DOI: 10.31915/NWS.2019.5

# Kézírásfelismerés Arany János levelein

Bobák Barbara
*Magyar Tudományos Akadémia Bölcsészettudományi Kutatóközpont*
*Irodalomtudományi Intézet*
bobak.barbara@btk.mta.hu

Gábori Kovács József[1]
*Magyar Tudományos Akadémia Bölcsészettudományi Kutatóközpont*
*Irodalomtudományi Intézet*
kovacs.jozsef@btk.mta.hu

Arany János 1865-1869 között titoknoki, majd 1870–1877 között főtitkári minőségben segítette az MTA munkáját. Az író tehát csaknem másfél évtizeden keresztül látott el fontos tudományszervezési feladatokat.[2] Életművének kutatása ezért elképzelhetetlen hivatali iratainak, levelezésének feltárása és publikálása nélkül. E dokumentumok vizsgálata kiegészítheti olvasmányainak listáját, tisztázhatja, hogy milyen mértékben vett részt az Akadémia kiadványainak elkészítésében, valamint segítséget nyújthat annak felismerésében, hogy a hivatali foglalatosság milyen, eddig nem ismert kapcsolatokat teremthetett az ő költészetével. Továbbá ezek a dokumentumok kulcsszerepet játszhatnak az Akadémia szervezeti rendszerének leírásában, az 1870-es átszervezés okainak, az Akadémia hazai és külföldi intézményes kapcsolatainak feltárásában; reprezentációs szerepének, illetve tudományszervezésben elfoglalt pozíciójának meghatározásában. Az intézmény sokfelé ágazó tudományszervezési feladatait alapvetően meghatározta az a tény, hogy az Akadémia számos tudományterület képviselőit – nyelv- és széptudományok, bölcseleti tudományok, törvénytudományok, történeti tudományok, matematikai tudományok és természettudományok – fogta össze a korszakban. E tudományterületek külső – hazai és külföldi – intézményekkel és kutatókkal való kapcsolattartása pedig rendszerint a mindenkori titoknok, illetve az általa vezetett Titoknoki, később Főtitkári Hivatal közvetítésével történt, így Arany János hivatali levelezésének kiadása az említett tudományterületek és a hozzájuk a korszakban sorolt tudományágak történetének, fejlődésének, hazai és külhoni intézmény- és kapcsolatrendszerének vizsgálatához is nélkülözhetetlen adatokat nyújthat.

1 Gábori Kovács József a tanulmány írásának ideje alatt MTA Prémium Posztdoktori Ösztöndíj támogatásban részesült.

2 Gergely Pál. „Bevezetés". In: Arany János. „Hivatali iratok 2.: Akadémiai évek (1859–77)". szerk., jegyz. Gergely Pál. Bp. Akadémiai K. (1964) (Arany János összes művei, 14.): 602–604. ; Arany akadémiai feladatairól lásd még Gergely Pál. „Arany János és az Akadémia." Bp. MTA Irodalomtudományi Intézete – Akadémiai K. (1957). (Irodalomtörténeti füzetek, 11.) ; Keresztury Dezső. „Csak hangköre más: Arany János 1857–1882." Bp. Szépirodalmi K. (1987): 468–480. ; Voinovich Géza. „Arany János életrajza 1860–1882." Bp. Magyar Tudományos Akadémia (1938): 205–214.

Korábban, az *Arany János összes művei* című kritikai kiadás XIV. kötetének összeállítása során Gergely Pál elkészítette Arany akadémiai iratainak jegyzékét. Ő körülbelül 2500 dokumentumot sorolt fel, ám ennek csak kisebb részét – 717 – tette közzé. [3] Ez a tény és a kötet megjelenése óta eltelt több mint ötven év eredményei szükségessé tették Gergely jegyzékének felülvizsgálatát és kiegészítését. Főként mivel a kötet összeállítása során a szerkesztő szinte csak az Arany által írott hivatali iratok közül válogatott, míg az Akadémia titoknokához írott levelek közül csak a legfontosabbak kerültek a gyűjteménybe. [4]

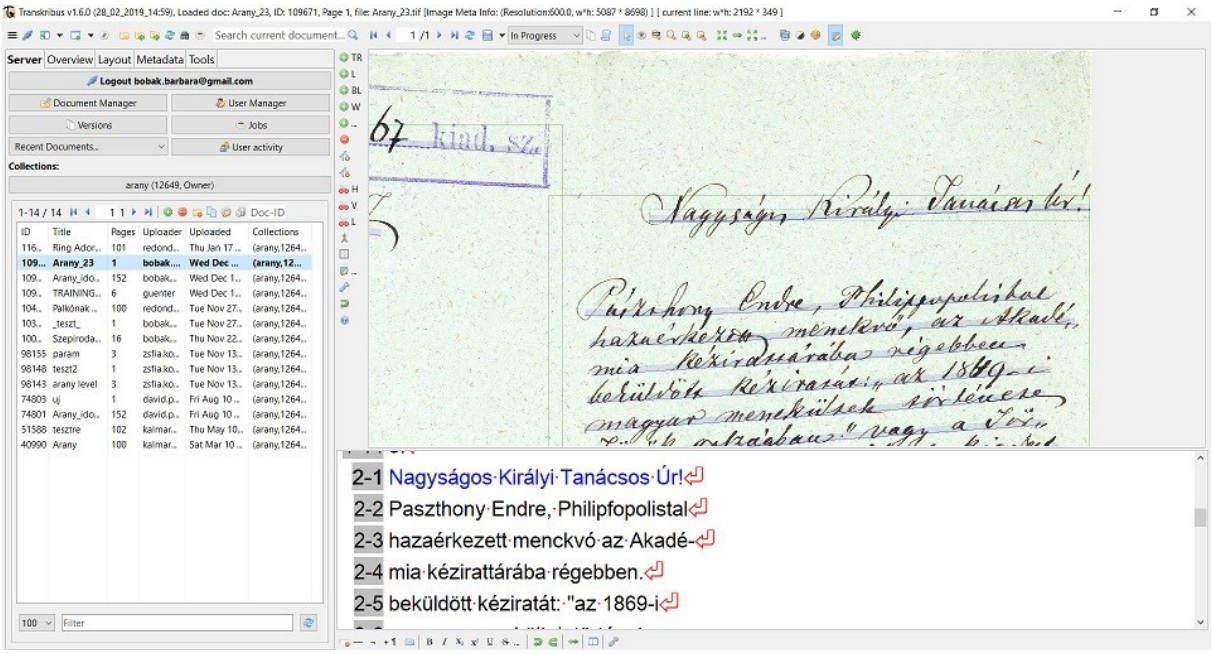
Mindezeket szem előtt tartva vált indokolttá, hogy Arany János hivatali levelezésére essen a választás a gépi kézírásfelismerés és átírás tesztelésére, valamint fejlesztésére. A papíralapú dokumentumok digitalizálásának fontossága kétségbevonhatatlan. A tartós megőrzés, az egyszerűbb tárolhatóság, a könnyebb és sokrétűbb feldolgozhatóság mind indokolja azt. Függetlenül attól, hogy a humán tudományok területén munkálkodó kutatók a mai napig print kiadásokat, továbbá kéziratokat használnak munkájukhoz, egyre több írásos anyag rendelkezik elektronikus verzióval is, az újonnan születő művek jelentős hányada pedig már eleve digitális formában jön létre. A technika fejlődésével és térhódításával tehát megváltozott a tudásanyag megszerzésének – és létrehozásának – a módja, megváltoztak az információhoz való hozzáférésre vonatkozó igények, ami szintén a digitalizálás elkerülhetetlensége felé mutat.

Egy szöveges dokumentum digitalizálásának legfőbb lépései a szkennelés és a karakterfelismerés, legyen szó akár nyomtatott, akár kézzel írt szövegről. Nyomtatott szövegek esetén az optikai karakterfelismerés (OCR), írott szövegek esetén pedig a kézírásfelismerés (HTR) technológiáját alkalmazzuk.

## Az OCR

A digitalizálási folyamat első lépése a szkennelés, amikor a papíralapú dokumentumokról jó minőségű, magas pontsűrűségű (min. 600 dpi[5]) szkennelt képek készülnek. Ezt követően a képeken valamilyen OCR program kerül futtatásra, amellyel további információt tudunk kinyerni a szkennelt anyagból. Az optikai karakterfelismerés (Optical Character Recognition – OCR) olyan technológia, amely felismerhetővé teszi egy képen vagy szkennelt dokumentumon szereplő karaktereket, számokat és központozási jeleket. A folyamatban nehézséget jelentenek azok a *zaj*nak nevezett elemek, amelyek egy nyomtatott dokumentumban előfordulhatnak, például egy folt vagy gyűrődés a papíron, homályos háttér, tintafoltok stb. A karakterfelismerő teljesítményét tanulással lehet fejleszteni, így idővel képessé válik olyan karakterek és minták azonosítására is, amelyekkel korábban nem találkozott.[6]

3 Arany János. „Hivatali iratok…"
4 Lásd pl. A Magyar Földhitelintézet pénzügyi osztálya t. c. Igazgatóságának a MTA elnöke. In: Arany János. „Hivatali iratok…": 589.
5 dots per inch – pont per hüvelyk
6 Optikai karakterfelismerés. Hozzáférés 2019. 03. 10. http://szovegbanyaszat.typotex.hu/content/PDF/ch+ocr.pdf




1. kép A Transkribus felhasználói felülete

Az OCR alkalmazása által tehát a képekből szövegfájl formátumok jönnek létre, amelyek szerkeszthetők és kereshetők, vagyis alkalmasak a feldolgozásra.

## A HTR

Ha a digitalizálni kívánt dokumentum nem nyomtatott, hanem kézírásos, abban az esetben kézírásfelismerő (Handwritten Text Recognition – HTR) programra van szükség. A cél ugyanaz, mint a nyomtatott szövegek esetében, azaz a kéziratról készült szkennelt képfájlból feldolgozásra alkalmas szövegfájl formátum létrehozása. A folyamatot itt is nehezítik mindazok a zajelemek, amik a nyomtatott dokumentumokban is előfordulhatnak, kézírásnál azonban a személyiségjegyekkel is számolni kell, hiszen ahány ember, annyiféle kézírás létezik. Az OCR technológiájához hasonlóan a HTR is tanulási fázison megy keresztül, ami alkalmassá teszi a szöveget alkotó karakterek felismerésére, később pedig új minták azonosítására is.

Az OCR és a HTR térhódításával olyan számítógépes alkalmazások és online szolgáltatások is megjelentek, amelyek képesek mindkét technológiát alkalmazni a feltöltött dokumentumokon.

## Transkribus[7]

A Transkribus online szolgáltatás, amellyel nyomtatott szövegek és kéziratok felismertetését és átírását végezhetjük, jelentős mennyiségű munkaórát megspórolva. 2016-ban az Innsbruck Egyetem Digitalizálás és Digitális Archiválás

7 Transkribus. Hozzáférés: 2019. 03. 10. https://transkribus.eu/Transkribus/

kutatócsoportja (DEA)[8] hozta létre a READ[9] projekt részeként, és azóta is folyamatosan fejlesztés alatt áll. A szolgáltatás, melynek jelentős része nyílt forráskódú, mindenki számára ingyenesen elérhető. A Transkribus, a szolgáltatás honlapjáról egy regisztrációt követően letölthető és telepíthető, használatához azonban internetelérés szükséges.

Az alkalmazás felülete rendkívül felhasználóbarát, logikusan felépített és tagolt, az egyes funkciók könnyen elérhetők. A kialakításnál láthatóan fontos szempont volt, hogy azok is könnyen boldogulhassanak vele, akik a technológiára kevésbé fogékonyak. A könnyen kezelhetőség célkitűzését az is bizonyítja, hogy a szolgáltatás wiki oldalán számos, részletesen leírt és képekkel illusztrált .pdf, valamint videó formátumú használati útmutató tölthető le angol és német nyelven.[10]

| Első modell | Második modell |
|---|---|
| f0 .lf-én -<br>re. oldy F.<br>ezennele<br>Tidlobb<br>zklá sz.<br>zoalaáádáéb<br>eMI.<br>MM865.. .<br>186 186j éé aa<br><br>zos. m á a a<br>s<br>Nagysags Králiyi Taneilcr<br>aszthoy Endre, Phdifyopolisbol<br>hazaészezett. menckvő, az Akadé-<br>mia kezrdttárába régebben.<br>beküldött keziratati ,"az 1869-i<br>magyar menekültet története<br>Török országban vagy a For-<br><br>ténelmi Bizottsag últal kiadat-<br>m vagy, högy ő maga kiadhasa<br>mg<br>s ez autal sányorú helyzeten<br>valamit könyéthessem, el nmn-<br>fogadás esetere visszakerven.<br>a folyó hó 9-én tartott<br>ostályűles határozatából<br>szerencsém van. e dolyozatot<br>aztenni a mondott Bizott<br>sághozs véleményada<br>vegon .A menmngrvea jedig elő<br>re lavható volaa, hogy e kez | 5.<br>730 1867 kiad. sz.<br>720 Toldy J.<br>ezennel<br>millobb<br>iejellilölböz<br>t i ibi.<br>1é851771<br>MMAÉL.7. 1885<br>136 /1867 levélt. szám.<br>G<br>99t175 d. rérééj8 tá.<br>3<br>Nagyságos Királyi Tanácsos Úr!<br>aszthony Endre, Philipgopolisból<br>hazaérkezett, ménckvő, az Akadé-<br>mia kézirdttárába régebben<br>beküldött kéziratát: "az 1869-i<br>magyar menekültek története<br>Török országban" vagy a Tor-<br>.<br>ténelmi Bizottság által kiadat-<br>m vagy, hogy ő maga kiadhatsa<br><br>s ez által sanyára helyzeten<br>valamit könyethessen, el nem,<br>fogadás esetére visszak érvén;<br>a folyó hó 9-én tartott<br>osztályáles határozatábal<br>szerincsém vann e dolgozatat<br>áttenni a mondott Bizott-<br>sághozg véleményada<br>algosy. A nenyren pedig elő |

2. kép Az első és a második modellel felismertetett levélrészlet

8 Digitalisierung & Elektronische Archivierung (DEA). Hozzáférés: 2019. 03. 10. https://www.uibk.ac.at/germanistik/einrichtungen/dea.html

9 Recognition and Enrichment of Archival Documents. Hozzáférés: 2019. 03. 10. https://read.transkribus.eu/about/

10 Transkribus – How to Guides. Hozzáférés: 2019. 03. 10. https://transkribus.eu/wiki/index.php/How_to_Guides

Az átírásra szánt dokumentumokat *kollekció*kba kell gyűjteni, az egyes kollekciók így számos dokumentumot tartalmazhatnak, mennyiségükre nincs megkötés.

A Transkribus – lehetőségeiből és funkcióiból adódóan – tartalmaz optikai karakterfelismerő és kézírásfelismerő modult is. Egy feltöltött dokumentum feldolgozásának első lépése a megfelelő szegmentálás, azaz a képfájlon lévő szöveg sorainak a beazonosítása és kijelölése. Megadható, hogy a dokumentum egészén vagy csak bizonyos oldalain menjen végbe a szegmentálás. A végeredmény szabad

| | Token | Karakter |
|---|---|---|
| **Első modell** | 128/200 | 276/954 |
| | 64% | 29% |
| **Második modell** | 48/200 | 116/954 |
| | 24% | 12% |

1. táblázat A modellek tesztelésének hibaszázalékai

kézzel javítható, vagy teljes egészében pótolható, amennyiben az automatikus azonosítás nem ment végbe megfelelően. Következő lépésként a kijelölt szegmentumokon futtatható az OCR vagy HTR, a feltöltött dokumentum jellegétől függően. Kézírásfelismerés esetén szükséges egy *modell* kiválasztása is, amely az adott nyelv karakterkészleti sajátosságainak felismerésére lett létrehozva, és ami tanítható további feltöltött és hibátlanul átírt dokumentumok révén, ezzel növelve a hatékonyságát. A Transkribus angol és német nyelvre tartalmaz beépített modellt, azonban kérhető bármely más nyelvre is, amennyiben megfelelő mennyiségű (min. 15 000 átírt szó)[11] és minőségű átírt dokumentum áll rendelkezésére. Az átírt szöveg végül számos formátumban exportálható: kétrétegű PDF, TEI, DOCX, TXT, stb.

## A korpusz

Arany János hivatali iratainak összegyűjtése és feldolgozása az Arany János munkái kritikai kiadáshoz kapcsolódva 2014 szeptemberében vette kezdetét. A munka során eddig hol két, hol egy munkatárs tevékenysége révén kb. 2000 dokumentum betűhív, nyers átirata készült el, nagyjából 2400 oldal terjedelemben, részben .docx formátumban, részben a LyX dokumentum előkészítő rendszer kritikai kiadás készítésére alkalmassá tett, Hegedüs Béla által fejlesztett verziójában. Eközben befejeződött a hivatali iratok túlnyomórészt az MTA Könyvtár és Információs Központ Kézirattárában található anyagának szkennelése is, amely kb. 9200 dokumentumról nagyjából 30000 felvételt eredményezett. A Transkribus teszteléséhez ebből a jelentős mennyiségű anyagból került felhasználásra – ezen szöveg megírásáig – 200 oldalnyi kézirat az Eötvös Loránd Tudományegyetem Bölcsészettudományi Kar Digitális Bölcsészet Központ kezdeményezésére és együttműködésében.

11 Transkribus – Questions and Answers. Hozzáférés: 2019. 03. 10. https://transkribus.eu/wiki/index.php/Questions_and_Answers

## Tesztelés

A teszteléshez két lépésben, 100-100 oldalnyi kézzel átírt – pontosabban a HTR eredményén javított – levelet biztosítottunk a szolgáltatás fejlesztőinek, a magyar nyelvre vonatkozó modell létrehozására és tanítására. Bár a Transkribus hivatalos leírása szerint már kb. 75 oldalnyi szöveg is elegendő lehet a tanulási fázishoz, a magyar nyelv esetében az első 100 oldalt követően csekély sikerrel tudott a modell boldogulni a kézírással. Teljesítménye összességében körülbelül 30%-ra volt tehető, ha nem vesszük figyelembe például a leveleken szereplő pecsétek szövegeit. Ezeket a modell szintén próbálta kiolvasni, gyakorlatilag eredménytelenül, amiben közrejátszott a pecsétek elhelyezkedése és árnyalata is.

A második körben biztosított további 100 oldalnyi átirattal azonban számottevően javult a modell teljesítménye. Nagyságrendileg csupán egy hibásan átírt szó fordult elő soronként. Fontos ugyanakkor megjegyezni, hogy mindegyik alkalommal azonos személytől származó kéziratok átirataival ment végbe a fejlesztés. A cél viszont az, hogy a modell képessé váljon bármilyen magyar nyelvű kézírás felismerésére a lehető legkisebb hibázási aránnyal. Ennélfogva a tesztelés következő szakaszában egy másik személytől származó, de időben ugyanakkor keletkezett kéziratok átiratát szolgáltatjuk a tanulási fázishoz.

Az 1. táblázat a két modell teljesítménybeli különbségeit mutatja számszerű és százalékos formában is. Egy tetszőlegesen kiválasztott levélen le lett futtatva mindkét modell, a végeredményül kapott átiratokból pedig véletlenszerűen ki lett emelve ugyanaz a 200 db token[12]. A táblázatból kiolvasható, hogy a modellek mennyi hibát ejtettek a tokenek szintjén, valamint a tokeneken belül a karakterek szintjén. Míg az első modell a mintának több mint felét (64%) elrontotta, a karaktereknek pedig majd harmadát (29%) nem tudta felismerni, addig a második modellnél ezek a hibaértékek valamivel több mint felére csökkentek. Fontos azonban megjegyezni, hogy ezek az arányok nagyon kis méretű mintát jelképeznek, nagyobb méretű, vagy akár a teljes átiratok összehasonlítása ezeket az arányokat valamelyest módosíthatják.

## Konklúzió

Az eddig összegyűjtött 9200 hivatali irat kiadása, nyomtatott kötetenként 700 dokumentummal számolva, 13-14 kötetben valósulhatna meg, ami a kritikai kiadások elkészítésének szokásos időtartamával – 8-10 év – kalkulálva, minimum 104 munkaévet venne igénybe. Ezt az időt pedig lerövidíthetné egy kéziratátíró-program használata, még akkor is, ha a program által elkészített nyers átiratok ellenőrzése, a hivatalos átírási elveknek való megfeleltetése, valamint a kritikai és magyarázó jegyzetek elkészítése továbbra is a textológusok feladata lenne. A kézzel írt karaktereket tehát némi betanítás után nagy hatásfokkal felismerni képes program jelentős segítséget nyújthat a szövegkiadások elkészítésében, ám ez főként akkor lehet igaz, ha a szövegkiadó az eredeti írásképet leginkább megőrző betűhű

12 Szóköztől szóközig tartó karaktersorozat

átiratok készítésére törekszik, illetve, ha az átíró program felhasználói felületén a gép által készített nyers átiraton a kívánt változtatások rögtön el is végezhetők, azaz az említett szövegszerkesztő alkalmas kritikai szöveg készítésére is.

Mindazonáltal a jelenleg – ingyenesen – elérhető online kézírásfelismerő és átíró szolgáltatások közül a Transkribus bizonyul a leghasznosabbnak a szövegfeldolgozás területén. Szabadon hozzáférhető, a lényeges eszközöket magába foglaló, nagyfokú teljesítményjavulásra képes szolgáltatás, amellyel még úgy is jelentős mennyiségű munkaórát tud megspórolni egy kutató, hogy az automatikus átírás végeredménye továbbra is ellenőrzést igényel. Felhasználóbarát kialakításának köszönhetően pedig egyedülálló a maga kategóriájában, szemben azokkal a szövegfeldolgozó alkalmazásokkal, amelyek informatikai és/vagy programozási háttértudást igényelnek a működtetésükhöz, valamint grafikus felülettel sem rendelkeznek.

**Bibliográfia**

Arany János. „Hivatali iratok 2.: Akadémiai évek (1859–77)". szerk., jegyz. Gergely Pál. Bp. Akadémiai K. (1964) (Arany János összes művei, 14.)

Digitalisierung & Elektronische Archivierung (DEA). Hozzáférés: 2019. 03. 10. https://www.uibk.ac.at/germanistik/einrichtungen/dea.html

Gergely Pál. „Arany János és az Akadémia." Bp. MTA Irodalomtudományi Intézete – Akadémiai K. (1957). (Irodalomtörténeti Füzetek, 11.)

Keresztury Dezső. „Csak hangköre más: Arany János 1857–1882." Bp. Szépirodalmi K. (1987)

Optikai karakterfelismerés. Hozzáférés 2019. 03. 10. http://szovegbanyaszat.typotex.hu/content/PDF/ch+ocr.pdf

Recognition and Enrichment of Archival Documents. Hozzáférés: 2019. 03. 10. https://read.transkribus.eu/about/

Transkribus. Hozzáférés: 2019. 03. 10. https://transkribus.eu/Transkribus/

Transkribus – How to Guides. Hozzáférés: 2019. 03. 10. https://transkribus.eu/wiki/index.php/How_to_Guides

Transkribus – Questions and Answers. Hozzáférés: 2019. 03. 10. https://transkribus.eu/wiki/index.php/Questions_and_Answers

Voinovich Géza. „Arany János életrajza 1860–1882." Bp. Magyar Tudományos Akadémia (1938)