# Morfologiškai ir sintaksiškai anotuotų tekstynų rengimo iššūkiai: nuo žodžio sampratos iki redagavimo įrankių pasirinkimo

Erika Rimkutė
VDU SITTI

CLARIN

# Projektas

- Europos Sąjungos NextGenerationEU projektas „Morfologiškai ir sintaksiškai anotuotų tekstynų modeliai apmokymui (auksiniai standartai)“ (Nr. 02-098-K-0001).
- Projektas finansuojamas Ekonomikos gaivinimo ir atsparumo didinimo plano „Naujos kartos Lietuva“ ir Lietuvos Respublikos valstybės biudžeto lėšomis.
- 2024–2026 m.

Finansuoja Europos Sąjunga NextGenerationEU

NAUJOS KARTOS LIETUVA

CLARIN

# Projekto tikslas

- Parengti 10 mln. žodžių morfologiškai ir sintaksiškai anotuotų tekstynų modelius kaip auksinį standartą įvairiems įrankiams (tiek pagrįstiems įprastomis technologijomis, tiek panaudojant dirbtinį intelektą) apmokyti.
- Išsamiau žr. https://sitti.vdu.lt/morfologiskai-ir-sintaksiskai-anotuotu-tekstynu-modeliai-dirbtinio-intelekto-apmokymui/.

# Projekto darbai (1)

- Apsibrėžti tekstynų mato vienetą (žodį, *tokeną*):
• žodžių formos, simboliai, skyrybos ženklai.
- Aprašyti tipiškus ir netipiškus tekstyno vienetus (*tokenus*), pvz.:
• 1-asis, 10 000, :-), *iš tikrųjų, McDonald's*;
• *Jonaitytė-Petraitienė*, 12:15, Omega-3, 2%.
- Rasti terminą tekstyno mato vienetui (*tokenui*) pavadinti.

# Projekto darbai (2)

- Apsibrėžti tekstyno sandarą (4 dalys: moksliniai, grožiniai, publicistiniai tekstai, dokumentai).
- Išspręsti autorių teisių klausimus.
- Neanotuotų tekstų tvarkymo ypatumai: labiau (nei sudarant neanotuotus tekstynus) keičiamas pirminis tekstas (atsisakoma paveikslų, lentelių, užsienio kalbos citatų, literatūros sąrašų; ištaisomos korektūros klaidos).

# Projekto darbai (3)

- Pasirinkti morfologiniam ir sintaksiniam anotavimui bei redagavimui skirtus įrankius:

- žodžių ir sakinių segmentavimo įrankis (*tokenizatorius*),
- automatinės morfologinės analizės įrankis *Morfuoklis*,
- morfologiškai anotuotų tekstų redaktorius,
- sintaksiniam anotavimui, redagavimui *UDPipe, Conllu Editor*.

# Projekto darbai (4)

- Parengti vertinimui ir apmokymui reikalingus tekstynus (aukso standartus):
- morfologiškai anotuotas tekstynas MATAS (~1,7 mln. žodžių),
- sintaksiškai anotuotas tekstynas ALKSNIS (~60 000 žodžių).
- Pritaikyti tarptautinius anotavimo standartus – *Universal Dependencies* (UD).

# Projekto darbai (5)

- Tarptautinis anotavimo standartas *Universal Dependencies* (UD):

- kai kurių kalbos dalių, gramatinių kategorijų morfologinio anotavimo skirtumai (sutrumpinimai, santrumpos, įvardžiai, skaitvardžiai);
- sintaksiniam anotavimui reikalingų pažymų rinkinys (daugiau negu 40 pažymų ir jų variacijų), pritaikymas lietuvių kalbos sintaksei.

- Įvertinti anotavimo tikslumą.

# Projekto komanda

- Baltrūnaitė Sabina
- Bielinskienė Agnė
- Boizou Loic
- Brokaitė Kristina
- Dadurkevičius Virginijus
- Dereškevičiūtė Sigita
- Deveikis Viktoras
- Jancaitė-Skarbalė Laima
- Kalpokienė Julija
- Kamandulytė-Merfeldienė Laura
- Kovalevskaitė Jolanta
- Kurtinaitytė Ieva
- Milčiuvienė Saulė
- Mingaudaitė Monika
- Ožeraitis Vytautas
- Razutytė Auksė
- Rimkutė Erika
- Stepšys Jonas
- Vaičenonienė Jurgita
- Žemrietė Miglė