Tantárgy neve/kódja:	BNY-DK-200 Bevezetés a nyelvtechnológiába
Tárgyfelelős:	A doktori iskola vezetője
Tantárgy oktatója:	Ligeti-Nagy Noémi, PhD
Kreditszám:	8 kredit
Óratípus:	szeminárium
A tantárgy céljának rövid ismertetése:	A kurzus gyakorlati áttekintést nyújt az empirikus nyelvészeti elemzésben és a természetesnyelv-feldolgozásban alkalmazott számítógépes módszerekről. Bemutatja a számítógépes nyelvészet, a nyelvtechnológia és az NLP központi fogalmait és technikáit, beleértve a reguláris kifejezéseket, a korpuszépítést, a nyelvi annotációt, a benchmarkokat, a számítógépes szemantikát, a gépi fordítást, a neurális nyelvi modelleket, a nagy nyelvi modelleket és a promptprogramozást. A kurzus az elméleti megbeszéléseket gyakorlati munkával ötvözi. A hallgatók kijelölt fejezeteket és tanulmányokat olvasnak, heti egyéni feladatokat készítenek, valamint csoportprojektben vesznek részt, amely során egy kis annotált korpuszt hoznak létre, és azt nyelvimodell-finomhangolási kísérletekhez használják. A kurzus végére a hallgatók megértik a számítógépes nyelvészet fő módszertani alapelveit, és képesek lesznek egy kis léptékű nyelvtechnológiai projekt megtervezésére, dokumentálására, annotálására, értékelésére és bemutatására.
Elsajátítandó elméleti ismeretanyag:	A hallgatók elméleti ismereteket szereznek a számítógépes nyelvészet és a természetesnyelv-feldolgozás főbb területeiről, beleértve: • a nyelvészet, a számítógépes nyelvészet, a nyelvtechnológia és az NLP kapcsolatát; • a reguláris kifejezéseket és szerepüket a szövegfeldolgozásban; • a korpuszokat, a korpusztervezést, az adatgyűjtést és a benchmarkokat; • az annotációelméletet, annotációs irányelveket és az annotátorok közötti egyezést; • a számítógépes szemantika alapfogalmait, beleértve a szójelentést, szemantikai hasonlóságot, szóértelmezési egyértelműsítést és szemantikai erőforrásokat; • a felügyelt és felügyelet nélküli tanulást a nyelvészeti elemzésben; • a gépi fordítást és a neurális megközelítéseket a nyelvfeldolgozásban; • a nyelvi modellezést, előtanítást, finomhangolást és a nagy nyelvi modelleket; • a prompt engineeringet és a generatív MI-eszközök gyakorlati használatát; • az NLP és az empirikus nyelvtechnológia értékelési módszertani alapelveit.

Elsajátítandó gyakorlati ismeretanyag:	A hallgatók gyakorlati készségeket szereznek az alábbi területeken: • reguláris kifejezések használata szövegfeldolgozási feladatokhoz; • nyelvi adatok gyűjtése, rendszerezése és dokumentálása; • annotációs kategóriák és annotációs irányelvek tervezése; • pilot annotáció végrehajtása és az irányelvek fejlesztése az eredmények alapján; • annotációs konzisztencia és adatkészlet-minőség értékelése; • kis korpusz előkészítése számítógépes nyelvészeti feladathoz; • benchmark adatkészletek és értékelési módszerek értelmezése; • nyelvimodell-finomhangolási vagy kapcsolódó nyelvtechnológiai munkafolyamatok kipróbálása; • kiválasztott NLP-eszközök és online források használata; • rövid kutatási tanulmány írása, amely bemutatja egy nyelvtechnológiai projekt szakirodalmát, módszertanát, adatait és eredményeit; • projektelőrehaladás és végső eredmények szóbeli bemutatása.
A 2-4 legfontosabb kötelező irodalom felsorolása bibliográfiai adatokkal (szerző, cím, kiadás adatai, (esetleg oldalak), ISBN)	Jurafsky, Daniel, és James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* 3. kiadás, tervezet. Kijelölt fejezetek: 2. fejezet 2.1–2.2 szakasz; 3. fejezet 3.1–3.2.1 szakasz; 4. fejezet, Bevezetés; 6. fejezet 1–10. oldal. Online elérhető. Jurafsky, Daniel, és James H. Martin. *Speech and Language Processing.* Korábbi kiadás, 2014-es verzió. 1. fejezet. Wang, Alex, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy és Samuel R. Bowman. 2018. „GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding.” *Proceedings of the 2018 EMNLP Workshop BlackboxNLP*, 353–355. Wang, Alex, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy és Samuel R. Bowman. 2019. „SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems.” *Advances in Neural Information Processing Systems* 32.
A 2-4 legfontosabb ajánlott irodalom felsorolása bibliográfiai adatokkal (szerző, cím, kiadás adatai, (esetleg oldalak), ISBN)	Wissler, Lars, Mohammed Almashraee, Dagmar Monett és Adrian Paschke. 2014. „The Gold Standard in Corpus Annotation.” *Proceedings of the 5th IEEE Germany Student Conference.* Yeomans, Michael, Alejandro Kantor és Dustin Tingley. 2018. „The Politeness Package: Detecting Politeness in Natural Language.” *The R Journal* 10(2): 489–502. Arafa-Hilal, Marwa. 2023. „The Use of Politeness Strategies in Academic Conversations as Represented in a Corpus Linguistics MOOC.” *Journal of Pragmatic Research* 5(1): 85–106. DeepLearning.AI. *ChatGPT Prompt Engineering for Developers.* Rövid kurzus. Ajánlott a promptprogramozásról és a ChatGPT gyakorlati használatáról szóló órához.

Elmélet-gyakorlat aránya:	Elméleti óra óraszáma: 0%	Gyakorlati óra óraszáma: 100%
Az alkalmazott oktatási módszerek:	A kurzus szemináriumi formában zajlik, amely rövid előadásokat, irányított vitát, olvasmányalapú felkészülést, gyakorlati feladatokat és projektalapú tanulást ötvöz. Az alkalmazott módszerek: • bevezető előadások az alapfogalmakról és módszerekről; • kötelező olvasmányok és kiválasztott kutatási tanulmányok megbeszélése; • órai gyakorlati feladatok, beleértve reguláriskifejezés-feladatokat és kis léptékű NLP-tevékenységeket; • heti egyéni feladatok; • csoportmunka korpuszépítésen, annotáción és kísérletezésen; • pilot annotáció és annotációs irányelvek megvitatása; • hallgatói prezentációk benchmark adatkészletekről és projektelőrehaladásról; • vendégelőadás a gépi fordításról; • ChatGPT és prompt engineering irányított kipróbálása; • a projekt eredményeinek záró szóbeli bemutatása. A hallgatóknak notebookot/laptopot kell hozniuk az órákra, mivel egyes alkalmak gyakorlati számítógépes munkát tartalmaznak.	
Az értékelés módja:	Az értékelés folyamatos félévi munkán és egy záró csoportprojekten alapul. A hallgatók rendszeres egyéni feladatokat készítenek a szemeszter során, körülbelül heti egy feladatot. A záróprojekt egy kis korpusz létrehozásából áll egy kiválasztott számítógépes nyelvészeti feladathoz, az annotációs folyamat dokumentálásából, az adatkészlet és/vagy a modell eredményeinek értékeléséből, valamint a munka bemutatásából. A végső beadandó egy 4–6 oldalas tanulmány, amely összefoglalja a releváns szakirodalmat, a projekt módszertanát, az adat- és annotációs folyamatot, valamint az elért eredményeket. A hallgatók a szemeszter során projektelőrehaladási beszámolókat is tartanak, és a kurzus végén záró prezentációt adnak elő.	
Az értékelés kritériuma:	Osztályzás háromfokú skálán: - 80-100% = Kiválóan megfelelt (5) - 60–79% = Megfelelt (3) - 0–59% = Nem felelt meg (1)	
Miként járul hozzá a tantárgy a doktori iskola képzési programjában tanulási eredményként megjelölt MKKR 8. szintű tanulási eredmények eléréséhez.	Tudás: A tantárgy doktori szintű tudást nyújt a számítógépes nyelvészet és az empirikus nyelvtechnológia elméleti alapjairól és aktuális módszereiről. A hallgatók megtanulják, hogyan operacionalizálhatók a nyelvészeti kérdések számítási feladatokká, hogyan épülnek fel a korpuszok és benchmarkok, valamint hogyan értékelik a nyelvi modelleket és NLP-rendszereket. A kurzus megismerteti a hallgatókat a neurális nyelvi modellezés, finomhangolás és nagy nyelvi modellek legújabb fejleményeivel is, lehetővé téve számukra, hogy saját kutatásukat a kortárs számítógépes megközelítések kontextusában helyezzék el. Képességek: A kurzus fejleszti a hallgatók képességét egy kis léptékű empirikus nyelvtechnológiai projekt megtervezésére és	

	megvalósítására. A hallgatók megtanulják a nyelvi adatok gyűjtését és strukturálását, annotációs kategóriák kialakítását, annotációs irányelvek írását, pilot annotáció végzését, annotációs minőség értékelését, benchmark adatkészletek értelmezését, valamint módszertani döntések tudományos formában történő bemutatását. A zárótanulmány és prezentáció erősíti a kutatási terv, adatok, módszerek és eredmények világos tudományos kommunikációját. Attitűdök: A tantárgy kritikus és reflektív hozzáállást alakít ki a számítógépes módszerekkel kapcsolatban a nyelvészeti kutatásban. A hallgatókat arra ösztönzi, hogy értékeljék a korpuszok, annotációk, benchmarkok, nyelvi modellek és generatív MI-rendszerek megbízhatóságát, korlátait és torzításait. A kurzus támogatja az interdiszciplináris módszerek iránti nyitottságot is, ötvözve a nyelvészeti szakértelmet a számítógépes gondolkodással, empirikus értékeléssel és együttműködésen alapuló kutatási gyakorlattal. Felelősség és autonómia: A kurzus támogatja a doktori szintű autonómiát azáltal, hogy a hallgatóknak önálló döntéseket kell hozniuk az adatválasztásról, annotációtervezésről, módszertani döntésekről és az eredmények értelmezéséről egy felügyelt projektkeretben. A hallgatók felelősséget vállalnak munkájuk minőségéért, dokumentálásáért és reprodukálhatóságáért, miközben társaikkal együttműködnek egy közös korpuszépítési és annotációs feladaton. A heti előrehaladási jelentések és a zárótanulmány révén elsajátítják az átlátható, felelősségteljes és kutatásorientált projektmunka gyakorlatát.