, S. E., & Draschkow, D. (2019). Reading scenes: How scene grammar guides attention and aids perception in real-world environments. *Current Opinion in Psychology*, 29, 205–210.
- Vö, M. L.-H., & Wolfe, J. M. (2013a). Differential Electrophysiological Signatures of Semantic and Syntactic Scene Processing. *Psychological Science*, 24(9), 1816–1823.
- Vö, M. L.-H., & Wolfe, J. M. (2013b). The interplay of episodic and semantic memory in guiding repeated search in scenes. *Cognition*, 126(2), 198–212.
- Wiesmann, S. L., & Vö, M. L.-H. (2022). What makes a scene? Fast scene categorization as a function of global scene information at different resolutions. *Journal of Experimental Psychology: Human Perception and Performance*, 48(8), 871–888.
- Wolfe, J. M. (1994). Guided Search 2.0 A revised model of visual search. *Psychonomic Bulletin & Review*, 37.
- Wolfe, J. M., Vö, M. L.-H., Evans, K. K., & Greene, M. R. (2011). Visual search in scenes involves selective and nonselective pathways. *Trends in Cognitive Sciences*, 15(2), 77–84.
- Woo, S., Kim, D., Cho, D., & Kweon, I. S. (2018). Linknet: Relational embedding for scene graph. *Advances in neural information processing systems*, 31.
- Wu, K., Wu, E., & Kreiman, G. (2018). Learning scene gist with convolutional neural networks to improve object recognition. *2018 52nd Annual Conference on Information Sciences and Systems (CISS)*, 1–6.
- Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2010). SUN database: Large-scale scene recognition from abbey to zoo. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 3485–3492.
- Xu, D., Zhu, Y., Choy, C. B., & Fei-Fei, L. (2017). Scene graph generation by iterative message passing. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5410-5419).
- Yu, H., Yang, Z., Tan, L., Wang, Y., Sun, W., Sun, M., & Tang, Y. (2018). Methods and datasets on semantic segmentation: A review. *Neurocomputing*, 304, 82–103.
- Zellers, R., Yatskar, M., Thomson, S., & Choi, Y. (2018). Neural Motifs: Scene Graph Parsing with Global Context. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5831–5840.
- Zhang, P., & Chartrand, G. (2006). *Introduction to graph theory*. Tata McGraw-Hill.
- Zhang, M., Tseng, C., & Kreiman, G. (2020). Putting Visual Object Recognition in Context. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12982–12991.
- Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., & Oliva, A. (2014). Learning deep features for scene recognition using places database. *Advances in neural information processing systems*, 27.

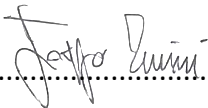
- Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., & Torralba, A. (2019). Semantic Understanding of Scenes Through the ADE20K Dataset. *International Journal of Computer Vision*, 127(3), 302–321.

Erklärungen

Erklärung zu bisherigen Promotionsverfahren

Ich erkläre hiermit, dass ich mich bisher keiner Doktorprüfung unterzogen habe.

Cremona, 27.10.2022



.....

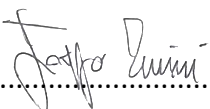
Jacopo Turini

Eidesstattliche Versicherung

Ich erkläre hiermit an Eides statt, dass ich die vorgelegte Dissertation über “The structure of the visual world in the human mind: measuring object statistics from large sets of real-world images and their impact on behaviour” selbstständig angefertigt und mich anderer Hilfsmittel als der in ihr angegebenen nicht bedient habe, insbesondere, dass alle Entlehnungen aus anderen Schriften mit Angabe der betreffenden Schrift gekennzeichnet sind.

Ich versichere, die Grundsätze der guten wissenschaftlichen Praxis beachtet, und nicht die Hilfe einer kommerziellen Promotionsvermittlung in Anspruch genommen zu haben.

Cremona, 27.10.2022



.....

Jacopo Turini

Erklärung der Punkte 1-7 der Kriterien für publikationsbasierte Dissertationen

1. Ich erkläre hiermit, dass die vorgelegte kumulative Dissertation über “The structure of the visual world in the human mind: measuring object statistics from large sets of real-world images and their impact on behaviour” 3 Schriften umfasst, die aus den letzten 5 Jahren stammen:
 - **Schrift 1:** Gregorová*, K., Turini*, J., Gagl, B., Vö, M. L.-H. (in revision). Access to meaning from visual input: Object and word frequency effects in categorization behavior. Manuscript in revision at *Journal of Experimental Psychology: General*. * = shared first authorship.
 - **Schrift 2:** Turini, J. & Vö, M. L.-H. (in revision). Hierarchical organization of objects in scenes is reflected in mental representations of objects. Manuscript in revision at *Scientific Reports*.
 - **Schrift 3:** Turini, J. & Vö, M. L.-H. (submitted). Scene hierarchy structures object representations while flexibly adapting to varying task demands. Manuscript submitted to *Psychological Science*.

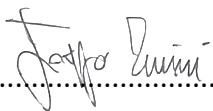
2. Die Schriften der vorgelegten kumulativen Dissertation entstammen einem im Wesentlichen zusammenhängenden Forschungsprogramm und die jeweils verfolgten Forschungsfragen sind sinnvoll zueinander in Beziehung zu setzen. In den 3 Schriften des Forschungsprogrammes “The structure of the visual world in the human mind: measuring object statistics from large sets of real-world images and their impact on behaviour”“ wurden die folgenden Forschungsfragen behandelt:

- **Schrift 1:** Beeinflusst die Häufigkeit von Objekten die kognitive Verarbeitung während der Objekterkennung? Können wir die Objekthäufigkeit aus einer großen Menge kommentierter Bilder messen?
 - **Schrift 2:** Sind Objekte in visuellen Szenen hierarchisch organisiert? Können wir diese Hierarchie anhand des gleichzeitigen Auftretens von Objekten in realen Bildern messen?
 - **Schrift 3:** Wird diese anhand realer Bilder gemessene Objekthierarchie durch verschiedene Aufgaben erweitert und reduziert?
3. Ich erkläre hiermit, dass ich alleinige Erstautorin der Schrift 2 und Schrift 3 bin, während ich gemeinsam mit Klara Gregorová Erstautorin der Schrift 1 bin.
4. Alle 3 Schriften sind in guten oder sehr guten englischsprachigen Zeitschriften mit PeerReview publiziert bzw. eingereicht worden:
- **Schrift 1:** Journal of Experimental Psychology: General (Impact Factor in 2021-22: 4.913)
 - **Schrift 2:** Scientific Reports (Impact Factor in 2021-22: 4.379)
 - **Schrift 3:** Psychological Science (Impact Factor in 2021-22: 7.029)
5. Die 3 Schriften der vorgelegten kumulativen Dissertation sind zumindest zur Veröffentlichung eingereicht:
- **Schrift 1:** in Revision (“in revision”) in *Journal of Experimental Psychology: General*
 - **Schrift 2:** in Revision (“in revision”) in *Scientific Reports*
 - **Schrift 3:** eingereicht (“submitted”) bei *Psychological Science*

6. Keine der 3 Schriften ist als Publikation in einem Lehrbuch oder Enzyklopädieband eingereicht.

7. Ich erkläre hiermit, dass die als Dissertation vorgelegte Abhandlung über die zusammengestellten Publikationen hinaus einen zusätzlichen Text enthält, in welchem eine kritische Einordnung der eigenen Publikationen aus einer übergeordneten Perspektive heraus vorgenommen wird. Hier werden die Fragestellungen theoretisch entwickelt und die empirischen Arbeiten und ihre Ergebnisse so dargestellt, dass sie auch ohne Lesen der Einzelarbeiten nachvollziehbar sind. Zudem ist eine Gesamtdiskussion enthalten, die die Fragestellungen beantwortet und den Erkenntnisgewinn der Arbeit herausstellt.

Cremona, 27.10.2022



.....

Jacopo Turini

Erklärung zu Punkt 2 der Regeln für die Eröffnung und Durchführung von Prüfungsverfahren: Erklärung über die Eigenleistung

Die vorliegende Dissertation "The structure of the visual world in the human mind: measuring object statistics from large sets of real-world images and their impact on behaviour" enthält zwei Schriften, die bei wissenschaftlichen Zeitschriften in Bearbeitung sind (Schrift 1 and 2), und ein Schrift, das bei einer wissenschaftlichen Zeitschrift eingereicht wurde (Schrift 3). Im Folgenden wird die Eigenleistung des Verfassers dargestellt:

Schrift 1: Gregorová*, K., Turini*, J., Gagl, B., Vö, M. L.-H. (in revision). Access to meaning from visual input: Object and word frequency effects in categorization behavior. Manuscript in revision at *Journal of Experimental Psychology: General*. * = shared first authorship.

Klara Gregorová (KG) und Jacopo Turini (JT) konzipierten die Studie gemeinsam unter der Leitung von Benjamin Gagl (BG) und Melissa Le-Hoa Vö (MLHV). KG und JT führten gemeinsam die Experimente durch, sammelten und analysierten die Daten. KG, JT, BG und MLHV diskutierten und interpretierten die Ergebnisse und verfassten gemeinsam das Manuskript.

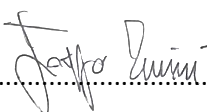
Schrift 2: Turini, J. & Vö, M. L.-H. (in revision). Hierarchical organization of objects in scenes is reflected in mental representations of objects. Manuscript in revision at *Scientific Reports*.

Jacopo Turini (JT) und Melissa Le-Hoa Vö (MLHV) haben die Studie gemeinsam konzipiert und gestaltet; JT führte das Experiment durch, sammelte und analysierte die Daten; JT und MV interpretierten die Ergebnisse und verfassten gemeinsam das Manuskript.

Schrift 3: Turini, J. & Vö, M. L.-H. (submitted). Scene hierarchy structures object representations while flexibly adapting to varying task demands. Manuscript submitted to *Psychological Science*.

Jacopo Turini (JT) und Melissa Le-Hoa Vö (MLHV) haben die Studie gemeinsam konzipiert und gestaltet; JT führte das Experiment durch, sammelte und analysierte die Daten; JT und MV interpretierten die Ergebnisse und verfassten gemeinsam das Manuskript.

Cremona, 27.10.2022




.....

Jacopo Turini

(Verfasser)

Rovereto, 27.10.2010



.....

Prof. Dr. Melissa Le-Hoa Vö

(Betreuerin)



Publiziert unter der Creative Commons-Lizenz Namensnennung (CC BY) 4.0 International.
Published under a Creative Commons Attribution (CC BY) 4.0 International License.
<https://creativecommons.org/licenses/by/4.0/>