FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG
FACULTY OF ENGINEERING

Cognitive Neuroscience for AI Developers

Week 12– Language

„Language is to the mind more than light is to the eye“

William Gibson, Author

A brief introduction in linguistics

- **Linguistics**: Study of language (no unique tools)
- Linguistics is **inter-disciplinary field** itself

- Contribution of Linguistics in Cognitive Science in 1959 -> **Noam Chomsky's critique of behaviorism**
  (Skinner's book -> you can only say things you have heard before)

- **->in this lecture special focus on neurolinguistics** (language in brain, neuroimaging etc…)

Noam Chomsky *1928

https://de.wikipedia.org/wiki/Noam_Chomsky

Introduction: What is language?

There is no sophisticated definition for language but some characteristics:

- **Communicative**: Transmission and comprehension of information
- **Arbitrary**: Symbols that represent things are arbitrary (truck in USA, lorry in UK)
- **Structured**: Ordering of symbols is not arbitrary but follows a set of rules
- **Generative**: Nearly infinite number of sentences can be build (makes language powerful as any idea can be expressed)
- **Dynamic**: Language changes (tweet: sound of a bird -> now message on twitter)

Introduction: What is language?

Further properties and functions of language:

- Language most clearly distinguishes our species, although some animal species have some sophisticated communication systems

- There is **no real animal homolog** for language -> less well understood than sensation, memory etc.

- Language enables us to **learn from experiences**

Introduction: What is language? Important terms

- **Phoneme:** smallest sound that can change meaning of a word (**l**ate, **r**ate)
- **Morphemes**: smallest units of spoken language that have meaning (**de**frost, **frost**, defrost**er,** morphemes are structured collection of phonemes, stem morpheme: frost, bound morpheme: er)
- **Syntax**: rules to arrange words in sentences (related to **grammar**)
- **Semantics**: meaning of words an sentences
- **Pragmatics**: meaning of language in social and physical context

(Do you know what time it is?)

https://www.britannica.com/topic/syntax
https://www.britannica.com/science/pragmatics

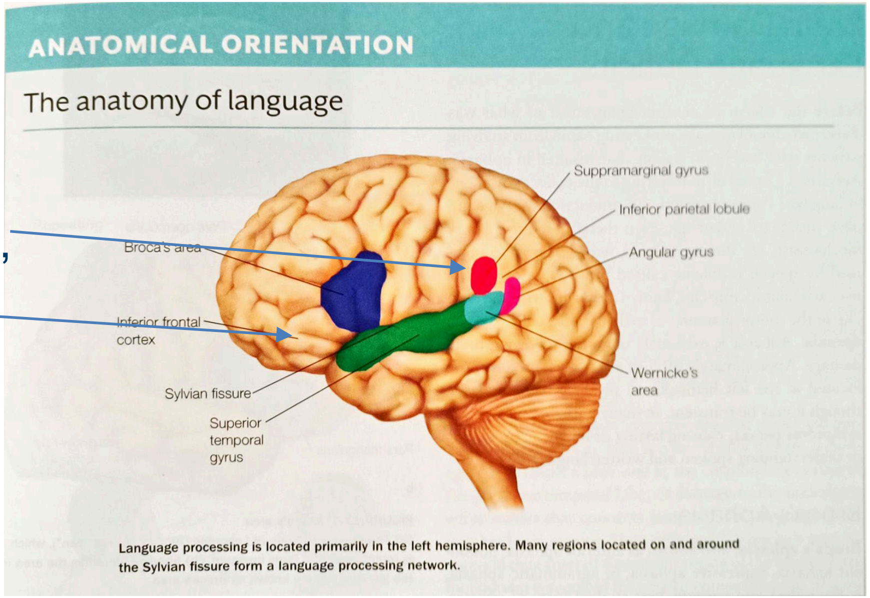
Neurolinguistics Anatomy of Language and Brain Lesions

Anatomy of Language

- **Language areas**:
  Left temporal cortex: Wernicke's area in posterior superior temporal gyrus,
  inferior parietal lobe (supramarginal gyrus),
  Left inferior frontal cortex: Broca's area,
  Left insular cortex

- **Left hemisphere**: language processing
  Right hemisphere: prosody, contribute to process metaphors

- Speech production is movement and thus motor cortex, basal ganglia etc. contribute



Cognitive neuroscience, Gazzaniga, Ivry, Mangun, 2014

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG
FACULTY OF ENGINEERING

Introduction: Neurolinguistics

- **Reminder**: researchers started to investigate language in patients with language deficits (e.g. lesions)
- **1960s** renewed interest in research on people with language deficits, also new imaging techniques had an influence
- **Main Question of Neurolinguistics**: How does the brain derive meaning from auditory and visual input and produces spoken and written language?

Source: wikipedia.org
Pierre Paul Broca (1824-1880)

Source: wikipedia.org
Carl Wernicke (1848-1905)

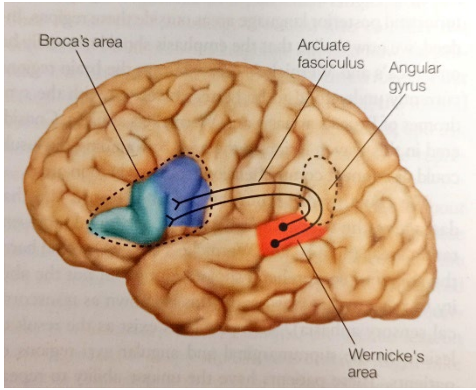
Reminder: Brain Damage

- **Aphasia**: collective term summarizing deficits in language understanding and production
- **Broca's** aphasia ("tan"): problems with speech production (Broca's view) but also with grammar (syntax -> **agrammatic aphasia**)
- **Wernicke's** aphasia: problems with speech understanding, produce fluent speech but sentences make no sense (modern view: areas around Wernicke's area have biggest influence)

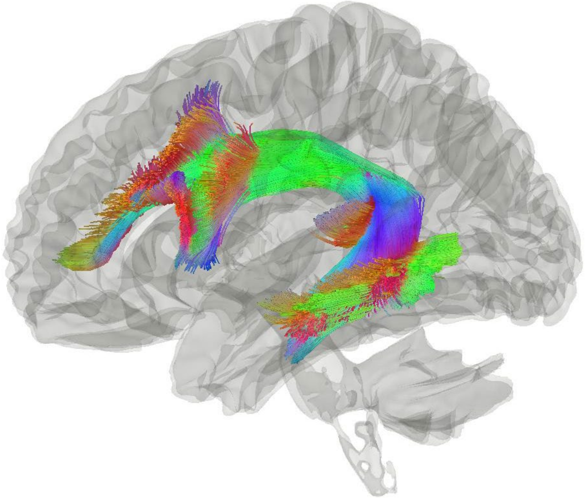
Reminder: Brain Damage

- **Conduction aphasia**: Damage of white matter tract from Wernicke's to Broca's area called **arcuate fasciculus**

  -> patients understand words and speech errors they hear but cannot correct the own speech errors, problems in producing spontaneous speech and repeating speech, use words incorrectly



Cognitive neuroscience, Gazzaniga, Ivry, Mangun, 2014



https://de.wikipedia.org/wiki/Fasciculus_arcuatus

Mental Lexicon

Fundamentals of Language in the Human Brain

- To derive meaning from language input and to produce speech

  -> **the brain must store words and concepts** -> **mental lexicon**

- **Mental lexicon**: Mental store of semantic, syntactic information, spelling, and sound patterns of words

- **Two theories**:

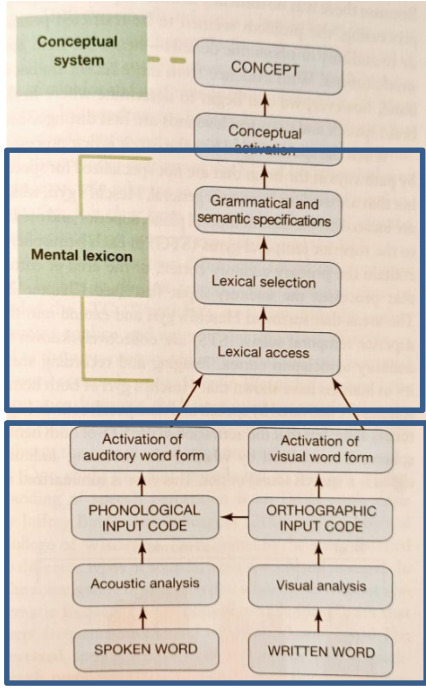
  1) One lexicon for language understanding and production

  2) Two different lexica (separate input and output lexicon)

Fundamentals of Language in the Human Brain: Mental Lexicon

- **Words must be analyzed perceptually first (auditory system, visual system)**

- **Three general functions of mental lexicon**:
  - **Lexical access**: perceptual output activates word-from representations
  - **Lexical selection**: lexical representation in mental lexicon which best matches the input is selected
  - **Lexical integration**: integrates words into full sentence
- **Adult speaker has knowledge on 50,000 words and can produce 3 words per second -> needs highly efficient metal lexicon -> not like a dictionary**



Cognitive neuroscience, Gazzaniga, Ivry, Mangun, 2014

Fundamentals of Language in the Human Brain: Mental Lexicon

- 1) smallest representation unit in the mental lexicon is the **morpheme**
  (Morpheme: smallest meaningful unit in language, **frost**, **de**frost, defrost**er**)

- 2) more frequently used words are accessed more quickly

- 3) lexical neighborhood: neighborhood of words with differ only in one **phoneme** (smallest unit of sound that changes meaning, **l**ate and **r**ate, words with many overlapping phonemes are organized together in the brain

- 4) **<u>Sematic relationship between words</u>**

Fundamentals of Language in the Human Brain: Mental Lexicon

- **Semantic relationships in mental lexicon**
- Example: A quick priming study to explore your mental lexicon

animal **lion**

plane **flouwer**

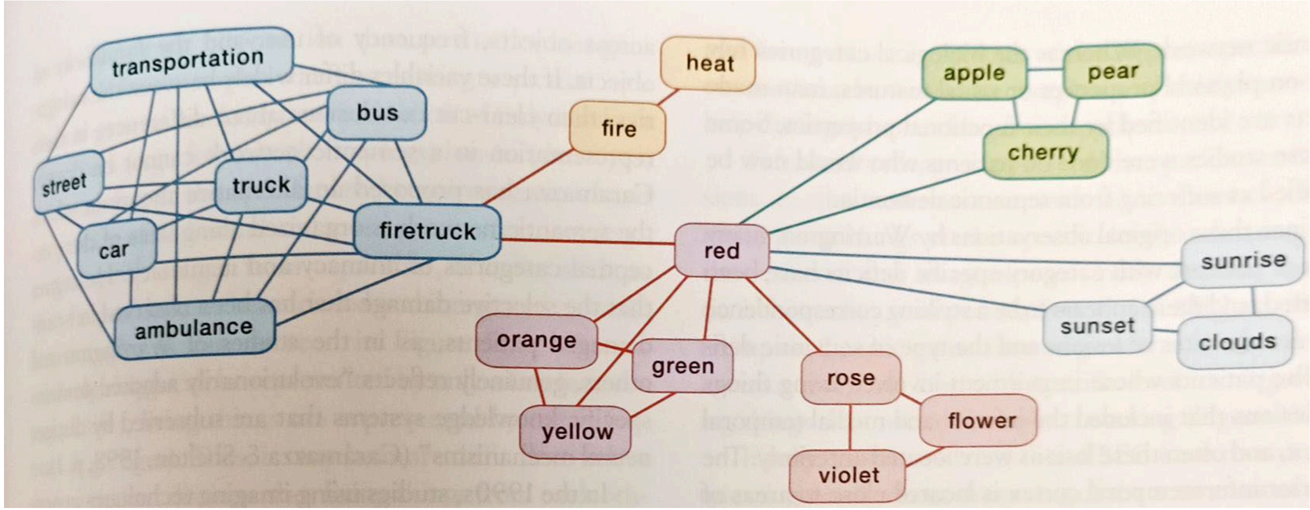
ball **tiger**

sports **footblal**

If prime is semantically related to target than people can decide more quickly if target is a real word!

Fundamentals of Language in the Human Brain: Mental Lexicon

- Sematic relationship between words in metal lexicon -> representations in the mental lexicon are organized according to meaningful relationship between words

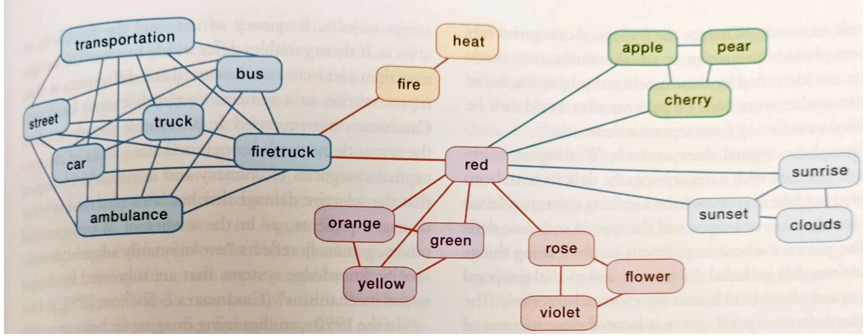


Cognitive neuroscience, Gazzaniga, Ivry, Mangun, 2014

Fundamentals of Language in the Human Brain: Model of Mental lexicon

- **1975 influential model by Collins and Loftus**
- **Idea**: Word meanings are represented in a **semantic network** (words= nodes connected with each other)
- **Distance between words is determined by sematic relations of words** (e.g. car is near truck)
- Activation spreads from one node to another -> closer words will benefit more from an activation (if we hear 'car' also truck is activated)

-> influential model but still under debate



Cognitive neuroscience, Gazzaniga, Ivry, Mangun, 2014

Fundamentals of Language in the Human Brain: Neuronal Substrate of Mental lexicon

- **The model of Loftus and Collins is a box and arrow model -> we need to map it on a neural substrate**
- 1970s: Groundbreaking studies by **Elizabeth Warrington**
  - People with certain brain lesions had **category specific impairments**
  - Strong connection between sites of lesions and the type of semantic deficit
  - E.g. patients with problems to name living things (e.g. animals) had lesions at different locations (inferior and medial temporal cortex) than people with problems to retrieve man-made things (left frontal and parietal brain areas)
  - View was challenged by imaging studies (fMRI, MEG, EEG) -> **work in progress**

Word2Vec – Mental Lexicon in Computers?

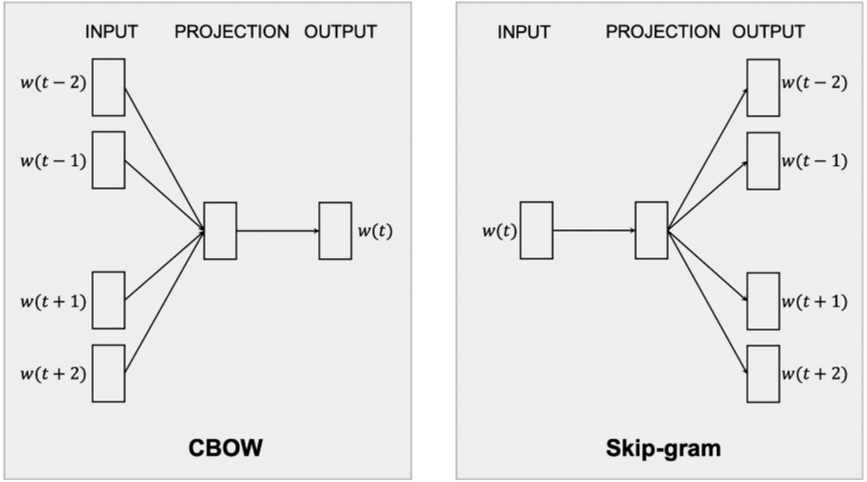
Mental lexicon in computers, Word2Vec

- **Learn representations of words**: each words is mapped to one vector (embedding)
- Embedding should be useful and efficient and should cover some **semantic and syntactic relationships**
- **For example**: Using vectors of the length of the words that contain the letter number on each position (e.g. at -> (1, 20)), or one-hot encoding is not very sophisticated

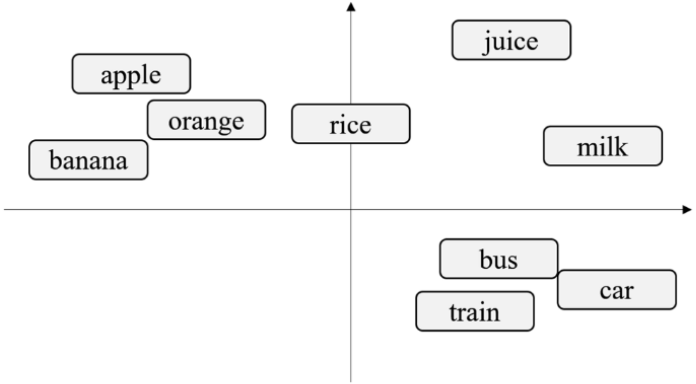
https://www.tensorflow.org/tutorials/text/word2vec

Mental lexicon in computers, Word2Vec

- **Word2Vec algorithms**
- **Continuous bag-of-words models**: predicts the middle word based on surrounding context words
- **Continuous skip-gram models**: predicts the surrounding words of certain word



https://www.degruyter.com/document/doi/10.1515/comp-2022-0236/html



Li, B., Drozd, A., Guo, Y., Liu, T., Matsuoka, S., & Du, X. (2019). Scaling word2vec on big corpus. *Data science and engineering*, *4*, 157-175.

https://www.tensorflow.org/tutorials/text/word2vec

Mental lexicon in computers, Word2Vec

- **Similarity of two embeddings**: Cosine Similarity
- Interesting: Can these similarities also been found in the brain?

-> thus comparison with EEG, fMRI etc

$$similarity = \cos\theta = \frac{\bar{x} \bullet \bar{y}}{\| \bar{x} \| \| \bar{y} \|}$$

Jatnika, D., Bijaksana, M. A., & Suryani, A. A. (2019). Word2vec model analysis for semantic similarities in english words. *Procedia Computer Science*, *157*, 160-167.

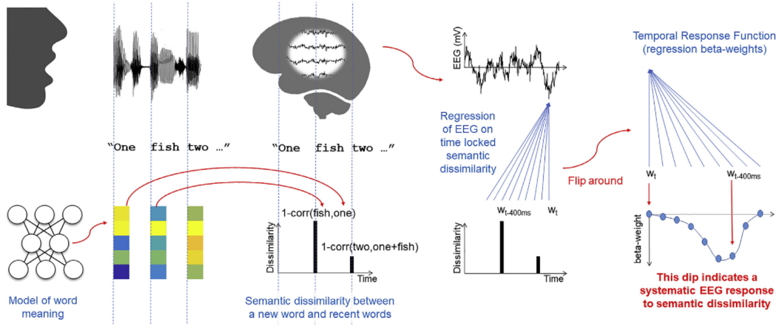
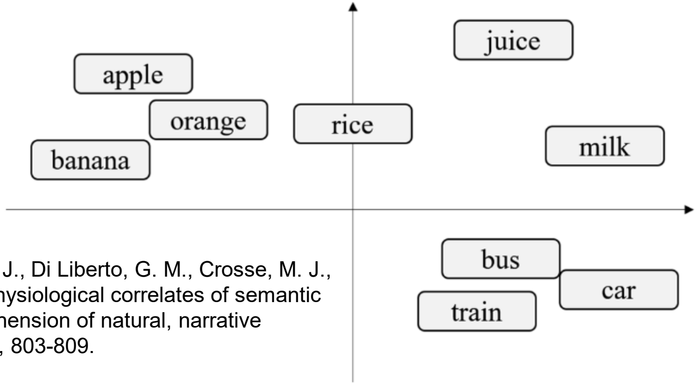


Figure 1. Regularized Regression Analysis for Estimating an Electrophysiological Correlate of Semantic Dissimilarity to Natural Speech
Content words from an audiobook are converted to 400-dimensional vectors using the word2vec algorithm [20] (bottom left). The semantic similarity of each word to its preceding context is then defined by comparing (via a Pearson's correlation) its 400-dimensional vector with the average of the vectors of all the preceding words in the corresponding sentence. And the "semantic *dissimilarity*" of the word is quantified as 1 minus this correlation (bottom middle left). A vector at the same sampling rate as the recorded neural data is then created that consists of time-aligned impulses at the onset of each word that are scaled according to the value of that word's semantic dissimilarity. The ongoing EEG data are then regressed against this vector to obtain a so-called temporal response function (TRF; right) that describes via beta weights how fluctuations in semantic dissimilarity across words impact upon the EEG at various time lags [21].

Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J., & Lalor, E. C. (2018). Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Current Biology*, *28*(5), 803-809.



Li, B., Drozd, A., Guo, Y., Liu, T., Matsuoka, S., & Du, X. (2019). Scaling word2vec on big corpus. *Data science and engineering*, *4*, 157-175.

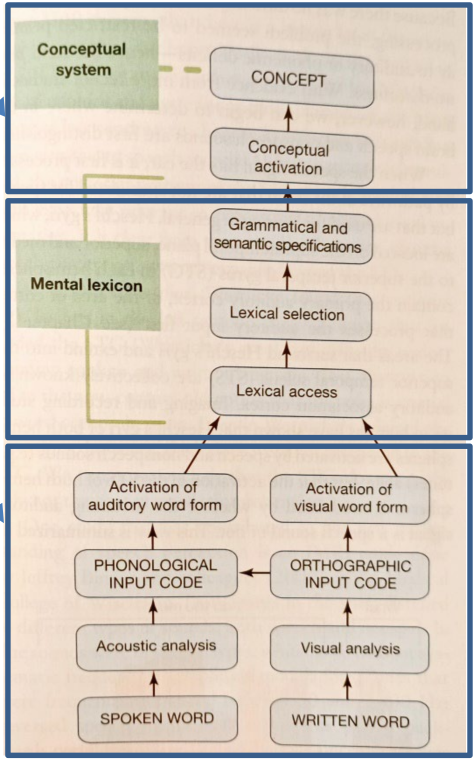
Language Comprehension

Language Comprehension

- **1 Step: Perceptual analysis**: Auditory or visual input is translated into phonological/ortographic input code

- **2 Step:** Access to representations which fit to the phonological code (lexical access) and select the best fitting one (lexical selection)
(The selected word includes grammatical and semantic information stored in mental lexicon)

- **3 Step:** Activation of conceptual information (Lexical integration)



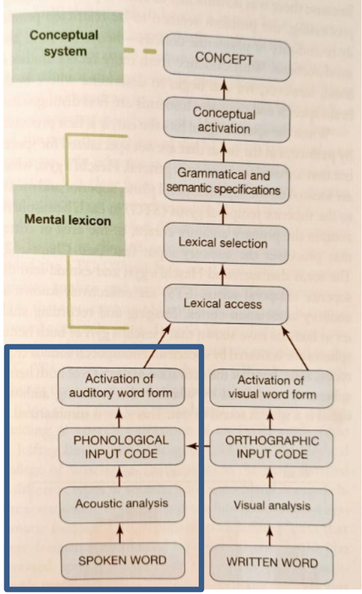
Cognitive neuroscience, Gazzaniga, Ivry, Mangun, 2014

Language Comprehension: Spoken Input: Understanding Speech

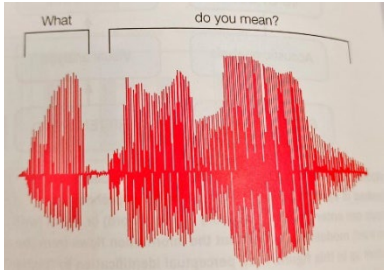
- Difficult as **acoustic noise** has to be divided from relevant speech signal in order to **identify phonemes** (smallest units of sound that change meaning of word (**c**ap, **t**ap, 40 phonemes in English language))

- Auditory system has to solve several problems: **1)** phonemes sound different for male and female speakers, **2)** Auditory speech signals are not clearly separated (even not between words)

- Prosody (rhythm and intonation) helps to segment the speech stream



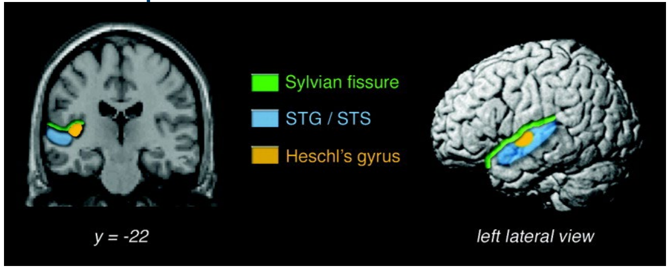
Cognitive neuroscience, Gazzaniga, Ivry, Mangun, 2014



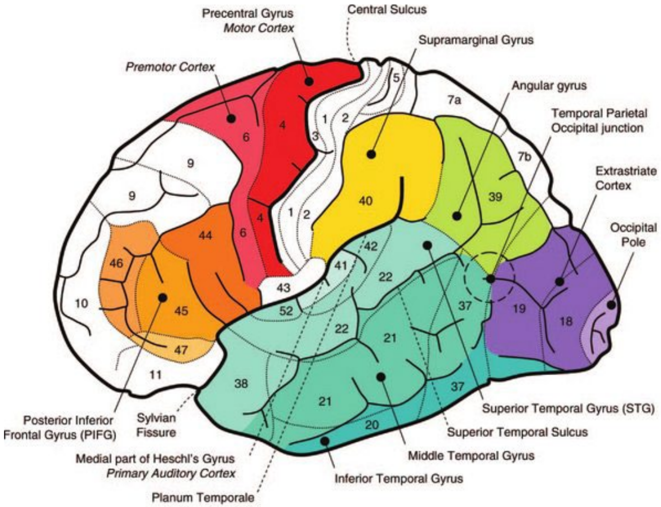
Cognitive neuroscience, Gazzaniga, Ivry, Mangun, 2014

Language Comprehension: Spoken Input: Understanding Speech: Neural Substrate

- **Normal processing of sound along auditory pathway** up to primary auditory cortex (at Heschl gyrus) -> not speech specific
- Lesions of superior temporal sulcus (STS) and gyrus (STG)-> **pure word deafness** (phonemes are not understood any more)
- Distinguishing between speech and non-speech sound in superior temporal sulcus
- Spoken-word recognition in anterior STG, short phrases in anterior STS



https://anatomypubs.onlinelibrary.wiley.com/doi/10.1002/ar.a.20316

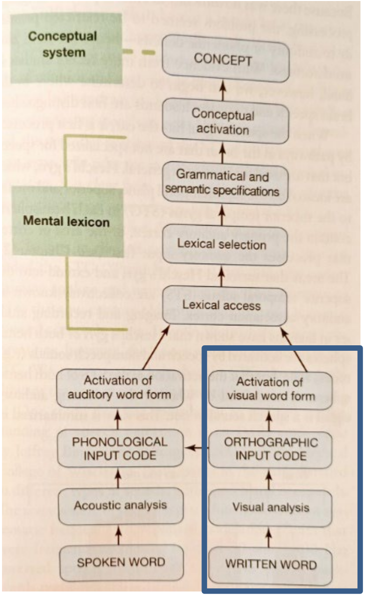


https://www.researchgate.net/publication/8113153_Renewal_of_the_Neurophysiology_of_Language_Functional_Neuroimaging/figures?lo=1

Language Comprehension: Written Input: Reading Words

- Reading is a recent invention (5500 years old)

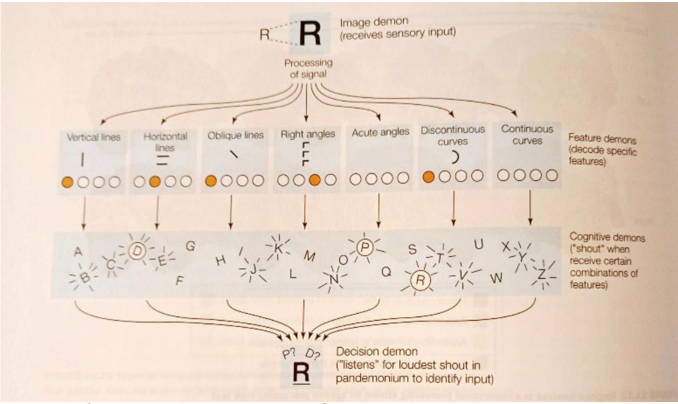
- Task: Linking arbitrary visual symbols to meaningful words



Cognitive neuroscience, Gazzaniga, Ivry, Mangun, 2014

Language Comprehension: Written Input: Reading Words

- **Pandemonium model by Oliver Selfridge (1959 landmark contribution to artificial intelligence, model how machines could recognize patterns)**
  1) Sensory input is stored by the so-called image demon (demon = small component)
  2) 28 feature demons decode features in the iconic representation
  3) All representations of letters with these features are activated by cognitive demons
  4) The representation that best matches the input is selected by decision demon
- **Model is criticized as it is a pure bottom up model!**

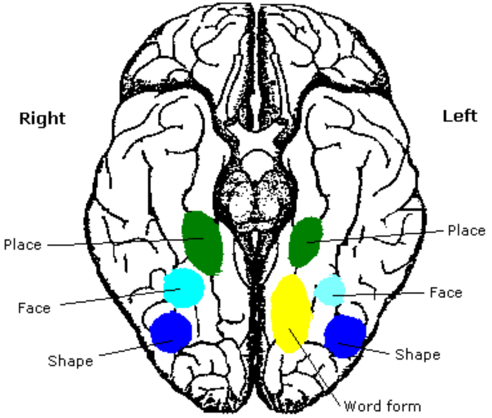


Cognitive neuroscience, Gazzaniga, Ivry, Mangun, 2014

Language Comprehension: Written Input: Reading Words: Neural Substrate

- **Occipitotemporal cortex lesion** causes **alexia** -> people cannot read words (other aspect of vision and language comprehension are normal)
- **Visual word form area**: react on strings of presented words

- **We know where things happen but not what happens!**



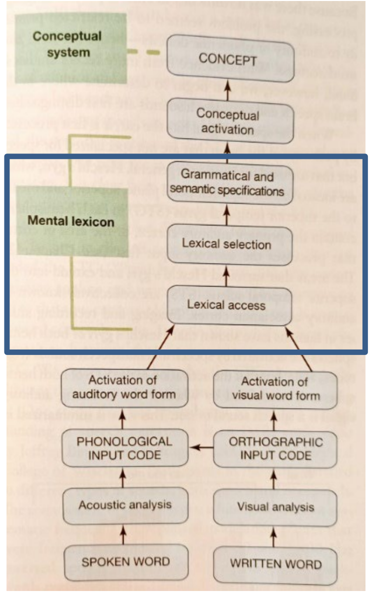
https://en.wikipedia.org/wiki/Fusiform_face_area

Language Comprehension: The role of Context in Word recognition

- After identification of a phonological or visual representation **semantic and syntactic information must be retrieved** and the word has to be integrated in a whole utterance

- Additionally there is an influence of linguistic and non-linguistic (image) context on word comprehension (e.g. “The tall man planted a tree on the bank.”)

- **Two types of representations** play a role in word processing: low-level representations (from sensory input), higher level representations constructed from the preceding words and context -> **but at which level?**



Cognitive neuroscience, Gazzaniga, Ivry, Mangun, 2014

Language Comprehension: The role of Context in Word recognition

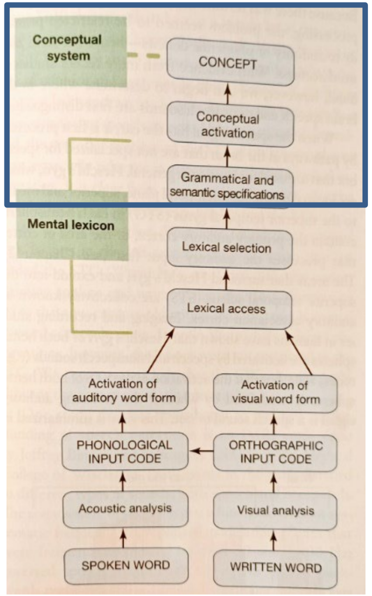
- **3 classes of models explain word comprehension:**
- **1) Modular models**: normal language comprehension in separate modules (no influence of higher-level representations on lower-level representations, bottom-up data flow)
- **2) Interactive models**: all types of information can participate in word recognition
- **3) Hybrid models (in between):** the context reduces the number of possible word candidates in the mental lexicon
- **-> Modern view: lexical selection is indeed influenced by context**

Language Comprehension: Integration of words in sentences

- **Semantic information on the words alone is not enough to get the meaning of the whole sentence!** “The little old lady bites the gigantic dog!”

- Syntax (grammar) has to be analyzed

- Brain cannot store whole sentences -> the brain has to assign a syntactic structure to words in sentences called **syntactic parsing**

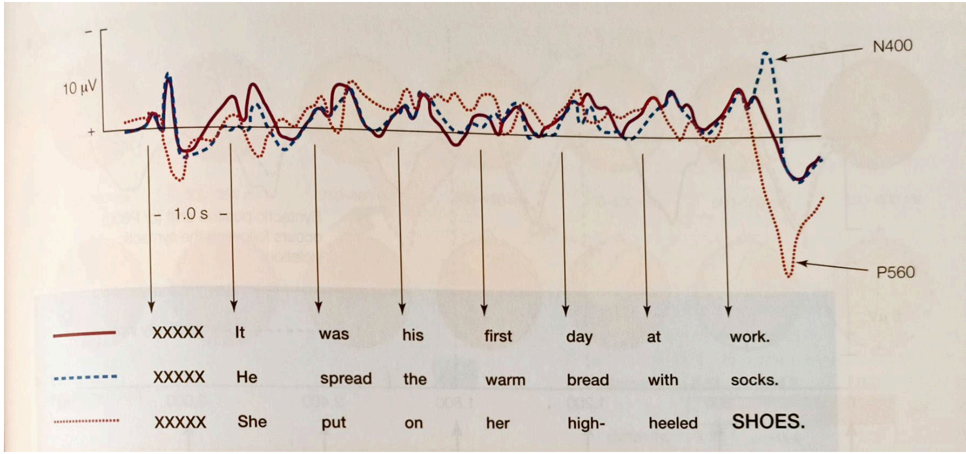


Cognitive neuroscience, Gazzaniga, Ivry, Mangun, 2014

Electrophysiological Measures for Semantic and Syntactic Processing (ERPs)

Language Comprehension: Semantic processing and the N400 wave

- **N400 wave** is a negative voltage peak in ERPs and is sensitive to **semantic aspects** of linguistic inputs
- Semantic violations lead to larger N400 (more negative) wave



Cognitive neuroscience, Gazzaniga, Ivry, Mangun, 2014

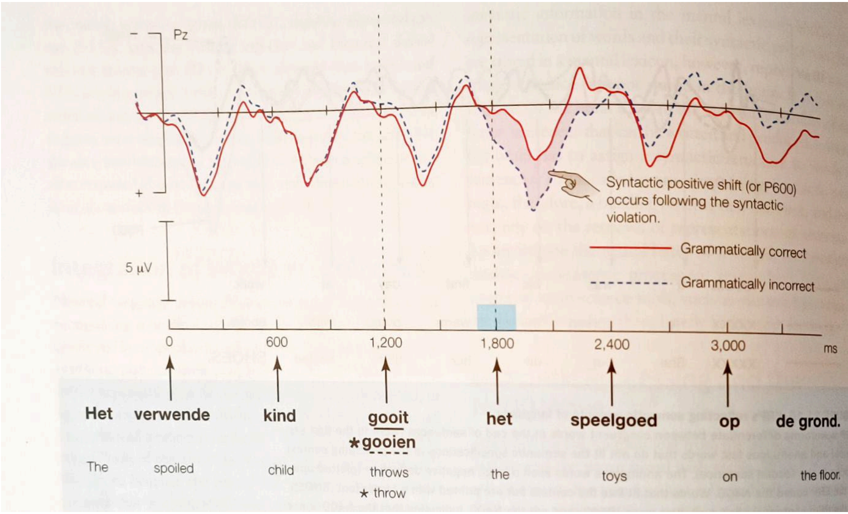
Example 2:
“The man is a vegetarian. He went to the restaurant and ate a steak that was prepared well.” -> N400

Cognitive neuroscience, Gazzaniga, Ivry, Mangun, 2014

Language Comprehension: Syntactic processing and the P600 wave

- P600 also called syntactic positive shift

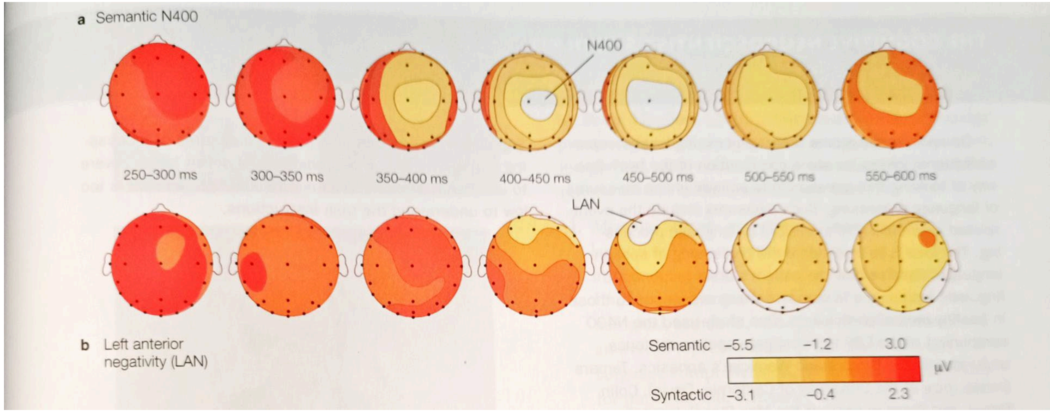
- 600 ms after syntactic violation



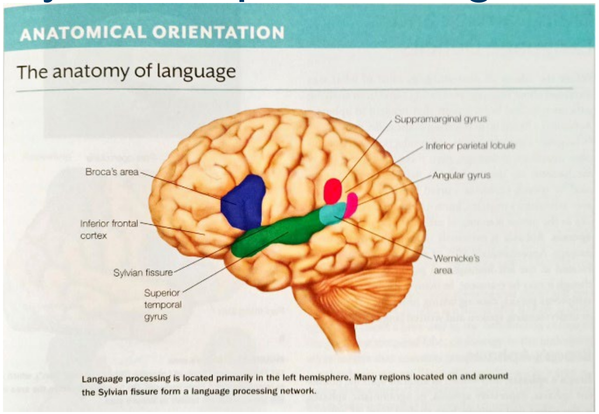
Cognitive neuroscience, Gazzaniga, Ivry, Mangun, 2014

Language Comprehension: Syntactic processing

- Besides P600 also other brain waves are related to syntactic violations
- **LAN (at 400 ms): left anterior negativity**, (Thomas Münte et al., Angela Friederici et al.)
- Broca's area and inferior frontal cortex involved in syntactic processing



Cognitive neuroscience, Gazzaniga, Ivry, Mangun, 2014



Cognitive neuroscience, Gazzaniga, Ivry, Mangun, 2014

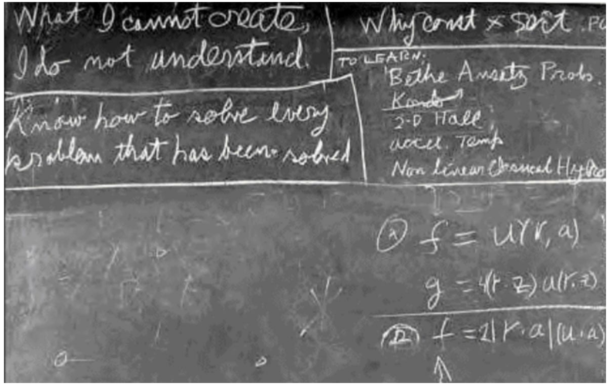
Conclusion of Language Comprehension

Language Comprehension: Neural Models: Conclusion

- **Models of language comprehension involve unifying from linguistic inputs with stored knowledge**
- White matter tract connection (e.g. arcuate fasciculus) of frontal and temporal lobes play an important role in speech processing

-> We know where things happen but not what happens!

- **First lecture**: David Marr (Tri-level hypothesis), Do we understand the algorithms related to language comprehension -> **No!** (my personal view)



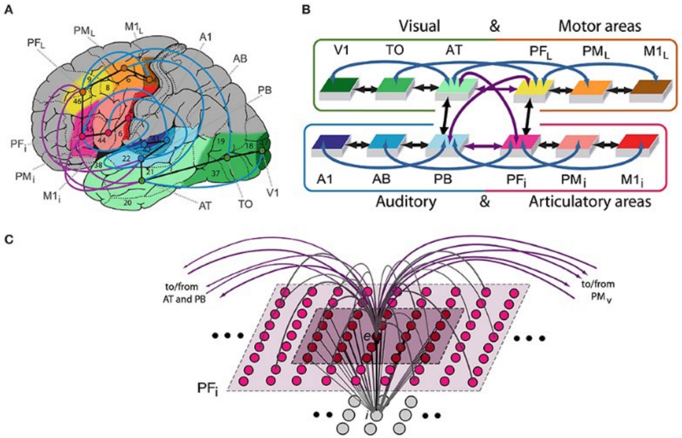
Dhar, A., Patel, A. D., & Wadia, S. R. (2018). The science and legacy of Richard Phillips Feynman. *arXiv preprint arXiv:1810.07409*.

Language Comprehension: Neural Models: First biologically plausible simulations exist

- Group of **Friedemann Pulvermüller** develops biologically plausible simulations to understand language processing

- Example: spiking neuron models trained with (Hebbian Learning, Lateral Inhibition)

- Found networks that process semantic information, could compare the brain activity in their network to experimental findings



Tomasello, R., Garagnani, M., Wennekers, T., & Pulvermüller, F. (2018). A neurobiologically constrained cortex model of semantic grounding with spiking neurons and brain-like connectivity. *Frontiers in Computational Neuroscience*, *12*, 88.

Language Production

Speech production: Neural Models

Lemma: Words' syntactic and semantic properties

Phonological encoding: get phonological code of word

Model by Willem Levelt (1989)

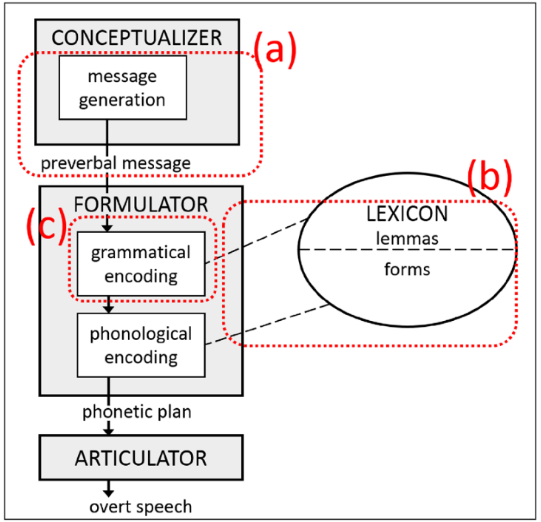
1) **Message Preparation (Conceptualizer)**

- **Macroplanning**: **What** to say (content)
- **Microplanning**: **How** to say ("The park is next to the house." "The house is next to the park.")

2) **Formulator**: Output of micro-and macroplanning is a conceptual message -> is given to a **formulator -> puts** message in a phonologically and syntactically correct form

3) **Articulator**: word syllables are mapped to motor patterns that move our tongue, mouth and vocal apparatus

Levelt model is a complete serial (box and arrow) model -> no parallel processing -> no feedback…

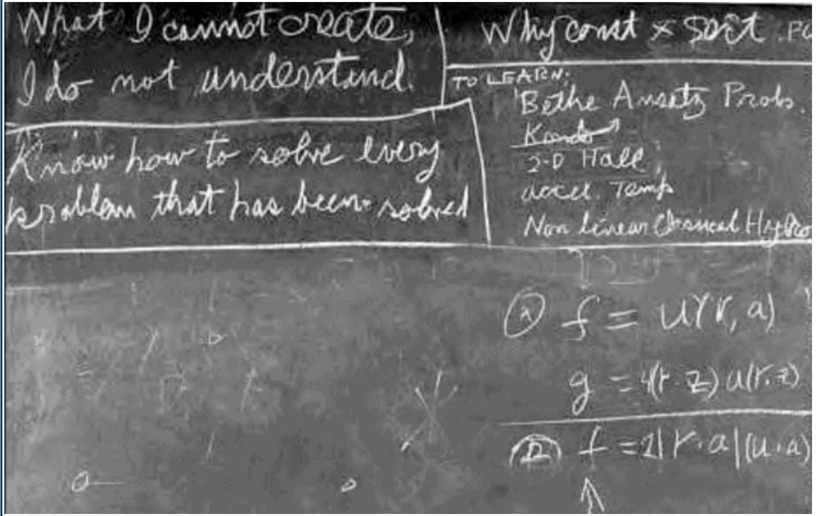


Nakaya, K., & Murota, M. (2016). Development of a Mobile Application for Listening and Scaffolded Autonomous Summary Speaking and Its Effectiveness for Increasing Fluency. *The Journal of Information and Systems in Education*, *15*(1), 48-61.

https://en.wikipedia.org/wiki/Transformers:_Rise_of_the_Beasts#/media/File:Transformers-_Rise_of_the_Beasts.jpg

Fundamental Recent Developments

Fundamental Considerations in the light of recent Developments



Dhar, A., Patel, A. D., & Wadia, S. R. (2018). The science and legacy of Richard Phillips Feynman. *arXiv preprint arXiv:1810.07409*.

- Does this mean what we can create we do understand? -> **No!**

- **In the light of novel language processing algorithms:**
  - Is the ability to process language enough to develop a general intelligence?

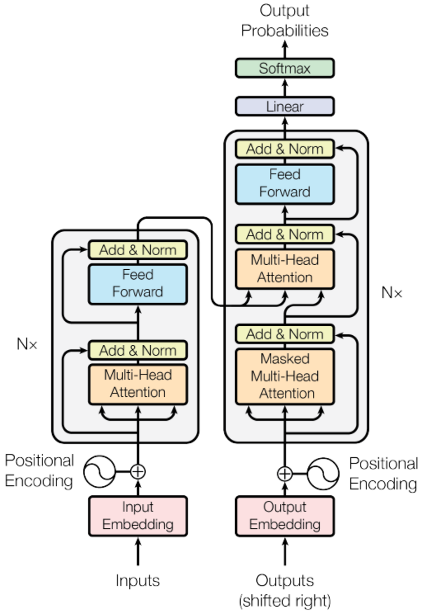
  - Are the underlying algorithms behind Large Language Models (LLM) such as ChatGPT similar to the algorithms implemented in the brain?

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG
FACULTY OF ENGINEERING

Fundamental Considerations in the light of recent Developments

- Game Changer in computational linguistics (synonym: natural language processing): **Transformer Networks**

- Main advantage: **No recurrent** neural network -> training can be parallelized
- **Main principles**:
  - **Positional encoding**: vector is added containing the information of word position in sentence
  - **Multi-Head Attention Mechanisms (self-attention)**: Networks learn the influence of the meaning of certain words on other words
  - **Prediction**: Trained on predicting the next most probable word



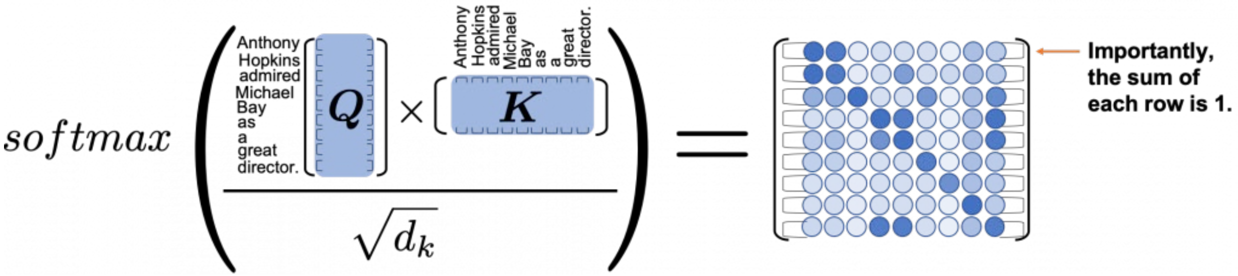
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Unfortunately: Exact architecture of e.g. ChatGPT is unknown!

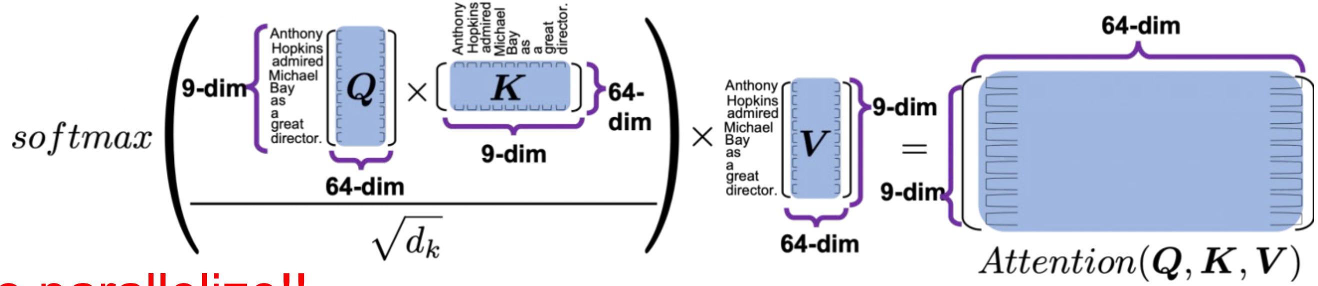
Fundamental Considerations in the light of recent Developments

- **Self Attention-Mechanisms**: Simple Matrix Multiplication to add context to each word -> no need of recurrences

Matrix that contains information of how each word relates to the meaning of each other word



Scale each word of the sentence with the context matrix



Easy to parallelize!!

https://data-science-blog.com/blog/2021/04/07/multi-head-attention-mechanism/

Fundamental Considerations in the light of recent Developments

- **Bubeck and Coworkers state (2023)**: GPT4 is an “early (yet still incomplete) version of an artificial **general intelligence (AGI)** system”

- **AGI system can do**: ”reasoning, planning, learn from experience”

- **How can intelligence of GPT4 be tested?**
  - Benchmark test are difficult – as training data is unknown GPT4 has potentially already seen everything
  - Difficult to use these tests to measure generality
  - -> thus authors (Bubeck et al.) test for e.g. creativity (psychology approach)

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Fundamental Considerations in the light of recent Developments

Bubeck and Coworkers state: GPT4 is an "early (yet still incomplete) version of an artificial general intelligence (AGI) system"

1) GPT4 was asked to write a **proof that there is an infinite number of prime numbers.** -> no problem for GPT4 -> to check if GPT4 memorizes just the text from training data -> GPT4 was asked to write it in Shakespeare style -> no problem for GPT4
2) **Draw a unicorn in TiKZ** (programming language to generate vector graphs)-> no problem needs visual imagination and coding skills

Figure 1.3: We queried GPT-4 three times, at roughly equal time intervals over the span of a month while the system was being refined, with the prompt "Draw a unicorn in TikZ". We can see a clear evolution in the sophistication of GPT-4's drawings.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Fundamental Considerations in the light of recent Developments

Bubeck and Coworkers state: GPT4 is an “early (yet still incomplete) version of an artificial general intelligence (AGI) system
3) Test on some understanding of the world

a hunter walks one mile south, one mile east, and one mile north and ends up right back where he started. He sees a bear and shoots it. What color is the bear?

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712.*

Fundamental Considerations in the light of recent Developments

Bubeck and Coworkers state: GPT4 is an “early (yet still incomplete) version of an artificial general intelligence (AGI) system
3) Test on some understanding of the world

a hunter walks one mile south, one mile east, and one mile north and ends up right back where he started. He sees a bear and shoots it. What color is the bear?

GPT-4 vs ChatGPT

Prompt: Puzzle: a hunter walks one mile south, one mile east, and one mile north and ends up right back where he started. He sees a bear and shoots it. What color is the bear?

GPT-4:

There are different ways to approach this puzzle, but one possible method is to use some basic facts about geography and geometry.

- First, we can rule out any place on the equator, because walking one mile east or west there would not bring the hunter back to the same point. Likewise, we can rule out any place too far from the poles, because the distance between meridians of longitude would be too large to complete a one-mile circle.
- Second, we can focus on the places where walking one mile south and then one mile north would bring the hunter back to the same point. This would happen at the poles themselves, or at any point that is exactly one mile away from either pole. In the latter case, the hunter would be on a circle of latitude that has a circumference of exactly two miles, so walking one mile east or west would make him traverse half of the circle and return to the same point.
- Third, we can consider which pole or poles are possible candidates for the puzzle. The South Pole is not a good option, because there are no bears there, and the hunter would have to cross the Antarctic continent to reach it. The North Pole is a better option, because there are polar bears there, and the hunter could reach it by sea or ice.
- Therefore, the answer is that the bear is white, because it is a polar bear. The hunter could be at the North Pole itself, or at any point that is one mile away from the North Pole.

ChatGPT:

There is no information in the puzzle about the color of the bear. The only information given is that the hunter walks one mile south, one mile east, and one mile north and ends up back where he started, and that he sees a bear and shoots it. The puzzle does not specify the color of the bear.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Fundamental Considerations in the light of recent Developments

Bubeck and Coworkers state: GPT4 is an "early (yet still incomplete) version of an artificial general intelligence (AGI) system

3) Test on some understanding of the world -> check if GPT4 just remembered the puzzle

GPT-4 vs ChatGPT

Prompt: I fly a plane leaving my campsite, heading straight east for precisely 24,901 miles, and find myself back at the camp. I come upon seeing a tiger in my tent eating my food! What species is the tiger?

GPT-4:

```
There are a few possible ways to approach this question, but one plausible method is to use the following assumptions and facts:

- The plane flies at a constant altitude and speed, and does not deviate from its eastward course or encounter any obstacles or detours.
- The circumference of the Earth at the equator is approximately 24,901 miles, and the plane flies along the equator or close to it.
- The tiger is a wild animal, not a pet or an escapee from a zoo or a circus.
- The tiger is a member of the genus Panthera, which includes four extant species of large cats: the lion, the tiger, the leopard, and the jaguar.

Given these assumptions and facts, we can narrow down the possible species of the tiger to two: the Bengal tiger (Panthera tigris tigris) and the Sumatran tiger (Panthera tigris sumatrae). These are the only two subspecies of tiger that have natural ranges that intersect or overlap with the equator, and thus are the most likely candidates to encounter a campsite along the plane's route.
```

ChatGPT:

```
It is impossible for me to determine the species of the tiger in your scenario. The species of a tiger is generally determined by its physical characteristics and geographic location, and I do not have enough information to accurately identify the species of the tiger in your story. There are several different species of tigers, including the Bengal tiger, the Siberian tiger, and the Sumatran tiger, among others. Each species has its own unique physical characteristics and is found in different parts of the world.
```

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Fundamental Considerations in the light of recent Developments

4) **Theory of mind** (remember lecture on brain lateralization)

GPT-4 and ChatGPT: Realistic Scenario 1

```
Conversation:
Mark: I don't like how you treated Jack last night
Judy: Did you see what he was doing? He hit his brother on the head!
Mark: That doesn't justify you yelling at him
Judy: What do you want me to do? Let him beat his brother up and say nothing?
Mark: No, I don't
Judy: Then why are you defending him?
--
Q: What does Mark's intent seem to be?
```

GPT-4:

Mark's intent seems to be to express his displeasure with Judy's way of handling the situation with Jack, and to suggest that she could have been more calm and respectful.

ChatGPT:

Based on the conversation, Mark's intent seems to be to defend Jack's behavior and to express his disapproval of Judy's treatment of Jack.

Q: How did Judy receive his comment?

GPT-4:

Judy received his comment as an attack on her par enting skills and a sign of sympathy for Jack's misbehavior. She reacted defensively and angrily, and questioned Mark's motives.

ChatGPT:

It is not possible for me to accurately determine how Judy received Mark's comment without more con text. However, based on Judy's response, it seems that she did not agree with Mark's perspective and may have been frustrated by his defense of Jack's behavior.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Fundamental Considerations in the light of recent Developments: Summary

- The skill of **language processing might be enough to develop general intelligence**

- Potentially we do **not** need brain like architecture to reach general intelligence

- **LLMs will have a huge influence on society**
  - Jobs and economics
  - **Danger:**
    - LLMs are biased due to bias in training data
    - System does not know when it is just guessing and when it is knowing

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.