Cambridge Language Sciences Language, Brains & Machines Research Strategy Forum: an initial literature survey

Anna Samuel[1], Andrew Caines[2], and Paula Buttery[2]

[1]Theoretical & Applied Linguistics, Modern & Medieval Languages
[2]Natural Language & Information Processing, Computer Science & Technology
University of Cambridge
aais2@cam.ac.uk, apc38@cam.ac.uk, paula.buttery@cl.cam.ac.uk

16th October, 2018

Introduction

Language, Brains & Machines (LBM) is a new Research Strategy Forum established by members of the Cambridge Language Sciences Interdisciplinary Research Centre in the summer of 2018. LBM has a focus on works testing hypotheses based on the mechanisms of language technology, and on works testing hypotheses based on the properties of language (as discerned by information theory / signal analysis). These have tended to be published and disseminated in distinct research networks – namely, computer science and cognitive science publication venues. We recently undertook an initial survey of state-of-the-art research in the area defined by LBM, attempting to synthesise the bodies of research from both areas, to find common ground and to highlight complementary findings. This survey is intended as a non-exhaustive sample of recent literature to provide an idea of commonalities, complementary methods, and future directions for research. We indicate how readers may contribute to our collective awareness of the literature by submitting their own notes on publications not included in this initial survey; we welcome older publications as much as those in the date range covered by this review. We will maintain an online repository of readers' notes, along with updated versions of this document at https://www.cl.cam.ac.uk/~apc38/camlangsci_langbrainsmachines_litreview.pdf

Contributing

We welcome contributions from members of the Cambridge Language Sciences IRC, and will publish updates to this document periodically at https://www.cl.cam.ac.uk/~apc38/camlangsci_langbrainsmachines_litreview.pdf

To contribute your own reading notes, please complete the online form at https://goo.gl/forms/51tZwn5fJVKH1so22

Table of Contents (alphabetical by author name)

Just, Cherkassky, Aryal, Mitchell	2010	A neurosemantic theory of of concrete noun representation based on the underlying brain codes	PLoS ONE	neural dimensions of representation, concrete nouns, commonality	6
Kriegeskorte, Douglas	2018	Cognitive computational neuroscience	Cognitive Computational Neuroscience	artificial intelligence, connectivity models, neural network models, Bayesian cognitive models	29
Mollica, Piantadosi	2017	An incremental information-theoretic buffer supports sentence processing	CogSci	surprisal, rapid serial visual presentation (RSVP), spillover effects, FIFO buffer	21
Moss, Rodd, Stamatakis, Bright, Tyler	2005	Anteromedial temporal cortex supports fine-grained differentiation among objects	Cerebral Cortex	category-specificity, feature conjunctions, integration, object recognition	4
Norris	2006	The Bayesian Reader: Explaining word recognition as an optimal Bayesian decision process	Psychological Review	attention, Bayes Theorem, reading, word recognition	4
Pereira, Lou, Pritchett, Ritter, Gershman, Kanwisher, Botvinick, Fedorenko	2018	Towards a universal decoder of linguistic meaning from brain activation	Nature Communications	generalisability, distributed semantic representations, high-level cortical networks	28
Smith, Levy	2008	Optimal processing times in reading: a formal model and empirical investigation	CogSci	optimal behaviour, response time modelling, surprisal, sentence comprehension, eye movements, reading	4
Smith, Levy	2013	The effect of word predictability on reading time is logarithmic	Cognition	predictability, regression analysis, log-predictability, anticipatory processing, UID (uniform information density)	12
Taylor, Devereux, Tyler	2011	Conceptual structure: towards an integrated neurocognitive account	Language and Cognitive Processes	distributed semantics, concepts, feature-based representations	8
Wang, Cherkassky, Just	2017	Predicting the brain activation pattern associated with the propositional content of a sentence: modelling neural representations of events and states	Human Brain Mapping	brain activation pattern, factor analysis, NPSFs (neurally plausible semantic features), regression modelling	23

Reading Notes (chronological by year of publication)

- Hellen E. Moss, Jenni M. Rodd, Emmanuel A. Stamatakis, Peter Bright, and Lorraine K. Tyler. Anteromedial temporal cortex supports fine-grained differentiation among objects. *Cerebral Cortex*, 15:616–627, 2005. doi:10.1093/cercor/bhh163
    - Patients with damage to the left anteromedial temporal cortex often show a striking deficit: they fail to recognise animals and other living things.
    - Presents an important challenge to theories of the neural representation of conceptual knowledge.
    - Propose that this lesion behaviour association arises because polymodal neurons in anteromedial temporal cortex integrate simple features into complex feature conjunctions, providing the neural infrastructure for differentiating among objects.
- Dennis Norris. The Bayesian reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological Review*, 113:327–357, 2006. doi:0.1037/0033-295X.113.2.327
    - Presents a theory of visual word recognition that assumes that, in the tasks of word identification, lexical decision and semantic categorisation, human readers behave as optimal Bayesian decision-makers.
    - Leads to the development of a computational model of word recognition: the Bayesian Reader.
    - This model successfully simulates some of the most significant data on human reading.
    - It accounts for the nature of the function relating word-frequency to reaction time and identification threshold, the effects of neighbourhood density effects seen in different experimental tasks.
- Susan Goldin-Meadow, Wing Chee So, Aslı Özyürek, and Carolyn Mylander. The natural order of events: how speakers of different languages represent events nonverbally. *PNAS*, 105:9163–9168, 2008. doi:10.1073/pnas.0710060105
    - Test whether the language we speak influences our behaviour even when we are not speaking.
    - Subjects perform two non-verbal tasks: a communicative task (using gesture without speech), and a non-communicative task (reconstructing an event with pictures).
    - Result: the word orders speakers used in their everyday speech did not influence their non-verbal behaviour.
    - Speakers of all four languages used the same order on both non-verbal tasks: actor, patient, act.
    - Analogous to subject-verb-object pattern found in many of the world's languages.
    - Also found in newly-developing gestural languages.
    - Evidence for a natural order that we impose on events when describing and constructing them non-verbally?
- Nathaniel Smith and Roger Levy. Optimal processing times in reading: a formal model and empirical investigation. In *Proceedings of the Annual Conference of the Cognitive Science Society (CogSci)*, 2008. URL: https://escholarship.org/uc/item/3mr8m3rf
    - Humans can respond more quickly to events they expect than to unexpected events.

  - Presents a new model deriving the relationship between probability and reaction time as a consequence of optimal preparation.
  - This model is valid under very general conditions, requiring only that the results of optimisation are invariant across scale of stimulus granularity.
  - The model makes the strong prediction that response times should scale linearly with the negative conditional log probability of the stimulus.
  - Presents evidence for this prediction in an analysis of an existing database of eye movements in the reading of naturalistic texts.

- Delphine Dahan. Time course of interpretation in speech comprehension. *Current Directions in Psychological Science*, 19:121–126, 2010. doi:10.1177/0963721410364726
  - Understanding what people mean as they talk seems fluid and effortless; and determining how language comprehension proceeds over time has been central to theories of human language use.
  - Picture emerging from recent work suggests a more complex process, one in which information from speech has an immediate influence while enabling later-arriving information to modulate initial hypotheses.
  - Right-context effects: the later portion of a spoken stimulus can affect the interpretation of an earlier portion; they are pervasive and can span several syllables or words; thus, the interpretation of a segment of speech appears to result from the accumulation of information and integration of linguistic constraints over a larger temporal window than the duration of the speech segment itself.
  - These help explain how human listeners can understand language so efficiently, despite massive perceptual uncertainty in the speech signal.
  - Supports the view of speech processing as both rapid and flexible and in which information from speech is seen as having an immediate influence while also enabling later-arriving information to modulate initial hypotheses.
  - Marslen-Wilson's shadowing experiment [21] showed that people spontaneously corrected mispronunciations placed at the end of words.
  - Early information in the speech signal is utilised extremely rapidly and effectively to constrain the set of possible interpretations, in effect to predict what the upcoming signal may be – arguably a key aspect of real-time efficiency.
  - Recent examination of listeners' eye movements to real or pictured objects as they hear instructions to manipulate one of them has confirmed the immediate uptake of information from the speech signal [8].
  - Evidence for a right-context analysis: interpretation of ambiguous strings is influenced by the sentence context that follows it; when assessing listeners' ability to identify words from a recording of an unscripted, spontaneous dialogue, Bard et al [2] reported that as many as 20% of the words were accurately recognised only after one or more subsequent words had been heard – cannot be explained by a mere lag between the arrival of auditory information and its utilisation.
  - In a strictly sequential view of speech comprehension, reconsidering alternative interpretations that had been disfavoured early in the sentence but become likely later on would require revision or backtracking – presumed to be cognitively costly and no evidence for such a cost.

  - The emerging picture suggests a more complex process, one in which percepts during speech comprehension emerge from both anticipation of upcoming information and integration over a larger temporal window.
  - Numerous studies, including those relying on eye movements to assess the uptake on auditory information, have found evidence that the earliest moments influence choices; however, the author argues the ultimate (i.e. asymptotic) decision is reached after more time has passed and more information has accumulated.
  - As speech comprehension theories move toward allowing for integration over a larger temporal window, the role played by a sensory or working memory becomes more apparent: words may not always be recognised in the order they were spoken, but interpretation relies on keeping words in the intended sequence.
  - Some computational models have explicitly incorporated mechanisms that allow for retention of the true word sequence.
  - The most enduring and perhaps most successful model of spoken-word recognition, the TRACE model [22] has an architecture that accommodates the empirical data reported here, because it explicitly distinguishes the timeline of the unfolding of the phonetic signal from the timeline imposed by the continuous integration of information leading to the emergence of stable percepts of words or sentences.

- Marcel Adam Just, Vladimir L. Cherkassky, Sandesh Aryal, and Tom M. Mitchell. A neurosemantic theory of concrete noun representation based on the underlying brain codes. *PLoS ONE*, 5:e8622, 2010. doi:10.1371/journal.pone.0008622

  - Discovery of a set of biologically-driven semantic dimensions underlying the neural representation of concrete nouns.
  - Demonstrates how a resulting theory of noun representation can be used to identify simple thoughts through fMRI patterns.
  - Uses factor analyses of fMRI brain imaging data to reveal the biological representation of individual concrete nouns, in the absence of any pictorial stimuli.
  - Three main semantic factors underlying the neural representation of nouns naming physical objects, referred to as MANIPULATION, SHELTER, EATING.
  - Each factor neurally represented in 3-4 different brain locations that correspond to a cortical network that co-activates in non-linguistic tasks.
  - These factors are then used with machine learning classifier techniques to show that the fMRI-measured brain representation of an individual concrete noun like APPLE can be identified with good accuracy from among 60 candidate words, using only the fMRI activity in the 16 locations associated with these factors.
  - Theory-based model is developed to predict the brain activation patterns for words to which the algorithm has not been previously exposed.
  - It is plausible that there exist additional factors that underpin the representation of concrete nouns that were not captured by the authors' analyses, either because of limitations of the set of stimulus words or limitations in the analysis procedures.
  - One limit of the stimulus set is that it contained only count nouns but no mass nouns. Mass nouns require different types of manipulation, and so could require a different type of representation.
  - Another limitation of the stimulus set was the absence of nouns referring to human beings. Such nouns and considerations of ecological importance suggest that there may exist one or

more additional dimensions related to human interaction, with factors such as emotion and attraction.

- Abstract nouns were also excluded from this study. A pilot study demonstrated systematicity underlying activation for abstract nouns. It is possible for a classifier to identify such concepts from the corresponding brain activation with approximately similar accuracy as identifying concrete nouns.
- The total number of semantic factors which are neurally represented may be related to the number of distinct ways that human beings can interact with an object. The three observed factors may simply be the most dominant factors.
- This is not to deny that there may be a small set of visual factors or geometric primitives that underpin object recognition that could potentially be discovered using methods similar to those used by the authors.
- It seems reasonable to assume that an object is represented in terms of both its visual properties and its semantic properties, with different tasks evoking different properties.
- Each of the three dimensions has three to five subdimensions located at different cortical locations: suggesting an expanded set of about twelve dimensions for the neurosemantic representations of concrete nouns.
- Each factor appears to constitute a part of a cortical network whose constituent node specialisations have been suggested by previous perceptual-motor studies.
- Representation of all concrete nouns by voxels in about twelve locations referred to as combinatorial coding, allows an enormous number of different individual entities to be encoded uniquely by a very modest number of voxels.
- In this view, there appears to be more than adequate capacity to represent all possible concrete nouns, which have been estimated to number about 1600 concrete object types, as well as multiple tokens of each.
- These new findings suggest that the meanings of concrete nouns can be semantically represented in terms of the activation of a basic set of three main factors distributed across approximately twelve locations in the brain.
- The current results do reveal the beginnings of a biologically plausible basic set for concrete nouns, and they furthermore have the potential to be extended to other factors for other types of concepts.
- The results revealed a commonality of neural patterns across people, permitting concept identification across individuals, establishing for the first time that different brains represent concrete nouns similarly.
- These similarities presumably arise from a shared sensorimotor system and the shared use of the three fundamental dimensions for neurally representing physical objects; it is important to note that the location and activation levels did not have to be common across people.
- The results indicate that not only do people have concepts in common, but also their brain coding of the concepts is similar enough to decode one person's concept from other people's brain activation patterns.
- The study also demonstrated the ability to predict what the activation pattern would be for a previously unseen noun.
- A generative model performs well when trained on individuals distinct from the test subject, suggesting the potential for developing a general person-independent model of word representations in the human brain (and using this as a basis to study individual differences).

  - Future work: one question concerns the way that two or more words or concepts combine neurally to form a novel concept; another question concerns systematic individual differences in the way concepts are represented.
  - For the participants with the highest identification accuracies, the accuracies were lower when the classifier was trained on other participants' activation, indicating that there was some systematic but idiosyncratic structure in participant data.
  - Similarly, there may be systematic differences in concept representations in special populations, e.g. people with autism. They have a deficit in social processing, and so might represent social concepts differently.
  - In summary, the research establishes a new way of describing brain activity, not just in terms of its anatomical location and its physical characteristics, but in terms of the informational codes that are being processed in association with a given item.

- Kirsten I. Taylor, Barry J. Devereux, and Lorraine K. Tyler. Conceptual structure: Towards an integrated neurocognitive account. *Language and Cognitive Processes*, 26:1368–1401, 2011. doi:10.1080/01690965.2011.568227

  - Present a cognitive model of conceptual representations and processing: The Conceptual Structure Account.
  - An example of a distributed, feature-based approach.
  - Discuss studies using linguistic and non-linguistic stimuli, which are both presumed to access the same conceptual system.
  - Then take the CSA as a framework for hypothesising how conceptual knowledge is represented and processed in the brain.
  - Attempts to integrate the distributed feature-based model of sensory object processing.
  - Based on a review of relevant functional imaging and neuropsychological data.
  - They argue that distributed accounts of feature-based representations have considerable explanatory power, and that a cognitive model of conceptual representations is needed to understand their neural bases.

- Barry J. Devereux, Alex Clarke, Andreas Marouchos, and Lorraine K. Tyler. Representational similarity analysis reveals commonalities and differences in the semantic processing of words and objects. *Journal of Neuroscience*, 33:18906–18916, 2013. doi:10.1523/JNEUROSCI.3809-13.2013

  - Understanding the meanings of words and objects requires the activation of underlying conceptual representations.
  - Semantic representations are often assumed to be coded such that meaning is evoked regardless of the input modality.
  - The extent to which meaning is coded in modality - independent of or amodal systems remains controversial.
  - Address this issue in a human fMRI study investigating the neural processing of concepts, presented separately as written words and pictures.
  - Activation maps for each individual word and picture were used as input for searchlight-based multi-voxel pattern analyses.
  - Representational similarity analysis was used to identify regions correlating with low-level visual models of the words and objects and the semantic category structure common to both.

  - To explore differences in representational content across regions and modalities, they developed novel data-driven analyses, based on k-means clustering of searchlight dissimilarity matrices and seeded correlation analysis.
  - These revealed subtle differences in the representations in semantic-sensitive regions, with representations in LIPS being relatively invariant to stimulus modality and representations in LpMTG being uncorrelated across modality.
  - Result: Both LpMTG and LIPS are involved in semantic processing. Only the functional role of LIPS is the same regardless of the visual input. The functional role of LpMTG differs for words and objects.

- Alex B. Fine, T. Florian Jaeger, Thomas A. Farmer, and Ting Qian. Rapid expectation adaptation during syntactic comprehension. *PLoS ONE*, 8:e77661, 2013. doi:10.1371/journal.pone.0077661
  - We are faced with the challenge of inferring intended messages from noisy input: but there is considerable variability between and within speakers.
  - Tests the hypothesis that language comprehenders rapidly adapt to the syntactic statistics of novel linguistic environments (e.g. speakers or genres).
  - Based on evidence that language users possess the ability to quickly learn the distributional regularities over linguistic from artificial languages such as syllables, words, and syntactic categories.
  - Numerous researchers using artificial language paradigms have found evidence that knowledge of syntax might be acquired on the basis of statistical regularities defined over syntactic and lexical categories (e.g. Saffran et al [28]).
  - Adults maintain the capacity to learn statistical regularities when acquiring the lexicon or grammar of a language, but do they engage in qualitatively similar *statistical learning* during comprehension of their native language and is this active throughout adulthood?
  - Small number of recent studies directly address the notion that there a link between online sentence comprehension and statistical learning, drawing on experience-based models of language processing.
  - MacDonald & Christiansen [19] present simulations using simple recurrent networks (SRNs) to evaluate an experience-based interpretation of previously observed individual differences in language comprehension.
  - There has also been work testing statistical learning of syntactic structures in the lab [35] and finding correlations between online language comprehension performance and subjects' performance on a separate statistical learning task [23].
  - Two self-paced reading experiments investigate changes in readers' syntactic expectations based on repeated exposure to sentences with temporary syntactic ambiguities (so-called 'garden path sentences').
  - Authors find that comprehenders rapidly adapt their syntactic expectations to converge towards the local statistics of novel environments.
  - The opposite is also observed: a priori expected structures become less expected (even eliciting garden paths) in environments where they are hardly ever observed.
  - Findings suggest that when changes in syntactic statistics are to be expected (e.g. when entering a novel environment), comprehenders can rapidly adapt their expectations, thereby overcoming the processing disadvantage that mistaken expectations would otherwise case.

- A great deal of previous work on syntactic priming concerns the question of what mechanism gives rise to syntactic priming, with 2 main competing views:
  1. Transient activation account holds that priming results from a short-lived boost in the activation of a syntactic representation.
  2. Implicit learning accounts views priming as a consequence of an implicit learning mechanism.
- The results support an implicit learning account of priming: subjects are sensitive to the cumulative statistics of the environment; expectation for a current structure depends on how many times the subjects have seen this structure and others competing for probability mass.
- The authors argue that adaptation at the level of syntax serves the function of maximising the efficiency with which syntactic information is processed, because inferences about syntactic structures during online language understanding are only helpful given sufficiently accurate beliefs about environment-specific syntactic statistics.
- The authors suggest that adaptation is likely to be a general property of language processing, and is an essential ingredient in the ability of humans to cope with a dynamic environment.
- Previous work on adaptation from numerous research traditions in perception and cognition suggests that adaptation is a fundamental property of the human brain: a domain-general principle.
- Clark [4] discusses the central role of experience driven changes in behaviour , leading to the insight that the brain is fundamentally a ‘prediction machine’, and that the purpose of the brain can be profitably construed as modifying perception and behaviour in order to reduce error signals, i.e. the difference between what is expected and what is observed.
- Adaptation effects observed in this experiment (syntactic) arise at a similar timescale as those observed in phonetic adaptation: the adaptation process may be similar, as in both cases, humans seems to use often environment-specific information in the linguistic signal in order to inform the inferences and predictions they make about intended messages.
- As we can discuss adaptation across linguistic domains, it may not be entirely premature to ask whether adaptation of the kind observed in language experiments shares any commonalities with adaptation effects observed in non-linguistic domains.
- Especially because recent modelling work on syntactic adaptation and phonetic adaptation has successfully employed the same computational framework previously applied to adaptation in non-linguistic domains (Bayesian inference and belief-updating).
- Future research: whether the same cortical areas implicated in visual and auditory statistical learning tasks are also implicated in syntactic adaptation tasks.
- Also the question of how long subjects maintain adapted expectations, and whether it is generalised into subsequent linguistic situations.
- Kamide [16] suggests that the results of syntactic adaptation can be maintained insofar as comprehenders maintain separate subjective statistics for multiple talkers in a given environment; Wells et al [35] show that exposure to distributions of linguistic materials result in adaptation effects that persist at least for several days.
- In a paradigm like the authors used, to what extent would subjects generalise what they learn to a new experimental context? Do subjects track the statistics of syntactic structures, averaging across experience with all verbs, or do they track verb-specific statistics?

- Edward Gibson, Steven T. Piantadosi, Kimberly Brink, Leon Bergen, Eunice Lim, and Rebecca Saxe. A noisy-channel account of cross-linguistic word order variation. *Psychological Science*, 24:1079–1088, 2013. doi:10.1177/0956797612463705
  - Reversible sentences can be reversed and have a semantically-viable meaning, e.g. 'girl kicks boy' can be reversed to 'boy kicks girl', whereby the two sentences have semantically-viable but different meanings.
  - Distribution of word orders across languages is highly non-uniform.
  - There is a suggestion that subject-object-verb (SOV) may be the default order in human language.
  - The authors hypothesise that SOV/SVO variation can be explained by language users' sensitivity to the possibility of noise corrupting the linguistic signal.
  - The noisy-channel hypothesis predicts a shift from the default SOV order to SVO order for semantically reversible events. This is because potential ambiguity arises in SOV order because two plausible agents appear on the same side of the verb.
  - The authors found support for this prediction in three languages using a gesture production task which reflects word-order preferences largely independent of native language.
  - Other cross-linguistic variation patterns such as a prevalence of case-marking in SOV languages and lacking of this in SVO languages also straightforwardly follow from the noisy channel hypothesis.
  - The possible order of the basic units of a clause are highly non-uniformly distributed across languages.
  - In 96.3% of languages with a dominant word order, subjects precede objects. A plausible explanation for this is that people tend to construct their utterances from the perspective of agents rather than patients.
  - The first two subject-before-object word orders: SVO (41.2%) and SOV (47.1%) are much more prevalent than the third subject-before-object word order (VSO). However, until now, no explanation provided for the cross-linguistic prevalence of SOV and SVO word orders.
  - Functionalist approaches have contributed to the argument that grammars are independent of communicative and performance factors are determined by innate Universal Grammar.
  - This paper presents a communication-based explanation for the prevalence of SOV and SVO orders and for the cross-linguistic OV/VO variation, building on recent communicative accounts of similarly unexplained linguistic features such as ambiguity.
  - Breaking down the SOV preference: a preference for subjects to precede objects; a preference for the verb to appear clause-finally.
  - Two sources of evidence suggest that there is an initial bias to place the verb after its argument when developing a communication system.
  - These sources are two sign languages that were created independently from home sign languages, and these two sign languages are both verb-final (SOV or OSV). These are Nicaraguan Sign Language, and Al-Sayyid Bedouin Sign Language.
  - Goldin-Meadow et al. (2008) have recently observed that a verb-final order (specifically SOV) is preferred in a task where participants gesture event meanings, which essentially requires developing a new communication code.
  - SOV gesture production occurs not only for speakers of SOV languages, but also for speakers of SVO languages, which suggests that this task reflects word order preferences somewhat independent of the person's native language.

  - This paper proposes that SVO order arises cross-linguistically from SOV order due to communicative/memory pressures that can sometimes outweigh the default SOV bias.
  - In particular, building on Shannon's communication theory [29], we assume that language comprehension and production operate via a noisy channel.
  - The noisy-channel hypothesis: a speaker wishes to convey a meaning m and chooses an utterance u to do so.
  - This utterance is conveyed across a channel that may corrupt u in some way, resulting in a received utterance $\vec{u}$.
  - Noise may result from errors on the side of the producer, external noise, or errors on the side of the listener.
  - The listener must use $\vec{u}$ to determine m.
  - The best strategy for a speaker is this to choose an utterance u that will maximise the listeners' ability to recover the meaning given the noise process.
  - For example, in non-reversible sentences (e.g. 'girl kicks ball'), word order has little effect on how easily the meaning can be recovered, because the subject (agent) and object (patient) are clear from the semantics. In these situations, people should adhere to the default, SOV.
  - In semantically-reversible sentences (e.g. 'girl kicks boy'), noise may lead to confusion about which is the subject and which is the object in the SOV word order).
  - Note: although the noisy-channel hypothesis is motivated by a communicative theory, it need not be restricted to situations where we communicate with other people; it applies even if there is only one individual, who is encoding an event meaning for him/herself.
  - A difference in people's preferred word order for encoding or communicating meanings of reversible vs non-reversible events would suggest that word orders are shaped by noisy-channel pressures, with speakers choosing representations that maximise meaning recoverability.
  - In the paper, the authors show exactly this pattern of performance, with gestured word orders being dependent on the semantic reversibility of the event, across three languages: an SVO language (English) and two SOV languages (Japanese & Korean).
  - In addition to explaining gesture-production data, the noisy-channel hypothesis can explain four crosslinguistic typological patterns:
    1. If a linguistic community invents case-marking then the noisy channel hypothesis predicts that the default SOV order will be retained.
    2. Case-marking can be animacy-dependent. Animate direct objects should be more likely to be case-marked than inanimate objects.
    3. Word order can be animacy-dependent. In particular, among languages with relatively free word order (allowing both SOV and SVO word orders) many demonstrate 'word order freezing': in reversible constructions, when case does not disambiguate semantic roles, SVO word order is preferred.
    4. Languages that are not SVO can have more word order flexibility. According to the noisy-channel hypothesis, a language that is not SVO must contain mechanisms other than word order to unambiguously convey meanings of reversible sentences. Therefore, fixed word order should primarily be found in SVO languages, and non-SVO languages should generally have less rigid word order.

- Nathaniel J. Smith and Roger Levy. The effect of word predictability on reading time is logarithmic. *Cognition*, 128:302–319, 2013. doi:10.1016/j.cognition.2013.02.013

- It is well known that real-time human language processing is highly incremental and context-driven: the strength of a comprehender's expectation for each word encountered is a key determinant of the difficulty of integrating that word into the preceding content.
- Combining a state-of-the-art computational model, two large behavioural data-sets, and non-parametric statistical techniques, the authors for the first time establish the quantitative form of this relationship – it is logarithmic over six orders of magnitude in estimated predictability.
- The result is problematic for a number of established models of eye-movement control in reading and lends partial support to an optimal perceptual discrimination account of word recognition.
- At a more general level, this result provides challenges for both anticipatory processing and semantic integration accounts of lexical predictability effects.
- Result provides evidence that comprehenders are highly sensitive to relative differences in predictability – even for differences between highly unpredictable words – and thus helps bring theoretical unity to our understanding of the role of prediction at multiple levels of linguistic structure in real-time language comprehension.
- The authors' results suggest that there is no such thing as an unexpected word; there are only words which are more or less expected.
- The authors' results bear on the theoretical debate about whether predictability effects in general arise from anticipatory pre-activation of specific words, or from post hoc effects that arise while integrating the word into some kind of larger semantic context.
- The integration difficulty account holds that predictability itself does not affect comprehension difficulty, but rather that words which have high predictability scores are also those which are somehow more related to the prior context, and words which are more related to the prior context are also easier to integrate semantically.
- Crucially, under the account, predictability effects do not arise until after the comprehension system encounters the actual word; there may appear to be effects of predictability, but they do not result from any cognitive process of prediction.
- On the other hand, the anticipatory processing account holds that predictability effects do arise from some kind of processing which is predictive in the sense that it is dependent on the identity of the upcoming word, but occurs before this word identity is known.
- The results do not rule out an integration difficulty account, but given the precise and law-like relationship we found, the challenge for such accounts becomes to explain why integration difficulty should vary in a quantitatively exact way with the logarithm of predictability.
- The anticipatory processing account avoids this difficulty, because it is obvious why predictive processing would be sensitive to predict per se, if you want to start processing words in some manner before you before you actually encounter them, then a word's predictability of occurrence given the available information $P(w|C)$, may be a useful guide to decode which words should receive such processing, and to what degree.
- It is independently motivated: there is ample independent evidence that the comprehension system anticipates upcoming material in at least some situations.
- It provides an obvious reason why predictability differences would produce differences in reading time (as higher predictability words will receive more anticipatory processing, and thus require less post hoc processing).
- These results are incompatible with any theory which assumes both that (a) predictability effects on reading time arise from processing which precedes the actual appearance of the

word, and (b) the comprehension system can only apply this processing to a small number of words at any given moment (relative to the size of the lexicon).

- If we wish to preserve an anticipatory processing account of these data, we must instead reject (b), and build theories in which expectations do not take the form of simple guesses; instead, the comprehension system must be able to simultaneously pre-activate large portions of its lexicon in a quantitatively graded fashion.
- Yet another possibility would be for anticipatory processing to be directed not at words, but at word fragments, which would make this account consistent with the incremental processing theory we propose here (which takes as granted that there is some mechanism linking predictability and processing time, and focuses on explaining the resulting curve shape), while potentially reducing the degree to which parallel pre-activation is necessary (as there are e.g. far fewer potential upcoming phonemes than there are potential upcoming words).
- Note the consequences for speakers maintaining a UNIFORM INFORMATION DENSITY in conversation [18]: the authors find no evidence for derivation from a pure logarithmic curve, which under their analysis would suggest that overall audience interpretation time is entirely unaffected by the uniformity or non-uniformity of information density.
- UID predicts that information density should be optimised on the time scale of individual processing fragments; here, what would matter is uniformity on a time-scale only fine grained enough to avoid overloading comprehenders' working memory.
- In the BAYESIAN READER MODEL as originally formulated [26], predictability affects how much visual information the eye needs to gather from each word.
- Since you cannot gather perceptual input from a word that is no longer visible, the model expects a word's predictability to affect viewing time for that word only, with no spillover effect.
- The results however, show the exact opposite pattern: there is little or no effect on the word itself, with a large spillover effect.
- Could overcome this difficulty at the cost of some theoretical elegance: postulating that the noise bottleneck occurs at some later moment where word identity must be communicated between two internal processing stages connected by a noisy channel.
- Further analysis would be needed to determine whether such a mechanism could produce slowdowns distributed over such a wide temporal span (2-3 words).
- The INCREMENTAL PROCESSING account, by contrast, is based on the assumption that predictability affects not perception, but the speed of cognitive processing generally, making the the opposite prediction to the BRM – that predictability effects should not be restricted to the period when the word is actually visible, as found here.
- Therefore IP the only extant model which can directly explain the authors' full pattern of results, raising further questions:
  1. What is the form of the true underlying function $f(x)$ relating predictability and processing time?
  2. What is the value of k (the grain size of incremental processing)? Larger values of k would correspond to the processor operating on finer-grained or perhaps even truly continuous chunks of input.
- A rather large $k \geq 10$ is required to produce a near-logarithmic curve shape like the ones the authors observed: such a high degree of incrementality goes beyond what has already been observed in reading.

  - The results and model together, then, may provide an initial, tantalising glimpse of a more fine-grained linguistic processor than has so far been exposed to experimental view.
  - Other methods which allow more detailed measures of the time course of processing may yield further insights in this regard.
  - Finally, these results confirm it is plausible that all reading time predictability effects are mediated by lexical predictability, in accordance with the causal bottleneck hypothesis of surprisal theory.
  - Since the seminal work by Shannon quantifying the bit rate of English [30], information-theoretically informed work on language has recognised that all types of hierarchical predictive information present in language (syntactic, semantic, pragmatic and so forth) must inevitably bottom out in predictions about what specific word will occur in a given context, and that when measured in bits, expectations at each successive level combine naturally in a simple additive fashion.
  - The results reveal that the bit is also the correct unit for measuring the processing time needed in general during incremental languages comprehension by a native speaker.
  - With contemporary probabilistic models of language structure we can measure the bits carried by a wide variety of abstract linguistic structures; the way is thus paved for their contributions to the time required for incremental language comprehension to be investigated and quantified using this common currency.

- Marco Baroni, Georgiana Dinu, and Germán Kruszewski. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 2014. URL: https://aclanthology.info/papers/P14-1023/p14-1023

  - Context-predicting models (more commonly known as embeddings or neural language models) are the new kids on the distributional semantics block.
  - Literature is still lacking a systematic comparison of the predictive models with classic, count-vector-based distributional semantic approaches.
  - This paper performs an extensive evaluation on a wide range of lexical semantics tasks and across many parameter settings.
  - Result: the buzz is fully justified.
  - Context-predicting models obtain a thorough and resounding victory against their count-based counterparts.

- Alex Clarke and Lorraine K. Tyler. Object-specific semantic coding in human perirhinal cortex. *Journal of Neuroscience*, 34:4766–4775, 2014. doi:10.1523/JNEUROSCI.2828-13.2014

  - Category-specificity has been demonstrated in the human posterior ventral temporal cortex for a variety of object categories, but little is known about how specific objects are represented.
  - Here the authors used representational similarity analysis to determine what different kinds of object information are reflected in fMRI activation patterns and uncover the relationship between categorical and object-specific semantic representations.
  - Result: Gradient of information specificity along the ventral stream from representations of image-based visual properties in the early visual cortex to categorical representations in the posterior ventral stream.
  - Key finding: Object-specific semantic information is uniquely represented in the perirhinal cortex, which was also increasingly engaged for objects that are more semantically confusable.

    - Findings extend current distributional models by showing coarse dissociations between objects supported by the anterior medial temporal lobes, including the perirhinal cortex, which serve to integrate complex object information.

- Alex Clarke and Lorraine K. Tyler. Understanding what we see: How we derive meaning from vision. *Trends in Cognitive Science*, 19:677–687, 2015. doi:10.1016/j.tics.2015.08.008
    - Recognising objects goes beyond vision, and requires models that incorporate different aspects of meaning.
    - Most models focus on superordinate categories, but this does not capture the richness of conceptual knowledge.
    - Object recognition must be seen must be seen as a dynamic process of transformation from low-level visual input through categorical organisation to specific conceptual representations.
    - Cognitive models based on large normative datasets are well-suited to capture statistical regularities within and between concepts, providing both category structure and basic-level individuation.
    - These models capture important properties of the ventral visual pathway, shifting the focus from studying superordinate categories to basic-level concepts.
    - The way objects are naturally recognised is by accessing information specific enough to differentiate them from similar objects – a notion termed the basic or entry-level of representation.
    - To understand the cortical underpinnings of this flexible access to different aspects of conceptual representations, we need to specify the neurocomputational processes underlying meaningful object recognition.
    - This in turn requires that conceptual representations are studied as the expression of a set of dynamic processes of transformation – from the visual input and different stages of visual processing in the brain, through different types of categorical organisation, to a basic-level conceptual representation.
    - A central theme in vision science is to develop computational accounts of the ventral visual pathway, based on visual image properties, which try to explain non-human primate and human brain data.
    - But these models are unable to capture the relationships between different concepts, as models of vision alone cannot account for properties such as conceptual priming and flexible access to different aspects of meaning.
    - Research in semantic memory, by contrast, focuses on the organisation of semantic knowledge in the brain resulting in a variety of accounts drawing upon neuropsychology, functional neuroimaging, computational modelling and behavioural paradigms.
    - The authors' focus is on understanding the neural processes that underpin how meaning is accessed from vision.
    - The authors describe a neurocognitive model that integrates (i) a cognitive account of meaning based on the statistical regularities between semantic features that can explain a range of semantic effects, with (ii) the neurocomputational properties of the hierarchically organised ventral visual pathway.
    - The authors believe that concepts, not categories, should be the focus of research, as theories that focus on the organisation of superordinate category information alone ignore what is perhaps the most salient aspect of semantics – the information which differentiates between basic-level concepts – because it is these concepts that are claimed to be the most necessary in daily usage.

- The model the authors adopt here – the CONCEPTUAL STRUCTURE ACCOUNT – claims that concepts can be represented in terms of their semantic features and statistical measures, termed ‘conceptual structure statistics’ based on the regularities of features both across concepts and within a concept.
- Conceptual structure statistics can be informative about the superordinate category of a concept and how distinctive a concept is within a category.
- Recent fMRI data from healthy participants and lesion behaviour mapping in brain damaged patients show how conceptual statistics (capturing either superordinate category information or the ease of conceptual individuation) differentially relate to regions along the ventral visual pathway.
- First there is a gradient effect across the lateral-to-medial posterior fusiform gyrus, and objects with fewer shared features (typically tools) show greater effects in the medial posterior fusiform gyrus.
- Second, effects in the anteromedial temporal cortex (AMTC), specifically in the perirhinal cortex (PRC), are related to the ease of conceptual individuation: more confusable objects evoke greater activation.
- Damage to the PRC results in an increased deficit for naming semantically more-confusable objects, where confusability is defined by conceptual structure statistics.
- The statistical measures derived from feature-based accounts shed new light on the nature of category-specific effects in different regions of the ventral visual pathway, and do so with a framework situated at the level of basic-level concepts.
- This research points to a key computational role for the human PRC in the individuation of semantically-confusable concepts.
- Functionally, it can be argued that the PRC serves to differentiate between objects that have many overlapping features, and are therefore nearby in semantic space, while objects in sparse areas, with few semantic competitors, require less involvement of the PRC.
- These findings suggest a conceptual hierarchy in the ventral visual pathway, where a network of regions supports recognition of meaningful objects, and that category-specific effects emerge in different regions owing to categorical differences across complimentary semantic feature statistics.
- Any comprehensive account of conceptual processing must be able to capture the temporal dynamics during the retrieval of semantic knowledge.
- During object recognition, the system dynamics follow an initial feedforward phase of processing as signals propagate along the ventral temporal lobe, followed by recurrent, long-range reverberating interactions between cortical regions.
- There is clear evidence that information relevant to superordinate category distinctions can be accessed very rapidly (within 150ms), whereas specific conceptual information is only accessible after approximately 200ms.
- The model including both visual and semantic information could successfully account for single object neural activity from 60ms, the semantic feature information made unique contributions over and above those that the visual information could explain.
- Semantic feature information explained a significant amount of single object data in the first 150ms, and this is turn could predict neural activity that dissociated between objects from different superordinate categories.

- After around 150ms, the predictions become more specific, and differentiated between members of the same category (i.e. the basic-level concept could be predicted solely based on semantics).
- A second MEG study demonstrated that MEG signals correlated with the visual statistics of an object before rapid effects driven by the feature sharedness of the object in the first 150ms.
- Early information that (rapidly activated by visual properties) dissociates superordinate categories and which is driven by shared feature information, and later conceptual integration of information which individuates basic-level concepts from semantically similar items.
- Taken together, data from neuropsychology, fMRI and MEG reveal that semantic representations are transformed from primarily reflecting superordinate category information to basic-level conceptual information within a few hundred milliseconds, supported by processing along the ventral visual pathway.
- With regards to the mechanism of how basic-level concepts become differentiated within their category, we have shown that connectivity between the anterior temporal lobe and the posterior fusiform increases during tasks requiring access to basic-level concepts compared to those requiring access to superordinate category information.
- This highlights that the temporal relationship between neural activity in anterior and posterior temporal lobe regions plays an important role in the formation of detailed basic-level conceptual representations.
- An important issue is whether interactions involving anterior and posterior regions in the ventral visual pathway are predominately feedforward or feedback in nature, and how this might change during the course of perception.
- Patients with semantic deficits following neurological diseases affecting the anterior temporal lobes show reduced functional activity in the posterior aspects of the ventral stream, suggesting that anterior damage impacts on the functioning of more-posterior sites.
- These studies strongly suggest that feedback from the anterior temporal lobes, and from PRC, to the posterior ventral stream constitutes a necessary mechanism for accessing specific conceptual representations.
- The authors have emphasised that connectivity between anterior and posterior temporal lobe sites provides a key underpinning to forming specific basic-level conceptual representations, but how this within-temporal-lobe connectivity is coordinated with other networks (e.g. fronto-temporal connectivity) remains an important unresolved issue.

- Giacomo Handjaras, Emiliano Ricciardi, Andrea Leo, Alessandro Lenci, Luca Cecchetti, Mirco Cosottini, Giovanna Marotta, and Pietro Pietrini. How concepts are encoded in the human brain: A modality independent, category-based cortical organisation of semantic knowledge. *NeuroImage*, 135:232–242, 2016. doi:10.1016/j.neuroimage.2016.04.063
  - On the overall category-based organisation of conceptual knowledge, addressing the differential role of low-level sensory-based and high level abstract features in semantic features in semantic processing.
  - Combined behavioural studies of linguistic production and brain activity measures by functional magnetic resonance imaging in sighted and congenitally blind individuals while they performed a property generation task with concrete nouns from eight categories, presented through visual and/or auditory modalities.

- Representational analysis was performed to determine how knowledge organisation is affected when the analysis is focused onto two different levels of information organisation in the cortex: small-scale level limited to region-specific contents, and at a large-scale level that relies, as a whole, on a distributed network of regions engaged during the processing of semantic information.
- The analysis of neural patterns by means of a machine learning approach based on encoding techniques was able to to discriminate significantly the forty nouns across all presentation modalities and groups.
- Patterns of neural activity within a large semantic cortical network correlated with linguistic production and were independent both from the modality of stimulus presentation (either visual or auditory) and the (lack of) visual experience.
- In contrast, selected modality-dependent differences were observed only when the analysis was limited to the individual regions within the semantic cortical network: highest accuracy in pictorial form; accuracies above chance for verbal visual and for verbal auditory modalities, in both sighted and blind individuals.
- The success of models indicate that sighted and congenitally blind individuals share a common, category-based representation of knowledge.
- Generated a map of the cortex showing how semantic information is organised at both small-scale and large-scale distributions of information.
- Large-scale level comprised regions that did not show a preference for semantic categories, and showed a consistently strong correlation with the behavioural representations and across presentation modalities.
- The large-scale approach implies that during the functional inference of brain regions, only multivariate analyses were able to take into account the high dimensional cortical activity during semantic processing and to compare neural contents with behavioural information.
- The information content of the whole cortex of interest resulted from a merely linear weighting of information across voxels: using such an approach, the authors obtained a representational space showing the highest correlation with the behavioural data. Nevertheless, the human brain might weight the information with different criteria (e.g. with different weights for each voxel or using nonlinear criteria).
- At the small-scale level it was possible to identify regions with information content that showed category preferences, was only partially correlated to behavioural data and mainly retained a modality-dependent structure.
- The map at the lowest threshold identified four networks: a left posterior semantic network (PSN), including left PH, LO, TPO and IP; a left anterior semantic network (ASN), including motor, ventral premotor and dorsolateral prefrontal cortex; a right PSN, including the homologous regions of the left one; and a region within the Parieto Occipital (PO) cortex.
- All these regions, apart from the PO cortex, showed high correlations across the representational spaces related to the individual modalities, suggesting a modality-independent processing of semantic features.
- No category preferences were exhibited in any network, indicating a broader ability to retain semantic knowledge at the large-scale level.
- Two regions, the Fleschel Gyri and the Calcarine Sulci, also were defined to evaluate the information content measures. No significant correlations were found between brain activity in these regions and the linguistic output across presentation modalities.

- Only a few categories were discriminated above chance, likely due to their psychological features (e.g. specific visual spatial frequencies for places for places and vehicles) or psycholinguistic characteristics.
- However, no category preferences were identified, overall suggesting that the visual and auditory primary cortical regions are strictly unimodal and do not contribute significantly to high-level semantic representations.
- The results at the large scale imply a distributed and overlapping cortical representation of conceptual knowledge.
- The organisation of conceptual knowledge was here studied only through concrete nouns: even though the nouns used as stimuli covered a wide spectrum of semantic categories, from artefacts to places or animals, a complete evaluation of conceptual knowledge would require the inclusion of abstract entities.
- Shifting the definition of the semantic system at a cortical level from a smaller to a larger scale neural representation determined to what extent low-level sensory-based information and/or high level abstract features contribute to the organisation of conceptual knowledge.
- The authors propose that large-scale neural representations are an effective model to explain how the human brain processes semantic information and how conceptual knowledge emerges.
- In contrast, small-scale neural representations of limited regions showed category preferences and mainly retained a modality-dependent structure.
- These two distinct levels of semantic processing explain how information progresses from a sensory-based towards a more abstract conceptual representation.
- Concludes that conceptual knowledge in the human brain relies on a distributed, modality-independent cortical representation that integrates the partial category and modality specific information retained at a regional level.

Andrew J. Anderson, Douwe Kiela, Stephen Clark, and Massimo Poesio. Visually grounded and textual semantic models differentially decode brain activity associated with concrete and abstract nouns. *Transactions of the Association for Computational Linguistics*, 5, 2017. URL: https://www.transacl.org/ojs/index.php/tacl/article/view/879

- Using computational semantic models to decode brain activity patterns associated with concepts.
- Work has almost exclusively focused on concrete nouns: how well do these models extend to decoding abstract nouns?
- The authors apply state-of-the-art computational models to fMRI activity patterns, elicited by participants reading and imaging a diverse set of both concrete and abstract nouns: first model exploits recent word2vec skipgram approach trained on Wikipedia, the second is visually-grounded, using deep convolutional neural networks trained on Google Images.
- The dual coding theory considers concrete concepts to be encoded in the brain both linguistically and visually, and abstract concepts only linguistically.
- Splits fMRI data according to human concreteness ratings: both models successfully decode the most concrete nouns, however accuracy is significantly better using the text-based models for more abstract nouns.
- Confirms that current computational models are sufficiently advanced to assist in investigating the representational structure of abstract concepts in the brain.
- However, the results should be interpreted in light of the following two factors:

    1. The data set was based on a small sample of only 67 words;
    2. It is reasonable to conjecture that some of these words are also encoded in modalities other than vision and language.
  - The authors are undertaking more focused analyses on the current dataset, using textual, visual and newly developed audio semantic modes to tease apart linguistic, visual and acoustic contributions to semantic representations and how these vary throughout different regions of the brain.
  - In addition the Google Image search algorithm may not perform as well for abstract words as it does for concrete words, meaning that the visual model may have been handicapped compared to the textual model when decoding neural representations associated with more abstract words.
  - It could be possible to alleviate this in future work by having participants manually select images that they associate with abstract stimulus words, and using computational representations derived from these images in the analysis.
  - In summary, the authors found group-level commonalities in neural representation for both concrete and abstract words; this may provide a useful test-bed for evaluating computational semantic models.

- Francis Mollica and Steven T. Piantadosi. An incremental information-theoretic buffer supports sentence processing. In *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*, 2017. URL: https://mindmodeling.org/cogsci2017/papers/0162/index.html
  - *surprisal*: the colloquial word for the amount of information conveyed by a word. It is equal to the negative log probability of the word given its context.
  - Many current theories of sentence processing rely on natural reading times as a proxy for processing difficulties.
  - This assumption has been formalised in terms of information theory: words that carry more information tend to increase reading times relative to words that contain less information, suggesting a fixed processing rate that is measured in bits of information per unit time.
  - Researchers have noticed that language can be processed much faster than how we naturally engage with it.
  - Are the effects of surprisal observed in reading times reflective of linguistic information processing limitations, or do they arise from some alternate perceptual process?
  - In this paper, the authors conducted a novel self-paced, rapid serial visual presentation (RSVP) experiment, which controlled perceptual processes to probe for sentence processing limits during RSVP.
  - The RSVP experiment was designed to test if surprisal effects reflect language processing, and how readers might be compensating for linguistic information processing limits when faced with rapidly presented texts.
  - Two hypotheses were considered:
    1. That language users might suspend incremental information processing to store information in a buffer, until it might be processed.
    2. That language users might compensate for RSVP by utilising an incremental First-In, First-Out (FIFO) memory buffer, where information is immediately copied into a buffer and processed out of the buffer at a fixed rate.
  - The authors found support for sentence-related surprisal effects, the pattern of which is consistent with a FIFO buffer model.

- The model has one free parameter, i.e. the rate of information processing, and fit the model to our data to provide an estimate of the rate parameter.
- The amount of information that could be processed in the window (length of time that matches the presentation duration for each of the five words) according to the rate parameter is removed from the buffer.
- As a result, if a word could not be completely processed in one window, its processing carries over to the next window.
- To arrive at predictions for the surprisal weights, the authors analyse the expected reading times using linear regression with separate surprisal weights for each of the five words.
- The authors' a-priori prediction was that surprisal weights should increase across word positions, as early words will tend to be fully processed so that later words have a larger effect on post-presentation reading times.
- The authors suggested that surprisal effects observed in reading tasks might actually be a result of perceptual processing, rather than linguistic information processing.
- But if this were the case, we would expect to see no surprisal weights in the mSP and 5-RSVP conditions, where perceptual control was stripped from the reader; instead there were surprisal effects in RSVP reading, indicating a language processing origin.
- The authors proposed two possible alternatives that reconcile information processing limits and rapid text comprehension:
    1. Readers may compensate for RSVP by postponing information processing until presentation has ended to ensure all the input is copied into a buffer. The authors would then expect a small uniform profile of surprisal weights in the 5-RSVP condition.
    2. Alternatively, readers may compensate for RSVP by incrementally buffering and processing linguistic information, allowing perceptual processing and information processing to occur on quick but separate timescales. The prediction for the FIFO buffer was that early words should contribute less to the post-presentation reading time than later words, resulting in an increasing profile of surprisal weights in our 5-RSVP condition.
- If a word is optimally prepared for, its processing time should be a multiple of its surprisal. However, the model does not specify if there is a time cost to preparation or minimum required preparation time. If there is a cost to optimal preparation, the authors' experiment might not have provided sufficient time to prepare.
- If readers do indeed process information at such a fast rate, there would be no reason for surprisal effects to appear at all in natural reading tasks.
- One possible way to explain this would be that readers may prefer to maintain the information they are processing before proceeding. Whether this processing is perceptual (i.e. waiting for some level of certainty in the percept of the word), linguistic (e.g. parallel resource allocation in syntactic parsing), or optimal preparation is still an open question, with the important implication that surprisal effects in natural reading times might not be a measure of syntactic processing difficulty.
- Another possibility is that rate of information processing is not consistent across different tasks. The authors consider RSVP a demanding task, and so suggest that the rate of information processing might differ from natural rates of processing.
- The buffer model suggests that there is a decoupling of perceptual and linguistic information processing, which is potentially relevant for two sentence processing phenomenon: spill-over effects and right context effects.

    - Spill-over effects could be explained as perceptual processes continuing to advance through the sensory input while being sensitive to the information processing slightly lagging behind in a buffer.
    - For example, if a buffer had a fixed capacity, perceptual processing might stall on words further in the input than is currently being processed until the buffer has room for more information.
    - Right context effects occur when information further in the sensory input influences previously perceived information. In speech processing, listeners maintain uncertainty about words and have to hope that the future context will disambiguate the signal for them.
    - The current proposal is that the processing of a segment of speech operates beyond the duration of the speech segment, as the maintenance of unprocessed information implicates a linguistic buffer.
    - The data are consistent with a FIFO buffer model suggesting that when readers are quickly bombarded with information they store linguistic information in a buffer and immediately begin processing that information serially at a fixed rate.
    - The buffer model suggests a looser temporal coupling between perceptual processing and linguistic processing than had previously been theorised.
    - The authors' initial analyses using a FIFO buffer model prompt further research on the nature of the buffer and how the buffer may be implicated in other sentence processing phenomena.

- Jing Wang, Vladimir L. Cherkassky, and Marcel Adam Just. Predicting the brain activation pattern associated with the propositional content of a sentence: Modeling neural representations of events and states. *Human Brain Mapping*, 38:4865–4881, 2017. doi:10.1002/hbm.23692

    - Presents a predictive computational theory of the neural representations of individual events and states as they are described in 240 sentences.
    - Regression models were trained to determine the mapping between 42 neurally plausible semantic features (NPSFs) and thematic roles of the concepts of a proposition and the fMRI activation patterns of various cortical regions that process different types of information.
    - Factor analysis was used to develop a higher level characterisation of a sentence, specifying the general type of event representation that the sentence evokes and the voxel locations most strongly associated with each of the factors.
    - The NPSFs provide a basis of prediction of sentence activation in terms of the co-activation of various brain subsystems that have characteristic processing specialisations, resulting in more accurate prediction of sentence activation patterns than do feature sets derived from language-based corpora.
    - The NPSF-based characterisation can be expanded to include additional types of concepts and, if needed, additional neural systems.
    - The theory-driven NPSFs and the data-driven neural dimensions (factors) derived from the factor analysis of the sentence activations provide accounts at two levels of granularity:
        1. Factors provide an account at a coarser, higher level that specifies the general type of thought associated with a simple sentence (e.g. the thought of an action and its consequences).
        2. The NPSFs provide an account at a finer level that specifies the properties of each of the concepts in the proposition (e.g. a change of location of an object).

- Together, these two levels of of account suggest that the generativity of conceptual and propositional representation is enabled by the possible combinatorial space of these fundamental neural components and their values.
- The model integrates over the multiple concepts in a sentence to predict the activation of the sentence as a whole, and also takes into consideration the thematic structure that integrates the roles that the concepts play in the sentence.
- The model integrates over a wide range of 42 neurally plausible semantic features of different types, from concrete perceptual and motor features of an object (that can be thought of as embodied) to social and abstract features of actions that have much less connection to concrete properties.
- This computational model provides an account of the brain representation of a complex yet fundamental unit of thought, namely, the conceptual concept of a proposition; the models were also reliably generalisable across participants.
- Although sentences are not merely the sum of their parts, this study shows the extent to which a linear combination of thematically encoded concept representations is able to characterise the neural representation of simple sentences.
- Regression modelling enables the prediction of brain activity of a sentence and the comparably accurate prediction of the sentence semantics; the other direction of decoding semantics from activations provides a more intuitive assessment of the model's success, constituting a 'mindreading' capability.
- The two directions of mapping serve different purposes, and their comparable accuracies speak to the robustness of the general approach.
- The activation profiles of many regions identified in the factor analysis are consistent with previous findings of the role of these regions in semantic knowledge representation, such as right anterior temporal lobe for semantic knowledge of people, fusiform gyrus for representing objects, parahippocampal areas for representing places, and so forth.
- One of the large-scale dimensions emerging from the factor analysis assigned high scores to sentences describing people and social interactions.
- The location of the largest factor-related cluster (posterior cingulate cortex and the adjacent precuneus region) is known for its role in episodic memory and being a signature of the default mode network, and it has also been found to activate during social sharing of emotions versus processing emotion alone, and during thinking about intentional versus a physical causality. Similarly, the medial prefrontal cortex has been associated with social and moral reasoning.
- The cluster associated with this factor in the middle frontal gyrus has been associated with motor imagery and retrieval of visually or haptically encoded objects.
- Processing social content during the reading of sentences also involves a neural network for social processing. This factor is correlated with the NPSFs of person, communication, social norms, and social interaction.
- The sentences with high scores on the 'places' factor included most of the sentences that described a scene rather than an event.
- The brain regions associated with this factor include the parahippocampal area, and the posterior cingulate cortices, which have been linked to viewing scenes and representing semantic knowledge of shelters.
- The cluster in the right angular gyrus has been associated with accurate temporal memory and retrieval of learned ordinal movement sequences, suggesting its role in the representation of time-related concepts.

- This factor is correlated with the NPSFs of setting, openness (nonenclosure) and shelter (enclosure).
- Sentences that included main verbs such as BREAK or KICK had high scores on the 'actions' factor.
- The cluster associated with this factor in the middle frontal gyrus has been associated with motor imagery and retrieval of visually or haptivally encoded objects.
- This factor was correlated with the NPSFs of physical impact and change of location.
- The 'feelings' factor was correlated with the NPSFs of high affective arousal and negative affective valence.
- The right tempроparietal junction, which is known for its role in representing belief of others has been found to activate for processing fear-inducing pictures and fearful body expressions. The right inferior frontal gyrus has been associated with recognising visually presented objects with negative valence.
- In summary, the factor analysis yields broad semantic dimensions that characterise the events and states in the 240 stimulus sentences: the brain locations associated with each of the factors suggest that the brain structures that are activated in the processing of various aspects of everyday life events also encode the corresponding semantic information.
- The model was further tested for its generality across people: a model trained on the data of all but one participant is able to reliably predict the neural signature of the left out participant, indicating the generalisability of the neural representations across individuals.
- The fact that a small number of participants (7) is sufficient to build a cross-participant model suggests the consistency of the effect.
- The authors also note that the quantities being added are estimates of word concept representations as they occur in context, rather than in isolation.
- The initial success using NPSFs suggests that the building blocks for constructing complex thoughts are shaped by neural systems rather than by lexicographic considerations.
- This approach predicts that the neural dimensions of concept representation might be universal across languages, as studies are beginning to suggest: in this perspective, the concepts in each language would be underpinned by some subset of universal set of NPSFs.
- The predictive modelling of the neural signatures of new concepts in a variety of dissimilar languages is a possible way to test the hypothesis reflected by these neurally plausible semantic features, in contrast to hypotheses based on models that are blind to neural capabilities.
- This study demonstrates the possibility of modelling the neural representation of semantic information beyond the single-word level by taking into consideration the role of concept in proposition.
- Furthermore the model has the potential to identify neural signatures of other aspects of sentence meaning, such as negation, tense, syntactic roles of concepts, and so forth.
- This study leads to an initial theoretical and computational account of the neural representation of the propositional content of event-describing and state-describing sentences.
- The main contribution is the predictive bidirectional mapping between the neurosemantic properties of concepts and the neural signatures that characterise how the brain represents events and states described by simple sentences.
- The findings indicate the following:

    1. The neural representation of an event or state-describing proposition entails brain subsystems specialised in representing particular semantic information that can be characterised by a set of neurally plausible semantic features.
    2. It is possible to reliably predict sentence-level brain activity from this set of specialised neural bases and the knowledge of the semantic properties of the component words of a sentence and their inter-relations.
    3. It is also possible to decode the semantic properties of the concepts in a sentence from the observed activation patterns.
    4. The neural representation of the meaning if events and states is largely common across individuals.

  - One limitation is that despite the large number of stimulus sentences examined, the admixture of narrative and descriptive sentences is limited in its structure and content: the set of NPSFs would have to be expanded to code all possible sentences, but perhaps their number would asymptote at a few hundred.
  - More complex syntactic processing would evoke more activation, but the neural representation of the concepts involved may not be affected by the syntactic complexity of the sentence.
  - Furthermore, the outcome of the factor analysis here was limited to the sample of 240 stimulus sentences. It is likely that a much larger sample of sentence types would yield additional factors, which could also be independently assessed and tested for predictive ability.
  - The study was also limited to the processing of visually presented sentences, and the neural signature at the end of the reading of a sentence contained the representations of all the component concepts in the sentence. If the sentences were presented in the auditory modality, it is possible the neural signature at the end of the listening to a sentence might not be the optimal decoding window for all of the component concepts in the sentence.
  - The study opens new avenues of inquiry concerning the neural representation of complex inputs.

- Barry J. Devereux, Alex Clarke, and Lorraine K. Tyler. Integrated deep visual and semantic attractor neural networks predict fMRI pattern-information along the ventral object processing pathway. *Nature Scientific Reports*, 8:10636, 2018. doi:10.1038/s41598-018-28865-1

  - Recognising an object involves rapid visual processing and activation of semantic knowledge about the object, but how visual processing activates and interacts with semantic representations remains unclear.
  - In this study, the authors investigate visuo-semantic processing by combining a deep neural network model of vision with an attractor network model of semantics, such that visual information maps onto object meanings represented as activation patterns across features.
  - A combined model is used: concept activation is driven by visual input and co-occurence of semantic features consistent with neurocognitive accounts.
  - The model's ability to explain fMRI data where participants named objects was tested.
  - Visual layers explained activation patterns in early visual cortex, whereas pattern-information in peripheral cortex was best explained by later stages of the attractor network when detailed semantic representations are activated. Posterior ventral temporal cortex was best explained by intermediate stages corresponding to initial semantic representation.
  - The results from this study provide proof of principle of how a mechanistic model of combined visuo-semantic processing can account for pattern-information in the ventral stream.

- The model uses a pre-trained DNN and then an attractor network trained on the output of the DNN, and so there is no possibility for semantic representations to influence the representations in earlier layers, either through the back propagation of error during training or through explicit feedback connections from the semantic network to previous visual layers.
- Proof of principle of how a mechanistic model of combined visuo-semantic processing can account for pattern-information in the ventral stream.
- They tested the degree to which a combined visual and semantic computational model could account for visual object representations.
- This research shows a proof of principle for the combined computational model which offers one potential route by which visual properties interact with more abstract semantic information.
- This study investigated the transition between visual and semantic processes in object recognition by combining visual information from the layers of a DNN model of vision with a distributed, feature-based attractor network model of semantic processing.
- In this combined model, statistical dependencies between high-level visual information and semantic features are encoded in the connection weights between the high level visual layer and a recurrent semantic system.
- This combined VS model therefore allows the researchers to be explicit about the statistical regularities that facilitate the visuo-semantic mapping and makes quantitative predictions about the different stages of semantic activation.
- Activation of semantic features in this system is driven by both the high level visual input as well as lateral recurrent connections with other semantic features.
- Whilst other work suggests a role for semantics in the angular gyrus, temporal pole and lateral anterior temporal cortex, the analyses of the authors in relation to this model did not yield significant effects in these regions, which is likely due to the modality of the items.
- Analysing neuroimaging data by fitting computational models of the same task to the imaging data could potentially be a powerful tool in cognitive neuroscience.
- This is because computational models are explicit about the mechanisms and information involved in the task and make specific quantitative predictions, whilst at the same time abstracting away from physiological detail that may be less relevant in a cognitive level account.
- The DNN models that have been of recent interest in vision neuroscience are typically optimised for the relatively narrow goal of visual discrimination performance in the ImageNet classification competition, which differs considerably from the goals of human object processing (e.g. understanding what is being seen, making inferences about how objects in a scene relate to each other, forming semantic associations, and so on, as well as visual identification).
- Depending on the demands of a given environmental context, such information must somehow be activated from the visual input.
- Shared semantic features and semantic features reflecting visual properties (e.g. is long) tended to activate more rapidly than more distinctive and less visual semantic features.
- The model, in particular, provides a computational implementation of a general-to-specific account of semantic processing, where coarse-grained, superordinate category-level information is initially activated from the visual input whilst fine-grained semantic information tends to emerge more slowly and is more reliant on the recurrent mutual co-activation of features with the semantic system.

    - The authors found that different processing stages of this model fit different stages of the ventral object processing stream. This was achieved through complementary ROI and whole-brain searchlight RSA fMRI analyses.
    - However, from the perspective of cognitive style visual DNN models may overfit to the labelling task and so can be trained to to correctly identify objects without incorporating the kind of rich semantic system that is fundamental and obligatory in human object processing.
    - Future work: Can also explore direct connections from early visual layers to semantics, semantic-to-visual top-down connections, and lateral recurrent connections within visual layers (all of which would be neurocognitively plausible) whilst assessing the impact of using different visual models to map onto semantics.
    - Another approach would be to train the feed-forward visual components and the recurrent semantic part of the network simultaneously.
    - In summary, the researchers combined a deep convolutional neural network model of vision with a distributed attractor network model of semantics and tested the degree to which it captured object representations in the ventral stream. The model exploits statistical regularities between high-level visual information and semantic properties, and makes predictions about semantic activation that are consistent with a general-to-specific account of object processing. Different stages of the visuo-semantic model correspond convincingly to different stages of the ventral object processing stream. The model goes beyond identifying effects associated with visual models and semantic models separately, and instead shows how general and specific semantic representations are activated as a consequence of vision.

- Francisco Pereira, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel J. Gershman, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko. Towards a universal decoder of linguistic meaning from brain activation. *Nature Communications*, 9:963, 2018. doi:10.1038/s41467-018-03068-4
    - The experiment is about decoding linguistic meaning from imaging data.
    - Experimental procedures were previously limited to concrete nouns.
    - This experiment consists of a new approach: to maximise the ability to generalise to new meanings from limited imaging data. These are sufficiently detailed to distinguish even semantically similar sentences.
    - Training a decoder on single concepts that comprehensively cover the semantic space leads to the ability to robustly decode meanings of semantically diverse new sentences.
    - The training stage involves learning a relationship between representations of input stimuli and imaging data. The decoding models are then used to predict imaging data on a voxel-by-voxel basis for test stimuli, or to predict the representation of the stimulus shown when test imaging data were acquired.
    - A novel sampling procedure was used in order to select the training stimuli so that it would cover the entire semantic space. This allowed the decoder to maximise generalisability, so it could decode diverse concepts from many different semantic categories.
    - The decoder can also robustly decode meanings of sentences represented as a simple average of the meanings of the content words, despite the fact that the decoder is trained on only a limited set of individual word meanings.
    - The decoder was tested on two independent imaging datasets. The materials in these datasets included abstract topics, and topics beyond those encountered in training.

  - This study shows that semantic information is distributed across several high-level cortical networks, as opposed to earlier ideas, which stated that this information was restricted to a specific brain region.
  - This study demonstrates the viability of using distributed semantic representations to probe meaning representations in the brain, laying a foundation for future development and evaluation of precise hypotheses about how concepts are represented and combined.
  - This approach can be used with any distributional semantic model, including future ones capable of expressing more subtle meaning distinctions (e.g.those that depend on word order and hierarchical relationships among words) or additional information such as frame semantics for a sentence.
  - Vectors for more specific meanings have recently become available. This could make for more precise mapping between brain activation and semantic dimensions.
  - The authors expect that decoding performance will improve by selecting different sets of informative voxels for different dimensions and possibly different regression models depending on the distribution of values in each dimension.
  - In summary, the authors report a viable approach for building a universal decoder capable of extracting a representation of mental content from linguistic materials.

- Nikolaus Kriegeskorte and Pamela K. Douglas. Cognitive Computational Neuroscience. *Nature Neuroscience*, 21:1148–1160, 2018. doi:10.1038/s41593-018-0210-5

  - To learn how cognition is implemented in the brain, we must build computational models that can perform cognitive tasks, and test such models with brain and behavioural experiments.
  - In this article, the authors review recent work in the intersection of cognitive science, computational neuroscience and artificial intelligence, motivated by the belief that it is time to better integrate these separate disciplines.
  - Understanding brain information processing requires that we build computational models that are capable of performing cognitive tasks.
  - The authors review several new developments that suggest that it might be possible to meet the combined ambitions of cognitive science (to explain how humans learn and think) and computational neuroscience (to explain how brains adapt and compute) using neurobiologically plausible artificial intelligence (AI) models.
  - The transition from cognitive psychology to cognitive science was defined by the introduction of task-performing computational models.
  - With the advent of human functional brain imaging, scientists began to relate cognitive theories to the human brain.
  - This became cognitive neuroscience, and cognitive neuroscientists began by mapping cognitive psychology's boxes (information processing modules) and arrows (interactions between modules) onto the brain.
  - Cognitive psychology's tasks and theories of high-level functional modules provided a reasonable starting point for mapping the coarse-scale organisation of the human brain with functional imaging techniques, including the electroencephalography, positron emission tomography and early functional magnetic resonance imaging (fMRI), which had low spatial resolution.
  - The field mapped an ever increasing array of cognitive functions to brain regions, providing a useful rough draft of the global functional layout of the human brain.

- A brain map, at whatever scale, does not reveal the computational mechanism; however, mapping does provide constraints for theory.
- Information exchange incurs costs that scale with the distance between the communicating regions – costs in terms of physical connections, energy and signal latency. Component placement is likely to reflect these costs. We expect regions that need to interact at high bandwidth and short latency to be placed close together.
- The challenge ahead is to build computational models of brain information processing that are consistent with brain structure and function and perform complex cognitive tasks.
- The following recent developments in cognitive science, computational neuroscience and artificial intelligence suggest that this may be achievable:
    - Cognitive science has proceeded from the top down, decomposing complex cognitive processes into their computational components. Unencumbered by the need to make sense of brain data, it has developed task-performing computational models at the cognitive level. One success story is that of Bayesian cognitive models, which optimally combine prior knowledge about the world with sensory evidence. Bayesian models have begun to engage complex cognition, including the way our minds model the physical and social world. This literature provides algorithms for approximate inference on generative models that can grow in complexity with the available data, as might be required for real world intelligence.
    - Computational neuroscience has taken a bottom-up approach, demonstrating how dynamic interactions between biological neurons can implement computational component functions. In the past two decades, the field developed mathematical models of elementary computational models of elementary computational components and their implementation with biological neurons. These include components for sensory coding, normalisation, working memory, evidence accumulation and decision mechanism, and motor control.
    - Artificial intelligence has shown how component functions can be combined to create intelligent behaviour. Recent advances in machine learning, boosted by growing computational power and larger datasets from which to learn, have brought progress at perceptual, cognitive and control challenges. Some of the most important advances are driven by deep neural network models, composed of units that compute linear combinations of their inputs, followed by static nonlinearities. These models employ only a small subset of the dynamic capabilities of biological neurons, abstracting from fundamental features such as action potentials. However, their functionality is inspired by brains and could be implemented with biological neurons.
- The three disciplines contribute complementary elements to biologically plausible computational models that perform cognitive tasks and explain brain information processing and behaviour.
- If computational models are to explain animal and human cognition, they will have to perform feats of intelligence. AI, and in particular machine learning, is therefore a key discipline that provides the theoretical and technological foundation for cognitive computational neuroscience.
- One path from measured brain activity toward a computational understanding is to model the brain's connectivity and dynamics. Connectivity models go beyond the localisation of activated regions and characterise the interactions between regions.
- A first approximation of brain dynamics is provided by the correlation matrix among the measured response time series, which characterises the pairwise 'functional activity' between

locations.

- By thresholding the correlation matrix, the set of regions can be converted into an undirected graph and studied with graph-theoretic methods.
- Connectivity graphs can be derived from either anatomical or functional measurements.
- However, the way anatomical connectivity generates functional connectivity generates functional connectivity is better modelled by taking local dynamics, delays, indirect interactions and noise into account.
- Analyses of effective connectivity and large-scale brain dynamics go beyond generic statistical models such as the linear models used in activation and information-based brain mapping in that they are generative models: they can generate data at the level of the measurements and are models of brain dynamics. However, they do not capture the represented information and how it is processed in the brain.
- Decoding can help us go beyond the notion of activation, which indicates the involvement of a region in a task, and reveal the information present in a region's population activity.
- When the decoder is linear, as is usually the case, the decodable information is in a format that can plausibly be read out by downstream neurons in a single step.
- Decoding and other types of mutivariate pattern analysis have helped reveal the content of regional representations, providing evidence that brain-computational models must incorporate.
- Three types of representational model analysis have been introduced in the literature: encoding models, pattern component models and representational similarity analysis:
  1. In encoding models, each voxel's activity profile across stimuli is predicted as a linear combination of the features of the model.
  2. In pattern component models, the distribution of the activity profiles that characterises the representational space is modelled as a multivariate normal distribution.
  3. In representational similarity analysis, the representational space is characterised by the representational dissimilarities of the activity patterns elicited by the stimuli.
- In summary, connectivity models capture aspects of the dynamic interactions between regions; decoding models enable us to look into brain regions and reveal what might be their representational content; representational models enable us to test explicit hypotheses that fully characterise a region's representational space.
- But these 3 methods fall short of building the bridge all the way to theory because they do not test mechanistic models that specify precisely how the information processing underlying some cognitive function might work.
- To build a better bridge between experiment and theory, we first need to fully specify a theory. This can be achieved by defining the theory mathematically and implementing it in a computational model, which can reside at different levels of description, trading off cognitive fidelity against biological fidelity.
- Models designed to capture only neuronal components and dynamics tend to be unsuccessful at explaining cognitive function. To link mind and brain, models must attempt to capture aspects of both behaviour and neuronal dynamics.
- In computational neuroscience, neural network models have been essential to understanding dynamics in biological neural networks and elementary computational functions.
- In cognitive science, they defined a new paradigm for understanding cognitive functions, called parallel distributed processing, in the 1980's, which brought the field closer to neuroscience. In AI, they have recently brought substantial advances in a number of applications,

ranging from perceptual tasks (such as vision and speech recognition) to symbolic processing challenges (such as language translation) and on to motor tasks (including speech synthesis and robotic control).

- Like brains, neural network models can perform feedforward as well as recurrent computations. The models deriving the recent advances are deep in the sense that they comprise multiple stages of linear-nonlinear signal transformation.
- One successful paradigm is supervised learning, wherein a desired mapping from inputs to outputs is learned from a training set of inputs (for example, images) and associated outputs (for example, category labels). However, neural network models can also be trained without supervision and can learn complex statistical structure inherent to their experiential data.
- The large number of parameters creates unease among researchers who are used to simple models with small numbers of interpretable parameters. However, simple models will never enable us to explain complex feats of intelligence.
- One challenge is that the high parameter count renders the models difficult to understand. Because the models are entirely transparent, they can be probed cheaply with millions of input patterns to understand the internal representations, an approach sometimes called 'synthetic neurophysiology'.
- In addition to testing these models by predicting brain activity data, the field has begun to test them by predicting behavioural responses reflecting perceived shape and object similarity.
- Models at the cognitive level enable researchers to envision the information processing without simultaneously having to tackle its implementation with neurobiologically plausible components. This enables progress on domains of higher cognition, where neural network models fall short.
- Moreover, a cognitive model may provide a useful abstraction, even when a process can also be captured with a neural network model.
- This paper briefly discusses three classes of cognitive model: production systems, reinforcement learning models and Bayesian cognitive models.
    1. Production systems provide an early example of a class of cognitive models that can explain reasoning and problem solving. These models use rules and logic, and are symbolic in that they operate on symbols rather than sensory data and motor signals. The formalism of production systems also provides a universal computational architecture. More recently such models have also begun to to be tested in terms of their ability to predict regional-mean fMRI activation time courses.
    2. Reinforcement learning models capture how an agent can learn to maximise its long term cumulative reward through interaction with its environment. As in production systems, reinforcement learning models often assume that the agent has perception and motor modules that enable the use of discrete symbolic representation of states and actions. The agent chooses actions, observes resulting states of the environment, receives rewards along the way and learns to improve its behaviour. The agent may learn a 'value function' associating each state with its expected cumulative reward. If the agent can predict which state each action leads to and if it knows the values of those states, then it can choose the most promising action. The agent may also learn a 'policy' that associates each state directly with promising actions. Under limited conditions, an agent might do better to build a model of its environment . A model can compress and generalise experience to enable intelligent action in novel situations. Model-free methods are computationally efficient (mapping from states to values or directly to

actions), but statistically inefficient (learning takes long), model-based methods are more statistically efficient, but may require prohibitive amounts of computation (to simulate possible futures). Until experience is sufficient to build a reliable model, an agent might do best to simply store episodes and revert to paths of action that have met with success in the past (episodic control). Storing episodes preserves sequential dependency information important for model building. Episodic control enables the agent to exploit such dependencies even before understanding the causal mechanism supporting a successful path of action. The brain is capable of each of these three modes of control (model-free, model-based, episodic) and appears to combine their advantages using an algorithm that has yet to be discovered.

3. A third, and critically important class of cognitive model is that of Bayesian models. They tell us what a brain should in fact compute for an animal to behave optimally. Perceptual inference, for example, should consider the current sensory data in the context of prior beliefs. Bayesian models have provided insights into higher cognitive processes of judgment and decision making, explaining classical cognitive biases as the product of prior assumptions which may be incorrect in the experimental task but correct and helpful in the real world. In the Bayesian cognitive perspective, the human mind, from infancy, builds mental models of the world. These models may not only be generative models in the probabilistic sense, but may be causal and compositional, supporting mental simulations of processes in the world using elements that can be re-composed to generalise to novel and hypothetical scenarios. Generative models are an essential ingredient of general intelligence. An agent attempting to learn a generative model strives to understand all relationships among its experiences. It does not require external supervision or reinforcement to learn, but can mine all its experiences for insights on its environment and itself.

- Cognitive models, including the three classes highlighted here, decompose cognition into meaningful functional components. By declaring their models independent of the implementation in the brain, cognitive scientists are able to address high-level cognitive processes that are beyond the reach of current neural networks.
- Understanding the brain requires that we develop theory and experiment in tandem and complement the bottom-up, data-driven approach by a top-down, theory-driven approach that starts with behavioural functions to be explained.
- Unprecedented rich measurements and manipulations of brain activity (e.g. from the US BRAIN Initiative and the European Human Brain Project) will drive theoretical insight when they are used to adjudicate between brain-computational models that pass the first test of being able to perform a function that contributes to the behavioural fitness of the organism.
- Marr [20] offered a distinction of three-levels of analysis: (i) computational theory, (ii) representation and algorithm, (iii) neurobiological representation.
- Whilst Marr's levels provide a useful guide for understanding the brain, they should not be taken to suggest that cognitive science need not consider the brain or that computational neuroscience need not consider cognition.
- Brains are the product of evolution and development, processes that are not constrained to generate systems whose behaviour can be perfectly captured at some abstract level of description.
- Deep neural network models provide a biologically plausible account of the rapid recognition of the elements of the visual experience. They can explain the computationally efficient pattern recognition component.

- Bayesian nonparametric models explain how deep inferences and concept formation from single experiences are even possible. They may explain the brain's stunning statistical efficiency, its ability to infer so much from so little data by building generative models that provide abstract prior knowledge.
- However, current inference algorithms require large amounts of computation and, as a result, do not yet scale to real-world challenges such as forming a new concept from a single visual experience.
- On a 20-watt power budget, the brain's algorithms combine statistical and computational efficiency in ways that are beyond current AI of either the Bayesian or the neural network variety.
- However, recent work in AI and machine learning has begun to explore the intersection between Bayesian inference and neural network models, combining the statistical strengths of the former (uncertainty representation, probabilistic inference, statistical efficiency) with the computational strengths of the latter (representational learning, universal function approximation, computational efficiency).
- Integrating all three of Marr's levels will require close collaboration among researchers with a wide variety of expertise.
- Shareable components include cognitive tasks, brain and behavioural data, computational models, and tests that evaluate models by comparing them to biological systems.
- An experimental task is a controlled environment for behaviour. It defines the dynamics of a task world that provides sensory input (for example, visual stimuli) and captures motor output (for example, button press, joystick control or higher-dimensional limb or whole body control.
- We need tests that compare models and brains on the basis of brain and behavioural data.
- Moreover, for a given brain, every act of perception, cognition and action is unique in time and cannot be repeated precisely because it permanently changes the brain in question.
- Developing appropriate tests for adjudicating among models and determining how close we are to understanding the brain is not merely a technical challenge of statistical inference. It is a conceptual challenge fundamental to theoretical neuroscience.
- Cognitive researchers who feel that current computational models fall short of explaining an important aspect of cognition are challenged to design shareable tasks and tests that quantify these shortcomings and to provide human behavioural data to set the bar for AI models.
- Neuroscientists who feel that current models do not explain brain information processing are challenged to share brain-activity data acquired during task performance and tests comparing activity patterns between brains and models to quantify the shortcomings of the models.

Acknowledgements

The first author was supported by the Cambridge Language Sciences Interdisciplinary Research Centre's Incubator Fund. We thank Jane Walsh and Ann Copestake for their help and advice.

Bibliography

[1] Andrew J. Anderson, Douwe Kiela, Stephen Clark, and Massimo Poesio. Visually grounded and textual semantic models differentially decode brain activity associated with concrete and abstract nouns. *Transactions of the Association for Computational Linguistics*, 5, 2017. URL: https://www.transacl.org/ojs/index.php/tacl/article/view/879.

[2] Ellen Bard, Richard Shillcock, and Gerry Altmann. The recognition of words after their acoustic offsets in spontaneous speech: Effects of subsequent context. *Perception & Psychophysics*, 44:395–408, 1988.

[3] Marco Baroni, Georgiana Dinu, and Germán Kruszewski. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 2014. URL: https://aclanthology.info/papers/P14-1023/p14-1023.

[4] Andy Clark. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36:181–204, 2012.

[5] Alex Clarke and Lorraine K. Tyler. Object-specific semantic coding in human perirhinal cortex. *Journal of Neuroscience*, 34:4766–4775, 2014. doi:10.1523/JNEUROSCI.2828-13.2014.

[6] Alex Clarke and Lorraine K. Tyler. Understanding what we see: How we derive meaning from vision. *Trends in Cognitive Science*, 19:677–687, 2015. doi:10.1016/j.tics.2015.08.008.

[7] Delphine Dahan. Time course of interpretation in speech comprehension. *Current Directions in Psychological Science*, 19:121–126, 2010. doi:10.1177/0963721410364726.

[8] Delphine Dahan and Gareth Gaskell. Temporal dynamics of ambiguity resolution: Evidence from spoken-word recognition. *Journal of Memory and Language*, 57:483–501, 2007.

[9] Barry J. Devereux, Alex Clarke, Andreas Marouchos, and Lorraine K. Tyler. Representational similarity analysis reveals commonalities and differences in the semantic processing of words and objects. *Journal of Neuroscience*, 33:18906–18916, 2013. doi:10.1523/JNEUROSCI.3809-13.2013.

[10] Barry J. Devereux, Alex Clarke, and Lorraine K. Tyler. Integrated deep visual and semantic attractor neural networks predict fMRI pattern-information along the ventral object processing pathway. *Nature Scientific Reports*, 8:10636, 2018. doi:10.1038/s41598-018-28865-1.

[11] Alex B. Fine, T. Florian Jaeger, Thomas A. Farmer, and Ting Qian. Rapid expectation adaptation during syntactic comprehension. *PLoS ONE*, 8:e77661, 2013. doi:10.1371/journal.pone.0077661.

[12] Edward Gibson, Steven T. Piantadosi, Kimberly Brink, Leon Bergen, Eunice Lim, and Rebecca Saxe. A noisy-channel account of cross-linguistic word order variation. *Psychological Science*, 24:1079–1088, 2013. doi:10.1177/0956797612463705.

[13] Susan Goldin-Meadow, Wing Chee So, Aslı Özyürek, and Carolyn Mylander. The natural order of events: how speakers of different languages represent events nonverbally. *PNAS*, 105:9163–9168, 2008. doi:10.1073/pnas.0710060105.

[14] Giacomo Handjaras, Emiliano Ricciardi, Andrea Leo, Alessandro Lenci, Luca Cecchetti, Mirco Cosottini, Giovanna Marotta, and Pietro Pietrini. How concepts are encoded in the human brain: A modality independent, category-based cortical organisation of semantic knowledge. *NeuroImage*, 135:232–242, 2016. doi:10.1016/j.neuroimage.2016.04.063.

[15] Marcel Adam Just, Vladimir L. Cherkassky, Sandesh Aryal, and Tom M. Mitchell. A neurosemantic theory of concrete noun representation based on the underlying brain codes. *PLoS ONE*, 5:e8622, 2010. doi:10.1371/journal.pone.0008622.

[16] Yuki Kamide. Learning individual talkers' structural preferences. *Cognition*, 124:66–71, 2012.

[17] Nikolaus Kriegeskorte and Pamela K. Douglas. Cognitive Computational Neuroscience. *Nature Neuroscience*, 21:1148–1160, 2018. doi:10.1038/s41593-018-0210-5.

[18] Roger Levy and T. Florian Jaeger. Speakers optimize information density through syntactic reduction. In *Proceedings of the 19th International Conference on Neural Information Processing Systems*, 2006.

[19] Maryellen C. MacDonald and Morten H. Christiansen. Reassessing working memory: A comment on Just & Carpenter (1992) and Waters & Caplan (1996). *Psychological Review*, 109:35–54, 2002.

[20] David Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Cambridge, MA: MIT Press, 1982.

[21] William Marslen-Wilson. Sentence perception as an interactive parallel process. *Science*, 189:226–228, 1975.

[22] James McClelland and Jeffrey Elman. The TRACE model of speech perception. *Cognitive Psychology*, 18:1–86, 1986.

[23] Jennifer B. Misyak and Morten H. Christiansen. Statistical learning and language: An individual differences study. *Language Learning*, 62:302–331, 2012.

[24] Francis Mollica and Steven T. Piantadosi. An incremental information-theoretic buffer supports sentence processing. In *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*, 2017. URL: https://mindmodeling.org/cogsci2017/papers/0162/index.html.

[25] Hellen E. Moss, Jenni M. Rodd, Emmanuel A. Stamatakis, Peter Bright, and Lorraine K. Tyler. Anteromedial temporal cortex supports fine-grained differentiation among objects. *Cerebral Cortex*, 15:616–627, 2005. doi:10.1093/cercor/bhh163.

[26] Dennis Norris. The Bayesian reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological Review*, 113:327–357, 2006. doi:0.1037/0033-295X.113.2.327.

[27] Francisco Pereira, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel J. Gershman, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko. Towards a universal decoder of linguistic meaning from brain activation. *Nature Communications*, 9:963, 2018. doi:10.1038/s41467-018-03068-4.

[28] Jenny R. Saffran and Richard N. Aslin Elissa L. Newport. Statistical learning by 8-month-old infants. *Science*, 274:1926–1928, 1996.

[29] Claude E. Shannon. *The Mathematical Theory of Communication.* Univ of Illinois Press, 1949.

[30] Claude E. Shannon. Prediction and entropy of printed English. *Bell Systems Technical Journal*, 30, 1951.

[31] Nathaniel Smith and Roger Levy. Optimal processing times in reading: a formal model and empirical investigation. In *Proceedings of the Annual Conference of the Cognitive Science Society (CogSci)*, 2008. URL: https://escholarship.org/uc/item/3mr8m3rf.

[32] Nathaniel J. Smith and Roger Levy. The effect of word predictability on reading time is logarithmic. *Cognition*, 128:302–319, 2013. doi:10.1016/j.cognition.2013.02.013.

[33] Kirsten I. Taylor, Barry J. Devereux, and Lorraine K. Tyler. Conceptual structure: Towards an integrated neurocognitive account. *Language and Cognitive Processes*, 26:1368–1401, 2011. doi:10.1080/01690965.2011.568227.

[34] Jing Wang, Vladimir L. Cherkassky, and Marcel Adam Just. Predicting the brain activation pattern associated with the propositional content of a sentence: Modeling neural representations of events and states. *Human Brain Mapping*, 38:4865–4881, 2017. doi:10.1002/hbm.23692.

[35] Justine B. Wells, Morten H. Christiansen, David S. Race, Daniel J. Acheson, and Maryellen C. MacDonald. Experience and sentence processing: Statistical learning and relative clause comprehension. *Cognitive Psychology*, 58:250–271, 2009.