VALÓS TÉRBEN – AZ ONLINE TÉRÉRT

Networkshop 31: országos konferencia

2022. április 20–22.
Debreceni Egyetem

Szerkesztette: Tick József, Kokas Károly, Holl András

HUNGARNET Egyesület
Budapest, 2022

NETWORKSHOP 2022

A kötet megjelenését támogatta az
Energiaügyi Minisztérium

Szerkesztette: Tick József, Kokas Károly, Holl András

Tipográfia és tördelés: Vas Viktória

Networkshop
2022. április 20–22. Debreceni Egyetem, konferencia előadásainak közleményei

ISBN 978-615-82243-0-7
DOI: 10.31915/NWS.2022

Kiadja a HUNGARNET Egyesület
az MTA Könyvtár és Információs Központ közreműködésével
Budapest
2022

Borítókép: freepik.com

TARTALOMJEGYZÉK

DOI: 10.31915/NWS.2022.42

Kéziratos források transzformációinak lehetőségei a közgyűjteményekben
Transformations of manuscripts in public collections

Szűcs Kata Ágnes
Digitális Bölcsészeti Központ
Országos Széchényi Könyvtár
szucs.kata@oszk.hu

Abstract

Providing digital accessibility for manuscript resources in public collections is feasible at various levels. Though what does digital processing of manuscripts mean? Filling in record fields of an online catalogue, uploading scanned images into Transkribus, and digital publishing a TEI XML file can all line up behind the concept. In my paper, I discuss the changes a manuscript undergoes and the multiple media formats it becomes interpretable and provided for the public. From a practical point of view, I demonstrate how the use of Transkribus is transforming a workflow of a public collection. The paper also investigates methods for building HTR models by comparing two basic theoretical strategies in practice. The first approach is to add the previously generated Ground Truth (GT) as a Base Model (BM). The other is to train an entirely new model by merging the old and new GTs without a base model.

Keywords: htr, handwritten text recognition, Transkribus, digital processing, digital edition, manuscript, public collection

1. Bevezetés

A Digitális Bölcsészeti Központ Kiss József-projektje egyrészt mint digitális forráskiadás, másrészt mint kéziratok feldolgozását és közzétételét kidolgozó közgyűjteményi feladat és végül mint automatikus kézírásfelismerést szorgalmazó vállalkozás is fontos célokat tűzött ki maga elé és valósított meg.

A dHUpla[1] – Digital Humanities Platform – digitális szövegkiadásokat és a szövegfeldolgozáson alapuló kreatív tartalmakat tesz közzé. Tehát egy olyan publikálási környezet, amely a közgyűjtemények szöveges forrásainak digitális megjelentetésére szolgál. Az elmúlt két évben, a Móricz-kutatócsoport[2] munkája mellett, ez a közgyűjteményi pilot projekt segített feltérképezni és kialakítani a megjelenítéstől kezdve az xml publikáció tagkészletén át, a kép és szöveg összekapcsolásának lehetőségeit.

https://dhupla.hu/collection/kiss-jozsef-levelezes

A Kiss József-projektben továbbá kialakult egy olyan workflow, amely bármely kézirat esetében alkalmazható lépéseket ír elő a feldolgozáshoz és nem utolsó sorban a publikálás elkészítéséhez, másrészt pedig a szoftverhasználatra is létrejött egy részletesen kidolgozott, mégis rugalmas ajánlás. Ez főképp a Transkribus platformot,[3] amely történelmi dokumentumok

1 *Digital Humanities Platform*, www.dhupla.hu hozzáférés: 2022. június 23.

2 A munkáról bővebben a *Móricz-műhely* jelenleg költöztetés alatt álló oldalán lehet olvasni, https://pim.hu/hu/digitalis-bolcseszeti-kozpont/moricz, hozzáférés: 2022. június 23.

3 READ-COOP, Transkribus, https://readcoop.eu/transkribus hozzáférés: 2022. június 23.

szövegfelismerésére, képelemzésére és szerkezetfelismerésére alkalmas, és az Oxygen tei-xml szerkesztő szoftvert[4] helyezi előtérbe.

A közgyűjteményekben fellelhető kéziratos források digitális szolgáltatása különböző szinteken lehetséges. A kéziratok mediális transzformációja ebben a projektben az alábbi, egymáshoz is szervesen kapcsolódó állomások köré szerveződött:

- Fizikai példány
- Digitalizált másolat(ok)
- Transkribus dokumentum
- Kétrétegű PDF
- XML-file
- Online publikáció (dHUpla)

A kéziratok minden formájának megvan a maga jelentősége és önmagában olyan információkat hordoz, amelyeket más változatok nem. A Transkribus dokumentum már egy átírással rendelkező, szegmentált képet, illetve xml dokumentumot tartalmazó objektum, amely ebben a formában lehetőséget teremtett nem csak a digitális forráskiadás létrehozására, hanem automatikus kézírásfelismerő modell építésére is. A továbbiakban esettanulmány jelleggel az e téren elért újabb eredményekről lesz szó.

2. Automatikus kézírásfelismerés (Handwritten Text Recognition – HTR)

A szoftver a gyakorlatban is alkalmasnak bizonyult a közgyűjteményekben fellelhető különböző kéziratok kezelésére és publikálására. A jelenlegi workflowban egyrészt az átírások elkészítésére és ezek különböző formátumokba történő exportálásra használjuk, illetve folytattuk az automatikus kézírásfelismertetést is a Kiss József-levelezésen. A Transkribus szervezet[5] és a körülötte kialakult felhasználói közösség[6] pedig több téren is segítette a kézírásfelismerésben végzett munkánkat.

Egy korábbi publikációmban[7] bemutattattam az általánosságoktól a konkrét kimenetelig a Transkribus program használatát, a mesterséges intelligencián alapuló kézírásfelismerő-modell építésének módját, illetve a Kiss József kézírásán tanult első modellek eredményeit is. Az egy kéz által írt leveleken tanult eddigi legjobb modellünk 6,94-es hibaszázalékkal rendelkezik.

Hosszútávú terveink közé tartozik egy minél általánosabb, a 19–20. századi magyar írók kézírását könnyen felismerő modell létrehozása. Efelé tett következő lépés az volt, hogy a Kiss József-levelezés másik felét, azaz a 19. századi magyar költőnek, *A Hét* című hetilap szerkesztőjének címzett leveleket is bevontuk a modellépítésbe. Az így keletkezett vegyes kézírásos korpusz már elegendő mennyiségű adattal rendelkezett egy általánosabb magyar

4 *Oxygen XML Editor*, https://www.oxygenxml.com/ hozzáférés: 2022. június 23.

5 A kezdeményezés 2016 óta létezik és az Európai Unió 2020-as *Horizon* programja által finanszírozott kutatási projektből nőtte ki magát. 2019 július 1-jétől Read COOP néven és immár vállalati alapokon folytatja megkezdett munkáját. European Comission CORDIS research results, *Recognition and Enrichment of Archival Documents*, hozzáférés: 2022. június 23., https://cordis.europa.eu/project/id/674943. Vö.: READ-COOP, *Our Story*, hozzáférés: 2022. június 23., https://readcoop.eu/our-story/.

6 „Transkribus users | Facebook", hozzáférés: 2022. június 23., https://www.facebook.com/groups/614090738935143.

7 Szűcs Kata Ágnes, „Automatikus kézírás-felismertetés Kiss József levelezésén", in *Online térben az online térért: Networkshop 30: országos online konferencia. 2021. április 6-9. Eötvös Loránd Tudományegyetem* (Networkshop, HUNGARNET Egyesület, 2021), 73–80, https://doi.org/10.31915/NWS.2021.8.

nyelvű kézírást felismerő modell létrehozásához, amely azóta nyilvánosan is hozzáférhető a Transkribus oldalán[8] és az asztali alkalmazáson keresztül is.

Neve: *Transkribus Hungarian Handwriting 19th–20th Century*
Epoch-ok száma: 250
Szavak száma: 74.862
Sorok száma: 16.630

2.1. Vegyes kéziráson tanult modellek

A jelenlegi modellhez felhasznált kéziratok a Petőfi Irodalmi Múzeumban találhatók, melyek között szerepelnek borítékok, képeslapok, hagyományos és fejléces levelek, névjegykártyák. A levélírók Kiss József és családja, illetve a századforduló írói, újságírói és művészei. Ez nagyságrendileg 300 darab változó hosszúságú és minőségű levelet jelent. A levelezés további kéziratai az Országos Széchényi Könyvtárban találhatók, amelyekkel terveink szerint a modell tanítókorpusza a későbbiekben bővülni fog.

A vegyes kézírásmodell építéséhez alapvetően több adat szükséges, mivel nagyobb változatosság mutatkozik az egyes írásképek között, mint egyetlen ember esetében. Például, a Kiss József kézírásán tanult modellnél körülbelül 1/3 mennyiségű adat állt rendelkezésre, mégis 2,27%-kal jobb eredményt sikerült elérni.[9]

Az egyik legfontosabb mérőszám a HTR modell építésekor, a Character Error Rate (CER) on Validation Set,[10] azaz, az ellenőrző korpuszon ejtett karakterhibaérték aránya. A jelenlegi legjobb eredmény 9,19% – és bár akadnak olyan levelek a korpuszban, amelyeket szinte tökéletesen képes átírni, azt nem lehet elvárni ettől a modelltől, hogy ezt bármilyen magyar nyelvű kézirattal megtegye.

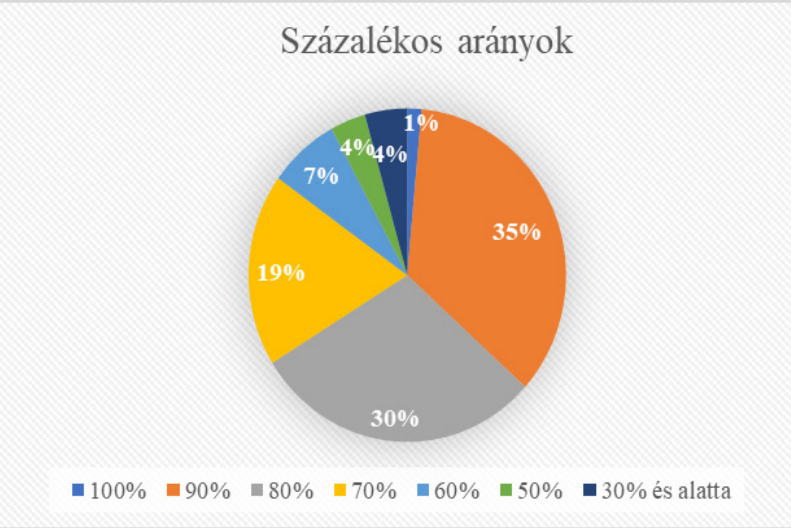
Egyrészt azért, mert a modell egy adott korpuszon, Kiss József-levelezésen tanult (1265 oldal), és így jelenleg ennek a felismerésére a legalkalmasabb. A 9,19-es érték pedig egy előjelzett átlag, amelyet a Validation Set nevű elkülönített ellenőrzőkorpuszon (138 oldal) számol ki az algoritmus.

Tehát a modell ezen a tanulási folyamatból kimaradó kézirathalmazon átlagosan közel 90%-os pontossággal meg tudja állapítani, hogy milyen karakterekből áll össze egy levél szövege, amely jó indikátora annak, hogy hogyan teljesítene egy új, ismeretlen kéziraton. Ez az eredmény egyes leveleknél valóban 90-, vagy akár 100%-os pontosságot jelent, de más, nehezebb kézírások/írásképek vagy oldaltükrök esetében 60-, 20- vagy akár 0%-os eredményt is produkálhat. Ezeket az olvasatokat átlagolja az algoritmus a CER meghatározásakor. Egy általános, a magyar nyelvű kéziratokat széleskörűen felismerő modellhez a jelenleginél sokkal több adatra van szükség.

8 READ-COOP, *Transkribus Hungarian Handwriting 19th–20th Cent.*, hozzáférés: 2022. június 23., https://readcoop.eu/model/hungarian/.

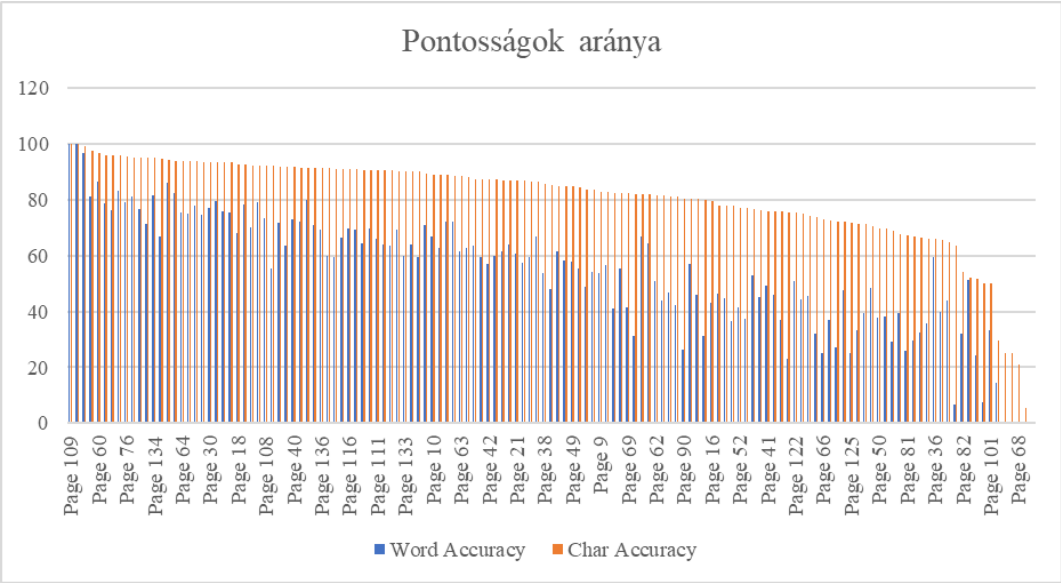
9 CER on VS: 6,94%
Epoch-ok száma: 200
Szavak száma: 74.862 25.348
Sorok száma: 16.630 5.634

10 Vö.: Elisabeth Heigl, *CER? Don´t Worry!*, hozzáférés: 2022. június 23., https://rechtsprechung-im-ostseeraum.archiv.uni-greifswald.de/cer-dont-worry/.



1. ábra A CER on Validation Set százalékos eloszlása: 100% tökéletes, átírás 30% és alatta, jelentős hibákat tartalmazó átírás.

Egy másik mérőszám, a Word Error Rate (WER) a szavak hibaarányát jelzi. A helyesen átírt szavak megoszlása a karakterekhez képest nagyobb kilengéseket mutat, mert egy karakter hibás felismerésekor az adott szót is, amelyben az szerepel automatikusan hibásnak fogja értékelni a program. Ez abban az esetben is igaz, ha csak egy nagybetű-kisbetű tévesztésről, esetleg a szó végén lévő írásjel elhagyásáról van szó.



2. ábra Szavak pontosságának az alakulása a karakterpontossághoz képest.

2.2. Modellek egymásba építése

A vegyes kéziratkorpusz folyamatos feldolgozása új kérdéseket vetett fel a modellek egymásba építésével kapcsolatban. Mivel a modellépítéshez szükséges átírt kéziratok folyamatosan készülnek, ezért egy-egy nagyobb mennyiségnél fontossá vált, hogy milyen módon tudjuk azokat beépíteni a már meglévő modellbe. Ezáltal be tudtunk kapcsolódni egy aktív nemzetközi diskurzusba, melyből kiderült, hogy ezek a kérdések nemcsak a Transkribus használói, hanem a fejlesztői számára sem egyértelműek.

Két elméleti lehetőség van, de az, hogy melyik a jó megoldás az a kéziratos korpusztól is függ. Eddigi tapasztalataink szerint befolyásolja az eredményeket, hogy mennyire homogén a kézírás, időben közel vagy távolabb vannak-e egymástól az egyes dokumentumok, műfajuk megegyezik-e stb.

Az első opció, hogy az új anyagba alapmodellként építjük be az eddigi kéziratokon tanult modellt és így fejlesztjük tovább.[11] Ebben az esetben a modell minden egyes tanulási ciklusa egy már meglévő modellre, azaz alapmodellre (Base Modellre – BM) épül. Az alapmodellek „emlékeznek" arra, amit már megtanultak, ezért elméletileg minden egyes új tanulási ciklus javítja a modell minőségét. Az új modell tanul az elődjéből, és így egyre jobbá válik.[12]

A második, hogy az elkészült új anyagot Ground Truth-ként (GT) adjuk hozzá a régihez és ebből hozunk létre új modellt. A GT vagy alapigazság egy statisztikából ismert alapfogalom, amely egy adott kérdéssel kapcsolatos igazság ismeretére vonatkozik, ez az ideális elvárt eredmény. Kéziratok esetében ez egy dokumentum 100 %-ban helyes gépelt példányát jelenti, amely a Mesterséges Intelligencia betanítására szolgál.[13]

Mindegyik modell HTR+ technológiával készült és a modellépítésnél be volt kapcsolva az *omit line by tags* funkció, melynek hatására az algoritmus automatikusan figyelmen kívül hagyja azokat a sorokat, ahol bizonytalan olvasatot vagy olvashatatlan szövegrészt jelöltünk. Ez minimum 2–3% javulást eredményezett a modelleknél, annak ellenére, hogy ezáltal kb. 1000 sorral kevesebb lett a GT. Az első lépésben elkülönítettünk egy kisebb Tesztkorpuszt (TK). Ez az anyag egyáltalán nem vett részt a modellépítésben, így befolyásoltság (bias) nélküli tesztelésre adott lehetőséget a modellépítés során.

2.3. A Base Modellel tanított modellek

Az **első** vegyeskézírásból készült modellünk 484 oldalnyi szövegen tanult (6.716 sor), alapmodell nélkül 10,38%-os lett. Ebbe építettük be a Kiss József kézírásán tanult modellt (CER on VS: 6,94%), amellyel 0,27%-ot javult a teljesítménye.

A **második** vegyeskézírásból készült modell 298 új oldalnyi szövegen tanult és azt a modellt (ID: 34757 Vegyes_kézírás_13) használtuk Base Modellként, amelybe Kiss József kézírása is be volt építve alapmodellként. Amikor az első vegyeskézírás korpuszon épült modellt, (ID: 34600 Vegyes_kézírás_11) használtuk alapmodellként, 0,1%-kal kaptunk rosszabb eredményt. Ebből az látszik, hogy egy alapmodell alapmodelljeként beépített – jelen esetben a Kiss József kézírásán tanult – modell jelenléte vagy hiánya alig okozott teljesítmény béli különbséget. Harmadik lépésként az összevont vegyes kéziráskorpuszhoz építettük be a Kiss József-kézírása alapmodellt, amely a legjobb eredményt hozta az első fázisban (9,92%). Tehát nagyobb hatása van az alapmodellnek, ha nincs annyira „eltemetve."

11 Vö.: Dirk Alvermann, *Use Case: "Model Booster"*, hozzáférés: 2022. június 23., https://rechtsprechung-im-ostseeraum.archiv.uni-greifswald.de/use-case-model-booster/.

12 READ-COOP, *Base Models*, hozzáférés: 2022. június 23., https://readcoop.eu/glossary/base-models/.

13 READ-COOP, *Ground Truth*, hozzáférés: 2022. június 23., https://readcoop.eu/glossary/ground-truth/.

	50 epoch	100 epoch	150 epoch	200 epoch
BM 34757 Vegyes kézírás_13	10,05%	10,28%	10,35%	0%
BM: 34600 Vegyes kézírás_11	10,68%	10,15%	10,42%	10,20%
BM: 31617 Kiss József kézírása + Összevont GT	12,27%	10,28%	9,92%	10,30%

3. ábra. Második vegyes kézírásmodellek alapmodellel.

Az alapmodellel történő modellépítés egyik vitathatatlan előnye, hogy viszonylag kevés epoch alatt éri el a maximumát, ami idő- és energiatakarékos megoldás lehet bizonyos esetekben. Az epoch a Transrkibus terminológiájában azt az iterációs ciklust jelenti, ahányszor az algoritmus a tanulási folyamatot elvégzi az adott korpuszon.

2.4. Összevont Ground Truth alapú modellek

A második szakaszban az összevont GT-s modellekkel folytattuk a tesztelést.

Először kombináltuk az első és második vegyes kézírásmodell anyagát, és BM nélkül az algoritmus 10,34%-os CER értéket produkált a korpuszon. Ehhez hozzátettük a Kiss József kézírásával készült levelek korpuszát, ezen már 9,31%-os eredményt értünk el.

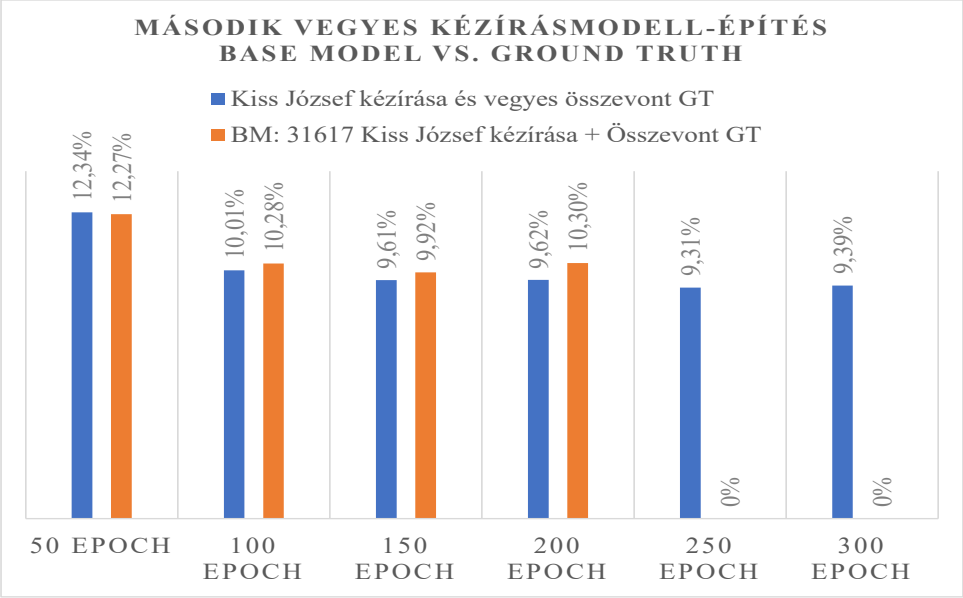
A különböző tesztek lefuttatása után a TK-t is beépítettük a modellbe és így értük a vegyes korpusz eddigi legjobb, 9,19%-os eredményét.

	50 epoch	100 epoch	150 epoch	200 epoch	250 epoch	300 epoch
Vegyes összevont GT	12,71%	10,96%	10,72%	10,55%	10,36%	10,34%
Kiss József kézírása és vegyes összevont GT	12,34%	10,01%	9,61%	9,62%	9,31%	9,39%
Kiss József kézírása és vegyes összevont GT + Teszt Korpusz	12,16%	10,14%	9,89%	9,19%	9,19%	0%

4. ábra. Második vegyes kézírásmodellek összevont alapigazsággal.

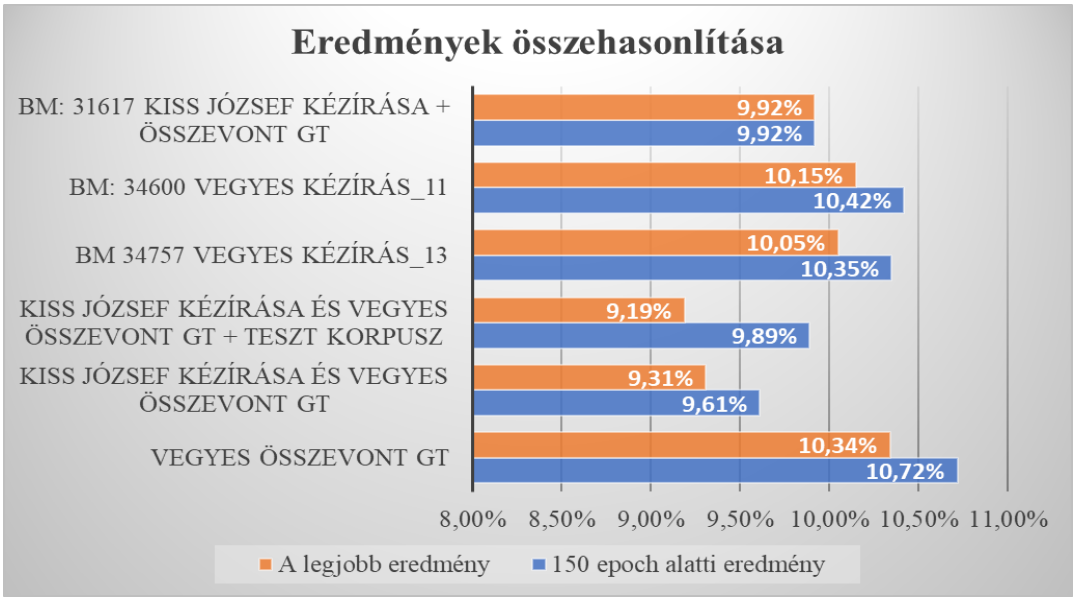
2.5. Az eljárások összehasonlítása

A Base Modelles és Ground Truth-os eljárásokat nehéz összehasonlítani, mert mechanizmusukból fakadóan más-más esetben hoznak jó eredményt. A diagrammon is látszik, hogy az alapmodellel rendelkező modell kevesebb epoch alatt hozott jó eredményt és egy bizonyos mennyiségű ismétlés (150 epoch) felett már túltanulta magát, így nőni kezdett a hibaszázalék aránya. A GT-s modellek épp ellenkezőleg, eleinte sokkal rosszabb eredményeket hoztak és csak 150 epoch után kezdett az értek 10% alá csökkenni.



5. ábra. Az eljárások összehasonlítása.[14]

150 epoch-nál még mindegyik modell értékelhető eredményeket hozott és összehasonlítva ezeket már körvonalazódnak a végső eredmények is. A Base Modellel rendelkező modell itt érte el a maximumát, a Ground Truth-os modellek közül a végén a második és az első helyet cserél, ugyanis 250 epochon 0,12%-kal jobban teljesít a TK beépítése utáni adattöbblet miatt.



6. ábra. Az hibaszázalékok összehasonlítása 150 epoch-nál.

14 250 epoch-nál 0%-os érték szerepel, mert az eggyel korábbi szakaszon is romló tendenciát mutatott a modell, így utána nem futtattuk le a betanítást.

3. Kitekintés

Terveink közé tartozik, hogy folytatjuk az automatikus kézírásfelismerés kutatását és tesztelését. A következő lépés az, hogy az OSZK-ban található Kiss József-levelezés elemeit is beépítsük a modellbe, hogy minél inkább rátanítsuk azt a vegyes kézírás felismerésére.

A legjobb eredmény érdekében a különböző magyar nyelvű projekteknek, amelyek automatikus kézírásfelismertetést használnak, a jövőben össze kell fogniuk, hogy a saját korpuszokon betanított modelleket egymásba építve egyre általánosabb érvényű automatikus kézírásfelismertető eszköz jöjjön létre a magyar nyelvű források kutatható közzétételéhez. Meg kell találni azokat az eszközöket, amelyek segítségével a közgyűjteményekben rejtőző kéziratos kincsek a digitális térben hozzáférhetővé, feldolgozhatóvá, kutathatóvá válnak. A most nyilvánosságra hozott első magyar kézírásfelismerő modell ennek a folyamatnak volt fontos mérföldköve.

4. Felhasznált irodalmak

Alvermann, Dirk. *Use Case: "Model Booster."* Hozzáférés: 2022. június 23. https://rechtsprechung-im-ostseeraum.archiv.uni-greifswald.de/use-case-model-booster/.

European Comission CORDIS Research Results. *Recognition and Enrichment of Archival Documents* | READ Project | Fact Sheet | H2020. Hozzáférés: 2022. június 23. https://cordis.europa.eu/project/id/674943.

Heigl, Elisabeth. *CER? Don´t Worry!*, Hozzáférés: 2022. június 23. https://rechtsprechung-im-ostseeraum.archiv.uni-greifswald.de/cer-dont-worry/.

READ-COOP. *Base Models*. Hozzáférés: 2022. június 23. https://readcoop.eu/glossary/base-models/.

READ-COOP. *Ground Truth*. Hozzáférés: 2022. június 23. https://readcoop.eu/glossary/ground-truth/.

READ-COOP. *Our Story*. Hozzáférés: 2022. június 23. https://readcoop.eu/our-story/.

READ-COOP. *Transkribus Hungarian Handwriting 19th–20th Cent*. Hozzáférés: 2022. június 23. https://readcoop.eu/model/hungarian/.

Szűcs, Kata Ágnes. „Automatikus kézírás-felismertetés Kiss József levelezésén". In *Online térben az online térért: Networkshop 30: országos online konferencia*. 2021. április 6-9. Eötvös Loránd Tudományegyetem, 73–80. HUNGARNET Egyesület, 2021. https://doi.org/10.31915/NWS.2021.8.

Transkribus users | Facebook. Hozzáférés: 2022. június 23. https://www.facebook.com/groups/614090738935143.