Lehet-e automatikus családfaépítő programot készíteni?

Naszódi Mátyás

MorphoLogic KFT., e-mail: naszodim@morphologic.hu

Kivonat
Ha WEB-es tartalmak alapján akarunk családfát készíteni, számos buktató kerül az utunkba: a források megbízhatatlansága, az OCR bizonytalansága, a nevek változatossága, azonos nevű különböző személyek elkülönítése. Ezeket a gondokat járja körbe a cikk, és keres megoldásokat, továbbá ezeknek a hibáknak csökkentésére tesz kísérletet a családfa építésének során. Módszerében felhasználja a szokásos szövegkorrekciót, a szintakszist, adatbázisok elemeinek összevetését. A konkrét példák specifikusak, de metodika általános és hatékony a névelem témakörében. Genealógiai adatbázis segít a források korrekciójában, és megfordítva, a források javítása segíti a családkutatást.
Kulcsszavak: karakterfelismerés, szövegkorrekció, nyelvtan, névelem-egyértelműsítés, helyesírás-ellenőrző, genealógia

1. Családfakutatás

1.1. Névelemek és egyedek

Személyeket a tulajdonnevük szerint különböztetjük meg, de a név nem azonosítja a személyt (Gulás és mtsai, 2021). Fontosabb a családi kapcsolatok hálózata, vagyis, hogy el tudjam helyezni a származási hálóban. Nem nevezem családfának, mert a kapcsolatrendszer gráfelméleti szempontból nem fastruktúrájú.[1]

1.2. Személyek unifikációja

A különböző adatbázisokban azonos személyek szerepelnek. Álatlában ezek egyértelműsítésére törekednek, de jelen esetben ennél többre van szükség, azonosításra, vagyis egy adatbázis elemére való egyértelmű hivatkozásra. (Nguyen és mtsai, 2016) Az unifikáció jelen esetben azt jelenti, hogy egy egyedként regisztrálhatjuk a genealógiai hálózatban. Ez a háló úgy azonosít, mint a szavak elhelyezése egy szemantikai hálóban. Az egyértelműsítést a néven kívül egyéb adatok segítik: szülei neve, gyermekei, születési, halálozási adatai.

[1] Azért nem fa, mert létezik rokonházasság. Emiatt két ember közt több rokoni szál is lehetséges. A genealógiai struktúra egy ciklusmentes irányított páros gráf, ahol az egyik csoport a családokat, a másik az egyéneket tartalmazza. Az egyéntől akkor irányul él a családhoz, ha szülőről, míg családtól az egyénhez akkor mutat él, ha gyermekről van szó. Csak akkor lehetne ebben ciklus, ha valaki saját maga elődje: *Oidipusz esete*.

1.3. Genealógiai adatbázisok, standardizálásuk

A genealógiai adatbázisoknak kialakult egy formája. A struktúrában kétfajta objektum van: *személy*ek és *család*ok. A személyeknek tulajdonságai lehetnek: *nem, családi név, keresztnév, becenév, születési hely, születés ideje...* Ezek elég rugalmas mezők a leírásban, de vannak olyan korlátok, melyek ma már elavulhattak, de az általam vizsgált korban és területen nem okoznak gondot. Ilyen a többnejűség mohamedánoknál, az azonos nemű pár, a nemváltás... Az adatbázisok leírására kialakult egy (nem jól szabványosított) nyelv (LDS, 1984), emiatt aki kommunikálni akar ebben a témában, ehhez alkalmazkodik.

1.4. A források sokfélesége

A weben hozzáférhrtő forrásanyagok között azok jönnek számításba, melyek nagy mennyiségű genealógiai adatot tartalmaznak. A következők hasznosak:
– **Családregények, memoárok**: Automatikus feldolgozásuk nehézkes, mivel a lényegi információ egy természetes nyelvű szövegben van elszórva.
– **Családfák**: A konkrét családra vonatkozóan általában pontos adatokat tartalmaznak. Automatikus feldolgozásuk könnyű, de minden családfának egyedi a formája.
– **Lexikon jellegű források**: Neves embereknél hatékony. A megbízhatósága lexikonfüggő. Automatikus feldolgozásuk könnyű, a lényeges genealógiai adatok jól szeparálhatók.
– **Szabad szövegű adathalmazok**: A legjellemzőbb erre a kategóriára a gyászjelentések, nekrológok. Ezek könnyen találhatók és szűrhetők folyóiratokból. Nagy mennyiségben találhatók az OSzK gyűjteményében. (OSzK, 2015) A szövegek aránylag könnyen kezelhetők lennének, ha az OCR minősége megfelelne.
– **Kötött formátumú adathalmazok**: Ilyenek a digitalizált telefonkönyvek, lakcím-nyilvántartások. A kötött formájú adatok könnyen feldolgozhatóak.
– **Mezőkbe rendezett adatbázisok**: Ezek feldolgozása lenne a legkönnyebb. Legnagyobb a Salt Lake City-ben (LDS, 2000) összegyűjtött adathalmaz (anyakönyvi kivonatok). Az adott nyelvet nem ismerők kódolták, ezért tele van elírással. A hibák öröklődnek, mert erre hivatkozik számos genealógiai szolgáltatás: a GENI (GEN, 2006), a MyHeritage (MyH, 2003), a JewishGen (Jew, 1987) ...

1.5. Adatok egységes formája

A feldolgozás érdekében minden forrásanyagot GEDCOM formátumúvá alakítom:

özv. **Wenzel Mihályné** szül. **Ulber Rozina** asszony

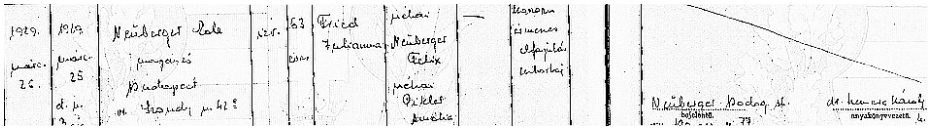
magánzónő,

f. évi szeptember hó 15-én reggel ½7 órakor rövid szenvedés után életének 86. évében történt gyászos elhunytát.

A boldogult földi maradványai f. hó 17-én délután 3 órakor fognak az oroszvári evang. temetőben örök nyugalomra helyeztetni.

Oroszvár, 1913. évi szeptember hó 15-én.

```
0 @I1@ INDI
1 NAME Mihály /Venzel/
1 SEX M
1 FAMS @F1@
1 DEAT
2 DATE BEF 1913
0 @I2@ INDI
1 NAME Rozina /Ulber/
1 SEX F
1 BIRT
2 DATE ABT 1827
1 DEAT
2 DATE 15 Sep 1913
2 PLAC Oroszvár
0 @F1@ FAM
1 HUSB @I1@
1 WIFE @I2@
```



```
0 @I3@ INDI
1 NAME Ede /Neŭberger/
1 SEX M
1 FAMS @F2@
1 FAMC @F1@
1 BIRT
2 DATE ABT 1866
1 DEAT
2 DATE 25 MAR 1929
2 CAUS koszoru ér meszes elfajulás cukorbaj
0 @I4@ INDI
1 NAME Julianna /Fried/
1 SEX F
1 FAMS @F2@
1 DEAT
2 DATE AFT 1929
0 @I2@ INDI
1 NAME Amália /Pickler/
1 SEX F
1 FAMS @F1@
1 DEAT
2 DATE BEF 1929
0 @I1@ INDI
1 NAME Felix /Neuberger/
1 SEX M
1 FAMS @F1@
1 DEAT
2 DATE BEF 1929
0 @F1@ FAM
1 HUSB @I1@
1 WIFE @I2@
1 CHIL @I3@
0 @F2@ FAM
1 HUSB @I3@
1 WIFE @I4@
```

1.6. Ideális eset

Tételezzük fel, hogy a forrásainkban az adatok pontosak, egyértelműek.

1.6.1. Adat (rekord) információtartalma

Shannon értelmezésében egy adathalmaz információtartalma nem más, mint az adategyüttes valószínűségének reciproka (illetve ennek logaritmusa). Ha a részadatok függetlenek, akkor a teljes információtartalom a részadatok információtartalmának összessége (szorzata vagy összege, aszerint, hogy milyen skálázást alkalmazunk). A személyi adatok nem függetlenek, a valószínűségek nem ismertek, de relatív gyakorisággal jól közelíthetők. Egy név önmagában csak akkor döntő, ha nagyon ritka. Egy születési évnek 365-öde az információtartalma, mint egy napra pontos dátumnak. A független részek információtartalma összeszorzódik.[2].

1.6.2. Forrásanyag információtartalma

Egy adat információtartalma becsülhető az elemei relatív gyakoriságának reciprokával. Hiányzó mező esetén 1 a relatív gyakoriság, mivel a hiány bármit jelenthet.

$$Inf(Record) = \prod_{Field \in Record} 1/Q_{Field} \tag{1}$$

[2] Ha Shannon-ban számolnánk, akkor ennek logaritmusa lenne, de akkor nem szorzódnak, hanem összeadódnak az információ-részmennyiségek.

Név (*családi, kereszt-, becenév...*) esetén a név relatív gyakoriságának reciproka az adott korban. *Születési dátum* esetén az adott időszakban született emberek számának reciproka. Ha annyit tudunk, hogy *Nagy Pista 1837-ben született Nagyatádon*, akkor az összes *1837-ben Nagyatádon születettek* számát kell elosztani az *1837-ben Nagyatádon született Nagy Pisták* számával.

1.6.3. Új adat bizonyossága

Azt vetjük össze, hogy az új adatok mennyire illeszkednek a már meglevőkhöz.[3] A bizonyosság mértéke számítható az összevethető adatelemek információtartama alapján.

$$Cert(Unit) = \prod_{Field} Inf(Field) \tag{2}$$

1.6.4. Új adat felvétele, ütközések

Ha egy forrásadat bizonyossága elér egy mértéket, felvehetjük a hálózatunkba. Ha bármi ütközik a már felvett adatokkal, meg kell vizsgálni, hogy a regisztrált adatnak vagy a felveendőnek nagyobb a bizonyossága, esetleg adatok alternatívájáról van szó. Ha pontosak az adatok, az ütközések kizárják a beépítést.

1.6.5. Munkamenet

Mindezek alapján a családfaépítés menete hasonló egy puzzle játékhoz:

1. **Kiindulási családfa felvétele:** Ez lehet akár egy személy, de lehet egy ismert család, netán egy már ismert családfa adathalmaza.
2. **Jelöltkeresés:** Keresünk a forrásanyagban egy elemet. Mivel a forrásanyag kimeríthetetlen, érdemes szűrőket alkalmazni, például névre rákeresni.
3. **A forrásanyag formalizálása:** A rekord átalakítása GEDCOM adattá.[4]
4. **Bizonyosságszámítás:** Megnézzük, az adatok alapján hogyan lehet a legnagyobb bizonyossággal beépíteni az új elem(ek)et, illetve több jelölt esetén melyiknek nagyobb a bizonyossága.
5. **Ütközésfeloldás, beépítés:** Ha adatütközés van, fel kell oldani azt: vagy a régi adat módosításával, egy korábbi elem törlésével, vagy az új elvetésével.
6. **Iteráció:** Folytathatjuk a 2. lépéstől.

Az algoritmusban lényeges, hogy a forrás általában nem zárt halmaz abban az értelemben, hogy nem ismerjük teljes terjedelmében vagy azért, mert túl nagy, vagy azért, mert a hozzáférés csak lekérdezéseken keresztül lehetséges.

---

[3] Ha két személy apját, anyját ugyanúgy hívják, akkor nem gyakori nevek esetén feltételezhetjük, hogy testvérek.

[4] A formalizálásra talán a legjobb eszköz a MorphoLogic *Markup Wizard*-ja lenne alkalmas, de ez a program hibás adatok kezelésére nem megfelelő.

1.7. Nem ideális eset

Eddig feltételeztem, hogy a források megbízhatóak. A gyakorlatban azonos személy azonos jellegű adatai forrásonként eltérhetnek. Ennek több oka van.

1.7.1. Forrás adata nem egyértelmű

Még anyakönyvi kivonatokban, sírfeliratokon is találhatók eltérések. Erre példa lehet az utólagos anyakönyvezés. Elírások, változatok gyászjelentésekben is megjelennek, de a sírfeliraton is bizonytalanság, ha a születési évet becsüljük: *élt x évet* vagy *életének x-edik évében*. Ha a halál ideje napra pontos, ez egy év intervallumnak számít, de ha csak az év volt feltüntetve, ez kétesztendős intervallumot jelent.
Nem csak személynév, hanem helységnév is lehet többértelmű.[5]
Egy ember neve változhat élete során, másrészt a neveknek változatai léteznek. A névváltozatok kezelésének egy részére a GEDCOM lehetőséget ad. Jelölhető, hogy a *Bódog*, a *Felix*, a *Felice* ugyanannak a magyar, német, illetve olasz változata. A keresztnevek halmaza zárt, emiatt bokrokba rendezhetem őket, mint a szinonimaszótárakban. Praktikus, ha a nevek között egy távolságot számítunk.[6]
A családnevek halmaza nem feltérképezhető, emiatt célravezető a helyesírás-ellenőrzőknél használt, de a célra hangolt távolságszámítás.
A mezők távolságával módosítható a bizonyosság számítása:

$$Cert(Unit) = \prod_{Field} \frac{Inf(Field_{source})}{|Field_{data} \quad Field_{source}|} \tag{3}$$

1.7.2. A forrás megbízhatósága

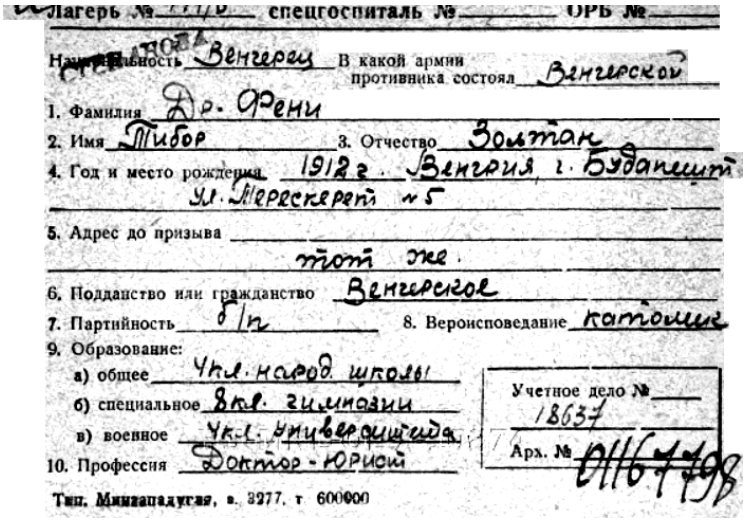
A forrás minden átírási lépésnél torzulhat. Ezt a **2.** fejezetben taglalom. A forrásállomány megbízhatóságával súlyozhatom az adat beépíthetőségét. (A mezők távolsága a változás valószínűségének reciproka.)

$$Cert(Unit) = \prod_{Field} \frac{Cert(Field_{source}) * Inf(Field)}{|Field_{data} \quad Field_{source}|} \tag{4}$$

1.7.3. A forrásállomány javítása

A hibák pár százalékos csökkentése a családfaépítést nagyságrenddel javíthatja, mivel a rokoni kapcsolatok távolsága 10-nél is nagyobb is lehet.
A Hadifoglyok állományban a családi nevek nagy százaléka sérült (Sass és mtsai, 2021). Ha az egyéb adatok megegyeznek a rekordban, akkor korrigálhatjuk a forrásanyagot.

[5] Nagyapámról tudtam, hogy *Telcs*en született, de csak további kutatásból derült ki, hogy nem az ismertebb cseh *Telč*-en, hanem a román területen levő *Telciu*-n.
[6] A *Katalin, Kata, Katica, Katerina...* közel vannak. Hasonlóan a *Telcs, Tetsch, Telch* családi nevek is.


Лагерь № спецгоспиталь № ОРБ №
Национальность Венгерец В какой армии противника состоял Венгерской
1. Фамилия Др. Фени
2. Имя Тибор 3. Отчество Золтан
4. Год и место рождения 1912 г. Венгрия г. Будапешт ул. Терескерёй № 5
5. Адрес до призыва тот же.
6. Подданство или гражданство Венгерское
7. Партийность б/п 8. Вероисповедание католик
9. Образование:
а) общее Чкл. народ. школы
б) специальное 8 кл. гимназии
в) военное Чк-л. университета
10. Профессия Доктор-юрист
Учетное дело № 18637
Арх. № 0116479
Тип. Минзападугля, з. 3277, т. 600000

	Orosz szöveg:	Az adatbázisban:	A valóságban:
Vezetéknév	Др. Феник	Fenyik	Dr. Fenyő
Utónév	Тибор	Tibor	Tibor
Apai utónév	Золтан	Zoltán	Zoltán
Állampolgárság	Венгерское	magyar	magyar
Születési hely	Будапешт	Budapest	Budapest
Születési év	1912	1912	1912

2. Forráskorrekció

A másodlagos (digitálisan kereshető) források esetenként az adatok 10-30 százalékban hibásak, mely erősen csökkenti a megbízhatóságot.
Vannak olyan eszközök, melyekkel jobb eredmény érhető el, mint általános célú nyelvi korrekciókkal vagy akár neuronhálós módszerekkel (Gulás és mtsai, 2021). Három korpuszt vizsgálok részletesebben:

	Hadifoglyok[7]	Öröklét[8]	Gyászjelentések[9]
szerkezet	record/mező	record/mező	folyó szöveg
méret	~700 000 rekord	~260 000 rekord	~600 000 rekord
kép	elérhető	nem elérhető	elérhető

[7] Az orosz levéltárban őrzött II. világháborús hadifogoly, illetve "málenkij robot" nyilvántartása. Kézzel, cirill betűkkel írt rekordok. (HAD, 2021)
[8] Boross Lajos projektje, mellyel igyekezett kereshetővé tenni magyarországi zsidó sírokat. (Boross, 2010)
[9] Eredetileg a mormonok a II. Világháborút követő projektjének magyarországi lecsapódása az OSzK gondozásában. (OSzK, 2015)

2.1. Az előfeldolgozás

A korpuszokat vagy egy humán lejegyzéssel, vagy OCR-rel digitalizálták. Ezt követően számos előfeldolgozásra lehet szükség: kódkonverzió, transzkripció, helyesírás-ellenőrzés, humán vagy gépi lektorálás, formai korrekció..., melyek egy részét már elvégezték, mire szabadon hozzáférhető anyaggá vált.

- A **Hadifoglyok** korpusza annyiban inhomogén, hogy különböző képességű emberek vették fel az adatokat, és különböző minőségben kódolták az egyes fázisokban. A magyar feldolgozás a digitalizált cirill betűs szövegeknél kezdődött, melynek kódolása egyértelmű.
- A **Gyászjelentések** nyomtatott betűkkel írt korpuszát OCR programmal digitalizálták. Mind nyelve (magyar, német, szlovák...), mind betűtípusa változó.
- Ezektől a lényegesen különbözik az **Öröklét** honlap adatbázisa, ahol a származás kézzel írt nyilvántartás, de van sírfelirat-feldolgozás is. Nyelve többnyire magyar, de sok német, esetenként horvát, spanyol... Egy részét OCR programmal digitalizálták, más részét humán erőforrással vitték gépre. Nem láthatjuk az eredeti képet sem.

Más szabadon elérhető forrásokat is vizsgáltam: lexikonokat, régi telefonkönyveket..., de a tanulság szempontjából a példáim elegendőek.

2.2. A javítás elve

Adott egy erősen rongált d kódsorozat, melyet a lehetséges de nem hozzáférhető helyes S kódsorozatokkal kell összevetni, és amelyikhez legközelebb áll, azt kell elfogadni eredménynek. $s : s\epsilon S| \min(|s \;\; d|)$
Ha figyelembe vesszük a helyes szövegek valószínűségét:

$$s : s\epsilon S| \min(p(s) * |s \;\; d|)$$

A legegyszerűbb esetben Hamming távolságot vesznek. Ez nem alkalmazható, ha a transzformációk között kódkonverzió is volt. A távolság számítása pontosabb, ha annak a valószínűségét nézzük, milyen karakterekből milyen valószínűséggel milyen karakter keletkezhet a transzformációk során. Így egyrészt nem kell a két karaktersorozatnak azonos kódban íródnia, másrészt könnyebben lehet alternatívákat elfogadni:

$$s : s\epsilon S|(p(s) * \prod_i |s_i \;\; d_i|) < K \tag{5}$$

A helyesírás-ellenérzők a legközelebbi találatokat választják ki. A végső döntés a felhasználóé. Az optikai és a beszédfelismerők mindig elfogadják a vélt megoldást. Most ez nem járható, mert nagy a hibaszázalék. Az (5) képlet alapján dolgoznak a Levenstein algoritmus javított változatai, pl. Viterbi algoritmus, amely a forrással lineáris sebességű algoritmus. A karaktertávolság alkalmazásonként eltér, mert figyelembe kell venni, mi mivé torzulhat a processzálás alatt. Az is gond, hogy az S forrástér meghatározása nehézségbe ütközik.

2.3. Karakterszintű távolság

Az említett források alkalmatlanok arra, hogy statisztikai alapon (öntanulással) határozzuk meg a karakterek távolságát, mert kicsi a korpusz ahhoz, hogy betanítsuk. Emiatt érdemes elemezni, mi mivé torzulhatott a forrásanyag kialakulása közben. Figyelembe kell venni a forrás kialakulásának lépéseit, és ehhez alkalmazkodó javítómintákat kell építeni. Erre kiváló példa a Hadifoglyok projekt és az Öröklét adatbázis.

2.3.1. Karakterszintű távolság idegen karakterek esetén

Ha a forrás keletkezésénél transzkripció is szerepel, akkor a transzkripció előtti és utáni karakterek viszonyát kell leírni. Ez történt a Hadifoglyok projektben is. A latin-cirill átírás inverze adja meg az elsődleges közelséget. Maga az átkódolás nem egyértelmű. Ezt részletesen tárgyalja az erről szóló cikk. (Sass és mtsai, 2021)

2.3.2. Kódnormalizálás

Az Öröklét adatbázisban vegyesen szerepel különböző kódkészlet. A kilógó (nem helyes kódban használt) karakterek többségét automatikusan helyre lehet állítani, esetleges ligatúrákat kifejteni, de ahol ez nem egyértelmű, ott az alternatívák jelentik a közeli karaktereket.[10] Bár a feloldás semmiben sem különbözik a helyesírás-ellenőrzőknél használtaktól, érdemes ezeket először korrigálni.
Egyszerű karakterstatisztikával ki lehetett szűrni az idegen elemeket. A korpuszban a következő kérdéses kódok szerepeltek a hozzájuk tartozó előfordulással:

char	freq	mi helyett
_	10008	Nem felismert betű (ált. ü vagy más ékezetes magánhangzó)
µ	2678	a legtöbb esetben á (852-es kód)
£	509	a legtöbb esetben ú (852-es kód)
ˇ	330	a legtöbb esetben í (852-es kód)
™	59	valószínűleg egy EŐ ligatúra (nincs UNICODE-ja)
;	16	é (a billentyűt elfelejtették angolról magyarra állítani)
=	15	ált. ó (a billentyűt elfelejtették magyarra állítani)
à	9	a legtöbb esetben Ő (852-es kód)
Ô	8	többnyire Ő
š	3	a legtöbb esetben ű vagy ü (852-es kód)
ä	3	ö (tipikus OCR hiba, ha magyarra van állítva)
*	3	-
‰	2	ä
[	2	ő (a billentyűt elfelejtették magyarra állítani)
Š	1	Ő (852-es kód)
}	1	Ű (a billentyűt elfelejtették magyarra állítani)
#	1	p

[10] Az Öröklét korpusz általában 1250-es kódkészlettel íródott, de volt ettől eltérő kódolás is. Ezeket fel kellet ismerni, átkonvertálni. Ennél cifrább esetekkel is találkoztam. PDF anyagokban egyes ékezetes karaktereket lokálisan olyan kódra konvertálnak, melyek előre nem becsülhetők. Más esetekben lebegő (prefix) ékezeteket használnak, és gyakoriak a ligatúrák (karakterösszeolvadások.) Néhány OCR program a bizonytalan felismeréseket speciálisan kódolja.

2.3.3. Karaktertávolságok

A forrásanyagokban nem jelzik a felismert karakterek bizonytalanságát. Emiatt feltételezni lehet hibás találatot ott is, ahol nincs. A hibák és azok valószínűsége OCR esetén közismertek. Nyomtatott szöveg feldolgozásánál a *c, e, o* nagyon közeliek, hasonlóan az *m, rn...* Kézzel írt szövegeknél a *C, S...* Cirill írásmódban egész más karaktereknek hasonló az alakja. ш-т, и-п-н

Ha ismert lenne, hogy az OCR mennyire biztos a felismert kódokban, akkor csak azokon a helyeken kéne alternatívát keresni, ahol ez egy szint alatt van. Ilyen próbálkozás létezett (Prószéky és mtsai, 2002). Sőt, a mai OCR programok ezt neuronhálós korrekcióval javítják, aminek sok haszna és sok a kára is.

A helyesírás-ellenőrzők nem számolnak karaktertávolságokkal. Helyette prioritási sorrendben bejegyeznek pár cseremintát, és azt optimalizálják, hogy minél kevesebb cserével elérjék az egyik karakterláncból a másikat.

Példaként a Helyeske korrekciós formalizmusa magyar OCR-javításra:

```
aáeéoóöőuúüűiítlsrznkgmdvbjfycphxwq'AÁEÉOÓÖŐUÚÜŰIÍTLSRZNKGMDVBJFYCPHXWQ-0123456789.%.$€
daéeóoőöúuűüiif1cn2rKGnauBytjsPbXvaáÁAÉEÓOŐÖÚUŰÜÍIIlsP2Hk6HOW8jEySRNKVO-OlzBASGZB    SE
áécáöőÖŐnÚŰóltrideZuhhMárnJlve nku9 ÄHFéoóöőuúüűiíF CBSMBCNOvEIPKOFMxwq  iZ
eósóOÓOŐóóŐőtÍltaRsmxjw yhvigo k mQ aaeËOOOönŐÖÜ11l Zr7nEOWdUbYf Gph UO  í
säëëOoooU öpflT1zf N y  w6  YC H W   ÄB ŐŐ ÓNŐUOll  5   ngm u y  cG  M
n E e  áv uu IF S    9  V              C  OV   T         w          m
```

eddig volt a karakter→karakter elírás függőlegesen prioritás rendjében

```
y  ij          al d          m  rn          di  ch
ij y           d  al         fi A           lj  b
h  ln          d  ol         fl A           SUh Sch
ln h           rn m          Ji N
```

ezek pedig a gyakori sztring→sztring elírás szabályai voltak.

Hasonlóan hangolható több ismert helyesírás-ellenőrző is (Naszódi, 2017), de a Hadifoglyok projektben is ilyen eszközhöz folyamodtak. (Sass és mtsai, 2021)

A Hadifoglyok adatbázisának elemzésénél kiderült, hogy az írott cirill betűs szöveget diktálással kódolták gépi szöveggé, tehát egyik ember olvasta fel a kézírást, egy másik pedig begépelte a hallottat. Ez azt jelentette, hogy a humán OCR hiba mellett a humán hangfelismerés is ejthetett hibát. Orosz kézzel írt szövegben gyakori az и-п-н felcserélés, a ш-т csere... Ezeket a Hadifoglyok projektben nem vették figyelembe. Ráadásul a humán hangfelismerés (nem ismerve a magyar nyelvet) okozhatta az **1.7.3.**-ban említette hibát: a Фени nevet kiegészítették egy к-val, hisz az orosz nyelvtől idegen hogy и-re végződjék egy név, viszont a -ник gyakori főnévvégződés.

Mivel a korábbi fázisok már nem elérhetők, az elemi cseréknél a teljes folyamat lehetséges torzításinak eredőjét kell figyelni.

2.4. Szószintű javítás

A vizsgált korpuszok mérete nagy, és az anyag erősen torzult. Ráadásul a nevek és a nem mai nyelvezet miatt az általános szószintű ellenőrzés nem lenne célravezető. A statisztikai (neuronhálós) módszer sem alkalmas automatikus javításra ilyen térben. A pontosabb munka érdekében érdemes a célnak megfelelő szövegellenőrzést kialakítani.

2.4.1. Ha a forrás mezőkre osztott rekordokból áll

A temetői adatbázis és a Hadifoglyok adatbázis ilyen. Az egyszerűbb eset a Hadifoglyok adathalmaza: vezetéknév, családnév, foglalkozás... mind olyan kategória, melyek lehetséges elemei felsorolható halmazt jelentenek. A lakcím ennél kicsit összetettebb, de ez sem jelent lényeges gondot. Ami viszont igazi probléma, a családi nevek halmaza.
Az Öröklét adatbázisa egy fokkal nehezebb. A személynevek mezőben nincs elválasztva a családi és keresztnév, sőt, a nevek írásmódja változatos: *Tarr Gézáné, Márkus Ildikó – Tarrné Márkus Ildikó...* Ennek ellenére a mezőkben szerepelő lehetséges szavak jellegzetesek és elvileg összegyűjthetőek.
Ha mezőnként egy-egy külön szókészletből álló halmazt készítünk, mezőnként a forrás lehetséges terét lényegesen lecsökkentjük ahhoz, hogy a szavak távolsága nagyobb legyen, a javítási algoritmus hatékonnyá váljék.

2.4.2. Ahol némi szintakszisra is szükség lehet

Ha összetettebb egy-egy mező, akkor egyszerű reguláris kifejezéssel jól le lehet fedni a szintakszist. Ilyen a lakcím. Lefedésen azt értem, hogy funkcionálisan megtalálják a részeket, de ezeknek nem kell karakterpontosan helyesnek lenniük. Ide tartozik az is, amikor mezőket tévesztenek a bejegyzéseknél: A Hadifoglyok adatbázis esetén a felekezeti hovatartozás és a nemzetiség kitöltése időnként elcsúszik. Az Öröklét esetén adatok át-átcsúsznak a következő mezőbe.
Ha a mezők a helyükön vannak, akkor a szavankénti ellenőrzés az előzőek alapján már lehetséges.

2.4.3. Ahol komolyabb szintakszisra van szükség

A Gyászjelentések természetes nyelvű korpusz. A javítás szempontjából itt is fontos szűkíteni, hogy egyes pozíciókban milyen szó jöhet számításba. A szintakszis a nevek helyének (és minősítésének) felismerését határozza meg, de a névelemek felismerésén kívül a mondatok szerkezete támpontot ad a családi kapcsolatokra. Családregények, memoárok kiértékeléséhez általános célú elemzőre lenne szükség, de ezzel nem foglalkozom, ez talán egy másik projekt lesz.

2.5. Mondatszint a Gyászjelentések feldolgozásában

Ha eltekintünk a szótári bejegyzésektől, akkor pár száz szabállyal olyan nyelvtant kreáltam, mellyel a korpusz mondatai elemezhetőek, és a nyelvtan alapján generált mondatok nagyobb hányada helyes.[11]
A szintakszisban részletesebb szótani kategóriákat használok a szokásosnál, és a különböző helyeken csak az alkategóriának megfelelő szót engedem meg. Helységnevek és személynevek klasszikusan a tulajdonnevek kategóriájába tartoznak, de a szintaktikai szerepük eltér. Ebből a szempontból lehet külön szófaj az orvosi

[11] A mondatelemzők túlgenerálnak, mert elemzésre készülnek. Még a helyesírás-ellenőrző mondatszintű megvalósítása is túlgenerál abban az értelemben, hogy sok hibát nem vesz észre, illetve szemantikailag lehetetlen mondatokra ritkán riaszt.

szövegekben a *testrész, orvosság*. A mi esetünkben a *keresztnév*, a *vezetéknév*, az *asszonynév* mind más, más szófaj.

2.5.1. A szintakszis

A gyászjelentésekben a mondatok korlátos sémáknak felelnek meg. A mondatokat funkciójuk szerint külön kategóriákba sorolom. A gyászjelentés fő mondatának, az `enounce`-nak a szabálya AGFL-ben (Koster, 1991):

```
RULE enounce(LANG): [SUBJ(LANG)], [INTROM(LANG)],                    # Enouncer(s)
                    [[SORROWLY(LANG)],[BUT(LANG),[sorrowcontr(LANG)]]],
                    INFORM(LANG), [OPTTHAT(LANG)],                   # verbal part
                    mainperson(LANG, GENDER, CASE),                  # Dead
                    cirumstances(LANG),                              # Time, place,...
                    died(LANG, CASE) .                               # verbal part
```

A fő mondat kötelező része a gyászjelentésnek, de a bejelentő személye opcionális. A mondat fő igéje az `INFORM` részben van, míg a mellékmondat igéjét a `DIED` egység foglalja magába.

A `cirumstances` kötelező:

```
RULE cirumstances(LANG): [town(LANG,CASELOC),[","]],date(LANG),[[","],time(LANG)],#of death
                         [age(LANG)],      # in the time of dead
                         [fage(LANG)],     # years in marriage
                         [CAUSE(LANG)] .   # of death
```

de ebben csak a halál időpontja, ami nem hiányozhat a szövegből.

2.5.2. ... és mi van a mondaton túl?

A mondat csak szintakszison belüli kategória. A mondatokra való tagolás beleolvad a dokumentum formális megfogalmazásába. A teljes gyászjelentés a következő szabállyal adható meg:

```
RULE abitoury: [cite(LANG)],                 # some liric pe. from the Bible
               enounce(LANG),                # Enouncement of the person's deth
               [burrial(LANG)],              # Burrial
               [misa(LANG)],                 # Celebration
               DATE(LANG),                   # date of enoucement
               [FARWELL(LANG)],              # Farwell words
               [personlist(LANG)],           # list of family members
               [maddress(LANG)],             # adress to reply
               [publish(LANG)], [PRESS(LANG)] # Publisher and print
```

Németre, angolra is megalkottam a szabályokat. Az elemzés szerkezete nem tér el a magyartól. A felhasznált eszköz alkalmas keverni a robusztus elemzést a pontos találatokkal, így lexikális hibák esetén is kiadhatja az elemzést feltüntetve a kérdéses szavakat. Ez nagy segítség a nem ismert tulajdonnevek keresésében és a korrekciók helyének megtalálásánál.

2.5.3. A javítás menete

Több ellenőrző eszköz van, amellyel hibák korrigálhatók. Minden szűrés után javul a forrás, módosul a szótár, esetleg a szintaktikus leírás, emiatt a korábbi fázisokat újra lehet futtatni. Az ellenőrzési szintek a következők:

– Idegen karakterek szűrése
– Kódnormalizálás
– Jellegzetes karakterhibák javítása
– Szószintű ellenőrzés
– Szintaktikai ellenőrzés / mezőkre bontás
– Kivonatolás / konvertálás GEDCOM formátumra

3. Tapasztalatok

A rendszer nincs kész, de az elért eredmények választ adnak a címben feltett kérdésre. Hiányzik az adatbázisok lekérdezőrendszerének automatizálása. Ezt kézzel pótoltam, alkalmazkodva a forrás elérési lehetőségeihez.
A kidolgozott eszköz jelenleg nem helyettesíti az emberi munkát, de sokban segít. *Telcs Edé*nek 9 generációs ismert családfájából (690 személy) egy-két generációs szigeteket tártam fel így. Viszont a család által összerakott fában is leltem hibára, amit a gépi módszer tisztázott. Más neveknél (*Korányi, Neuberger*) ennél jobb eredményt értem el.
Kísérletképpen kiindultam Kornai János gyászhíréből. A standard források biztos adatai alapján 40 családtagot szedtem össze. Enyhítve a megbízhatóságon 100 családtagra derített fényt nagy megbízhatósággal az algoritmus. Kicsit lejjebb véve a megbízhatósági paramétert újabb 70 személyt helyeztem el a hálóban, de az így kapott családfában előfordulhat tévedés is.
A családkutatás a jelenlegi forrásállománynál gyakori nevek esetén eleve kudarcra van ítélve, ha nincs mellettük jobban kereshető rokon. Az eredményesség attól is függ, hogy milyen nagy választékban találunk forrást. A XIX. századot megelőzően többnyire csak nemesekre lehet lelni. A nyomtatás elterjedése és a kötelező regisztrációk miatt a XIX. század közepétől várhatók jó eredmények. Amennyiben valóban megvalósul Magyar Nemzeti Levéltár 2023-ra tervezett digitalizálása (Biszak és mtsai, 2017), a teljes automatizálásra is sor kerülhet.
A források korrigálása általában 10-20 százalékkal növeli a fellelhető személyek számát. Emiatt valóban érdemes a források javításával foglalkozni.
Pontosabb számszerű kiértékelésre nem ad módot a jelenlegi munka, de egyértelmű a haszna.

4. Tervek

Mindenképpen megérné, ha a karakterfelismerő és a nyelvi modul egyszerre venne részt a döntésben. (Naszódi, 2000)
Kísérleteim azt is igazolják, hogy a kép előfeldolgozása (szürkeség-, ferdeségkorrekció, a valós szöveg behatárolása, inicálé felismerése) javít az OCR minőségén.
A gyászjelentések hasábbeosztása hiányos, ezért a szósorrend sérül, ami a mondatszintű elemzést rontja el. Az OCR programok szokásos hasábolását, képfelismerését újságokra optimálták, pedig ez is lehetne témakörfüggő.
Ha a forrásban valamilyen rendezésben szerepelnek a rekordok, a rendezési elv is adhat támpontot a szövegek korrigálására. Ha a nevek szerint vannak listába szedve (telefonkönyv, összeírás), akkor a sorrendtől eltérő bejegyzésnél könnyen felismerhető a hiba. Ha anyakönyvi kivonat, akkor az esemény ideje adja meg a sorrendezést.
További lehetőség, ha hasonló, de más időben készült regisztrációkat vetünk össze. Erre tipikus példa a telefonkönyv, melynek egymás utáni kiadványai nem sokban különböznek. Ha az egyikben nem olvasható jól egy mező, de hasonlít a másikban levőhöz, triviális a korrekció, pontosabban több bizonytalan bejegyzés egymást korrigálhatja, kiegészítheti.

Hivatkozások

JewishGen: Zsidó történelmi és genealógiai adatok rendszerezett, kereshető tára. WWW (1987), https://en.wikipedia.org/wiki/JewishGen

MyHeritage: Családfaépítő és megosztó online szolgáltatás. WWW (2003), https://en.wikipedia.org/wiki/MyHeritage

GENI: Általános családfaépítő és megosztó szolgáltatás. WWW (2006), https://en.wikipedia.org/wiki/Geni.com

Szovjetúnióba elhurcoltak. AOL (2021), https://adatbazisokonline.hu/gyujtemeny/szovjetunioba-elhurcoltak

Biszak, S., Lakatos, A., Ádám, V.: Az egyházi anyakönyvek digitalizálásának lehetőségei - Módszertani tanulmány (2017), https://archivum.asztrik.hu/sites/default/files/letoltesek/akvi_tanulmany.pdf

Boross, L.: Öröklét (2010), http://www.oroklet.hu/

Gulás, M., Yang Zijian, G., Dömötör, A., Laki, L.J.: Automatikus hibajavítás statikus szövegeken. In: XVII. Magyar Számítógépes Nyelvészeti Konferencia előadásai. p. 243-251. Szegedi Tudományegyetem (2021), http://nlpg.itk.ppke.hu/sites/default/files/publications/MSZNY2021_paper_29.pdf

Koster, C.H.: Affix Grammars for Natural Languages. In: Melichar, B. (szerk.) Summer School on Attribute Grammars, Applications and Systems (1991), https://link.springer.com/chapter/10.1007%2F3-540-54572-7_19

LDS: GEnealogicalDataCOMunicaton. WWW (1984), https://en.wikipedia.org/wiki/GEDCOM

LDS: FamilySearch (2000), https://www.familysearch.org/hu/

Naszódi, M.: Nyelvi visszacsatolás karakter-, kézírás- és beszédfelismerő rendszerek számára. Tech. rep. (2000), http://www.nefmi.gov.hu/ikta/projektek/ikta3/20010228/i063/ossze1.html

Naszódi, M.: A magyar helyesírás-ellenőrzők mai állása. In: XIII. Magyar Számítógépes Nyelvészeti Konferencia. p. 347-354. Szegedi Tudományegyetem (2017), http://www.cs.bme.hu/~naso/langeng/SpellsSate20016.pdf

Nguyen, D.B., Theobald, M., Weikum, G.: Joint Named Entity Recognition and Disambiguation with Rich Linguistic Features. In: Transactions of the Association for Computational Linguistics, Volume 4. p. 215-2229 (2016), hhttps://aclanthology.org/Q16-1016

OSzK: Gyászjelentések. Pannon Digitális Egyesített Archívum (2015), https://dspace.oszk.hu/handle/20.500.12346/663648

Prószéky, G., Naszódi, M., Kis, B.: Recognition Assistance. In: Conference on Computational Linguistics. Taipei, Tajvan. p. 1263-1267 (2002), http://www.cs.bme.hu/~naso/langeng/C02-2014.pdf

Sass, B., Mittelholcz, I., Halász, D., Lipp, V., Kalivoda, A.: Magyar hadifoglyok adatainak orosz-magyar átírása és helyreállítása, és a szabadszöveges adatbázisok tulajdonságai. In: XVII. Magyar Számítógépes Nyelvészeti Konferencia előadásai. p. 39-51. Szegedi Tudományegyetem (2021), http://www.nytud.hu/oszt/korpusz/resources/sb_etal_hadifoglyok.pdf

Is it possible to create a family tree building program?

Mátyás Naszódi

MorphoLogic, e-mail: naszodim@morphologic.hu

Abstract.
While building a family tree from WEB contents, we face several challenges: the unreliability of the sources, the uncertainty of OCR, the variety of names, distinguishing different people with the same name... The article addresses these issues and seeks solutions. Also, during the construction of the family tree, it attempts to reduce the errors in the sources. Standard text correction tools, syntactical parsers, and comparison of elements of databases are used. The tools are tuned according to the purposes. Restricted vocabularies and restricted grammars are applied. The examples are specific, but the methodology is general and effective in the subject of named entities. Genealogical databases are useful for repairing source corpuses, and conversely, correction of source files improves the genealogical databases.
Keywords: OCR, text correction, grammar, NERD, spelling, genealogy