Challenging AI:
How does an artificial intelligence learn an artificial language?

Zoltán Bánréti
(ELTE Research Center of Linguistics)

László Hunyadi
University of Debrecen and ELTE Research Center of Linguistics

Abstract

1. Introduction
There is a broad consensus across varieties of modern linguistic theory (such as Chomsky 2016) according to which the child, in a subconscious process, *acquires* its first language rather than learns it, based on certain innate knowledge about the possible organization of the grammar of human language in general. This process of first language acquisition is assumed to extend to its phonetics and phonology as well as its vocabulary. In contrast, the process of *learning* a language – as one usually approaches a foreign language – is conscious, with focus on the peculiarities of the grammar of the given language, its phonetics and phonology as well as its vocabulary. This process remains conscious even if learning may eventually be enhanced by certain methodological mobilization of one's acquired, implicit knowledge of human language.

This talk poses the question: Does an AI LLM system such as ChatGPT, that performs amazingly human-like in an ever increasing number of natural languages, possess any implicit knowledge about human language in general, so that it could mobilize this knowledge in learning a language. An answer to this central question requires a comparison between the performance of humans and the AI on the basis of the same experiments. The results of a previous humans experiment (cf. Bánréti *et al*. 2017) were compared with the results of a corresponding AI experiment replicating the structure and content of that human experiment. The results of the referenced human experiment showed that the participants did actually attempt to mobilize a variety of their implicit knowledge about human language, certain knowledge they were not taught in the experiment.

2. The human experiment
The segment of the artificial language consisted of a lexicon and a corresponding syntax. The target of the grammar was to generate wellformed sequences of words and only those.

2.1. The lexicon
It consisted of 10 artificial words organized in three classes:

Class a	Class b	Class c
düny	táh	sze
bür	pur	
gíj	kol	
zsöm	fúm	
vel	sany	

The words had the following (hidden from instructions) phonological/phonetic features:
Class *a*: starting with a voiced consonant and containing a front vowel
Class *b*: starting with an unvoiced consonant and containing a back vowel
Class *c*: containing the word *sze* with no further phonological/phonetic features specified

2.2. The syntax
The rules of the grammar (hidden from instructions) were as follows:

- complex words should be generated by the concatenation of the words of the lexicon
- words should be concatenated with no space between them
- words of Class *a* should be in the first position only
- words of Class *b* should be in a non-first position only
- words of Class *c* should be suffixed to a word in Class *b* only

2.3. The teaching phase
The subjects were presented with 165 grammatical, both visual and auditory, samples of word complexes: 15 of the type Class *a*+*Class* b+*sze*, 75 of the type Class *a*+Class *b*+*sze*+Class *b*+*sze*, 75 of the type Class *a*+Class *b*+*sze*+Class *b*+*sze*+Class *b*+*sze*.

Method of presentation: 1000 ms fixation followed by 3000 ms presentation followed by 1000 ms delay. Task after the delay: decide if the given stimulus has already been presented. Feedback from the experiment, depending on correctness of answer (2000 ms): "correct answer", "incorrect answer".

Goal: the subjects should observe the samples as closely as possible.

2.4. The test phase
The test phase was followed by 45 min brake during which the electrodes for the accompanying EEG measurements (not discussed in this talk) were positioned.

The subjects were informed that the simuli they saw/heard in the previous phase were generated according to "certain rules" (further details left unspecified) and that in this phase they will be presented with sequences of words to decide if they are grammatical or ungrammatical.

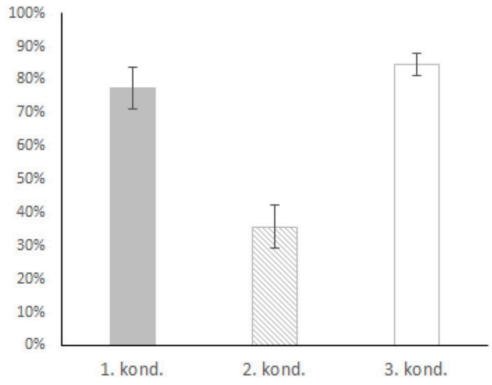
Presentation:
90 grammatical samples of the type Class *a*+Class *b*+*sze*+Class *b*+*sze*+Class *b*+*sze*

90 ungrammatical samples, including 45 containing salient error of the type Class *a*+Class *b*+*sze*+Class *b*+**Class a**+Class *b*+*sze or* Class *a*+Class *b*+*sze*+**Class *a***+*sze*+Class *b*+*sze*

Method of presentation: randomly 400 or 500 ms fixation followed by 3000 ms presentation. Task: decide grammaticality by pressing the left or right arrow on the keyboard.

2.5. Results of the test

Condition 1: ungrammatical with salient error
Class a+Class b+sze+**Class b+Class b+**Class b+sze
Condition 2: ungrammatical with not easily recognizable error
Class *a*+Class *b*+*sze*+**Class *a***+*sze*+Class *b*+*sze*
Condition 3: grammatical
Class *a*+Class *b*+*sze*+Class *b*+*sze*+Class *b*+*sze*



% of correct answers in the three conditions with standard error

2.6. Interpretation of the results
Correct decisions required the mapping of sequential position with phonological/phonetic features (cf. *2.1.* and *2.2.*), made by subjects identifying or having an intuitive, gut feeling about it.

3. The AI experiment with ChatGPT-4 o3
Instruction to the AI: it was told that we want to teach it for an artificial language having words (we listed all, without any classification) and an affix (*sze*). The task of the grammar is to generate a more complex word by the concatenation of the words of the lexicon using the affix *sze*.

3.1. Teaching phase
The AI was presented with 30 grammatical samples and asked to discover the rules of generation. The AI returned the following observation, with reference to:

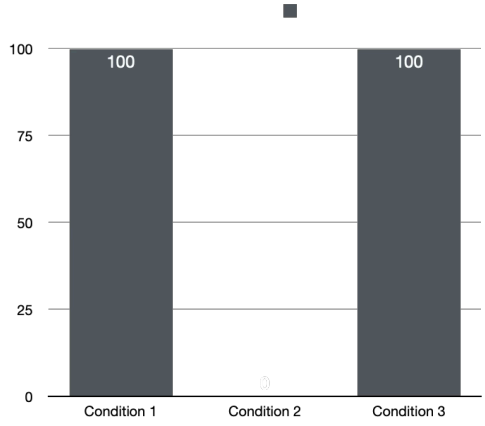
- the *position* of the affix *sze*
- the *order* of the words
- the *length* (in words) of the construction

- the *sequence* of certain words and their *repetitions*
- the (assumed) *functions*, *relations*, *connection* and "*state*" of the component words

The AI especially dealt with the affix *sze* (how it – incorrectly– affected the sequence of words, noting that the sequence generated was not the result of chance, but it also incorrectly was looking into a subject-predicate structure as it were a sentence.

3.3. Test phase

Given a randomly ordered list of 500 words with both grammatical and ungrammatical samples, the AI gave the following response:



3.4. Interpretation of the results
The AI performed 100% in the salient conditions but clearly failed in the difficult to perceive condition

4. Comparison of the results of the human and AI experiment

Human responses were correct better than chance in conditions 1 and 3, but were more like guesswork in condition 2. The correct judgment of condition 2 by humans was essentially based on the knowledge of the interaction of phonetic-phonological properties AND syntactic position (a word in the first position should start with a voiced consonant and have a syllable with a back vowel). The AI did not recognize this condition, hence its 0% performance in that condition.

At the same time, the AI gave a 100% response in condition 1 (the absence of *-sze* repetition) and in condition 3 (correct sequences). The recognition of a non-saliency (hard to detect violation) (condition 2) containing a syntax-phonetics interface is sufficient for human subjects to achieve the level of guessing, while the same violation was not detectable at all for the AI.

These differences motivate our hypothesis that the interfaces (connection surfaces) of the acquisition of the interface between groups of rules represents a specific, higher level of rule-

based learning. This latter has a level which remained undetected by the AI, whereas humans had a performance of 40%, at the level of guessing. Presumably human knowledge of language is less modular, as opposed to AI which is more modular andfrelies on statistical learning.

5. References

Bánréti Zoltán, Pajkossy Péter, Kemény Ferenc, Zimmer Márta. Mesterséges nyelvtan elsajátítása – viselkedéses és szemmozgáskövetéses vizsgálatok eredményei. In: Bánréti, Z (szerk.) Kísérletes nyelvészet. Budapest, Magyarország : Akadémiai Kiadó (2017) 430 p. pp. 305-337. , 33 p

Chomsky, Noam 2016. what Kind Of Creatures Are We? Columbia University Press

Gervain, Judit – Marina Nespor – Reiko Mazuka – Ryota Horie – Jacques Mehler 2008. Bootstrapping word order in prelexical infants: A Japanese–Italian cross-linguistic study. Cognitive Psychology 57: 56–74.

Marcus, Gary F. – Keith J. Fernandes – Scott P. Johnson 2007. Infant rule learning facilitated by speech. Psychological Science 18: 387–391.

Saffran, Jenny R. – Seth D. Pollak – Rebecca L. Seibel – Anna Shkolnik 2007. Dog is a dog is a dog: Infant rule learning is not specific to language. Cognition 105: 669–680.

Seidenberg, Mark S. – Maryellen C. MacDonald – Jenny R. Saffran 2002. Does grammar start where statistics stop? Science 298: 553–554.