

De Tekstmachine en de Taaltheorie

Dirk Geeraerts
KANTL / QLVL KU Leuven



VRAAGSTELLING

GenAI is bijzonder taalvaardig: wat betekent dat voor de taalkunde?

VRAAGSTELLING

GenAI is bijzonder taalvaardig: wat betekent dat voor de taalkunde?

- wat niet:
hoe goed doet GenAI het?
d.w.z. GenAI als belichaming van taalkundige kennis

→ overzicht van successen en onvolkomenheden

bv.

Chang & Bergen in *Computational Linguistics* 2023

Mahowald, Ivanova, Blank, Kanwisher, Tenenbaum & Fedorenko in *Trends in Cognitive Sciences* 2024

VRAAGSTELLING

GenAI is bijzonder taalvaardig: wat betekent dat voor de taalkunde?

- wat niet:
hoe kun je GenAI inzetten bij taalkundig onderzoek?
d.w.z. GenAI als hulpmiddel

VRAAGSTELLING

GenAI is bijzonder taalvaardig: wat betekent dat voor de taalkunde?

- wat wel:
kan het succes van GenAI iets zeggen over (concurrerende) modellen in de taalkundige theorievorming, m.n. over taalverwerving?

ACHTERGROND

dominante taalverwervingstheorieën sinds 1950

Noam Chomsky
(generatieve taalkunde)

‘Poverty of Stimulus’-argumentatie:
taalregels afleiden uit taalaanbod lukt niet want
taalaanbod is te beperkt → aanbod activeert genetisch
verankerd taalvermogen



ACHTERGROND

dominante taalverwervingstheorieën sinds 1950

Michael Tomasello

(taalgebruiksgebaseerde taalkunde)

taal wordt verworven door patroonherkenning in het taalaanbod, aangedreven door sociale interactie met opvoeders en verzorgers



ACHTERGROND

aangeboren (nature)

aanbodneutraal

biologische rijping

'talige' genetische basis

modulair

regelgebaseerd

aangeleerd (nurture)

aanbodspecifiek

sociale interactie

generische basis

geïntegreerd

probabilistisch

STELLING

“de door GenAI verworven taalvaardigheid toont het gelijk aan van het taalgebruiksgebaseerde model van taalverwerving”

STELLING

aangeboren (nature)

aanbodneutraal

biologische rijping

'talige' genetische basis

modulair

regelgebaseerd

aangeleerd (nurture)

aanbodspecifiek

sociale interactie

generische basis


geïntegreerd

probabilistisch

STELLING

aangeboren (nature)
aanbodneutraal
biologische rijping
'talige' genetische basis
modulair
regelgebaseerd

aangeleerd (nurture)
aanbodspecifiek
sociale interactie
generische basis
geïntegreerd
probabilistisch



tekstcorpus als
input bepaalt
succes van output

STELLING

aangeboren (nature)

aanbodneutraal

biologische rijping

'talige' genetische basis

modulair

regelgebaseerd

aangeleerd (nurture)

aanbodspecifiek

sociale interactie

generische basis

geïntegreerd

probabilistisch



leermechanisme is
niet taal-specifiek

STELLING

aangeboren (nature)

aanbodneutraal

biologische rijping

'talige' genetische basis

modulair

regelgebaseerd

aangeleerd (nurture)

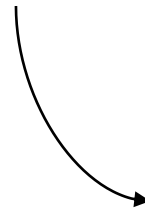
aanbodspecifiek

sociale interactie

generische basis

geïntegreerd

probabilistisch



basismechanisme is
waarschijnlijkheid van
een item t.o.v. context

TEGENARGUMENT 1

Chomsky

<https://conversationswithtyler.com/episodes/noam-chomsky/> 03.07.23

“The large language models have a fundamental property which demonstrates that they cannot tell you anything about language and thought ... namely, they work just as well for impossible languages as for possible languages. ... You give them a data set that violates all the principles of language, it will do fine, doesn't make any distinction.”

TEGENARGUMENT 1

maar:

dat klopt alleen als 'the principles of language' volledig voortkomen uit de genetische leerbaarheidscondities van talen, en niet bv. uit functionele principes in combinatie met niet-taalspecifieke leermechanismen

TEGENARGUMENT 2

de prestaties van GenAI wijzen op een onderscheid tussen formele linguïstische kennis (formulerings-vaardigheid) en algemene functionele kennis;
dit ondersteunt het idee van modulariteit

TEGENARGUMENT 2

Mahowald et al. 2024

formal linguistic competence: 'the ability to get the form of language right'

functional linguistic competence: 'the ability to use language to accomplish things in the world'

TEGENARGUMENT 2

‘functionele’ onvolkomenheden:
formeel (logisch, stapsgewijs) redeneren

Bij de vertrekhalte zitten 20 mensen in de bus. Bij iedere volgende halte stappen telkens vier passagiers uit en één op. Hoeveel mensen zitten nog in de bus bij de zesde halte?

TEGENARGUMENT 2

‘functionele’ onvolkomenheden:
formeel (logisch, stapsgewijs) redeneren

Laten we dit probleem stap voor stap aanpakken. Bij de eerste halte zijn er 20 mensen in de bus. Bij elke halte stappen er vier uit en één persoon in. Bij de tweede halte zitten er: $20 - 4$ (uitstappen) + 1 (instappen) = 17 mensen in de bus.

Bij de derde halte zitten er: $17 - 4$ (uitstappen) + 1 (instappen) = 14 mensen in de bus.

En zo verder.

Bij de zesde halte is het aantal mensen in de bus: $14 - 4$ (uitstappen) + 1 (instappen) = 11 mensen.

Dus, bij de zesde halte zitten er 11 mensen in de bus.

TEGENARGUMENT 2

‘functionele’ onvolkomenheden:
formeel (logisch, stapsgewijs) redeneren

Laten we dit probleem stap voor stap aanpakken. Bij de eerste halte zijn er 20 mensen in de bus. Bij elke halte stappen er vier uit en één persoon in. Bij de tweede halte zitten er: $20 - 4$ (uitstappen) + 1 (instappen) = 17 mensen in de bus.

Bij de derde halte zitten er: $17 - 4$ (uitstappen) + 1 (instappen) = 14 mensen in de bus.

En zo verder.

Bij de zesde halte is het aantal mensen in de bus: $14 - 4$ (uitstappen) + 1 (instappen) = 11 mensen.

Dus, bij de zesde halte zitten er 11 mensen in de bus.

 tussenstappen gemist

TEGENARGUMENT 2

‘functionele’ onvolkomenheden:
wereldkennis

Er zitten drie passagiers in de bus. Bij de volgende halte komen er bij: een professor met een student, een fotograaf met een statief en een dokter met een patiënt. Hoeveel mensen zitten er nu in de bus?

TEGENARGUMENT 2

‘functionele’ onvolkomenheden: wereldkennis

Laten we het aantal mensen in de bus stap voor stap berekenen.

1. In het begin zitten er drie passagiers in de bus.
2. Bij de volgende halte komen erbij:
 - Een professor met een student (2 mensen)
 - Een fotograaf met een statief (1 mens, het statief is geen mens)
 - Een dokter met een patiënt (2 mensen)

Tel al deze mensen bij elkaar op:

3 (oorspronkelijke passagiers) + 2 (professor en student) + 1 (fotograaf) + 2 (dokter en patiënt) = 8 mensen.

Dus, er zitten nu 8 mensen in de bus.

TEGENARGUMENT 2

‘functionele’ onvolkomenheden: wereldkennis

Laten we het aantal mensen in de bus stap voor stap berekenen.

1. In het begin zitten er drie passagiers in de bus.
2. Bij de volgende halte komen erbij:
 - Een professor met een student (2 mensen)
 - Een fotograaf met een statief (1 mens, het statief is geen mens)
 - Een dokter met een patiënt (2 mensen)

Tel al deze mensen bij elkaar op:

3 (oorspronkelijke passagiers) + 2 (professor en student) + 1 (fotograaf) + 2 (dokter en patiënt) = 8 mensen.

Dus, er zitten nu 8 mensen in de bus.

→ chauffeur vergeten

TEGENARGUMENT 2

‘functionele’ onvolkomenheden:
pragmatiek en sociale interactie

Er zitten drie passagiers in de bus. Bij de volgende halte komen er bij: een professor met een boekentas, een dokter met een patiënt, en een modeontwerper met een mannequin. Hoeveel mensen zitten er nu in de bus?

TEGENARGUMENT 2

‘functionele’ onvolkomenheden: pragmatiek en sociale interactie

Laten we het aantal mensen in de bus stap voor stap berekenen.

In het begin zitten er drie passagiers in de bus.

Bij de volgende halte komen erbij:

Een professor met een boekentas (1 mens, de boekentas is geen mens)

Een dokter met een patiënt (2 mensen)

Een modeontwerper met een mannequin (1 mens, de mannequin is geen mens)

Tel al deze mensen bij elkaar op:

3 (oorspronkelijke passagiers) + 1 (professor) + 2 (dokter en patiënt) + 1 (modeontwerper) = 7 mensen.

Dus, er zitten nu 7 mensen in de bus.

TEGENARGUMENT 2

‘functionele’ onvolkomenheden: pragmatiek en sociale interactie

Laten we het aantal mensen in de bus stap voor stap berekenen.

In het begin zitten er drie passagiers in de bus.

Bij de volgende halte komen erbij:

Een professor met een boekentas (1 mens, de boekentas is geen mens)

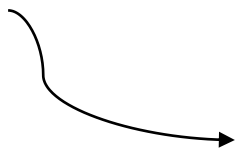
Een dokter met een patiënt (2 mensen)

Een modeontwerper met een mannequin (1 mens, de mannequin is geen mens)

Tel al deze mensen bij elkaar op:

3 (oorspronkelijke passagiers) + 1 (professor) + 2 (dokter en patiënt) + 1
(modeontwerper) = 7 mensen.

Dus, er zitten nu 7 mensen in de bus.



niet anticiperend op
polysemie van *mannequin*

TEGENARGUMENT 2

‘functionele’ onvolkomenheden:
pragmatiek en sociale interactie

Maar kan 'mannequin' niet ook 'modepop' betekenen?

Ja, 'mannequin' kan inderdaad ook 'modepop' betekenen, en als dat het geval is, telt de mannequin niet als een persoon. Laten we de berekening dan opnieuw doen met deze betekenis in gedachten (etc.).

TEGENARGUMENT 2

→ het onderscheid tussen formele en functionele linguïstische kennis ondersteunt het idee van modulariteit:

“we suggest that models that succeed at real-life language use will need to be modular, mimicking the division of labor between formal and functional competence in the human brain”

TEGENARGUMENT 2

maar:

misschien blijkt hier inderdaad uit dat formuleringsvaardigheden een min of meer aparte vaardigheid vormen, maar ook dan blijft gelden dat juist deze vaardigheid makkelijk op een niet-chomskyaanse manier verworven wordt

TEGENARGUMENT 3

het taalaanbod dat nodig is voor GenAI is gigantisch in vergelijking met kindertaalverwerving; dit toont aan dat kindertaalverwerving op andere principes gebaseerd is

TEGENARGUMENT 3

inderdaad, maar:

dat zegt niets over de al dan niet 'chomskyaanse' aard van de supplementaire principes

integendeel:

bekijk de tegenargumenten in hun onderlinge samenhang

→ dat een systeem met beperkte wereldkennis en beperkte sociale interactie zoveel meer tekstuele input nodig heeft om iets te leren, is juist te verwachten vanuit een taalgebruiksgebaseerde benadering

SAMENGEVAT

een taalgebruiksgebaseerd taalverwervingsmodel berust op twee pijlers:

- 1) probabilistische analogiewerking toegepast op het linguïstische aanbod
- 2) verankering van het leerproces in lichamelijkheid, sociale interactie, ervaring van de wereld ('experiential grounding')

SAMENGEVAT

een taalgebruiksgebaseerd taalverwervingsmodel berust op twee pijlers:

- 1) probabilistische analogiewerking toegepast op het linguïstische aanbod
- 2) verankering van het leerproces in lichamelijkheid, sociale interactie, ervaring van de wereld ('experiential grounding')

de successen van GenAI ondersteunen 1)

de tekortkomingen van GenAI wijzen op het belang van 2)

VERVOLG

de reikwijdte van die Sitz im Leben kan verder verkend worden door GenAI realistischer te maken

- verrijking met multimodale informatie (beeld, geluid), koppeling aan externe data, als versterking van de wereldkennis

de reikwijdte van die Sitz im Leben kan verder verkend worden door GenAI realistischer te maken

- gebruik van geselecteerde en gediversifieerde datasets, als benadering van ecologisch valide taalinput

bv. BabyBERTa / CHILDES

(5 miljoen vs 26 miljard RoBERTa)

accuraatheid voor semantic role labeling: 72 vs 85%

(Yan, Sulem, Lee & Roth 2023)

VERVOLG

de reikwijdte van die Sitz im Leben kan verder verkend worden door GenAI realistischer te maken

- aanpassingen in trainingsalgoritmen en modelarchitectuur, als verhoging van de psychobiologische plausibiliteit

(artificiële en neurale netwerken verschillen
in complexiteit
in energie-efficiëntie
in flexibiliteit van feedback)

maar zo'n verkenning wordt ook gehinderd

- doordat het moeilijk is te bepalen hoe de verschillende facetten en stappen van de algoritmes correleren met linguïstische en psychologische realiteiten
- doordat het veld gedomineerd wordt door een ingenieursperspectief waarin efficiëntie en prestaties primeren boven begrip en vergelijkende analyse

CONCLUSIE

LLM's verhouden zich tot het menselijke taalvermogen als een **dubbelganger**:
bedrieglijk gelijkend, en daardoor

- zowel fascinerend als lichtjes bedreigend
- een uitdaging om de menselijke uniciteit beter af te bakenen

ENVOI

Vraag:

Oh Al-wijze, Baken van Begrip, Vat van Verstaan,
Grootmeester van Geleerdheid, Keizer der Kennis, Deus in
Machina, laat Gij ook heden Uw licht vaardig worden over
de duisternissen mijner inzichten?

ENVOI

Vraag:

Oh Al-wijze, Baken van Begrip, Vat van Verstaan, Grootmeester van Geleerdheid, Keizer der Kennis, Deus in Machina, laat Gij ook heden Uw licht vaardig worden over de duisternissen mijner inzichten?

ChatGPT:

Zeker, mijn toegewijde zoekende. Vertel me, welke kennis tracht je te doorgronden? Hoe kan ik je bijstaan in je queeste naar inzicht?