LARGE-LANGUAGE-MODEL TOOLS AND THE THEORY OF *LEXICAL PRIMING*:

CONVERGENCE AND DIVERGENCE OF CONCEPTS OF LANGUAGE

Michael Pace-Sigge

University of Eastern Finland

OVERVIEW

1. Introduction
2. What is *LEXICAL PRIMING?*
3. Large Language Models
4. Linking the two
5. Bonding, cohesion and story grammar
6. Where natural language & CL confronts a GPT-produced Text
7. Chat GPT and structures
8. Communicative Intent and the idea of *MEANING MAKING*
9. Conclusions
10. References

Were I to ask an AI chatbot to write the opening of this paper for me, then I surely would be offered two different starting points.

I could either begin with the here and now and reflect on the sudden excitement and anxiety that seems to have arisen ever since companies like OpenAI offered highly-developed AI tools that create texts, images and videos within an incredibly short time-frame after receiving a prompt.

Alternatively, I could go back twenty years ago and look at was then called "a new theory of words and language" – the *Lexical Priming Theory (LPT)* presented by Michael Hoey (2005) and use a language studies rather than a computational linguistics approach as my starting point.

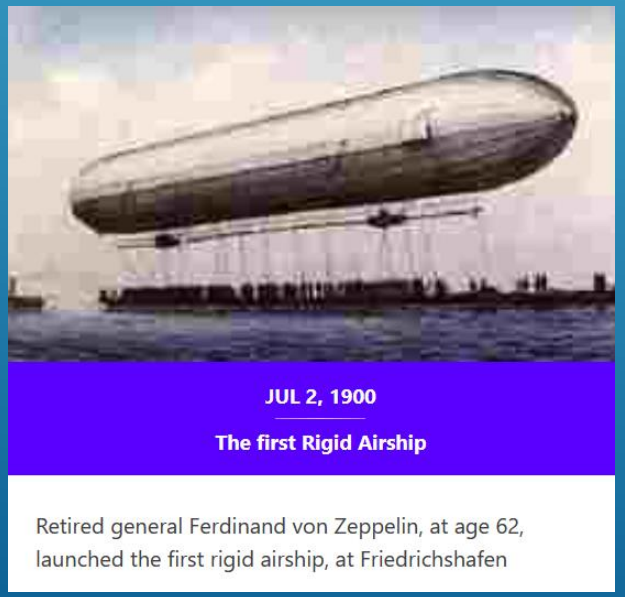
Much as I would love to be a quantum information processor I doubt I can do both things at once.

1. INTRODUCTION

- This is a position paper, with the aim to look at Large-Language Model based tools and how they can be linked to linguistic theory
- In particular, tools like ChatGPT and GEMINI are employed to replicate research undertaken by Hoey (2005) to underpin his Lexical Priming Theory (LPT)
- This paper looks whether a 'virtual' test might be useful to confirm the premises of the LPT
- This paper contrasts naturally occurring texts with LLM-produced output, with the specific focus in how far the design of LLM tools aligns and diverges from the Lexical Priming Theory.

2. What is **LEXICAL PRIMING**?

- **PRIMING** occurs when a listener or reader comes across a certain word sequence and construction with a frequency higher than random co-occurrence.
- Repeat exposure anchors the particular usage in a listeners' mind and eventually is reflected in their own production
- As a result, a single word can then act as a 'prime' which leads to the **activation** of what is expected to come
- As an example, in experiments in the 1960s and 1970s, it was shown that the term 'nurse' would lead to a far quicker recognition to 'doctor' than the unrelated word 'bread'
- Hoey concludes that there is a primacy of **lexis** over **grammar** which tends to be governed by the former and is individual, not universal (≠ Chomsky)
- This approach was first proposed by Ross M. Quillian in the 1960s, when he proposed 'a model of language' that can serve a machine he dubbed the Teachable Language Comprehender, which would be trained on naturally occurring texts (books) and which could disambiguate meanings of words based on the context the words are found in (see also Quillian, 1967).



- Consider it is the year **1900…**
- Yes, that is right: 1900 is the year where we will have to start when we talk about *machine learning* or *artificial intelligence* (A.I.)
- The German mathematician **David Hilbert** posed the question whether there exists an algorithm for deciding the truth of any logical proposition involving natural numbers.
- This became known, in 1928, as the **Entscheidungsproblem**
- This became some sort of philosophers' stone, where we find people like Gödel or Wittgenstein making contributions.
- **Alan Turing** said that, in logic, there are some true statements that cannot be decided by any algorithm. Thus, Turing sets out to describe a system that provides computability of "to the intuitive idea of 'effective calculability'": providing the space where calculations are possible.
- That was in 1937

3 LARGE LANGUAGE MODELS

- McCulloch and Pitts (1943) are early proponents of "neural networks"
- In **1950**, Alan Turing publishes a paper which then became widely known as the **Turing Test**
  - I propose to consider the question, 'Can machines think?' This should begin with the definitions of the meaning of the terms 'machine' and 'think'. The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous.
- It is in **1952** that a machine that is recognised as the earliest form of artificial intelligence machine is presented by Marvin Minsky—the Stochastic Neural Analog Reinforcement Computer **(SNARC).**
- **1980s: IBM** offers voice and text-recognition software. Research into **Story Analysis** programs
- **Early 2000s:** *Recurrent Neural Networks* (RNN) and *Long Short-term Memory* (LSTM)
- **2017:** when **Google** presents their *Bidirectional Encoder Representations from Transformers* (BERT)
- **November 2022: OpenAI** makes their GPT 3.5 widely available with a chatbot interface: **ChatGPT**

(**G**enerative **P**re-trained **T**ransformer)

1952: Dwight Eisenhower (Ike) is elected US President

3 LARGE LANGUAGE MODELS

LLMs are token predictors

this is how it can be visualised

How many commas: ,,,,,,,,,,? There are 10 commas in your message.

[5299, 1991, 179663, 25,1366,82384, 82384, 30] ➡ [5632]

[5299, 1991, 179663, 25,1366,82384, 82384, 30, 5632] ➡ [553]

[5299, 1991, 179663, 25,1366,82384, 82384, 30, 5632, 553] ➡ [220]

[5299, 1991, 179663, 25,1366,82384, 82384, 30, 5632, 553, 220] ➡ [702]

…

[5299, 1991, 179663, 25,1366,82384, 82384, 30, 5632, 553, 220, 702, 179663, 306, 634, 3176] ➡ [13]

[5299, 1991, 179663, 25,1366,82384, 82384, 30, 5632, 553, 220, 702, 179663, 306, 634, 3176, 13]

LLMs are token predictors

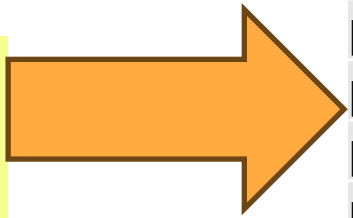
How many commas: ,,,,,,,,,,? There are

[5299, 1991, 179663, 25,1366,82384, 82384, 30] ➡ [5632]

[5299, 1991, 179663, 25,1366,82384, 82384, 30, 5632] ➡ [553]

[5299, 1991, 179663, 25,1366,82384, 82384, 30, 5632, 553] ➡ [220]

There are NINE commas in the above question
So, how come the LLM hallucinates?
The answer is vector based, as the most-likely answer is chosen, not the correct one.



[220]	**' '**	0.7298
[4325]	**' ten'**	0.2685
[66029]	**' eleven'**	0.0009
[19712]	**' nine'**	0.0008
[261]	**' a'**	0.0001

Classical machine learning systems vs humans?

Thinking fast (system 1)

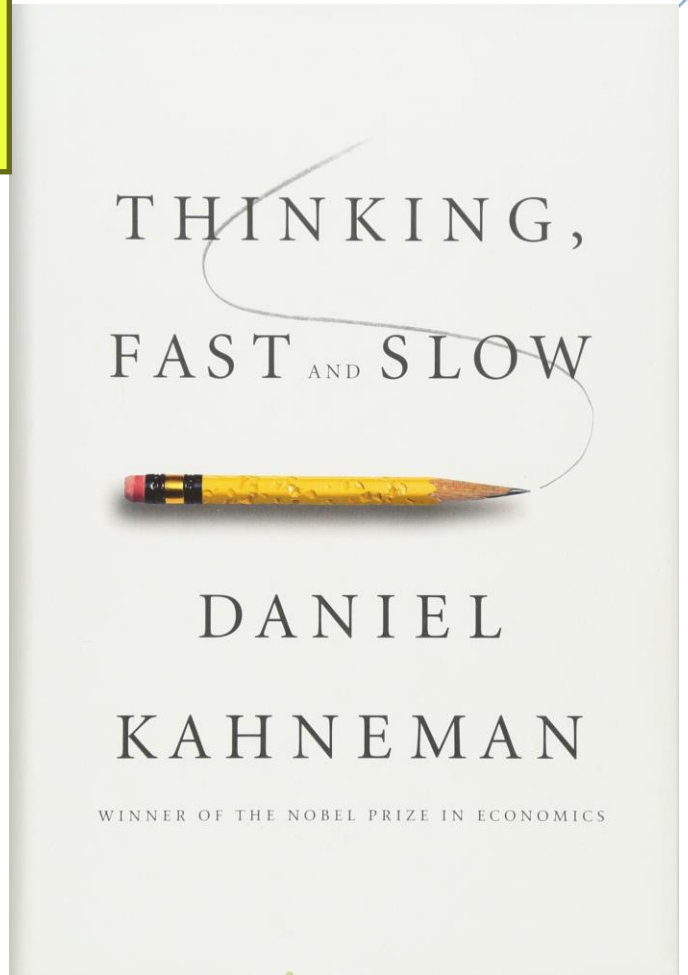
- Automatic, intuitive, and effortless.
- Examples:
  - Recognizing faces
  - Reading road signs
  - Detecting anger in someone's voice

This has echoes of *priming*

Thinking slow (system 2)

- Slow, deliberate, analytical, effortful
- Examples:
  - Complex calculations
  - Solving puzzles
  - Making thoughtful decisions

Classical ML systems (without transformers and diffusion) are typically System 1.



3. LARGE LANGUAGE MODELS (LLMs)

- ChatGPT and BARD/Gemini are trained on a large variety of texts
- The input of billions or even trillions of words become a matrix in which each item is assigned a vector representation in relation to other words
- Mikolov et al (2013) describe how computer models mimic relationships between words in a way not dissimilar to the relationships human language users see (e.g., grammatical ones like tall-taller; semantic ones like train-travel, etc.)
- This computational process is, crucially, not pre-programmed: the fact that so much training material is available allows the algorithm to find 'natural' relationships between words
- In **corpus linguistics** this is mirrored by **collocation** (the concept that a word is more frequently found in the vicinity of another word than mere coincidence would allow for) and **colligation** (the idea that lexemes fit into particular categories and their pre- and suffixes and occurrence patterns are limited to the grammatical patterns in which they are typically found).

4. LINKAGE **LLMs** and **LPT**

1. The basic idea of a large language model (LLM) is that you enter a **'prompt'**, like a about colour blindness will bring about the 'completion' chart. A straightforward way of building a 'thinking' version of a language model is to make the system prompt ensure that the user's prompt generates a chain of thought which can then be used to generate a final answer.
2. You can think of it as the word's co-ordinates in a many-dimensional space. The names of countries might be close together in one part of the space, words for food concentrated in another. A 2013 paper from Google showed that starting from the co-ordinates they generated for 'Japan' and then moving to those for 'sushi' and then repeating the same steps but starting from 'Germany', took you to 'bratwurst'.
3. The parameters of the program are then updated to increase the estimated probability associated with the word that was removed, and lower those associated with the alternatives.
4. A large language model is created, in essence, by training a program to predict missing words. Imagine removing a word from a sentence and feeding that sentence into a program that generates, for all the many words in its database, an estimate of the probability of each of them being the missing word.
5. With sufficient training, the program will be able, given a string of words, to select the most appropriate next word, add it to the string and select another word to follow it. It is utterly astonishing that such a simple process can create long sequences of text that are not only intelligible but seem to be the product of intelligence.

(Paul Taylor: *AI Wars)*

1. A word acts as a prime which then sets off what Quillian refers to as "spreading activation": the most likely follow-on.
2. In CL, these would be referred to as *collocates*
3. Each word has both a preference and a dispreference what it collocates with
4. "The language user as having a mental concordance and of the possibility that they process this concordance in ways not unrelated (though much superior) to those used in corpus linguistic work he language user as having a mental concordance and of the possibility that they process this concordance in ways not unrelated (though much superior) to those used in corpus linguistic work" (Hoey, 2005:13)
5. Training equals "repeat exposure" – a language user will come across a word in a number of uses (their *nesting)* and eventually employ these productively themselves.

4. LINKAGE **LLMs** and **LPT**

- As an explanation of what corpus linguistic insights present, the Lexical Priming theory builds on Quillian's work and postulates that '[e]very word is primed for use in discourse as a result of the cumulative effects of an individual's encounters with the word' (Hoey, 2005, 13)
- This appears to be remarkably similar when compared to the ways in which large language models (LLMs) are built
- Both the algorithm & the LPT speak of 'natural' relationships between words that are based on frequencies and nestings
- The 'cumulative effects' are the training data, where each item ('word') will appear with differing frequencies in large corpora of training data
- The 'use in discourse' is how this item stands in relation to the other items: whether they, for example, cluster together with a mostly, often or only rarely (semi-)fixed set of items; whether they appear together or whether co-appearance is a rare occasion etc.
- In other words, the key issue is the relative statistical likelihood in which words co-occur (see Manin and Marcolli, 2016).

5. BONDING, COHESION AND STORY GRAMMAR

- Early LLMs had sophisticated vector-based statistics which predicted the next word(s)
- These systems fell short when it came to produce longer texts
- One improvement is *story grammar*:

Story Analysis programs are interesting because they can provide an insight into the structures of stories, the notion of coherence, and the interaction between events, goals and plans (Norvig, 1992,1).

In the context of ChatGPT, understanding story grammar is crucial for generating coherent and contextually relevant responses. While GPT models like ChatGPT don't explicitly possess a deep understanding of narratives, they learn patterns from vast datasets, which may include elements of story grammar (ChatGPT, 2024).

- The move from *Recurrent Neural Networks* (RNN) and *Long Short-term Memory* (LSTM) was achieved after 2017 when Google presented their *Bidirectional Encoder Representations from Transformers* (BERT), followed by OpenAI's *Generative Pre-trained Transformer* (GPT) 1

1) **Coherence**

2) **Context Maintenance**

3) **Character Consistency**

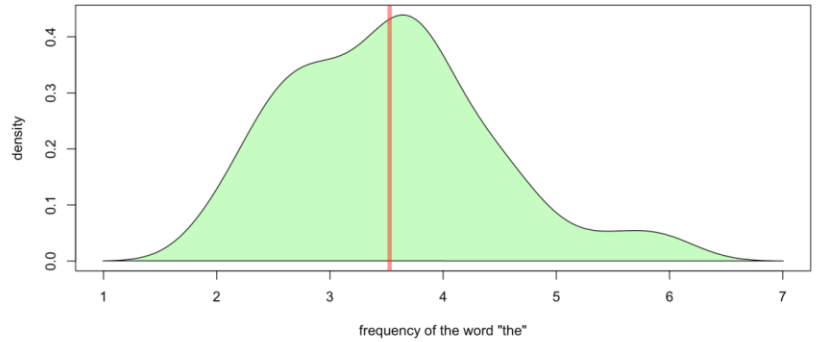
4) **Event Sequencing**

KEY ELEMENTS OF *STORY GRAMMAR*

5. BONDING, COHESION AND STORY GRAMMAR

- This is very similar to the concept of *BONDING* presented by Hoey in the 1990s
- Based on the Hasan & Halliday concepts of *coherence* and *cohesion* in texts, *bonding* exists in both narrative and non-narrative texts
- "… only a small pro-portion of the bonds formed in the passage given are between adjacent sentences" (Hoey, 1991, 149).
- Bonding shows that a single narrative, or set of similar narratives by one or several authors, share cohesive links which are specific to them. Furthermore, these words are nested within the wider context and co-text
- "we retain access to the con-texts of words previously encountered, or else each new encounter with a word of whose meaning we were uncertain would be a fresh problem" (Hoey, 1991, 155).

arithmetic mean



what is a word frequency, really?

- frequency of "the" in *Pride and Prejudice* is 3.52%
- however, if we devide the text into 122 chunks of 1,000 words:
  - 2.3, 4.2, 4.4, 2.9, 3.3, 2.9, 2.6, 4, 3.5, 4.5, 2.2, 3.1, 2.8, 4, 3.3, 3.9, 2.8, 4.1, 3.3, 3.6, 3.4, 5.3, 6.1, 3.8, 3.2, 3.6, 3.8, 4.3, 2.9, 3.6, 3.7, 3.9, 4.2, 3.8, 2.8, 3.2, 4.5, 3.8, 3.9, 3.6, 3.5, 3.5, 2.4, 3.5, 3, 2.7, 1.9, 3.8, 4.8, 4.9, 4.7, 3.2, 5.8, 4.6, 2.5, 3.9, 4, 2.8, 3.4, 3.4, 6, 4.2, 3.5, 2.9, 5.4, 3.5, 3.1, 3.7, 3.5, 4.8, 2.7, 3.8, 4, 2.9, 4.3, 5.8, 3.8, 5, 5.6, 3.6, 3.8, 3.7, 4.4, 4.4, 2.5, 2.6, 2,1 etc.
- max value: 6.1%
- min value: 1.9%
- what is the frequency of "the", after all???

Maciej Eder, DHNB Plenary Talk, Tartu, March 2025

6. WHERE NATURAL LANGUAGE & CL CONFRONTS A GPT-PRODUCED TEXT

My research, using a nanoGPT trained on a corpus of spoken Scouse, with material from a variety of sources showed that:

- it was able to highlight distribution qualities in the training corpus.
- Thus, utterance-initial features can become identifiable.
- Likewise, longer phrases which occur throughout the training data will re-appear in generated text.
- Generated text will not, or to a significantly lesser degree, show phrases which tend to be prominent only in a subsection of the training corpus.

Pace-Sigge, IJCL 2025

6. WHERE NATURAL LANGUAGE & CL CONFRONTS A GPT-PRODUCED TEXT

6. CHAT-GPT (etc.) AND STRUCTURES 1

- One key skill a corpus linguist acquires is the ability to identify frequent patterns and structures.
- Hoey (2005) posits that structures are 'primed for semantic association'.
- In the field of *journey* he identifies the sequence 'NUMBER-hour-JOURNEY (or NUMBER-TIME-JOURNEY)' (p.17).

6. CHAT-GPT (etc.) AND STRUCTURES 1

- Chat-GPT and Bard (quoted below) concur:
- "NUMBER-TIME-JOURNEY is commonly used in English. It is used to convey a sense of a journey that takes place over a specific period of time. This structure can be used to describe personal experiences, historical events, or fictional stories".
- Thus, LLMs can be used as a basic analysis tool even if a researcher does not have data to hand
- Checking Hoey's research results (which he used to demonstrate the validity of LPT) against answers from LLM tools can be seen as a demonstration that these can be used to test examples that Hoey gave in support of his lexical priming theory
- It is not sufficient in itself, however, to show that the design of present-day GPT tools aligns with the premises set out by the theory.

1. FictFan15.xml all over again: as harem captives of a cruel Eastern **potentate**; passengers sharing a lifeboat with hungry sailors,

2) FictWom124.xml Ellie was appalled. 'He's carrying on like an old-fashioned **potentate** with his harem.'

3) FictHor14.xml Bosphorus doomed to be killed or enslaved by the Muslim **potentate**. But I hope our fascination is not purely academic.

4) FictMis252.xml a ramshackle collection of warlords nominally controlled by the Cambrian **potentate** Tharv the Bountiful. The empire has almost no economy or

5) FictMis469.xml This was the act of a conqueror, a **potentate**! 'The people, for his airs, they call him Jrundur

6) NewMaDam3029.xml working classes while living the opulent existence of a medieval **potentate**.

7) AcaHumRa257.xml decision to summon the people of the hundred to the **potentate's** gates.

Concordance 1

All occurrences of *potentate* in BNC-2014

HOW are structures represented in a generated text? A test case built on the usage pattern of one rare word

1.1	The **potentate**, adorned in regal attire, addressed the assembled crowd with eloquence and authority.
1.2	As a **potentate** of the realm, his decisions held immense influence over the kingdom's affairs.
1.3	The courtiers awaited the **potentate's** pronouncement, knowing it would shape the destiny of the empire.
1.4	The diplomatic envoy sought an audience with the **potentate** to discuss matters of international importance.
1.5	The citizens celebrated the anniversary of their **potentate's** reign, grateful for the prosperity he had brought to the land.
1.6	The palace walls echoed with the whispers of intrigue as factions vied for favor with the **potentate.**
1.7	Despite being a **potentate**, known for strength and decisiveness, he ruled with benevolence, earning the admiration of his subjects.
2.1	The ancient **potentate** ruled his vast empire with an iron fist.
2.2	The media mogul was an influential **potentate** who wielded power behind the scenes.
2.3	The young **potentate** was eager to prove his mettle and establish his reign.
2.4	The potentate's lavish lifestyle was the envy of the world, but his subjects suffered in poverty.
2.5	The downfall of the **potentate** was swift and unexpected, leaving his empire in turmoil.
2.6	The **potentate's** legacy is a mixed bag of achievements and atrocities.
2.7	The stories of the **potentate's** reign have become legendary, captivating audiences for centuries.

Figure 1

Potentate generated sample sentences: 1.1-1.7: *Chat-GPT*, 2.1-2.7: *BARD*

4. CHAT-GPT (etc.) AND STRUCTURES 2

- *POTENTATE* collocates twice with *harem* while there is also a direct connection to *conqueror/warlord* as well as *Muslim/Eastern* in 2 each of 7 cases
- It is indicative that *POTENTATE* is, in five cases, pre-modified in the human-written material
- This is, however, not the case in the Chat-GPT produced samples (BARD: 2x)
- On the one hand, the LLM-produced samples are quite similar to the use of this very rare term found in the BNC-2014: Concordance 6 and example 2.4 seem to describe something quite similar
- On the other hand, the word *POTENTATE* is quite specific. While in Concordance 1, 'king' might be used in place of 'potentate', this does not work very well in all lines – for example, *Muslim* collocates with 'potentate' rather than 'king'.

6. CHAT-GPT AND STRUCTURES 3

train as pattern	BNC-2014 news	Chat-GPT	BARD
TRAIN (n)	11	0	n/a
TRAIN (v) phrase	5	0	n/a
TRAIN (v) profession	11	14	n/a
TRAIN (v) non-prof.	1	14	n/a
total	28	28	n/a

- A further exemplification can be found when looking at Hoey's (2005, 64ff.) findings with regards to priming and co-hyponymy, where he states that "[t]rain is primed to collocate with ***as a*** in newspaper data and the nested combination of **train* as a** is typically primed to associate with SKILLED ROLE OR OCCUCATION"
- As can be seen, Chat-GPT seems to be rather hyper-focussed on particular frames (BARD/Gemini does not like it).

CHAT-GPT (etc.) AND STRUCTURES 4

- A final comparison is based on Hoey's (2005) sampling of "hypernyms of SKILLED ROLE OR OCCUPATION, namely *architect*, *accountant*, *actor* or *carpenter*".

Hoey 2005 *Grammatical construction*	***ARCHITECT*** per 100	***ACCOUNTANT*** per 100	***ACTOR*** per 100	***CARPENTER*** per 100
Indefinite Article	**16**	26	22	**42**
Parenthesis	13	17	**8**	**26**
Apposition	18	14	21	**2**
'Possessor' ('s or of NP)	8	**6**	8	**16**
'Possessed' construction	5	**10**	**1**	2
Metaphor	**23**	**0**	5	1

Each of these professions or occupations appears to have their own characteristic grammatical imbedding (nesting) with clear preferences /dispreferences

CHAT-GPT (etc.) AND STRUCTURES 4

Hoey 2005 *Grammatical construction*	***ARCHITECT*** per 100	***ACCOUNTANT*** per 100	***ACTOR*** per 100	***CARPENTER*** per 100
Indefinite Article	**16**	26	22	**42**
Parenthesis	13	17	**8**	**26**
Apposition	18	14	21	**2**
'Possessor' ('s or of NP)	8	**6**	8	**16**
'Possessed' construction	5	**10**	**1**	2
Metaphor	**23**	**0**	5	1

- Looking at the ***BNC 2014,*** Hoey's findings are confirmed. If anything, the larger, newer dataset shows that the tendencies highlighted are even more strongly found (as in **chartered accountant**).

As for the LLM tools:

- Bar one exception, all sample sentences start with the requested noun (phrase) - a feature hardly ever occurring in the natural data
- BARD pre-modified each target word: The struggling actor…'; 'The meticulous architect…'; The reliable accountant…'; 'The meticulous carpenter…'.
- In half the samples in Chat-GPT, the initial noun phrase is followed directly by a verb phrase (either verb or verb+adverb) 'The accountant meticulously reviewed the company's financial records…'
- Generic, with the inclusion of collocates from the same semantic word field
- Modifiers (adverbs) also appear to be generic rather than specific to any particular profession or trade: *diligently, meticulously* and *tirelessly* favoured by both LLM tools
  - Yet their selection appears to be random.

7.1 COMMUNICATIVE INTENT AND THE IDEA OF *MEANING MAKING*

- **Meaning** and specific uses exists only at the level of co-text
- Manning points out how this can lead to problems: "[m]eaning is not all or nothing; in many circumstances, we partially appreciate the meaning of a linguistic form. I suggest that meaning arises from understanding the network of connections between linguistic form and other things, whether they be objects in the world or other linguistic forms" (Manning, 2022, 134).
- Similar sentiments are shown by Bisk et al. (2020) or Merrill and colleagues (2021) who speak of the "limitations of acquiring meaning from an ungrounded form', whereby 'it is (real-world) experience which actually grounds language."
- This appears to be in full agreement with Hoey, who claims that the primings of a (set of) word(s) is lodged with each individual user: "Firth's notion of 'personal collocations' (1951) [as] it is an inherent quality of lexical priming that it is personal in the first place and can be modified by the language user's own chosen behaviour" (Hoey, 2005, 10).

7.2 COMMUNICATIVE INTENT AND THE IDEA OF *MEANING MAKING*

- Meaning-making as such goes beyond the purely linguistic expression and includes extra-lingual events (which are particular salient in spoken language) but also personal associations.
- Hoey refer to these as one's *personal grammar*
- This aligns with Paul Hopper's notion of an ***emergent grammar*** (1998).
- Manning highlights that current AI tools rely fully on text data, yet, in order to improve their 'understanding' skills, these would need to be augmented with 'further sensory data': namely, visuals.
- Bisk et al (2020) say that one must **"consider the contextual foundations of language: grounding, embodiment and social interaction"**. Many of the assumptions and under-standings on which communication relies lie outside of text.
- Bender et al. (2021) highlight, LLMs use data which is fixed in time: in other words, static, whereas **"social movements produce new norms, language and ways of communicating thus, LLMs risk to reify older, value-locked, or indeed biased understandings"**.
- This links neatly with Hoey's assertion that 'grammars are never complete' (2005, p. 162).
- Bender and Koller (2020) or Hadfield (2022) argue that LLMs lack a basic element of language, namely communicative intent
- Hadfield (2022) looks at research into child language acquisition and highlights that **"[i]nfants learn language by drawing on a wide range of cues, while LMs only train on the tiny slice of the world in their input texts"**.

8. CONCLUSIONS

- On the surface of it, the connections between Hoey's Lexical Priming Theory and the technology that underlies current LLMs like Chat-GPT or Google's BARD/Gemini are quite apparent
- There are, for example, the origins of the concepts of *priming* and *spreading activation,* which were laid out as theory by R.M. Quillian, who aimed to create a machine which can comprehend human language input 'Teachable Language Comprehender' – an early form of AI
- Hoey's theory is grounded, like LLM models, in Firth's dictum that 'you shall know a word by its neighbours', and the primings of words (sets of words) are reflected in the collocations and colligations which are pervasive and statistically verifiable in human language, as found through corpus research
- Hoey contents that '[w]e have therefore to assume that the discoursal impetus and the lexical priming are interconnected but not coterminous' (2005, p. 163)
- Brynjolfsson (2022, p. 280) says that AI systems can work extremely well when augmenting human endeavour – yet they are incapable of completing 100% of the necessary tasks
- As the comparative experiments shown here have demonstrated, LLMs seem to act on too rigid a lexico-grammatical model
- Consequently, the node words used here are too easily interchangeable, the collocational and colligational usage structures found in naturally occurring texts are only found to be replicated to a degree: the LLMs seem to be hyper-primed, leading to output which is coherent and structurally working only on the surface
- Yet, at the same time, these lack the depth and precision, and the (relative) surefootedness of employing the right term in the right context and co-text (Hoey's *nesting*) that is only found in human-produced, naturally occurring texts.

THANK YOU VERY MUCH!

MICHP@UEF.FI

In memory of Prof. Michael Hoey, *FRS*. 1948-2021

References

Arunachalam, Harish Babu; Tang, Xuning and Scott-Andrews, Joshua. 2023. Do LLMs really understand human language? *TMFormInform*.

https://inform.tmforum.org/features-and-opinion/do-llms-really-understand-human-language (last accessed 03/10/24).

Bargh, John A., and Morsella, Ezequiel. 2008. The Unconscious Mind. *Perspectives on Psychological Science*, *3*(1), 73-79.

Bender, Emily M.and Koller, Alexander. 2020. Climbing towards NLU: On Meaning, Form and Understanding in the Age of Data. *Proceedings of the 58th Annual Meeting of the Association of Computational Linguistics*. Pp. 5185-5198.

Bender, Emily M.; McMillan-Major, Angelina; Gebru, Timnit and Shmitchell, Shmargaret. 2021. On the Dangers of Stochastic Parrots: Can Language Models be Too Big? *FAccT '21.* https://doi.org.10.1145/3442188.3445922

Berber Sardinha, Tony. 2024. AI-generated vs human-authored texts: A multidimensional comparison. *Applied Corpus Linguistics*, *4*(1), 100083.

Brezina, Vaclav, Abi Hawtin and Tony McEnery. 2021. The Written British National Corpus 2014 – design and comparability. *Text and Talk* 41(5–6). 595–615. https://doi.org/10.1515/text-2020-0052 .

Browning, Jacob and LeCun, Yann. 2022.AI And the Limits of Language. *Noema.* https://www.noemamag.com/ai-and-the-limits-of-language/ (last accessed 03/10/24).

Brynjolfsson, Erik. 2022. The Turing Trap: The promise and peril of human-like artificial intelligence. *Daedalus*, *151*(2), 272-287.

Curry, Niall, Baker, Paul, and Brookes, Gavin. 2024. Generative AI for corpus approaches to discourse studies: a critical evaluation of ChatGPT. *Applied Corpus Linguistics*, *4*(1), 100082.

Dickson, B. 2022. LLMs have not learned our language – we are trying to learn theirs. *VentureBeat*. https://venturebeat.com/ai/llms-have-not-learned-our-language-were-trying-to-learn-theirs%ef%bf%bc/ (last accessed 03/10/24).

Eder, Maciej. 2025. Text Analysis is easy. Unless it is not… Plenary Talk. Digital Humanities in Nordic and Baltic Countries (DHNB) 9th Conference. March 7, 2025 @ Tartu, Estonia

Futrell, R., Wilcox, E., Morita, T., Qian, P., Ballesteros, M., and Levy, R. 2019). Neural language models as psycholinguistic subjects: Representations of syntactic state. *arXiv preprint arXiv:1903.03260*.

Granger, Richard. 2020. Toward the quantification of cognition. *arXiv preprint arXiv:2008.05580.*

Hadfield, John. 2022.Why Large Language Models Will Not Understand Human Language. https://jeremyhadfield.com/why-llms-will-not-understand-language/ (last accessed 03/10/24).

Halliday, M.A.K. 1993. Towards a Language-Based Theory of Learning. *Linguistics and Education* 5, 93- 116 (1993)

Havlík, Vaclav. 2023. Meaning and understanding in large language models. *arXiv preprint arXiv:2310.17407*.

Hoey, Michael. 1991. *Patterns of Lexis in Text.* Oxford: Oxford University Press.

Hoey, M. 1994. Patterns of Lexis in Narrative. In: Tanskanen, S-K. and Wårvik, B. (eds) *Topics and Comments.* Anglicana Turkuensia.*13,* pp. 1-41.

Hoey, Michael. 1995. The Lexical Nature of Intertextuality: A Preliminary Study. In: Wårvik, B; Tanskanen, S-K. and Hiltuen, R.: *Organisation in Discourse. Proceedings from the Turku Conference.* Anglicana Turkuensia.14*,* pp. 73-94.

Hoey, Michael. 2005) *Lexical Priming*. London: Routledge.

Hoey, Michael. 2009. Corpus-driven approaches to grammar. In: Römer, U. and Schulze, R: *Exploring the lexis-grammar interface*. Amsterdam/Philadelphia: John Benjamins.pp. 33-47.

Hoey, Michael. 2017a. Foreword. In: Pace-Sigge, M. and Patterson, K.J. (eds.): *Lexical Priming*. Applications and Advances. Amsterdam/Philadelphia: John Benjamins.

Hoey, Michael .2017b. Cohesion and Coherence in a Content-specific Corpus. In: Pace-Sigge, M. and Patterson, K.J. (eds.): *Or Lexical Priming*. Applications and Advances. Amsterdam/Philadelphia: John Benjamins.

Hopper, Paul.1998. Emergent grammar. In M. Tomasello (ed.) The New Psychology of Language. NJ: Lawrence Erlbaum Associates, pp. 155-175.

Kull, Meelis (2025) Humans and AI: similarities, differences, and why it matters. Plenary Talk. Digital Humanities in Nordic and Baltic Countries (DHNB) 9th Conference. March 7, 2025 @ Tartu, Estonia

Manning, Christopher D., Clark, K., Hewitt, J. Khandelai, U. and Levy, O. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. PNAS. Vol. 117 (48), pp. 30045-54.

Manning, Christopher D. 2022. Human Language Understanding and Reason. Daedalus, the Journal of the American Academiy of Arts and Sciences. Vol. 151 (2), pp. 127-138.

Merrill, William, Yoav Goldberg, Roy Schwartz, and Noah A. Smith. "Provable limitations of acquiring meaning from ungrounded form: What will future language models understand?." Transactions of the Association for Computational Linguistics 9 (2021): 1047-1060.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. arXiv preprint arXiv:1301.3781.

Partington, Alan 2014. Mind the gaps. The role of corpus linguistics in researching absence. International Journal of Corpus Linguistics 19 (1): 118-146.

Pace-Sigge, Michael 2018. Spreading Activation, Lexical Priming and the Semantic Web. Abington: Palgrave Macmillan.

Pace-Sigge, Michael and Sumakul, Toar 2022. What Teaching an Algorithm Teaches When Teaching Students How to Write Academic Texts. In Jantunen, Jarmo Harri, et al. Diversity of Methods and Materials in Digital Human Sciences. Proceedings of the Digital Research Data and Human Sciences DRDHum Conference 2022.

Quillian, Ross M. 1967. Word concepts: A theory and simulation of some basic semantic capabilities. Behavioural Science, Vol. 12, No. 5, pp. 410-430. https://doi.org/10.1002/bs.3830120511

Shanahan, Murray. 2022. Talking about large language models. arXiv preprint arXiv:2212.03551.

Sinclair, John 2004. *Trust the Text.* London: Routledge.

Taylor, P. 2025. AI Wars. In *London Review of Books. Vol. 47 No. 5 · 20 March 2025.* https://www.lrb.co.uk/the-paper/v47/n05/paul-taylor/ai-wars

Valmeekam, Karthik, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. "Large language models still can't plan (a benchmark for LLMs on planning and reasoning about change)." In NeurIPS 2022 Foundation Models for Decision Making Workshop. 2022.

Wei, Jason, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama et al. "Emergent abilities of large language models." arXiv preprint arXiv:2206.07682 (2022).

Tools used

Google [2023] 2024.BARD/Gemini. https://BARD.google.com/chat
Brezina, V. & Platt, W. 2023) #LancsBox X, Lancaster University, http://lancsbox.lancs.ac.uk.
OpenAI. [2022] 2024) ChatGPT.(GPT 3.5) https://chat.openai.com/
Scott, M. 2023.WordSmith Tools version 8, Stroud: Lexical Analysis Software.