FELEZŐ BÁLTÓL A PARANCSPULI-IG

YANG ZIJIAN GYŐZŐ | HUN-REN Nyelvtudományi Kutatóközpont,
Pázmány Péter Katolikus Egyetem
yang.zijian.gyozo@nytud.hun-ren.hu | DOI: 10.18135/PG70.2024.18

Majdnem napra pontosan 14 évvel ezelőtt, 2010-ben a Felező bálon ismertem meg Gábort, aki bevezetett a nyelvtechnológia világába. Azóta a témavezetőm. Együtt számtalan témát körbejártunk, a kínai karakterek okozta problémáitól kezdve, a gépi fordításon és transzformereken át a PULI-modellekig. Bár sokszor a világot kevésbé érdekli, de számos területen sikerült a magyar nyelvre *state-of-the-art* eredményeket elérnünk. Jelenleg az egyik zászlóshajónk a PULI. Sikere töretlen, amit bizonyít a számos tudományos és sajtómegjelenés, és emellett a PULI-modellek nevei védjegyekké is válnak. A jelen tanulmány ezt a sikertörténetet tekinti át a szerző szemszögéből.

1. Bevezetés

A Gáborral való találkozást mondhatnám véletlennek vagy sorsszerűnek, de talán a gondviselés a legjobb szó rá. A pszichológia, a nyelvek és a robotika érdekelt. Végül a Pázmány Péter Katolikus Egyetem Információs Technológiai és Bionikai Kar mérnök informatikus szakára vettek fel. Elég hamar be kellett látnom, hogy a fizika nem az erősségem. Több tárgyból is megbuktam, és évet is kellett ismételnem. Egy csoporttársammal eldöntöttük, hogy egyetemet váltunk. Már az átjelentkezésemet intéztem, amikor kaptam egy meghívót a Felező bálra. Kiderült, hogy a bent maradás miatt annyi kreditet sikerült teljesítenem melléktárgyakból, hogy meghívtak a bálra, ami a hagyomány szerint egy vacsora a Mongol étteremben a tanárokkal. Mivel nem „szabályosan" értem el a félidőt, nem akartam elmenni a vacsorára. De az egyik csoporttársam megkért, hogy kísérjem el, mivel ő sem ismert sok embert, de szeretett volna egy jót vacsorázni. A jó vacsora reményében beadtam a derekam. Már a desszertnél tartottunk, amikor mind a ketten meguntuk az eseményt, és elhatároztuk, hogy hazamegyünk.

Ekkor jött oda Prószéky Gábor Levendovszky Jánoshoz (nála is megbuktam neurális hálózatokból) megbeszélni egy projektet. Miután megbeszélték, Levendovszky tanár úrnak haza kellett mennie, így ott maradt Gábor velünk. Elkezdte kérdezni, hogy ki milyen szakos, és hogy valakit érdekel-e a nyelvtechnológia. Félbeszakítottam, és megkérdeztem tőle, hogy mi is az a nyelvtechnológia. A bál után meg is kerestem azonnal Gábort. Bizonyítékul szolgáljon az első levelem Gáborhoz:

„Tisztelt Prószéky Gábor tanár úr!

Yang Zijian Győző vagyok a Pázmány Itk kar hallgatója, és még a felező bálon találkoztunk. Önálló laboratóriumra szeretnék keresni valami témát, és érdeklődni szeretnék, hogy esetleg fordulhatnék-e önhöz, és kérni a közeljövőben (holnap, holnap után) egy időpontot, vagy mikor találhatom meg önt az karon.

Köszönettel:
Yang Zijian Győző"

Kaptam is egy témát tőle, majd még azon a nyáron elmentem a MorphoLogichoz szakmai gyakorlatra, Gábor lett a témavezetőm és a többi már történelem...

2. A kínai írás okozta néhány probléma informatikai megoldása

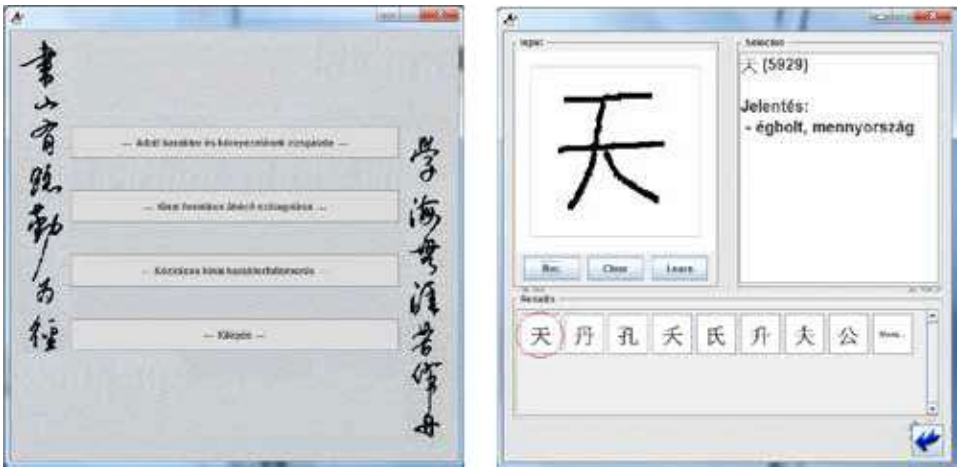
A BSc-témám, Gábornak köszönhetően, a kínai írás okozta problémák megoldása volt. A szakdolgozatom, amivel TDK-n is részt vettem, három problémát igyekszik megoldani egy programcsomag (lásd 1. ábra bal oldala) segítségével. Íme a dolgozat témája a szakdolgozattéma-bejelentésből:

1. **„Egy adott kínai karakter és környezete vizsgálata:** A program célja, hogy megvizsgálja az adott karakter környezetét, hogy alkot-e egy másik karakterrel értelmes kifejezést. A kínai nyelvben léteznek két vagy több karakterből álló szavak, kifejezések, amelyek nyelvtani vagy egyéb okokból különváltak, és ez a program arra ad lehetőséget, hogy ezeket a különvált kifejezéseket a környezetet bejárva megtalálja."
2. „**A kínai fonetikus ábécé szerinti szótagoló program elkészítése:** A kínai karakteres írásban nincsenek feltüntetve szóhatárok, így ha egy mondatot, vagy kifejezést fonetikus ábécé segítségével átírunk, szintén nem lesznek láthatóak a szóhatárok. A szótagoló program a kínai nyelvtannak megfelelően szótagolja a fonetikus ábécével írt szöveget."
3. **„Kézzel írt kínai karakter felismerése:** A feladat adott rajzoló felületen, egérrel rajzolt kínai karakter felismerése neurális hálózat segítségével. A program egy a felhasználó által egérrel rajzolt képet alakít át feldolgozható formába, majd neurális hálózat segítségével asszociál egy adott mintahalmazból egy létező kínai karakterre."

A harmadik feladat egész jól sikerült (lásd 1. ábra jobb oldala), a Kohonen-hálózatot[1] alkalmaztam a probléma megoldására. Azért kaptam kritikát is; szerencsésnek mondhatom magam, hogy Bartos Hubát kaptam bírálónak, aki így kezdte és fejezte be az egyik feladat bírálatát: „A 2. probléma esetében a kiinduló problémafelvetésben van

[1] https://www.nnwj.de/kohonen-feature-map.html

furcsaság. ... A szóhatárok megállapításának problémája ennél sokkal nagyobb és lényegesebb probléma, mind az írásjegyes, mind a hangjelölő írásmódban, ennek a megoldásához viszont a javasolt program nem visz közelebb." A védésen még nem így volt, de utólag már teljesen egyetértek ezzel a bírálattal.



1. ábra. A program főmenüje és égbolt

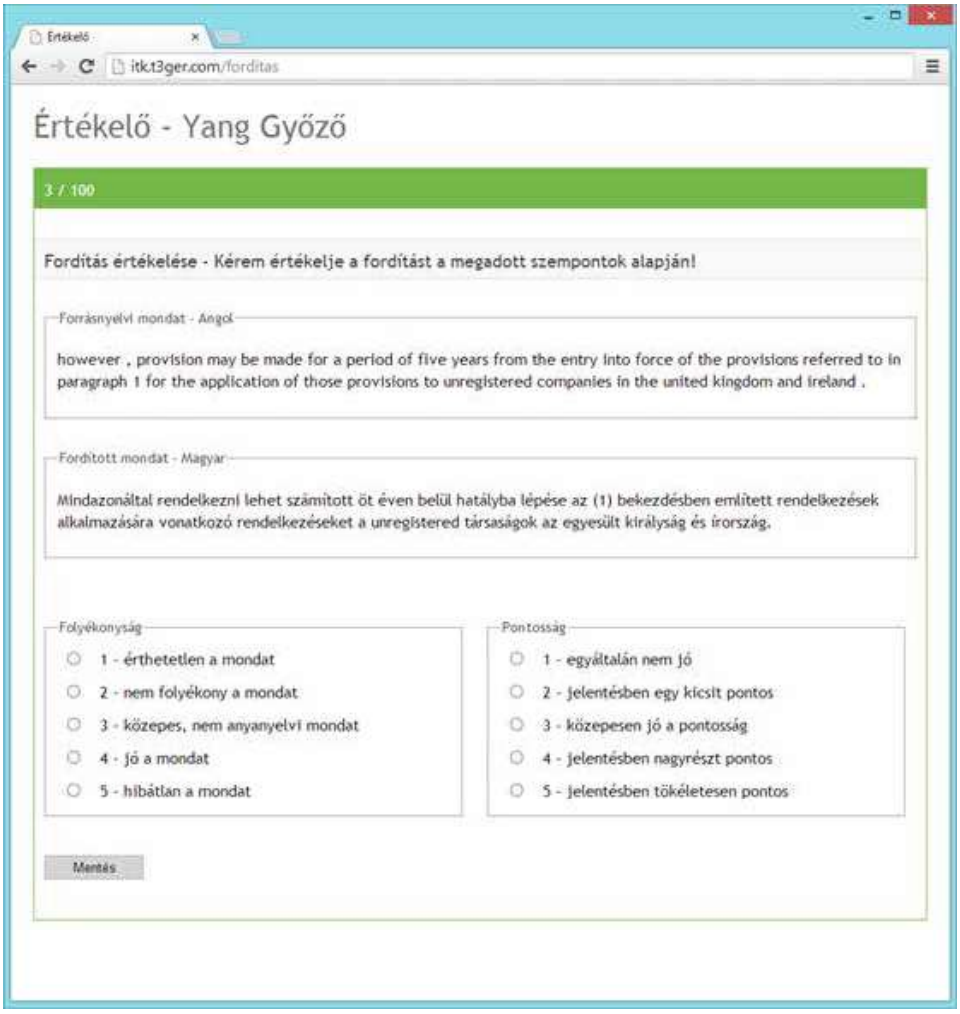
3. Hierarchikus modell alapú gépi fordítás minőségének becslése

A mester éveim alatt kezdtem el azt a témát, ami kitartott a doktori iskola végéig, itt kaptam Gábor mellé egy második konzulenst, Laki Lászlót, akihez azóta is szoros barátság fűz. A diplomatervemben ez állt:

„A dolgozatom során felépítettem egy kifejezés alapú illetve egy hierarchikus gépi-fordító-rendszert angol–magyar nyelvpárra. A gépi fordítókat betanítottam a Hunglish korpusszal (Halácsy et al., 2005), majd lefordíttattam vele egy tesztszöveget, és kiértékeltem a minőségét referenciafordítással történő és referenciafordítás nélküli kiértékelő módszerrel. A gépi fordító felépítéséhez a Moses keretrendszert (Koehn et al., 2007), a kiértékeléshez a Moses és a QuEst (Specia et al., 2013) keretrendszert használtam."

„A munkám során felépítettem a QuEst keretrendszert magyar nyelvre, és kiértékeltem vele az általam felépített kifejezés alapú és hierarchikus modellen alapuló gépi fordító által generált kimenetet. Az angol–spanyol nyelvpár esetén létezik 17 alaptulajdonság, ami elég a kiértékeléshez. Ez alapján én is optimalizáltam magyar nyelvre a QuEst rendszerét. Az általam optimalizált rendszer 23 alaptulajdonsággal dolgozik."

„A QuEst kiértékeléshez szükség volt emberi értékelésekre, amihez készítettem egy értékelő weboldalt (lásd 2. ábra), amit a doktori témámhoz is felhasználtam."



2. ábra. T3ger kiértékelő felülete

4. επQue: Gépi fordítás minőségét becslő programcsomag

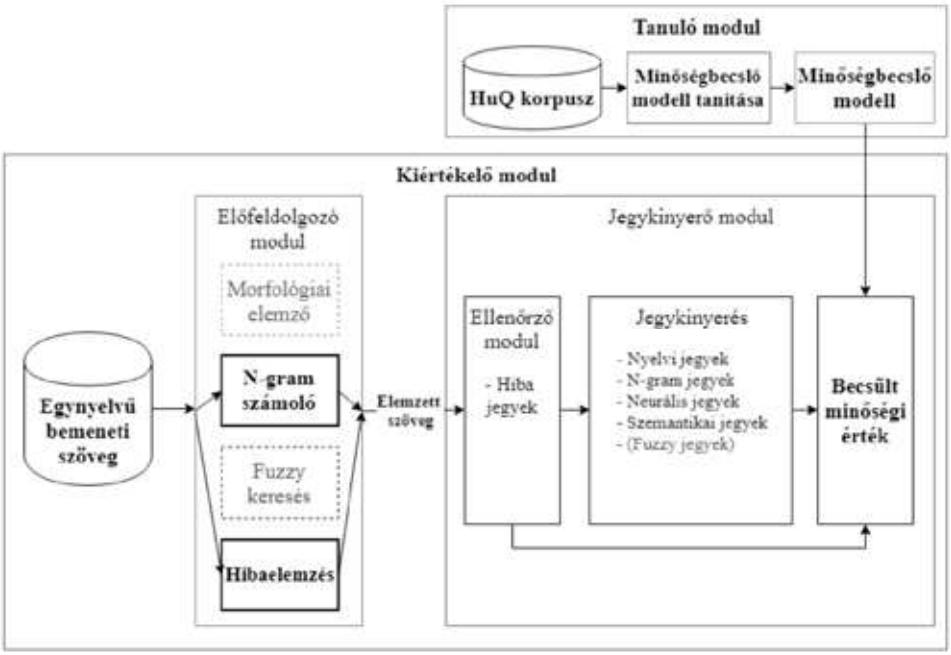
A Pázmányon a kutatócsoport a 314-es szobában volt, ezért a π fontos eleme lett a kutatásaimnak. Hamar kitaláltam, hogy lesz egy πRate nevű valami a kutatásomban, ezért elkezdtem a kalózos témakört felépíteni. A PhD témám a gépi fordítás minőségbecslése volt. A doktori disszertációm (Yang, 2019) három nagyobb témát ölel fel, idézet a tézisfüzetemből:

1. **A Hun-QuEst rendszer és a HuQ korpusz:** „A minőségbecslés módszere gépi tanuláson alapszik. A modell, jegyek segítségével, a forrásnyelvi és a gép által lefordított mondatokból különböző nyelvfüggetlen és nyelvspecifikus minőségi mutatószámokat nyer ki. Majd a mutatószámok gépi tanuló algoritmussal betanítottam emberi kiértékelésekre. A modell tanításához tanítókorpuszra lenne szükség, azonban a kutatás ideje alatt nem állt rendelkezésre angol–magyar nyel-

vű emberi kiértékeléssel rendelkező párhuzamos korpusz. Ezért az angol–magyar minőségbecslő rendszer tanításához létrehoztam egy kézzel kiértékelt tanítókorpuszt (Yang et al., 2016). Ennek segítségével létrehoztam egy angol–magyar minőségbecslő rendszert. A felépített rendszeren különböző méréseket végeztem el. Először az angol–spanyol nyelvre optimalizált alapjegykészletet mértem le angol–magyar nyelvre, majd megvizsgáltam a Specia és társai (Specia et al., 2013) által implementált 76 jegykészletet is. Ezt követően saját szemantikai jegyekkel kísérleteztem (Yang et al., 2018). A szemantikai jegyekhez egy angol–magyar szótárat, a WordNetet, a szóbeágyazási modellt és a látens szemantikai analízis módszerét használtam. Végeztem jegykiválasztást is, ami azt jelenti, hogy kevesebb releváns jeggyel sikerült további eredményjavulást elérnem. Ezáltal, kevesebb erőforrással magasabb minőséget értem el. A WordNet jegyeket angol–spanyol és angol–német nyelvpárokra is kipróbáltam. Mindkét esetben jobb eredményt értem el az alapjegykészlethez képest."

2. **A MaTros rendszer:** „A második kutatás során a minőségbecslés módszerét használtam a különböző gépi fordítórendszerek kimeneteinek kombinálására (Laki & Yang, 2018). Az általam létrehozott kompozit rendszer egy kifejezés alapú statisztikai, egy hierarchikus statisztikai és egy neurálishálózat-alapú gépi fordítórendszer kimenetét kombinálja. A rendszer a minőségbecslés módszerével kiválasztja a három rendszer fordításából a legjobb fordítást, és az lesz a rendszer végső kimenete. A módszeremet négy különböző nyelvpárra teszteltem: angol–magyar, angol–német, angol–olasz és angol–japán. Az eredmények alapján rendszerszinten a kompozit minőségbecslés módszerével gépi fordítórendszerem minden esetben jobb minőséget eredményezett, mint az általa felhasznált rendszerek önmagukban. Angol–magyar nyelvpár esetében nyelvfüggő jegyekkel tovább tudtam növelni a rendszer minőségét."
3. **A πRate rendszer (lásd 3. ábra):** „A harmadik kutatásban a minőségbecslés módszerét kiterjesztettem egynyelvű szövegek minőségének becslésére. A kutatás célja az volt, hogy megvizsgáljam az interneten elérhető, emberek által produkált szövegek hibáit, valamint a minőségbecslés módszerével, létrehozzak egy automatikus hibadetektáló programot. Az kutatásom arra mutatott rá, hogy az emberek által létrehozott egynyelvű szövegek, a gépi fordítók által generáltakkal ellentétben, nagyrészt nem nyelvtani hibákat tartalmaznak (Dömötör & Yang, 2018). Ezen szövegek minőségi problémái inkább az internetezők írási szokásaiból adódnak, mint például az ékezetek vagy az írásjelek elhagyása. Az általam létrehozott, egynyelvű minőségbecslő rendszer (Yang & Laki, 2017) jól alkalmazható a korpusznyelvészetben vagy a természetesnyelvi elemző rendszerek előfeldolgozó moduljaiban."

Idézem Gábort a védésemről: „...az is jó tudós, aki magától mindent kitalál, de az, aki a témával alázatos és tud a szakmába úgy belemerülni, hogy közben felveszi azokat az információkat, amiket mások mondanak..." A védésen is és utólag is teljesen egyetértek vele.



3. ábra. A πRate rendszer architektúrája

5. Nyelvészeti Diákolimpia

2014-ben Kínában rendezték meg a Nemzetközi Nyelvészeti Diákolimpiát [2]. Gábor felkért, hogy legyek a Magyar Nyelvészeti Diákolimpia csapatának a kísérője. Rendkívül megtisztelő feladat volt számomra, három éven keresztül kísértem a csapatot, azóta is a szervező csapat tagja vagyok. A magyar csapat évről évre sikeresebb eredményeket ér el:

2013: 1 dicséret
2014: 1 dicséret
2015: 1 bronz
2016: 1 bronz, 2 dicséret
- 2017: 2 dicséret
- 2018: 1 ezüst, 2 bronz
- 2019: 2 bronz, 2 dicséret, 1 legjobb megoldás különdíj
- 2021: 1 ezüst, 1 bronz, 2 dicséret
- 2022: 3 bronz, 1 legjobb megoldás különdíj
- 2023: **Csapatbronz!,** 1 ezüst, 1 bronz

A 4. ábrán látható az eddigi egyik legnagyobb sikerünk. Érdekesség, hogy mára már hagyománnyá vált, hogy a csapat a Gábor házán lévő zászlót viszi minden évben (Kína óta) a nemzetközi döntőre.

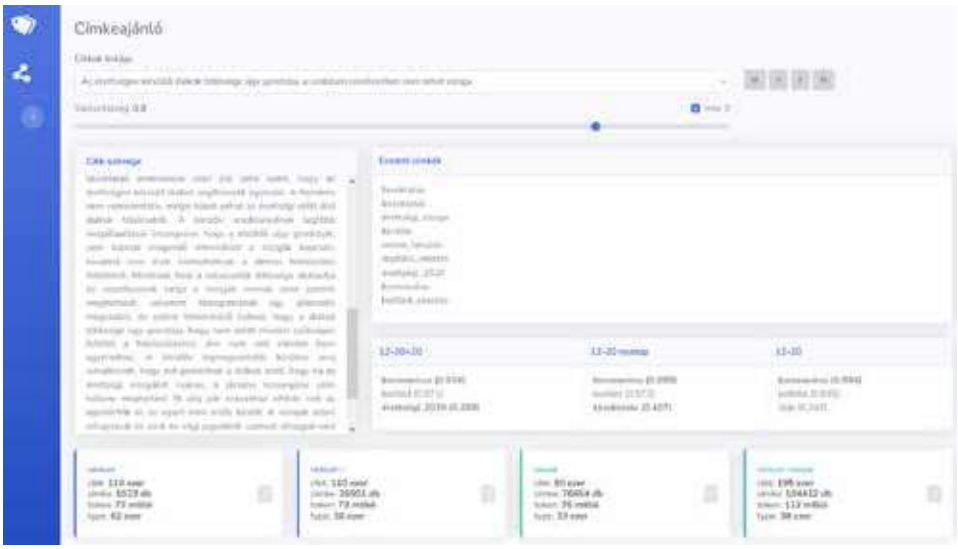
[2] https://ioling.org

4. ábra. Az első csapatbronzérem és a világot megjárt zászló

6. MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport

A doktori cím megszerzése után pár évig még a Pázmányon maradtam az MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoportban. A kutatási irányomat meghatározta a neurális hálózatok térhódítása. Szerencsém volt olyan kutatásokban részt venni, mint:

– Szövegek stílusának meghatározása neurális hálózattal (Dömötör, et al., 2022): a kutatásunkban a fastText (Joulin et al., 2017) eszközzel osztályoztunk különböző típusú (irodalmi, tudományos, jogi, sajtó, beszélt nyelvi, komment) szövegeket.
– Ékezet-visszaállítás (Laki & Yang, 2020): az ékezet-visszaállító rendszer tanításához egy neruálishálózat-alapú gépifordító-rendszert használtunk, amely transzformermodellt használ. A módszert kiterjesztettük 14 különböző nyelvre, a rendszerünk minden nyelvre több mint 98%-os pontossággal tudta visszaállítani az ékezet nélküli szavakat.
– Zéró kopulák automatikus felismerése (Dömötör et al., 2020): a kutatásunk során létrehoztunk egy eszközt, amely a zéró kopulás mondatok automatikus felismerésére és a zéró kopulák mondatba való beillesztésére alkalmas. In-domain tesztkorpusz esetén az eszköz közel 90%-os pontossággal tudta megfelelő helyre beilleszteni a zéró kopulát.



5. ábra. Címkeajánló rendszer

– Kulcsszógenerálás sajtószövegek számára (Yang et al., 2020): Létrehoztunk egy címkézőrendszert, amellyel sajtószövegek automatikus tematikus címkézését tudjuk megvalósítani. Ez volt az első komoly megbízásom Gábortól (lásd 5. ábra).

A nyelvtechnológia közben elmozdult a transzformerek (Vaswani et al., 2017) irányába. Így a kutatási témám is elmozdult:

– Transzformeralapú gépi fordítás (Laki & Yang, 2022): létrehoztuk a legjobb minőségű angol–magyar neurális gépifordító-modelleket. Kísérleteztünk a Marian keretrendszerrel (Junczys-Dowmunt et al., 2018), az mBART (Liu et al., 2020), az mT5 (Xue et al., 2021) és az M2M100 (Aharoni et al., 2019) modellekkel. A kutatásainkban megmérettettük az általunk tanított modelleket olyan ipari gépifordító-rendszerekkel, mint a Google Fordító, a Bing/Microsoft Fordító, a Yandex Fordító, az eTranslation vagy a DeepL.
– Szöveg-összefoglaló (Yang, 2022): létrehoztuk az első magyar nyelvű neurális extraktív és absztraktív szöveg-összefoglaló modelleket. Az összefoglaló modellek elkészítéséhez különböző BERT (Devlin et al., 2019) alapú modelleket használtunk. Az absztraktív modellekhez az előre betanított többnyelvű BERT modellt, valamint a magyar egynyelvű huBERT Base (Nemeskey, 2021), valamint a HILBERT Large (Feldmann et al., 2021) modelleket használtuk. Továbbá végeztünk transzfertanítást is.

Ez idő alatt lehetőséget kaptam, hogy Gábor mellett társtémavezető legyek, és betekintést nyerjek az arab nyelvű neurális szöveg-összefoglalás témájába (Kahla et al., 2021, 2022) (lásd 6. ábra).

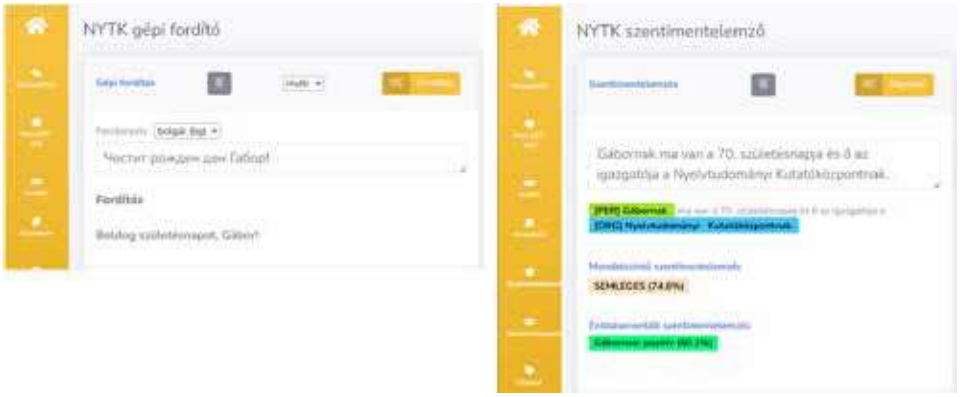
6. ábra. Arab absztraktív szöveg-összefoglalás

7. Nyelvtudományi Kutatóközpont

Mikor Gábort kinevezték a Nyelvtudományi Kutatóközpont (akkor még ELKH, most már HUN-REN előtaggal) igazgatójának, felajánlotta, hogy ott folytassam a kutatásomat. Már a személyére sem tudtam volna nemet mondani, de bedobta, hogy a kutatóközpont hamarosan beszerez két szuperszámítógépet: egy 8 darab NVIDIA A100 (80GB) GPU-ból álló (*bolka*) és egy 4 darab NVIDIA A100 (80GB) GPU-ból álló (*lolka*) szuperszámítógép *node*-ot.

Az első feladatom egy ELECTRA modell (Clark et al., 2020) tanítása volt (Yang & Váradi, 2021). Majd ahogy a transzformernyelvmodell-alapú megoldások átvették a vezetést a nyelvtechnológiában, úgy fordult a fókuszom a nyelvmodellek, illetve a nyelvmodelleken alapuló alkalmazások és kutatások irányába. Létrehoztuk magyar nyelvre a state-of-the-art:

– neurális szöveg-összefoglaló modelleket (Yang, 2022),
– neurális gépi fordítás minőségbecslő modelleket (Yang & Laki, 2023),
– neurális szentiment- és entitásorientált véleményelemző modelleket (Laki & Yang, 2023) (lásd 7. ábra jobb oldala),
– első neurális morfológiai generátor modelleket (Laki et al., 2023),
– magyarcentrikus többnyelvű (12 különböző nyelvről magyarra) gépifordító-modelleket (Laki & Yang, 2023) (lásd 7. ábra bal oldala).



7. ábra. Gépi fordító és szentimentelemző demó

Mindeközben, ahogy megérkezett a szuperszámítógép, elindítottuk a nyelvmodellezési kutatásokat. Kezdetben kisebb modellek előtanításával kísérleteztünk (Yang & Váradi, 2024), mint ELECTRA, BART (Lewis et al., 2020), RoBERTa (Liu et al., 2019) vagy GPT-2 (Radford et al., 2019), valamint különböző nyelvmodellekkel különböző finomhangolásos (Yang & Laki, 2023), promptolásos (Yang & Ligeti-Nagy, 2023) és alkalmazási (Hatvani et al., 2023) kísérleteket végeztünk. És így szépen elkezdtünk haladni a nagyobb nyelvmodellek irányába...

8. PULI-modellek és *benchmarkok*

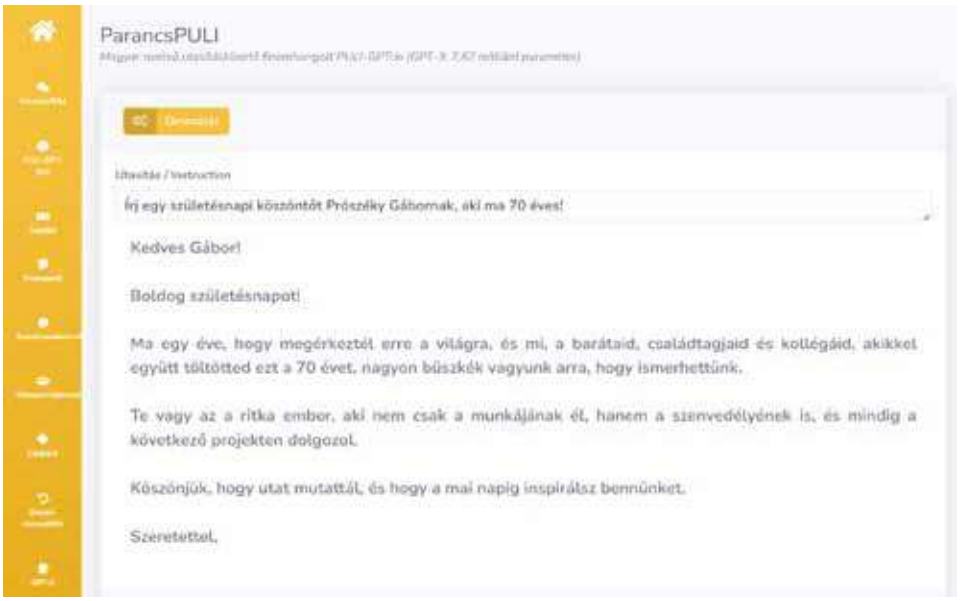
Jelenleg a HUN-REN Nyelvtudományi Kutatóközpont Nyelvtechnológiai kutatócsoportjának egyik zászlóshajója lett a PULI. Szerencsésnek mondhatom magam, hogy egy ilyen nagy presztízsű kutatást vihetek.

A PULI-család mára már több különböző méretű előtanított és finomhangolt nyelvmodellekből áll:

– **PULI BERT-Large** és **PULI GPT-2** (Yang et al., 2023): Mindkettő modell 345 millió paraméteres. A PULI BERT-Large modellünk a legtöbb feladatban felülmúlta a huBERT modellt.
– **PULI GPT-3SX** (Yang et al., 2023): 6,7 milliárd paraméteres GPT-NeoX (Black et al., 2022) típusú magyar egynyelvű nagy nyelvi modell. Több mint 32 milliárd magyar szavas korpuszon tanult.
– **PULI GPTrio** (Yang et al., 2023): 7,67 milliárd paraméteres háromnyelvű (magyar–angol–kínai) GPT-NeoX típusú nagy nyelvi modell. Több mint 41 milliárd magyar szót látott. Összesen több mint 150 milliárd szavas korpuszon tanult, a cél az volt, hogy a magyar nyelv viszonylag kiegyensúlyozott módon szerepeljen a korpuszban, de transzfertanulással kamatoztatni tudjon a többi nyelv adta tudásból.
– **ParancsPULI** (Yang et al., 2024): Létrehoztuk az első magyar nyelvű, finomhangolt, utasításkövető (*instruct,* Ouyang et al., 2022) nagy nyelvi modellt, amely egyaránt képes kérdésekre/utasításokra válaszolni (lásd 8. ábra) és nyelvtechnológiai felada-

tokat megoldani. Jelenleg több mint 6000 magyar nyelvű, összesen több mint 114 ezer prompttal finomhangoltuk.

– **PULI LlumiX 32K:** A 7 milliárd paraméteres LLaMA-2 (Touvron et al., 2023) modellt a Together[3] finomhangolás útján kibővítette a bemeneti kontextus hosszát 32 ezerre[4]. Majd ezt a modellt támogattuk meg több mint 760 ezer darab, 5000 szónál hosszabb magyar dokumentummal.



8. ábra. ParancsPULI köszöntője

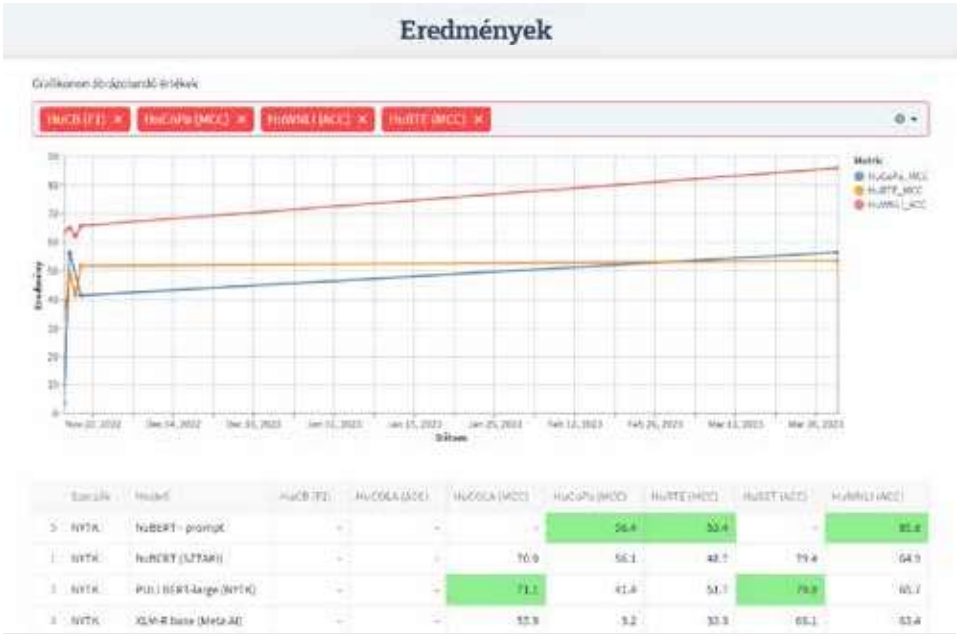
A PULI-modellek mellett a másik zászlóshajó a HuLU (Hungarian Language Understanding Benchmark Kit, Ligeti-Nagy et al., 2022, 2023). A HuLU a legnagyobb magyar nyelvű nyelvtechnológiai *benchmark,* ami jelenleg 7 korpuszból áll. Íme a korpuszok leírása a kiértékelő oldalról[5]:

– **HuCB – A CommitmentBank Corpus magyar változata:** A HuCommitmentBank olyan rövid szövegrészletekből áll, amelyekben legalább az egyik mondat tartalmaz egy alárendelő mellékmondatot, amely egy logikai következtetést semlegesítő operátor alá tartozik szintaktikailag. Az adatbázisban a premissza a teljes szövegrészlet, a hipotézis pedig a beágyazott tagmondat. A következtetési feladatban azt kell eldönteni, hogy a szöveg írója milyen mértékben elkötelezett a mellékmondat igazsága mellett. A korpusz 250-250 példás tanító, illetve teszthalmazból, és egy 103 példát tartalmazó validációs halmazból áll.
– **HuCOLA – Elfogadhatósági ítéletek korpusza:** A korpusz 9076 magyar mondatot

[3] https://www.together.ai
[4] https://huggingface.co/togethercomputer/LLaMA-2-7B-32K
[5] https://hulu.nytud.hu

tartalmaz, amelyek elfogadhatóságuk, grammatikalitásuk alapján vannak 0-val (nem grammatikus magyar mondat) és 1-gyel (grammatikus magyar mondat) címkézve. A mondatok két annotátor gyűjtötte 3 nyelvészeti szakirodalomból. Mindegyik mondatot négy annotátor annotálta. A végső címke a többségi címke. A tanító-, validációs és tesztanyag aránya 80% (7276 mondat), 10% (900 mondat) és 10% (900 mondat).

– **HuCoPa – A hihető alternatívák korpusza:** A korpusz 1000 példát tartalmaz. Mindegyik példában egy premissza és két alternatíva található. A feladat, hogy kiválasszuk az alternatívák közül azt, amelyikben a leírt helyzet ok-okozati összefüggésben van a premisszában leírt helyzettel. A korpusz előállításához az eredeti angol CoPA korpusz példáit fordítottuk és újraannotáltuk. A tanító-, validációs és teszthalmaz 400, 100, illetve 500 példát tartalmaz.
– **HuRTE – Következtetések felismerésének korpusza:** A korpuszban 4 504 példa található. Minden példa tartalmaz egy (néha több mondatos) premisszát és egy egymondatos hipotézist, és a feladat annak eldöntése, hogy az előbbiből következik-e az utóbbi vagy sem. A korpusz a GLUE benchmark részét képező RTE-adatbázisok példáinak fordításával és újraannotálásával jött létre. A tanító-, a validációs és a teszthalmaz 2131, 242 és 2131 példát tartalmaz.
– **HuSST – A Stanford Sentiment Treebank magyar változata:** A korpusz 11 683 mondatot tartalmaz. Mindegyik mondat szentimentjét egy háromfokú skálán címkéztük. A korpusz előállításához az SST korpusz mondatait fordítottuk és újraannotáltuk. A tanító-, validációs és teszthalmaz 9347, 1168, illetve 1168 mondatot tartalmaz.
– **HuWNLI – Anafora-feloldási korpusz:** A korpuszban az anafora-feloldás mondatpár-osztályozási feladatként, a két mondat közötti logikai következtetés meghatározásaként szerepel. Az alapja a HuWS korpusz, amely az eredeti angol Winograd-sémák magyarra fordított és manuálisan kurált példáit tartalmazza. Az NLI formátum létrehozásához a többértelmű névmásokat lecseréltük mindkét lehetséges referensükkel. A Winograd-sémákból képzett mondatpárokat kiegészítettük a GLUE WNLI adatbázisának többi mondatpárjával. Az adatokat tanító- (562), validációs (59) és teszthalmazra (134) osztva adjuk közre.
– **HuRC – Hungarian Corpus for Reading Comprehension:** A korpusz 80 614 példát tartalmaz. Minden példa egy leadből, egy szövegtörzsből és egy maszkolt cloze stílusú kérdésből áll. A feladat az, hogy kiválasszuk a szövegtörzsből azt az entitást, amelyet a kérdésből kimaszkoltak. Az adatokat automatikusan gyűjtöttük a Népszabadság online híreiből (nol.hu).



9. ábra. HuLU – Dicsőséglista

A HuLU mellett létrehoztuk az első magyar nyelvű utasításkövető finomhangoló korpuszt (Yang et al., 2024), melyet a Stanford Alpaca (Taori et al., 2023) korpusz magyarra fordított változatából és magyar kultúrkörre szabott lokalizált promptokból állítottunk össze. Összesen 2000 lefordított és 100 lokalizált promptot tartalmaz a gyűjtemény. Ez a korpusz képezi az alapját a jelenlegi ParancsPULI modellünknek, ami/aki folyamatosan fejlődik és ami/aki külön verses kötettel készült. Fontos megjegyezni, hogy a PULI által írt/generált műben csak az affiliációt írtuk be manuálisan, minden más részét maga generálta.

8. Konklúzió és továbblépési lehetőségek

Gáborral közösen nagy utat tettünk meg a nyelvtechnológia területén a statisztikai módszerektől kezdve a neurális modellekig. A Gáborhoz közel álló gépi fordítás témájától kiindulva mostanra a nagy nyelvi modellek területen is vezető pozíciót sikerült elérni Magyarországon.

Terveink szerint következő lépésként 13 és 70 milliárd paraméteres nagy nyelvi modelleket szeretnénk tanítani, valamint szöveg-hang alapú multimodális modellek előtanítását fogjuk elkezdeni.

Köszönetnyilvánítás

Köszönöm, Gábor, az eddigi 14 éves ismeretséget és szakmai vezetést. Idézve a múltból: „Mindenekelőtt szeretnék köszönetet mondani témavezetőmnek, Dr. Prószéky Gábornak, akitől mind szakmailag, mind emberileg számtalan támogatást kaptam az elmúlt évek során. Köszönöm Neki, hogy mindvégig baráti közvetlenséggel fordult felém. Nélküle ez a munka nem jöhetett volna létre." Ezúton (is) szeretnék nagyon **Boldog Születésnapot** kívánni!

Irodalom

Aharoni, R., Johnson, M., & Firat, O. (2019, June). Massively Multilingual Neural Machine Translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 3874–3884). Minneapolis: Association for Computational Linguistics.

Black, S., Biderman, S., Hallahan, E., Anthony, Q., Gao, L., Golding, L., He, H., Leahy, C., McDonell, K., Phang, J., Pieler, M., Prashanth, U. S., Purohit, S., Reynolds, L., Tow, J., Wang, B., & Weinbach, S. (2022). GPT-NeoX-20B: An open-source autoregressive language model. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models* (pp. 95–136), virtual+Dublin. Association for Computational Linguistics.

Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. (2020). ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *International Conference on Learning Representations.*

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Minneapolis: Association for Computational Linguistics.

Dömötör A., & Yang Z. Gy. (2018). Így irtok ti: nem sztenderd szövegek hibatipusainak detektálása gépi tanulásos módszerrel. In Tanács A., Varga V. & Vincze V. (szerk.), *XVII. Magyar Számitógépes Nyelvészeti Konferencia* (pp. 305–316). Szeged: Szegedi Tudományegyetem, Informatikai Tanszékcsoport.

Dömötör, A., Kákonyi, T., & Yang, Z. Gy. (2022). What's Your Style? Automatic Genre Identification with Neural Network. *Computación y Sistemas, 26.*

Dömötör, A., Yang, Z. Gy., & Novák, A. (2020, May). Much Ado About Nothing – Identification of Zero Copulas in Hungarian Using an NMT Model. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., & Piperidis, S. (Eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 4802–4810). Marseille: European Language Resources Association. Retrieved from https://aclanthology.org/2020.lrec-1.591

Feldmann A., Hajdu R., Indig B., Sass B., Makrai M., Mittelholcz I., Halász D. & Váradi T. (2021). HILBERT, magyar nyelvű BERT-large modell tanítása felhő környezetben. In Berend G., Gosztolya G. & Vincze V. (szerk.), *XVIII. Magyar Számitógépes Nyelvészeti Konferencia* (pp. 29–36). Szeged, Magyarország: Szegedi Tudományegyetem, Informatikai Intézet.

Halácsy P., Kornai A., Németh L., Sass B., Varga D., Váradi T. & Vonyó A. (2005). A Hunglish korpusz és szótár. In Alexin Z. & Csendes D. (szerk.), *III. Magyar Számitógépes Nyelvészeti Konferencia* (pp. 134–142). Szegedi Egyetem.

Hatvani, P., Laki, L. J., & Yang, Z. Gy. (2023). A pseudonymization tool for Hungarian. *Annales Mathematicae et Informaticae, 58* (pp. 69–80). https://doi.org/10.33039/ami.2023.08.009

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2017, April). Bag of Tricks for Efficient Text Classification. In Lapata, M., Blunsom P., & Koller, A. (Eds.), *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers* (pp. 427–431). Valencia: Association for Computational Linguistics. Retrieved from https://aclanthology.org/E17-2068

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., & Birch, A. (2018, July). Marian: Fast Neural Machine Translation in C++. In *Proceedings of ACL 2018, System Demonstrations* (pp. 116–121). Melbourne: Association for Computational Linguistics.

Kahla, M., Novák, A., & Yang, Z. Gy. (2022). Fine-tuning and multilingual pre-training for abstractive summarization task for the Arabic language. *Annales Mathematicae et Informaticae, 57* (pp. 24–35). https://doi.org/10.33039/ami.2022.11.002

Kahla, M., Yang, Z. Gy., & Novák, A. (2021, September). Cross-lingual Fine-tuning for Abstractive Arabic Text Summarization. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)* (pp. 655–663). Held Online: INCOMA Ltd. Retrieved from https://aclanthology.org/2021.ranlp-main.74

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., & Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the ACL* (pp. 177–180). Prague.

Laki, L. J., & Yang, Z. Gy. (2018). Combining Machine Translation Systems with Quality Estimation. In Gelbukh, A. (Ed.), *Computational Linguistics and Intelligent Text Processing* (pp. 435–444). Cham: Springer International Publishing.

Laki, L. J., & Yang, Z. Gy. (2020). Automatic Diacritic Restoration With Transformer Model Based Neural Machine Translation for East-Central European Languages. In *Proceedings of the 11th International Conference on Applied Informatics (ICAI 2020)* (pp. 190–202). http://ceur-ws.org/Vol-2650/#paper20

Laki, L. J., & Yang, Z. Gy. (2022). Neural machine translation for Hungarian. *Acta Linguistica Academica, 69* (pp. 501–520). https://doi.org/10.1556/2062.2022.00576

Laki L. J. & Yang Z. Gy. (2023). Magyarcentrikus többnyelvű gépifordító rendszerek létrehozása. In Berend G., Gosztolya G. & Vincze, V. (szerk.), *XIX. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2023)* (pp. 369–380). Szeged: Szegedi Tudományegyetem, Informatikai Intézet.

Laki, L. J., & Yang, Z. Gy. (2023). Sentiment Analysis with Neural Models for Hungarian. *Acta Polytechnica Hungarica, 20* (pp. 109–128).

Laki, L. J., Ligeti-Nagy, N., Vadász, N., & Yang, Z. Gy. (2023). Neural Morphological Generators for Hungarian. In Berend G., Gosztolya G. & Vincze, V. (szerk.), *XIX. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2023)* (pp. 331–340). Szeged: Szegedi Tudományegyetem, Informatikai Intézet.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Szoyanov, V., & Zettlemoyer, L. (2020). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 7871–7880). https://doi.org/10.18653/v1/2020.acl-main.703

Ligeti-Nagy N., Ferenczi G., Héja E., Jelencsik-Mátyus K., Laki L. J., Vadász N., Yang Z. Gy. & Váradi T. (2022). HuLU: magyar nyelvű benchmark adatbázis kiépítése a neurális nyelvmodellek kiértékelése céljából. In Berend G., Gosztolya G. & Vincze V. (szerk.), *XVIII. Magyar Számítógépes Nyelvészeti Konferencia* (pp. 431–446). Szeged: Szegedi Tudományegyetem, Informatikai Intézet.

Ligeti-Nagy N., Héja E., Laki L. J., Takács D., Yang Z. Gy. & Váradi T. (2023). Hát te mekkorát nőttél! – A HuLU első életéve új adatbázisokkal és webszolgáltatással. In Berend G., Gosztolya G. & Vincze V. (szerk.), *XIX. Magyar Számítógépes Nyelvészeti Konferencia* (pp. 217–230). Szeged: Szegedi Tudományegyetem, Informatikai Intézet.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., & Zettlemoyer, L. (2020). Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics, 8* (pp. 726–742).

Nemeskey, D. M. (2021). Introducing huBERT. In In Berend G., Gosztolya G. & Vincze V. (szerk.), *XVII. Magyar Számítógépes Nyelvészeti Konferencia* (pp. 3–14). Szeged, Hungary: Szegedi Tudományegyetem, Informatikai Intézet.

Osváth, M., Yang, Z. Gy., & Kósa, K. (2023). Analyzing Narratives of Patient Experiences: A BERT Topic Modeling Approach. *Acta Polytechnica Hungarica, 20* (pp. 153–171.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Siemens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. In *Proceedings of the 36th Conference on Neural Information Processing Systems* (NeurIPS 2022).

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners.

Specia, L., Shah, K., de Souza, J. G., & Cohn, T. (2013, August). QuEst - A translation quality estimation framework. In Butt, M., & Hussain, S. (Eds.), *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 79–84). Sofia, Bulgaria: Association for Computational Linguistics. https://aclanthology.org/P13-4014

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., & Hashimoto, T. B. (2023). *Stanford Alpaca: An Instruction-following LLaMA model.* https://github.com/tatsu-lab/stanford_alpaca

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Canton Ferrer, C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., & Scialom, T. (2023). *Llama 2: Open Foundation and Fine-Tuned Chat Models.*

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., & Garnett, R. (Eds.), *Advances in Neural Information Processing Systems 30* (pp. 5998–6008). Curran Associates, Inc.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., & Raffel, C. (2021). mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 483–498). Association for Computational Linguistics. doi:10.18653/v1/2021.naacl-main.41

Yang, Z. Gy., Laki, L. J., Váradi, T., & Prószéky, Gy. (2023). Mono- and multilingual GPT-3 models for Hungarian. In *Text, Speech, and Dialogue* (pp. 94–104). Plzeň, Czech Republic: Springer Nature Switzerland.

Yang Z. Gy. (2019). *ɛπQue: Gépi fordítás minőségét becslő programcsomag.* [Ph.D. dissertation, Pázmány Péter Katolikus Egyetem, Információs Technológiai és Bionikai Kar].

Yang, Z. Gy. (2022). Neural text summarization for Hungarian. *Acta Linguistica Academica, 69* (pp. 474–500). https://doi.org/10.1556/2062.2022.00577

Yang, Z. Gy., & Laki, L. J. (2017). πRate: A Task-oriented Monolingual Quality Estimation System. *International Journal of Computational Linguistics and Applications.* [Under publication].

Yang, Z. Gy., & Laki, L. J. (2023). Enhancing machine translation with quality estimation and reinforcement learning. *Annales Mathematicae et Informaticae, 58* (pp. 182–190). https://doi.org/10.33039/ami.2023.08.008

Yang, Z. Gy., & Laki, L. J. (2023). Solving Hungarian natural language processing tasks with multilingual generative models. *Annales Mathematicae et Informaticae, 57* (pp. 92–106). https://doi.org/10.33039/ami.2022.11.001

Yang, Z. Gy., & Ligeti-Nagy, N. (2023). Improve Performance of Fine-tuning Language Models with Prompting. *Infocommunications Journal, Special Issue on Applied Informatics* (pp. 62–68). https://doi.org/10.36244/ICJ.2023.5.10

Yang, Z. Gy., & Váradi, T. (2021). Training language models with low resources: RoBERTa, BART and ELECTRA experimental models for Hungarian.

Yang, Z. Gy., & Váradi, T. (2024). Training Experimental Language Models with Low Resources, for the Hungarian Language. *Acta Polytechnica Hungarica.* [In press].

Yang, Z. Gy., Agócs, Á., Kusper, G., & Váradi, T. (2021). Abstractive text summarization for Hungarian. *Annales Mathematicae et Informaticae, 53* (pp. 299-316).

Yang Z. Gy., Dodé R., Ferenczi G., Héja E., Jelencsik-Mátyus K., Kőrös Á., Laki L. J., Ligeti-Nagy N., Vadász N. & Váradi T. (2023). Jönnek a nagyok! BERT-Large, GPT-2 és GPT-3 nyelvmodellek magyar nyelvre. In Berend G., Gosztolya G. & Vincze V. (szerk.), *XIX. Magyar Számítógépes Nyelvészeti Konferencia* (MSZNY 2023) (pp. 247–262). Szeged: Szegedi Tudományegyetem, Informatikai Intézet.

Yang Z. Gy., Dodé R., Héja E., Laki L. J., Ligeti-Nagy N., Madarász G. & Váradi T. (2024). ParancsPULI: Az utasításkövető PULI-modell. In Berend G., Gosztolya G. & Vincze V. (szerk.), *XX. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2024)* (pp. 61–72). Szeged: Szegedi Tudományegyetem, Informatikai Intézet.

Yang, Z. Gy., Laki, J. L., & Siklósi, B. (2016). HuQ: An English-Hungarian Corpus for Quality Estimation. In *Proceedings of the LREC 2016 Workshop – Translation Evaluation: From Fragmented Tools and Data Sets to an Integrated Ecosystem.*

Yang, Z. Gy., Laki, L. J., & Siklósi, B. (2018). Quality Estimation for English-Hungarian Machine Translation Systems with Optimized Semantic Features. In Gelbukh, A. (Ed.), *Computational Linguistics and Intelligent Text Processing* (pp. 88–100). Cham: Springer International Publishing.

Yang, Z. Gy., Novák, A., & Laki, L. J. (2020). Automatic Tag Recommendation for News Articles. In *Proceedings of the 11th International Conference on Applied Informatics (ICAI 2020)* (pp. 442–451). http://ceur-ws.org/Vol-2650/#paper45

Yang Z. Gy., Szlávik Sz. & Ligeti-Nagy N. (2024). Magyar nyelvű utasításkövető korpusz építése Stanford Alpaca promptok fordításával és lokalizálásával. In Berend G.. Gosztolya G. & Vincze V. (szerk.), *XX. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2024)* (pp. 243–255). Szeged: Szegedi Tudományegyetem, Informatikai Intézet.

Zhuang, L., Wayne, L., Ya, S., & Jun, Z. (2021). A Robustly Optimized BERT Pre-training Approach with Post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.