

LLMs enhance linguistic research: two examples

Roland Meyer

2026-06-30

Humboldt-Universität zu Berlin, Institut für Slawistik und Hungarologie



First example: Processing minority variants

- Project *Slavic in a multilingual setting: Register and fused (hybrid) lects* (Meyer/Szucsich) in CRC (SFB) 1412 *Register*

💡 *Register:*

(socially recurring) intraindividual variation in linguistic behavior influenced by *situational and functional parameters* such as distance, social-role relationship, topic, purpose, formality, mode.

- In *multilingual* situations, the **alternation between languages (language mixing)** can work as a register
- two instances in Slavic: Ukrainian/Russian *surzhyk* and Czech/Polish *po naszymu*

Grasping language mixing

- **po-naszymu** or **mixed** stems with **Pol[ish]** or **Cz[ech]** endings:

(1) to mi **pojča-ł**-a ta Nela s **třeć-i** klasy
this me.dat lend-pst-f that N-f.nom from third-gen.sg grade
'This lent me that Nela from third grade.'

- **Goal:** Measuring the divergence of utterances/words/speakers ... from standard Cz or Pol
- We need a very close translation of the minority utterances into Cz and/or Pol, aligned word-by-word to the original.
- good task for a general LLM (in our case: Qwen3.6-27B-FP8, HU KI-tools):
 - can translate
 - tolerant towards variation in input, some dialect awareness, output standard
 - can align

Sample output

a	do	Polski	choźiće↑		jo	doś	čynsto~/	tak	na	přiklat	po	jakši	ubraña	abo	do	sklepu	po	jakši	mjyso	abo	po	taki	řečy↓
a	do	Polska	chodíte		jo	docela	čistě	tak	na	přiklad	po	nějaké	oblečení	nebo	do	obchodu	po	nějaké	maso	nebo	po	takové	věci
a	do	Polski	chodzicie		tak	dosyć	często	tak	na	przykład	po	jakieś	ubrania	albo	do	sklepu	po	jakieś	mięso	albo	po	takie	rzeczy
a do Polski choźiće↑					jo doś čynsto~/			tak na přiklat po jakši ubraña abo do sklepu po jakši mjyso abo po taki řečy↓															
a do Polska chodíte					jo docela čistě			tak na přiklad po nějaké oblečení nebo do obchodu po nějaké maso nebo po takové věci															
a do Polski chodzicie					tak dosyć często			tak na przykład po jakieś ubrania albo do sklepu po jakieś mięso albo po takie rzeczy															

Figure 1: Sample output of LLM-based translation & alignment

- translation utterance-by-utterance plus alignment (with some gaps)
- errors (e.g. Cz: *jo docela čistě* instead of *jo dost často*)
- sometimes misses out on Cz/Pol varieties, hand-correction desirable
- importantly, no other “traditional” language-processing tools available

Aggregate distances across utterances

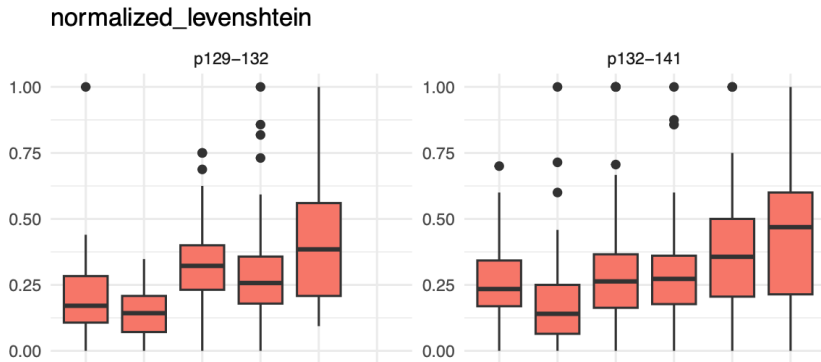


Figure 2: Speaker 0-6 in two sample conversations. Distance *po-naszymu*-Pol

- speaker 2 “control” (Pol native)
- speaker 1 always interviewer (*po-naszymu* native)

Second example: Generating picture stimuli for psycholinguistics

- **Task:** Evaluate an auditory stimulus in the minority language variant in a given situation.
- AI-generated pictures:



Figure 3: “formal” setting



Figure 4: “informal” setting

Conclusion

- LLMs excel over traditional methods (manual transcription, real scene photos)
- LLMs potentially induce biases and show no regard for researchers' control
- “flat” pictures, less individualized (= here: a feature)
- explicit control in prompt
- cooperation, human-in-the-loop