# Rec. et exp. aut. Abbr. mnyelv. KLIN. szöv-ben – rövidítések automatikus felismerése és feloldása magyar nyelvű klinikai szövegekben

Siklósi Borbála[1], Novák Attila[1,2]

[1] Pázmány Péter Katolikus Egyetem, Információs Technológiai és Bionikai Kar,
1083 Budapest, Práter utca 50/a,
[2] MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport
e-mail:{siklosi.borbala, novak.attila}@itk.ppke.hu

**Kivonat** Az orvosi szövegek feldolgozása ma a nyelvtechnológia egyik legaktívabban kutatott részterülete. Az általános szövegekre ma már jól működő eszközök helyes, normalizált bemenetet feltételeznek. Orvosi szövegek esetén ez a feltétel nem teljesül, ezért az ezekre jellemző nagy mennyiségű zaj miatt kész eszközök alkalmazása nem lehetséges. A normalizálás egyik lépése a rövidítések észlelése és feloldása. Ebben a cikkben egy nem felügyelt automatikus módszert mutatunk be rövidítéssorozatok feloldására magyar nyelvű klinikai dokumentumokban. Három módszert ismertetünk, melyek különböző mértékben támaszkodnak külső erőforrásokra, illetve magára a klinikai korpuszra.

## 1. Bevezetés

Az orvosi szövegek feldolgozása ma a nyelvtechnológia egyik legaktívabban kutatott részterülete. Olyan programcsomagokból, melyek orvosi szövegekből nyernek ki a felszínen nem elérhető információkat és összefüggéseket, angol nyelven számos jól működő megvalósítás létezik. Az orvosi protokollok helyi jellegzetességeire vonatkozó adatok, illetve a helyi közösségeket érintő járványügyi információk azonban csak az adott nyelven lejegyzett szövegekben fedezhetők fel. A magyar nyelvű klinikai szövegek feldolgozására alkalmas eszközök létrehozása tehát nemcsak érdekes kihívás a nyelvi nehézségek miatt, hanem szükséges feladat is.

Azok a dokumentumok, amelyek kórházi körülmények között, nyelvi ellenőrzés nélkül jönnek létre, jellemzően sok helyesírási hibát tartalmaznak, tele vannak magyar-latin szakkifejezésekkel, illetve következetlenül használt rövidítésekkel [1,2]. Ezeknek a rövidítéseknek a használata sok esetben követ valamilyen szabályrendszert, de legtöbbször mégsem felelnek meg a rájuk vonatkozó hivatalos szabályzatnak, nem is beszélve a nem szándékos, ámde gyakori elírásokról. Ezért a rövidítések feloldása nem oldható meg egy lépésben, azoknak egy lexikonra való egyszerű illesztésével. Továbbá, a klinikai körülmények között létrejött dokumentumokban sokszor hosszabb, sok szóból álló szerkezetek szinte minden szava rövidítve van, nem csupán elvétve találunk egy-egy rövidítést az

egyébként teljes szavakat tartalmazó mondatokban [3]. Az egy egységet alkotó kifejezések elhatárolása az ilyen rövidítéssorozatokban sokszor még emberi szakértők számára is kihívást jelent, eltekintve az adott szöveg szerzőjétől, akinek remélhetőleg teljesen érthetőek a saját maga számára készített feljegyzések. Ha a hosszabb rövidítéssorozatokban az egyes szavakat tokenenként próbálnánk feloldani az egyes rövidítéseket egymástól függetlennek tekintve, az nagyon nagy számú feloldási kombinációt eredményezne. Ez az ilyen kijelentések jelentésének egyértelműsítése helyett a feloldásból eredő zajt növelné a szövegekben. A klinikai dokumentumok feldolgozása tehát nem egyszerű feladat, melynek részeként a rövidítések feloldása minden további lépés előfeltétele.

A rövidítések feloldásához használt külső lexikonok használata azért sem vezet önmagában megoldáshoz, mert ilyen erőforrások magyar nyelvre csupán korlátozott mértékben és minőségben állnak rendelkezésre. A BNO-kódrendszer hivatalos leírása az egyik ilyen elérhető adatbázis, azonban ennek használatakor is külön feladat a rövidítésekből a megfelelő, a leírásokra illeszthető minták előállítása. Ez tehát korántsem alkalmazható olyan közvetlen módon, mint az angol nyelven elérhető UMLS (Unified Medical Language System) rövidítéstára, amely a legtöbb angol orvosi rövidítést, azok változatait és lehetséges feloldásait tartalmazza [4].

Ha lenne is magyar nyelven elérhető ilyen erőforrás, az csupán a lehetséges feloldási javaslatok kigyűjtésére lenne alkalmas. A javaslatok megfelelő rangsorolásához, melynek során a szövegkörnyezetben is helytálló feloldásnak kellene első helyezettként megjelenni, megfelelően egyértelműsített nyelvmodellre lenne szükség. Mivel azonban nincsen olyan orvosi korpusz, amiben a rövidítések helyett azok kifejtett formája szerepelne, ezért ilyen nyelvmodell sem áll rendelkezésre. Egy ilyen korpusz létrehozása pedig olyan nagy mennyiségű és drága szakértői munkát igényelne, ami jelen kutatás keretei között nem volt megvalósítható.

A bemutatott kutatás célja a több tokenből álló rövidítéssorozatok automatikus feloldása külső erőforrások és a rendelkezésünkre álló klinikai korpusz felhasználásával. Bemutatjuk, hogy ebben a folyamatban szükséges, de nem elégséges a kész lexikonok és az általunk készített kisebb, korpuszspecifikus lexikon használata. Módszerünk hatékonyságát ugyanakkor jelentős mértékben növelte az algoritmus kiegészítése egy olyan lépéssel, amelynek során a rövidítéseket a korpusz szövegére illesztve is keresünk feloldásjelölteket. Ezzel biztosítható továbbá a doménfüggetlenség, hiszen a korpuszt a maga nyers formájában használjuk fel, tehát módszerünk a nem felügyelt algoritmusok körébe tartozik.

## 2. Az orvosi korpuszban előforduló rövidítések jellemzői

A rövidítések sorozatából álló jegyzetelési stílus bevett szokás a klinikai jegyzetek, dokumentumok létrehozása során. Ez a tömörített írásmód számos hivatalos és egyedi rövidítést vagy jelölést tartalmaz, amelyek nagy részének használata csak az adott szakterületre, esetleg csak egy orvosra vagy asszisztensre jellemző. A rövidítések jelölhetnek az adott orvosi szakterület körében releváns fogalmakat, vagy olyan hétköznapi szavakat és kifejezéseket, amelyek a klinikai szövegekben

gyakran fordulnak elő, ezért bevett szokás a rövidített alak használata. A szakértő olvasó számára az ilyen rövidített alakok jelentése általában éppen annyira egyértelmű, mint a szabványos rövidítések esetén, hiszen kellő ismerettel és gyakorlattal rendelkezik, valamint tisztában van a szövegkörnyezet jelentésével is. Az 1. táblázatban látható néhány példa a különböző rövidítéstípusokra. Vannak közöttük elterjedt, gyakran használt, egyértelmű alakok, melyek általában latin eredetűek. Másoknak azonban még az orvosi szakterületen belül is több jelentése lehet.

1. táblázat. Példák a korpuszban előforduló rövidítésekre.

| Domén | Rövidítés | Feloldás | Magyarul |
|---|---|---|---|
| szabványos | o. d. | oculus dexter | jobb szem |
| | med. gr. | mediocris gradus | közepes fokú |
| doménspecifikus | o. (ophthalmology) | oculus | szem |
| | o. (general anatomy) | os | csont |
| általános szóhasználatú | sü | saját szemüveg | saját szemüveg |
| speciális kifejezések | fén | fényérzés nélkül | fényérzés nélkül |
| | n | normál | normál |
| általános szavak | köv | következő | következő |
| | lsd | lásd | lásd |

A folyó szövegekben található rövidítésekkel kapcsolatos első probléma azok felismerése. Mivel ezek a szövegek nem követik a helyesírási és központozási szabályokat a rövidítések jelölésének a területén sem, ezért ezek felismeréséhez nem elegendő a rájuk vonatkozó helyesírási szabályok formalizált alkalmazása. A rövidítést jelző pontok általában hiányoznak a rövidített szóalakok végéről, a rövidítésekben vegyesen szerepelnek kis- és nagybetűk, jellemző, hogy a betűszavakat is csupa kisbetűvel írják, valamint ugyanannak a szónak vagy kifejezésnek számtalan különböző hosszú rövidítése lehet. A következő formák például mind ugyanazt a fogalmat jelölik: *vvf*, *vvfény*, *vörösvfény* - ezek mind a "vörös visszfény" kifejezés rövidített alakjai.

A 600792 tokenből álló klinikai korpuszban 3154 különböző rövidítést azonosítottunk automatikus módszer alkalmazásával (l. a 4.2 bekezdést). Egy rövidítésnek tekintettük azokat a rövidítéssorozatokat is, amelyeket nem tör meg semmilyen teljes szóalak. Természetesen ezeknek a szekvenciáknak az egyes tagjai különálló rövidítések is lehetnek. A következő példamondatban tehát négy rövidítés(sorozat) található.

*Dg : Tu. pp. inf et orbitae l. dex. , Cataracta incip. o. utr. , Hypertonia.*

A rövidítések:

*Dg*,
*Tu. pp. inf*,
*l. dex.*,
*incip. o. utr.*.

A példában szereplő utolsó minta félrevezető, hiszen az *incip.* token az őt megelőző szóhoz (*Cataracta*) kapcsolódik szemantikailag, ami viszont nem része a mondatban felismert rövidítések halmazának. A kifejezések ilyen vegyes formában való leírása igen gyakori, továbbá változó a rövidített és a teljes alakban kiírt szavak megválasztása is.

A rövidítéssorozatok egyes tagjainak a jelentése a szövegkörnyezet figyelembevétele nélkül általában nem határozható meg egyértelműen az egyes rövidítések nagyfokú többértelműsége miatt. Így ha létezne is a magyar orvosi nyelvre vonatkozó rövidítések teljes és jól használható listája, az egyes rövidítések erre való illesztése nem oldaná meg a problémát, csupán javaslatokat tudna tenni a lehetséges feloldásokra. Sok esetben egyetlen önmagában álló rövidítésre nagyon nagy számú javaslat érkezhet. (További problémát jelent a klinikai dokumentumok keverék magyar-latin nyelvezete, ezért már a rövidítéseknél is fontos azok nyelvének megkülönböztetése.)

Annak ellenére azonban, hogy az egyes rövidítések önmagukban állva erősen többértelműek, gyakran fordulnak elő rövidítéssorozatok részeként, ahol biztosabban meghatározható az egyértelmű jelentésük. Például, az "o." rövidítés bármely o-val kezdődő magyar vagy latin szó rövidítése is lehet. Még az orvosi szaknyelvre szűkítve is igen nagy a lehetőségek száma. Az általunk vizsgált klinikai korpusz szemészeti részében azonban az "o." rövidítés csak elvétve fordul elő önmagában, sokkal inkább olyan szerkezetekben, mint például "o. s.", "o. d.", vagy "o. u.", melyek jelentése *oculus sinister* (bal szem), *oculus dexter* (jobb szem), illetve *oculi utriusque* (mindkét szem). Az ilyen összetételekben az "o." jelentése már egyértelműen meghatározható. Természetesen az ugyanazzal a jelentéssel bíró rövidített alakoknak is számos variációja előfordulhat, így az "o.s." gyakori változatai például az "o. sin.", "os", "OS" stb. A példában szereplő kifejezések változataira vonatkozó gyakorisági adatokat tartalmaz a 2. táblázat.

2. táblázat. Három gyakori kifejezés: *oculus sinister*, *oculus dexter* és *oculi utrisque* néhány rövidített alakja, azok korpuszbeli gyakoriságával.

| oculus sinister | freq | oculus dexter | freq | oculi utriusque | freq |
|---|---|---|---|---|---|
| o. s. | 1056 | o. d. | 1543 | o. u. | 897 |
| o.s. | 15 | o.d. | 3 | o.u. | 37 |
| o. s | 51 | o. d | 188 | o. u | 180 |
| os | 160 | od | 235 | ou | 257 |
| O. s. | 118 | O. d. | 353 | O. u. | 39 |
| o. sin. | 348 | o. dex. | 156 | o. utr. | 398 |
| o. sin | 246 | o. dex | 19 | o. utr | 129 |
| O. sin | 336 | O. dex | 106 | O. utr | 50 |
| O. sin. | 48 | O. dex. | 16 | O. utr. | 77 |

Elsődleges célunk az olyan rövidítéssorozatok felismerése volt, melyek egyben vizsgálva egyértelműen feloldhatóak. Mivel sok esetben teljes kijelentések,

vagy akár mondatok vannak csak rövidített alakokkal megfogalmazva, ezért az első lépés a hosszabb rövidítéssorozatok önálló jelentéssel bíró partíciókra való optimális felosztása. A fenti példában szereplő "incip. o. utr." sorozat optimális felbontását az "incip." és "o. utr." különválasztásával kapjuk, akkor is, ha az "incip." szó jelentése önmagában nem értelmezhető, azonban az nem része az "o. utr." tokenek által rövidített kifejezésnek sem.

## 3. Módszerek

A feladat során szemészeti osztályon keletkezett magyar nyelvű klinikai szövegekkel foglalkoztunk. Először a rövidítéseket azonosítottuk, majd három módszert alkalmaztunk a jelentéssel bíró egységek felismerésére és feloldására. Az utóbbi két problémát mindhárom módszer esetén egy lépésben oldottuk meg, ezzel érve el egyszerre az optimális lefedettséget és a jelentés meghatározását.

### 3.1. Rövidítések felismerése

A rövidítések azonosítása során nem támaszkodhatunk olyan felszíni tulajdonságokra, amik általános esetben egy token rövidítés mivoltára utalnának (pont a szó végén, csupa nagybetűs mozaikszavak, stb.). Ezért néhány heurisztikus szabályt alkalmaztunk, többek között a következő jellemzők figyelembevételével: pont jelenléte, vagy hiánya a szóalak végén; a szóalak hossza; magánhangzók és mássalhangzók aránya a szóalakon belül; a kis- és nagybetűk aránya a szóalakon belül; a HuMor morfológiai elemző [5,6] ítélete az adott szóalakról. A rövidítéseket azonosító algoritmus részletes ismertetésére ebben a cikkben nincs módunk. A rövidítések felismerésére alkalmazott módszerünkkel magasabb fedés és alacsonyabb pontosság garantálható, ami a további feldolgozás szempontjából előnyös. Célunk nem az egyes rövidített alakokat jelölő tokenek kinyerése, hanem a rövidítéssorozatok megtalálása, amiket később bontunk szemantikailag releváns részekre. Így, ha egy önmagában álló szót tévesen rövidítésként jelölünk úgy, hogy egyik szomszédja sem rövidítés, akkor az nem kerül a feloldandó rövidítések közé. Másrészt viszont, ha egy tokent tévesen beveszünk egy rövidítéssorozatba, akkor a feloldó algoritmus fogja biztosítani azt, hogy ne kerüljön feloldásra, hiszen nem tud majd rá optimálisan illeszthető feloldást találni. Például, az *Exstirp. tu. et reconstr. pp. inf. l. d.* sorozatban az *et* latin szó nem rövidítés. A sorozat feldarabolása során nem is lesz semmilyen szemantikailag összetartozó csoport része, sem az őt megelőző, sem a rákövetkező szóalakokhoz nem csatolható.

### 3.2. Rövidítések feloldása

**Feloldási lehetőségek keresése külső erőforrásokban.** Miután kinyertük a lehetséges rövidítéssorozatokat, egy maximális lefedést biztosító feloldási javaslatokat generáló rendszert alkalmaztunk. Az algoritmus a következő: egy rövidítéssorozat esetén annak összes lehetséges, nem átfedő felosztására reguláris

kifejezéseket generálunk, amiket aztán a rendelkezésünkre álló lexikonokra illesztünk. A reguláris kifejezések a rövidítés szabályai alapján jönnek létre, mint például minden egyes betű a rövidített kifejezés egy szavának kezdőbetűje, vagy többtagú rövidítések esetén, az egyes tagok felelnek meg egy-egy szó kezdetének. A 3. táblázat tartalmaz néhány ilyen szabályt leíró mintát. A létrejött minták száma és komplexitása arányos a vonatkozó rövidítéssorozat hosszával.

3. táblázat. Két rövid rövidítésből generált illesztendő reguláris kifejezések.

| rövidítés | regexp | illeszkedő feloldás | regexp | illeszkedő feloldás |
|---|---|---|---|---|
| o. s. | `o[^ ]* s[^ ]*` | oculus sinister | | |
| os | `os[^ ]*` | osteoporosis | `o[^ ]* s[^ ]*` | oculus sinister |

A reguláris kifejezések illesztésére használt lexikonok egyike a BNO kódrendszer szemészeti szekcióinak leírásaiból és az Orvosi helyesírási szótárból [7] készített, 3329 elemet tartalmazó szólista volt. A másik lexikon egy kisebb méretű, szakterületi szakértő segítségével kézzel készített doménspecifikus kifejezéslista volt. Ebbe a listába olyan kifejezések kerültek be, amelyek olyan hétköznapi kifejezések rövidítései, melyek a szemészeti leírásokban egyedi feloldással, jelentéssel bírnak (például: “mou”, azaz *méterről olvas ujjat*). A feloldási javaslatok rangsorolásánál figyelembe vettük, hogy a javaslat melyik lexikonból származik.

A feloldási javaslatok meghatározása után azok mindegyike egy pontszámot kap. A legnagyobb lefedettséget és a legjobb feloldást egyszerre előnyben részesítő pontszámot három tulajdonság alapján határozzuk meg: 1) a feloldott alak hány tokent fed le az eredeti rövidítéssorozatból, 2) hány tokenből áll a rövidítéssorozat feldarabolása során keletkezett legnagyobb partíció, 3) hány tokenből áll a rövidítéssorozat feldarabolása során keletkezett legkisebb partíció. A fenti példában szereplő *Exstirp. tu. et reconstr. pp. inf. l. d.* sorozat esetén, annak az *Exstirp. tu. – et – reconstr. pp. inf. – l. d.* felbontására vonatkozó három szám a sorrendnek megfelelően: 7, 3 és 2.

**Feloldás keresése a korpusz alapján.** A fent ismertetett módszer legnagyobb hátránya az, hogy csak a hivatalos leírásokra illeszthető rövidítések oldhatóak fel a segítségével. A klinikai szövegekben azonban szabadon rövidítenek mindent, az ezeknek megfelelő kifejezések pedig nem találhatóak meg a hivatalos erőforrásokból épített lexikonokban. A viszonylag szűk domén miatt azonban feltételezhetjük, hogy az ilyen rövidített kifejezéseknek (vagy azok egyes részeinek) a kifejtett alakjai is legalább egyszer szerepelnek a korpuszban. Ezért ebben az esetben is a rövidítéssorozatok összes lehetséges feldarabolása során kapott egységekből képzett reguláris kifejezéseket illesztjük magára a korpuszra. Azzal a különbséggel tesszük ezt, hogy az egytagú darabokra kapott eredményeket nem vesszük figyelembe (ezek az általános szavak miatt a feloldási javaslatok listájában csak a zajt növelnék), illetve a korpuszban adott gyakoriságnál ritkábban előforduló illeszkedő kifejezéseket sem vesszük hozzá az eredményekhez.

**Korpuszillesztés és külső erőforrások együttes alkalmazása.** A harmadik esetben a fenti két módszert együtt alkalmaztuk, ezáltal érvényesítve előnyeiket és egymás által pótolva hiányosságaikat. A rövidítéssorozatok összes lehetséges felbontására először a korpuszban való keresést végezzük el, majd az így megkapott, részlegesen feloldott sorozatokra a külső lexikonokban is elvégezzük a reguláris kifejezések illesztését. Ezáltal a korpusz alapján való illesztésből maradt "lyukak" pótolhatóak.

A korpusz ilyen módon való felhasználása nem felügyelt módszerrel történik, ezáltal bármilyen más olyan aldoménre alkalmazható, amire egy nyers korpusz rendelkezésre áll. Ahogy a kiértékelés során később részletezzük, a korpuszban való kereséssel a rendszer robosztusabbá tehető, a teljesítmény sokkal kisebb mértékben esik a kézzel készített lexikon méretének csökkentése esetén.

## 4. Eredmények

A klinikai korpuszból 23, előre tokenizált dokumentumot különítettünk el tesztelési célra (összesen 4516 token). Ebben a teszthalmazban az automatikus felismerő 323 különböző rövidítést, illetve rövidítéssorozatot azonosított. Ezek közül kézzel választottuk ki azt a 44 egyedi rövidítéssorozatot (összesen 140 token), melyre a kiértékelést végeztük. Ezek a sorozatok legalább kétszer előfordulnak a tesztkorpuszban, és legalább két token hosszúságúak. A 4 táblázatban ezek közül szerepel néhány példa, a különböző rendszerek által generált feloldásokkal és a szakértő segítségével meghatározott tényleges feloldással együtt. Az automatikus rendszereknél az első helyre rangsorolt javaslatot tekintettük a végleges feloldásnak.

A kiértékelést két szinten végeztük. Megvizsgáltuk az egyes rendszerek teljesítményét a teljes, többszavas feloldások szintjén, és az egyedi tokenek szintjén is. Az első esetben, egy sorozat feloldása akkor és csak akkor helyes, ha annak minden tagját sikerült helyesen meghatározni. A második esetben a helyesen feloldott tokenek számát mértük, ami nyilvánvalóan jobb mérési eredményeket adott minden rendszer esetén. A teljesítmény mérésére a fedés, pontosság és F-mérték metrikákat használtuk a következő definíciókkal: a pontosság a helyes feloldások száma osztva az összes, bármilyen minőségű feloldás számával (sorozat- vagy tokenszinten), a fedés a helyes feloldások száma osztva az összes elem számával (sorozat vagy token szinten). Az F-mérték pedig a kettő harmonikus közepe. Az 5. táblázatban szerepelnek az automatikus kiértékelés számszerű eredményei.

Az eredményekből világosan látszik, hogy a korpusz önmagában való használata nem kielégítő. Ez nem is meglepő, hiszen éppen a leggyakoribb rövidítések azok, amelyek sosem szerepelnek kifejtett formában a szövegekben. A külső erőforrásokra tehát szükség van, de a kizárólag ezekre építő rendszer teljesítménye is rosszabb, mint a korpuszt és a lexikonokat együtt használóé.

A lexikonokat használó rendszerek esetén külön megvizsgáltuk a kézzel készített szótár jelentőségét. Ehhez a kiértékelést elvégeztük ennek a lexikonnak egy csökkentett verziójának használata mellett is. Az eredetileg 97 rövidítést, és azok feloldását tartalmazó listát 70-re csökkentettük, ami 28%-os méretváltó-

4. táblázat. Néhány példa az egyes rendszerek által automatikusan feloldott rövidítéssorozatokra, összehasonlítva a szakértő által megadott tényleges feloldással.

| | **Cat. incip. o. utr.** |
|---|---|
| 1. módszer | cat. incip. oculi utriusque |
| 2. módszer | cat. incip. o. utr. |
| 3. módszer | cataracta incipiens oculi utriusque |
| gold standard | cataracta incipiens oculi utriusque |
| | **Myopia c. ast. o. utr.** |
| 1. módszer | myopia kritikus fúziós frekvencia ast. oculi utriusque |
| 2. módszer | myopia cum ast. o. utr. |
| 3. módszer | myopia cum astigmia oculi utriusque |
| gold standard | myopia cum astigmia oculi utriusque |
| | **myop. maj. gr. o. u.** |
| 1. módszer | myop. maj. gr. oculi utriusque |
| 2. módszer | myop. maj. grad. o. utr. |
| 3. módszer | myopia maj grad. oculi utriusque |
| gold standard | myopia major gradus oculi utriusque |
| | **med. gr. cum** |
| 1. módszer | med. gr. cum |
| 2. módszer | med. gr. cum |
| 3. módszer | med. gr. cum |
| gold standard | medium gradus cum |

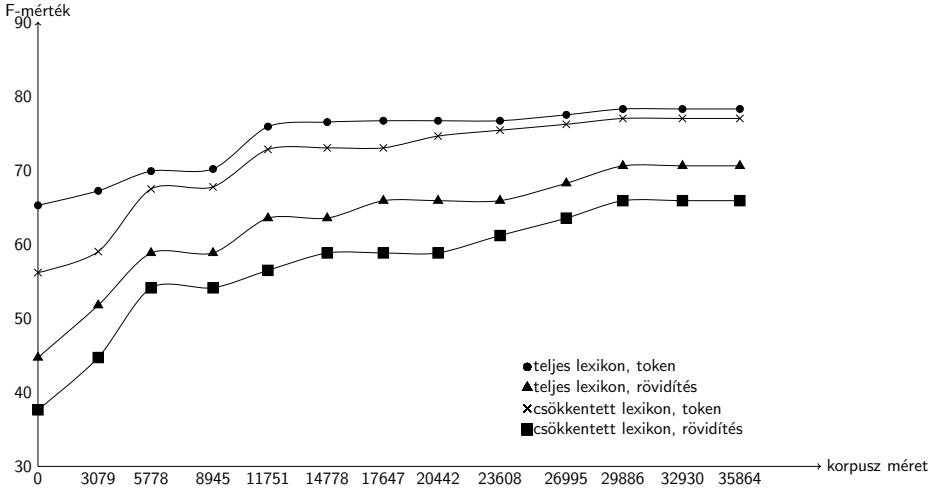
zást jelent. A másik lexikon mérete minden mérés során ugyanakkora volt. Bár a méretcsökkentés során mindegyik rendszer teljesítménye romlott, ez a romlás a korpuszt is felhasználó esetben sokkal kisebb mértékű. Ebben az esetben azokat a kifejezéseket, amelyeket a lexikonból töröltünk, a rendszer automatikusan pótolta a korpuszból. Az 1. ábrán a korpuszt használó rendszerek tanulási görbéje látszik a felhasznált korpusz méretének függvényében. Az $x$ tengely 0 pontja megfelel a csak lexikont használó rendszernek (0 méretű korpusz). Ebben a pontban a saját lexikon méretének csökkentésével járó teljesítménybeli különbség még jelentős mértékű, de ezt a lemaradást a felhasznált korpusz növelésével a rendszer automatikusan behozza.

## 5. Konklúzió

Bemutattuk, hogy a magyar nyelvű klinikai dokumentumokban található rövidítések feloldása során szükség van ugyan külső lexikai erőforrásokra, a feloldás minősége azonban jelentős mértékben javítható a korpuszra alapuló nem felügyelt tanulási algoritmus használatával. Ezáltal megspórolható a drága és nagy erőfeszítésekkel járó, szigorúan doménspecifikus lexikonok kézi összeállítása. A rövidítések többértelműségének problémáját pedig azok sorozatokban való kezelésével oldottuk meg, így elkerülhetővé vált az egytagú rövidítések lehetséges

5. táblázat. A kiértékelés eredményei tokenek és teljes rövidítések szintjén, teljes és csökkentett saját lexikon használata mellett

| | pontosság | | fedés | | F-mérték | |
|---|---|---|---|---|---|---|
| | abbr. | token | abbr. | token | abbr. | token |
| 1. módszer (teljes lexikon) | 46.34% | 78.57% | 43.18% | 55.79% | 44.70% | 65.25% |
| 1. módszer (csökkentett lexikon) | 39.02% | 68.04% | 36.36% | 47.82% | 37.64% | 56.17% |
| 3. módszer (teljes lexikon) | 73.17% | 86.08% | 68.18% | 71.73% | 70.58% | 78.26% |
| 3. módszer (csökkentett lexikon) | 68.29% | 85.08% | 63.63% | 70.28% | **65.88%** | **76.98%** |
| 2. módszer (lexikon nélkül) | 6.66% | 41.79% | 4.54% | 20.28% | 5.4% | 27.31% |




1. ábra. Az egyes rendszerek tanulási görbéje a felhasznált korpusz méretének függvényében

feloldásaiból kialakuló kezelhetetlen méretű keresési tér generálása. A mérési eredményekből kiderült, hogy bár van még lehetőség a minőségbeli javulásra a pontosság szempontjából, azonban látható az is, hogy a rendszer könnyen adaptálható bármilyen szűk domén rövidítéseinek feloldására.

## Köszönetnyilvánítás

Ez a munka részben a TÁMOP–4.2.1./B–11/2-KMR-2011-0002 és a TÁMOP–4.2.2./B–10/1-2010-0014 pályázatok támogatásával készült.

## Hivatkozások

1. Siklósi, B., Orosz, G., Novák, A., Prószéky, G.: Automatic structuring and correction suggestion system for Hungarian clinical records. In De Pauw, G., De Schryver, G.M., Forcada, M., M. Tyers, F., Waiganjo Wagacha, P., eds.: 8th SaLTMiL Workshop on Creation and use of basic lexical resources for less-resourced languages. (2012) 29.–34.
2. Siklósi, B., Novák, A., Prószéky, G.: Context-aware correction of spelling errors in Hungarian medical documents. In Dediu, A.H., Martin-Vide, C., Mitkov, R., Truthe, B., eds.: Statistical Language and Speech Processing. Volume LNAI 7978., Springer Verlag (2013)
3. Barrows, J.R., Busuioc, M., Friedman, C.: Limited parsing of notational text visit notes: ad-hoc vs. NLP approaches. Proceedings of the AMIA Annual Symposium (2000) 51–55
4. Liu, H., Lussier, Y.A., Friedman, C.: A study of abbreviations in the UMLS. Proceedings of the AMIA Annual Symposium (2001) 393–397
5. Novák, A.: What is good Humor like? In: I. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, SZTE (2003) 138–144
6. Prószéky, G., Kis, B.: A unification-based approach to morpho-syntactic parsing of agglutinative and other (highly) inflectional languages. In: Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. ACL '99, Stroudsburg, PA, USA, Association for Computational Linguistics (1999) 261–268
7. Fábián, P., Magasi, P.: Orvosi helyesírási szótár. Akadémiai Kiadó, Budapest (1992)