Језички модели, шта је то?

Михаило Шкорић

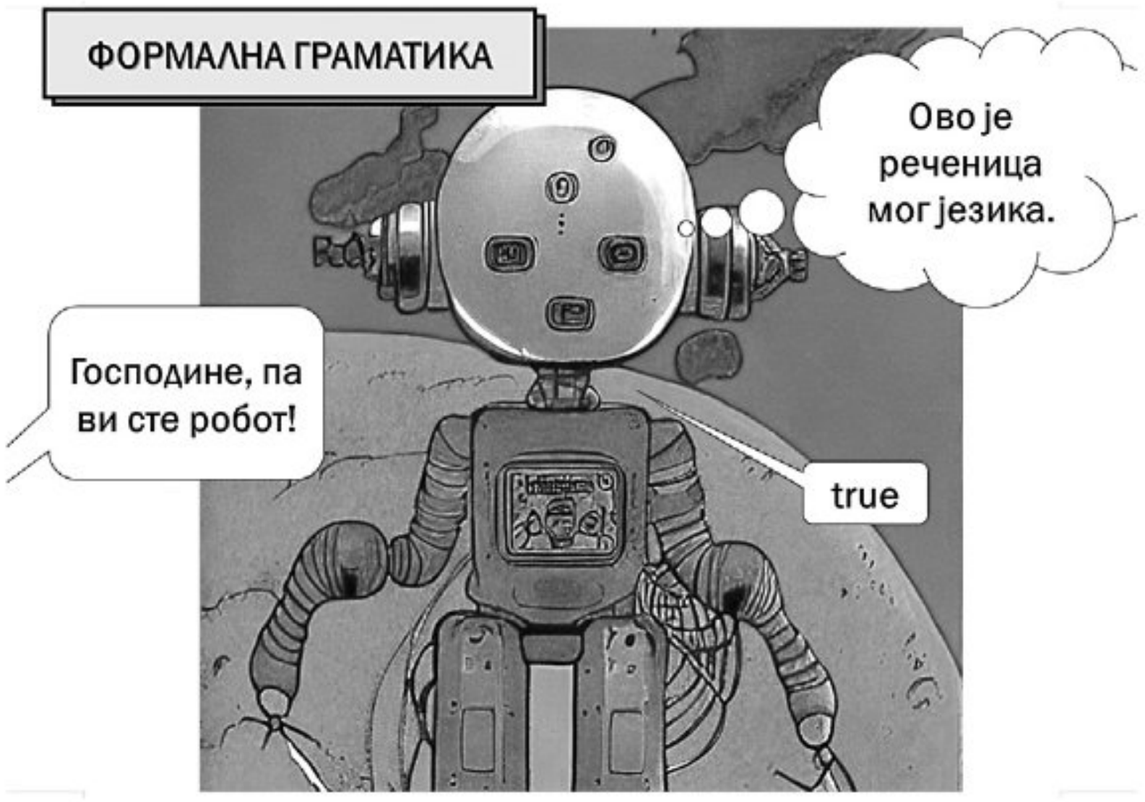
Михаило Шкорић

ЈЕЗИЧКИ МОДЕЛИ, ШТА ЈЕ ТО?

Појам *језичкоī модела* се, према уобичајеној дефиницији, односи на расподелу вероватноће над неким задатим текстом (Jurafsky–Martin 2018). Како бисмо на једноставнији начин описали његово порекло и улогу, одговор ћемо започети примером из формалног образовања – употребе граматике на часу српског језика. Неком приликом наставница пита *да ли је нека реченица исīправна*, при чему се од ђака очекује *буловски одīовор*: *јесīте* или *није* тј. *īачно* или *неīачно* (Слика 1).

Слика 1: Час срīскоī језика

Школска граматика српског језика функционише на исти начин као и *формална граматика* (Carroll–Long 1989). Формална граматика је (рачунарски) систем који нам даје одговор на питање *да ли неки улазни текст припада језику који та граматика покушава да опише/моделује*. Дакле, формалној граматици дате текст, а она вам каже *јесте/није* (*true/false*).

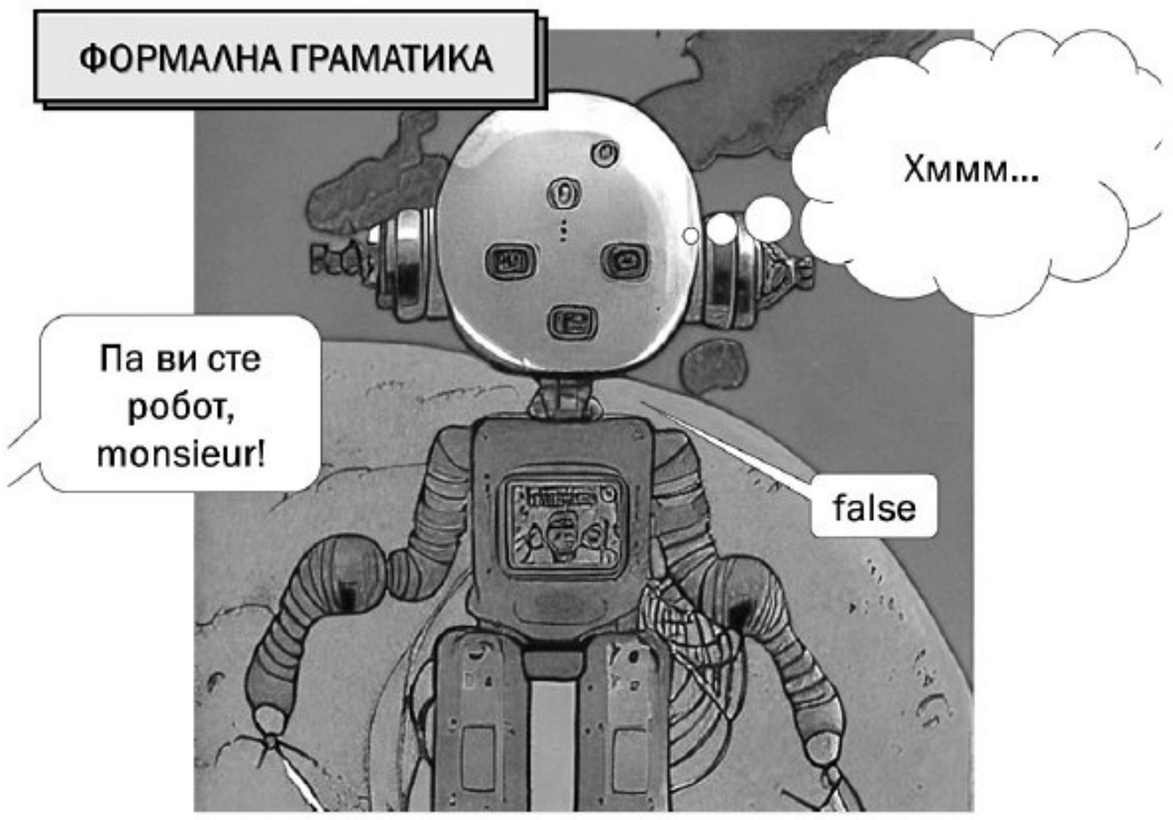


Слика 2: Формална граматика српског језика

Формална граматика српског језика нам говори да ли неки улазни текст припада српском језику (Слика 2), док формална граматика неког другог језика ради то исто, али за тај други језик, нпр. француски (Слика 3).

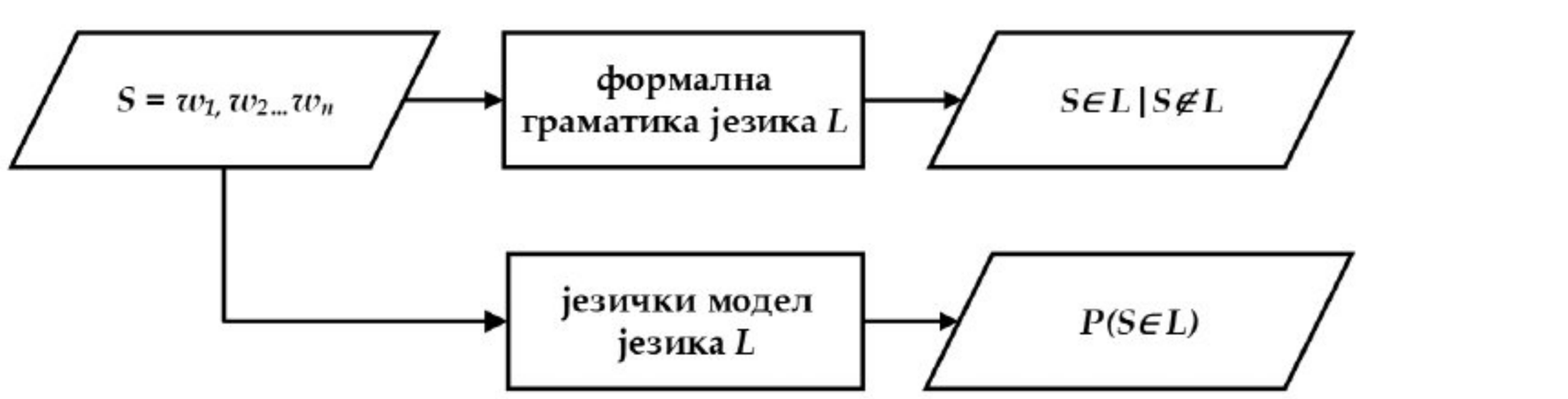
Слика 3: Формална граматика француског језика

Међутим, шта се дешава ако је улазни текст мешовитог порекла? Која ће га граматика препознати? Услед ограничења буловског одговора све граматике ће, вероватно, рећи да је текст неисправан и да није део њиховог језика (Слика 4).



Слика 4: Формална граматика српског језика

Као одговор на тај проблем, намеће се *језички модел*. Већ је речено да је језички модел расподела вероватноће над улазним текстом. Из тог разлога, језички модел одговор на питање припадности језику даје у виду *вероватноће*.

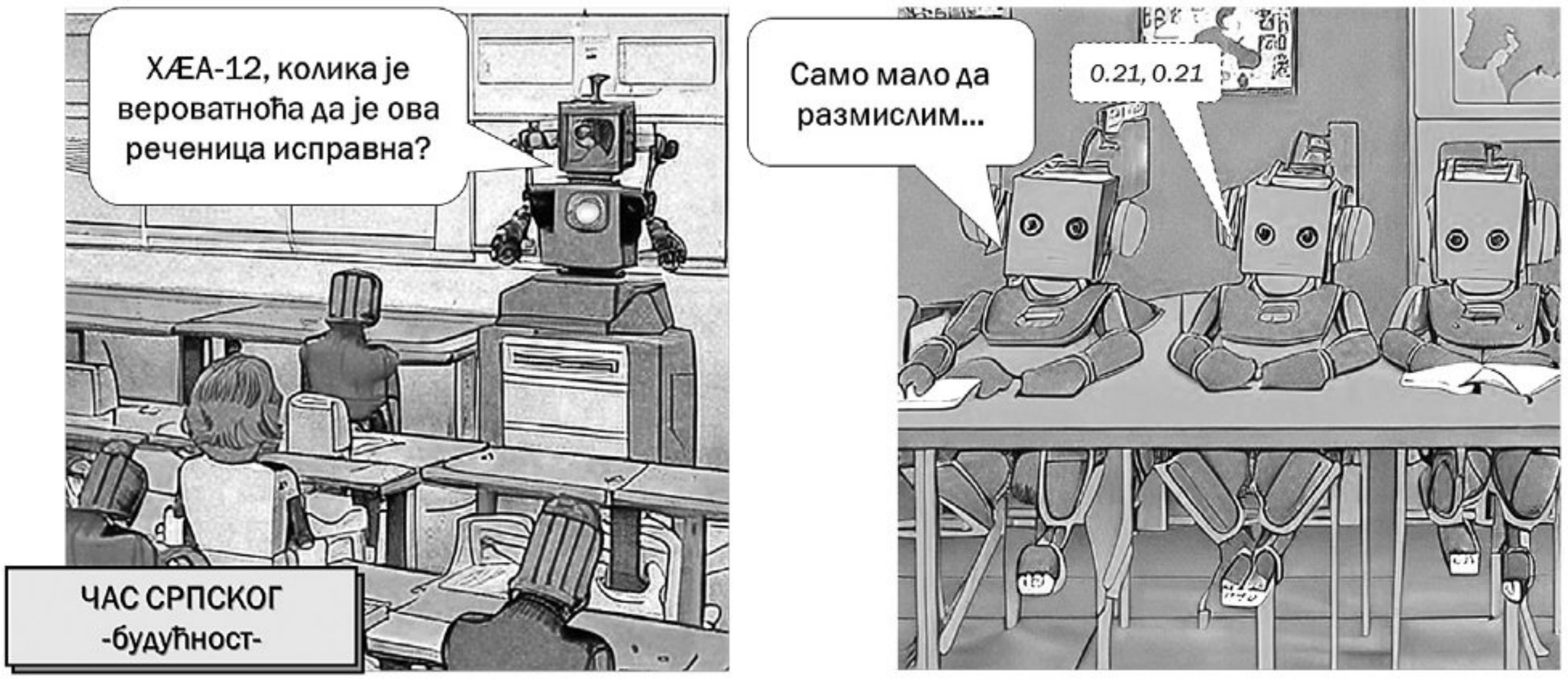


Слика 5: Поређење функционалности формалних граматика (горе) и језичких модела (доле), где S означава улазни текст, а P(S∈L) означава вероватноћу припадности неком језику L

За неуобичајен текст даје ниску вероватноћу, а за уобичајен, високу вероватноћу да је текст исправан, тј. да припада језику који се описује (Слика 6).

Слика 6: Језички модел српског језика

Прелазак са граматика на језичке моделе у формалном образовању би тако потенцијално закомпликовао часове српског у будућности. Од ђака би се очекивао одговор *на скали од 0 до 1*, што је много шири спектар од буловског одговора (Слика 7).

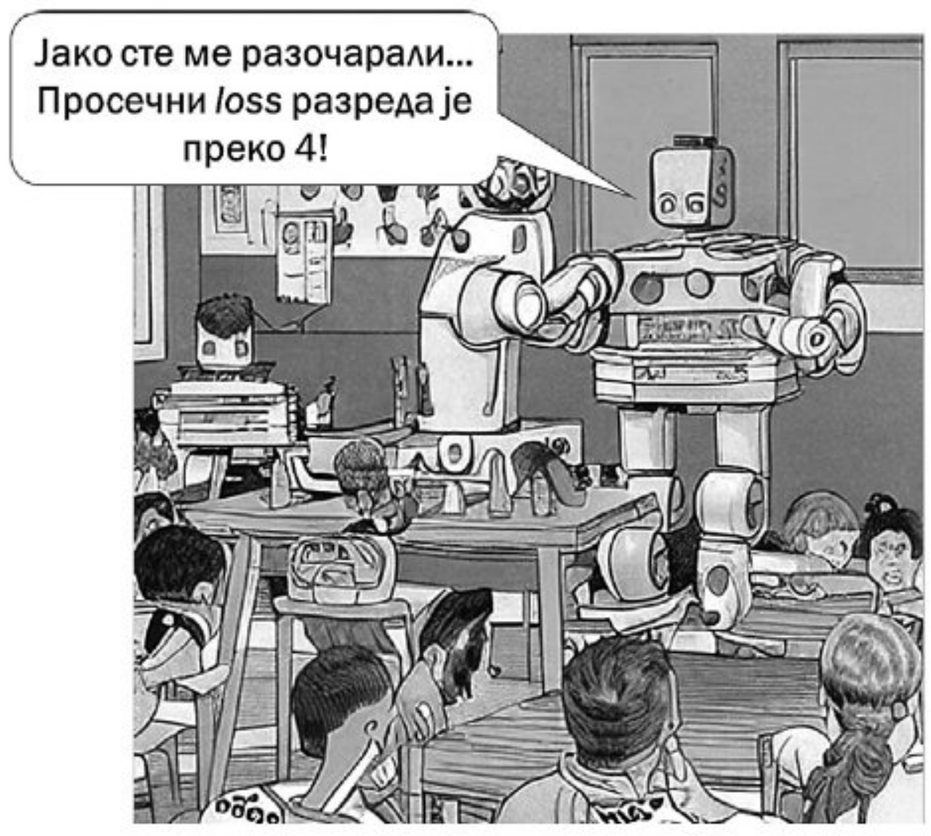


Слика 7: Час српског језика, будућност

У том случају би се за оцењивање могла дати следећа формула – Слика 8, при чему би нижи резултат сугерисао већу оцену – језички модели треба да теже томе да ентропија (*loss*) буде што нижа (Слика 9), што би указивало на то да се вероватноће реченица правилно израчунавају.

$$loss(x, y) = -\sum_{i=1}^{n} x_i \log y_i$$

Слика 8: Формула за израчунавање ентропије између два векторска низа



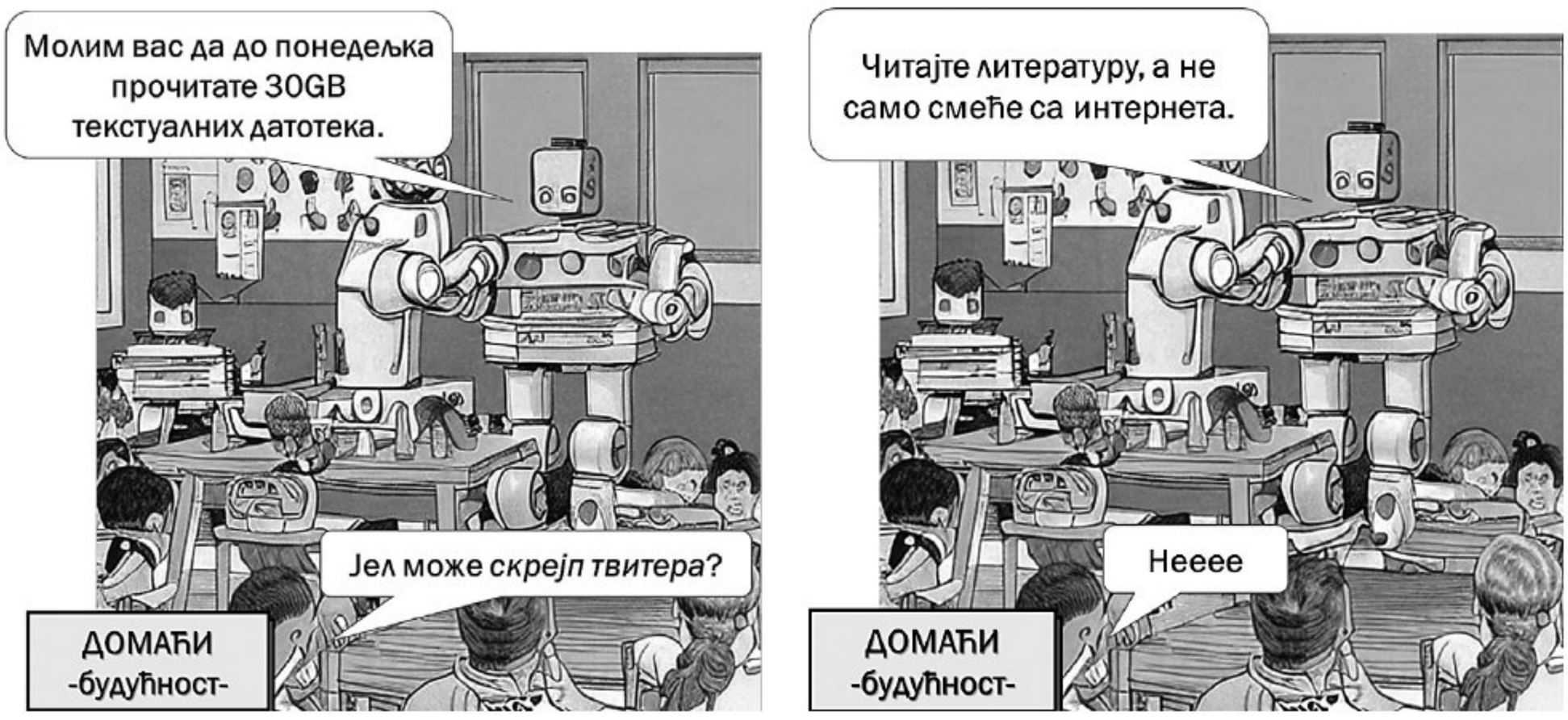
Слика 9: Час српског језика, будућност

Квалитетни језички модели омогућавају машинама да разумеју и/или користе природни, људски језик, што олакшава комуникацију између људи и машина или чак између више машина које користе језичке моделе.

Шта је потребно за обучавање квалитетних језичких модела?

Квалитетни језички модели заснивају се, пре свега, на корпусу квалитетног текста, који такође треба да буде што већег квантитета. Поред тога, неопходни су рачунарски ресурси (што веће процесорске моћи и што стабилнији), као и софтвер који ће омогућити генерализацију прикупљеног корпуса у рачунарски модел.

Када говоримо о корпусима, истина је да боље моделе узрокују већи корпуси (Brown et al. 2020), али исто тако и квалитетнији подаци, пре свега стручна и научна литература, као и лепа књижевност (Слика 10).



Слика 10: Домаћи задатак из српског, будућност

Што се тиче рачунарских ресурса, за обучавање језичких модела (па чак и савремених) може се користити било који рачунар, под условом да модел који желимо да обучимо и скуп за обучавање могу да стану у радну активну меморију рачунара. Ипак, *добра* графичка картица (са што више сопствене меморије) омогућиће да се обучавање оствари много брже (уколико се испуне одређени услови, попут инсталације неопходних *драјвера* и осталог потребног софтвера). Дакле, моћни рачунарски ресурси нису пресудни за обучавање модела, али јесу за брзину његовог обучавања.

Слика 11: Школски ӣрибор, будућносӣ

Коначно, када говоримо о софтверу, дефинитивни примат у остваривању најбољих резултата има *ӣрансформерска* архитектура (Vaswani et al. 2017). Употреба ове архитектуре, како за обучавање тако и за употребу модела, лако је доступна путем библиотеке *transformers*[1] за програмски језик *Пајӣон*[2], а уколико не желите да пишете програмски код од почетка, можете користити неки од бројних пројеката имплементације, на пример, *Scratch2LM*[3].

Цитирани извори:

Brown et al. 2020: Tom Brown et al. "Language Models are Few-Shot Learners." *arXiv: 2005.14165.*

Carroll–Long 1989: John Caroll and Darrell D. E. Long. *Theory of finite automata with an introduction to formal languages.* s. l.: s. n.

Jurafsky–Martin 2018: Daniel Jurafsky and James H. Martin. "N-gram language models." *Speech and language processing* 23.

Vaswani et al. 2017: A. Waswani et al. "Attention Is All You Need." *arXiv:1706.03762.*

Михаило Шкорић
Рударско-геолошки факултет, Београд
mihailo.skoric@rgf.bg.ac.rs

[1] https://huggingface.co/docs/transformers
[2] https://www.python.org/
[3] https://github.com/procesaur/Scratch2LM