Nem felügyelt módszerek alkalmazása releváns kifejezések azonosítására és csoportosítására klinikai dokumentumokban

Siklósi Borbála[1], Novák Attila[1,2]

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

Kivonat A kórházi körülmények között létrejövő klinikai dokumentumok feldolgozása a nyelvtechnológia egyik központi kutatási területévé vált az utóbbi időben. A más jellegű, általános nyelvezetű szövegek feldolgozására használt kész eszközök azonban nem alkalmazhatóak, illetve gyengén teljesítenek a speciális orvosi szövegek esetén. Továbbá számos olyan feladat van, amelyek során a szakkifejezések azonosítása és a közöttük lévő kapcsolatok meghatározása nagyon fontos lépés, azonban csak külső lexikai erőforrások, tezauruszok és ontológiák segítségével oldhatók meg. Az olyan kisebb nyelvek esetén, mint a magyar, ilyen tudásbázisok nem állnak rendelkezésre. Ezért a szövegekben lévő információk annotálása és rendszerezése emberi szakértői munkát igényel. Ebben a cikkben bemutatjuk, hogy statisztikai módszerekkel milyen módon alakíthatók át a nyers dokumentumok egy olyan előfeldolgozott, részben strukturált formára, ami ezt az emberi munkát könnyebbé teszi. A csupán a korpusz felhasználásával alkalmazott modulok felismerik és feloldják a rövidítéseket, azonosítják a többszavas kifejezéseket és meghatározzák azok hasonlóságát. Végül létrehoztuk a szövegek egy magasabb szintű reprezentációját, ahol az egyes kifejezések helyére a hasonlóságuk alapján kialakított klaszterek azonosítóját helyettesítve a szövegek egyszerűsíthetőek, a gyakran ismétlődő mintázatok általános alakja meghatározható.

1. Bevezetés

A klinikai rekordok olyan dokumentumok, amelyek kórházi körülmények között jönnek létre a mindennapi esetek, kezelések dokumentálása céljából. Ezeknek a szövegeknek a minősége messze elmarad az orvosbiológiai szövegekétől, amelyek feldolgozásával jóval több tanulmány foglalkozik éppen könnyebb kezelhetőségük miatt. Az orvosbiológiai szövegek általában angol nyelvűek, és tudományos folyóiratokban, könyvekben, kiadványokban jelennek meg, nyelvezetük követi a nyelvi és helyesírási normákat [1,2]. Ezzel szemben, a klinikai beteglapok olyan strukturálatlan szövegek, amelyek minden ellenőrzés nélkül jönnek létre. Ezért gyakran fordulnak elő bennük helyesírási hibák, elírások, és az egyedi szóalakok használata is igen jellemző ezekre a dokumentumokra. Nyelvezetük pedig

gyakran a helyi nyelv (a mi esetünkben a magyar) és a latin sajátos keveréke [3,4]. Jellemző rájuk továbbá a gyakran egyedi módon is használt rövidítések magas aránya, olyannyira, hogy akár teljes állítások, mondatok is csupán rövidített alakokból állnak.

A klinikai dokumentumok egy további jellemzője, hogy olvasóik általában maguk a lejegyzést készítő vagy diktáló orvosok, ezért az egyedi nyelvhasználat és rövidítési szokások nem okoznak információveszteséget, amikor az orvos újra elolvassa ezeket. Azonban az egyes betegek kórtörténetének tárolása mellett az ezekből a dokumentumokból kinyerhető információk az orvostudomány más területein is felhasználhatóak lehetnének. Ahhoz, hogy hozzáférjünk ezekhez a rejtett adatokhoz, a szövegekben található tények és állítások hatékony reprezentációjára van szükség.

Több kísérlet irányult már általános szövegekre működő eszközök klinikai dokumentumokra való alkalmazására, azonban ilyenkor teljesítményük jóval gyengébb, mint általános, jól formált szövegek esetén (pl. [5]). Azok az eszközök, amik pedig doménspecifikus szövegek feldolgozására alkalmazhatóak, általában valamilyen külső, kézzel készített lexikai erőforrásokat, ontológiákat használnak. Azon nyelvek esetén azonban, amelyekre kevés ilyen erőforrás áll rendelkezésre, ezek a módszerek nem alkalmazhatóak, az erőforrások létrehozása pedig jelentős emberi munkát igényel. Egy további lehetőség a klinikai dokumentumokban használt nyelv alnyelvként való kezelése [6]. Azonban ehhez is szükség van az alnyelv kifejezéseinek doménspecifikus kategorizációjára, ami szintén nem oldható meg teljesen automatikus módszerekkel [7,8,1].

A már meglévő eszközök adaptálhatósága és a strukturált erőforrások építésének támogatása céljából egy magyar szemészeti korpuszt vizsgáltunk meg. Majd különböző, nem felügyelt statisztikai módszerek alkalmazásával tettünk kísérletet többféle információ felfedezésére a nyers korpuszban. Bár az egyes modulok által létrejött eredmény önmagában nem tekinthető a dokumentumok információtartalmát teljesen lefedő reprezentációnak, azonban az ezekből a félig strukturált adatokból létrejövő csoportok felhasználhatóak a későbbi konstrukciók megalkotása során. Mindegyik modul magja a nyers korpuszból kinyert statisztika, csupán néhány ponton volt szükség alapvető nyelvi szabályok, illetve erőforrások bevonására.

2. A korpusz

Vizsgálataink során anonimizált, szemészeti osztályon keletkezett magyar nyelvű dokumentumokat használtunk. A rendelkezésünkre álló korpusz mérete 334.546 token (34.432 mondat). A korpusz a feldolgozás előtt [3]-ban ismertetett struktúrájú magas szintű xml formátumban volt. Ebben jelölve voltak a mondat- és tokenhatárok, illetve a szófaji egyértelműsítés eredménye is. A szegmentálás és szófaji egyértelműsítés automatikusan történt [9]-ben és [5]-ben bemutatott módszerekkel. Bár ezt a két előfeldolgozási lépést a legtöbb nyelv esetében általános nyelvű szövegekre megoldották már automatikus módszerekkel, az előbb említett tanulmányok kitérnek arra is, hogy esetünkben a teljesítmény jelentősen

elmarad az elvárthoz képest a klinikai szövegeken. Jelen munkánkban azonban, mivel ezek az előfeldolgozási lépések elengedhetetlenek, elfogadottnak tekintettük a korpusz ilyen minőségű kiindulási állapotát.

Egy általános nyelvű magyar korpuszhoz hasonlítva, számos jelentős különbség fedezhető fel a két domén között, ami magyarázatot ad arra, hogy az általános szövegeken akár bonyolultabb feladatok esetén is jól teljesítő eszközök miért nem alkalmazhatóak a klinikai dokumentumokra. A különbség a két szövegtípus között nem csak azok tartalmában nyilvánul meg, hanem már a nyelvtani szerkezetek és a szövegekben előforduló szóalakokban is. A két domén részletes összehasonlítása megtalálható [10]-ben és [11]-ben.

3. Az alkalmazott módszerek

Ebben a fejezetben négy módszert mutatunk be, amelyeket később egymással kombinálva alkalmaztunk a klinikai szövegekre, ami az egyes dokumentumok félig strukturált kivonatát eredményezte. Az első modul a rövidítések feloldásáért felelős, a második összetett szakkifejezések felismerését végzi, illetve rangsorolja ezeket, a harmadik szó-, illetve jelen esetben kifejezéspárok hasonlóságát határozza meg, a negyedik pedig fogalmi klasztereket hoz létre, illetve helyettesíti be ezeket az eredeti szövegekbe. Mind a négy modul működése során a korpusz statisztikai jellemzői a meghatározók.

3.1. Rövidítések feloldása

A rövidítésfeloldás feladatát többen a jelentés-egyértelműsítés (word sense disambiguation, WSD) egy speciális eseteként kezelik ([12]). A jelentés-egyértelműsítés megoldása során legjobb eredményt elérő módszerek felügyelt gépi tanulási algoritmusokat alkalmaznak. A magyar orvosi nyelvhez viszont sem kézzel előre annotált adatok, sem a lehetséges jelentések, azaz feloldási javaslatok adatbázisa nem áll rendelkezésre, ami a felügyelt tanulási módszerek alkalmazhatóságának előfeltétele ([13]). A szintén WSD feladatokra alkalmazott nem felügyelt módszerek két fázisból állnak. Az egyértelműsítés előtt meg kell határozni a lehetséges jelentések halmazát is (word sense induction, WSI). Kontextuális jellemzők alapján ugyan meghatározhatóak a lehetséges jelentések egy adott korpuszból, azonban ehhez nagy méretű korpuszra van szükség, különösen akkor, ha a többértelmű kifejezések (rövidítések) aránya olyan nagy, mint a klinikai szövegekben. Mivel kellően nagy méretű korpusz nem állt rendelkezésünkre, ezért ezt a megközelítést sem követhettük.

Így egy olyan korpuszalapú megoldást dolgoztunk ki a rövidítések automatikus feloldására, ami csupán kiegészítésként fordul a néhány, magyar nyelven elérhető, klinikai kifejezéseket is tartalmazó erőforrásokhoz. Mivel a csupán a korpuszra építő módszer nem teljesített kielégítően, ezért szükséges volt egy doménspecifikus lexikon létrehozása is. Az orvosi, klinikai kifejezéseket teljesen lefedő adatbázis helyett [14]-ben megmutattuk, hogy egy kisebb, doménspecifikus lexikon is elégséges, az ebben definiálandó rövidítések pedig a korpuszból

közvetlenül kinyerhetők. Miután ez a lexikon rendelkezésre áll, valamint a rövidítések azonosítása is megtörtént, a feloldást rövidítések sorozatára végeztük. Erre azért volt szükség, mert annak ellenére, 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 rövidítések feloldása során tehát olyan rövidítéssorozatokat tekintettünk kiindulási egységnek, melyek a szövegben egymást folytonosan, megszakítás nélkül követő rövidített alakok vagy rövidítések sorozata.

Továbbá, az olyan vegyes kifejezések egyben tartása végett, melyeknek nem minden tagja rövidített, a rövidítéssorozatokhoz azok adott méretű környezetét is csatoltuk. A feloldás során ilyen módon elérhető szövegkörnyezet a rövidítések jelentésének egyértelműsítésében is szerepet játszik, hiszen egy konkrét felszíni alakkal rendelkező rövidítés jelentése (feloldása) a környezetétől függően változhat.

Az így automatikusan felismert, majd a szövegkörnyezettel kiegészített rövidítéssorozatok feloldására először a korpuszból nyertünk ki lehetséges feloldásjelölteket, majd csupán az ebben a lépésben nem, vagy csak részlegesen feloldott rövidítéssorozatok feloldása során fordultunk a külső lexikonhoz. Az algoritmus részletei és eredményei megtalálhatóak [14]-ben és [15]-ben. Az eredmények alapján kimutatható, hogy bár a korpusz önmagában nem elégséges minden rövidítés feloldásához, ennek használatával pontosabb és helyesebb feloldások nyerhetők ki, mint csupán külső lexikon alkalmazásával.

A 1. táblázatban látható néhány automatikusan felismert rövidítés a hozzájuk tartozó feloldásokkal. A feloldásokhoz a magyar és latin változatot is megjelenítjük, ahol ezek a változatok elérhetőek és relevánsak (pl. *mindkét szem*; *oculi utriusque*), illetve több latin, vagy több magyar feloldás is szerepelhet (pl. *szemfenék*; *fundus oculi*; *fundus*).

3.2. Többszavas szakkifejezések azonosítása

A klinikai nyelvben (bármely más szaknyelvhez hasonlóan) gyakoriak az olyan többszavas kifejezések, melyek együtt jelölnek egy fogalmat. Mivel olyan releváns információk jelenhetnek meg ilyen formában, mint a betegségek, kezelések, testrészek neve, ezért fontos ezek azonosítása. Az ilyen kifejezések azonosítására nem elegendő egy általános lexikon, hiszen vannak olyan kifejezések, melyek az általános nyelvben nem feltétlenül tartoznak össze. Például a *szem* szó, mint testrész, a szemészeti szövegekben önmagában nem sok információt tartalmaz, viszont a

1. táblázat. Néhány rövidítés és a hozzá tartozó feloldások

Rövidítés	Feloldások
mydr	mydrum
mksz	mindkét szem; oculi utriusque
V	visus
D	dioptria
mou	méterről olvas ujjat
ünj	üveg nem javít
o. u	oculi utriusque; mindkét szem
F	szemfenék; fundus oculi; fundus
j.o.	jobb oldal

bal szem, *jobb szem*, *mindkét szem* kifejezések már konkrétan meghatározzák a dokumentumban leírt jelenségek pontos helyét. Éppen emiatt a szemészeti korpuszban a szem szó önmagában nem is gyakran fordul elő. Az ilyen kifejezések felismerésére tehát jól alkalmazható a kölcsönös információ (mutual information) és a kollokációk vizsgálatán alapuló módszerek, melyek éppen a korpuszbeli előfordulások alapján definiálhatóak. Ezeknek a módszereknek a többszavas szakkifejezések felismerésére való alkalmazását [16] foglalja össze, majd az egymásba ágyazott kifejezések problémájára is megoldást nyújtó c-value módszert ismerteti.

Ennek a c-value algoritmusnak egy módosított változatát használtuk. Először egy nyelvi szűrőt alkalmaztunk annak érdekében, hogy a kifejezésjelöltek listáján csak nyelvtani szempontból is helyes kifejezések szerepeljenek. A megengedett kifejezések formája a következő:

$$(FN|ADJ|IGE\_OKEP|IGE\_MIB)^{+}FN$$

Ez a minta biztosítja, hogy egyrészt csak főnévi csoportok legyenek a jelöltek között, másrészt kizárja a gyakori kifejezéstöredékeket is. Természetesen más jellegű kifejezések, mint például igei csoportok, is relevánsak lehetnek. Azonban, ahogy a 2. fejezetben bemutattuk, az igék gyakorisága alacsony a klinikai szövegekben. Ezért egy viszonylag kis méretű korpuszból nem építhetőek pontos statisztikai modellek az ilyen, ritkábban előforduló kifejezésekre.

Miután az összes, a fenti mintára illeszkedő n-gramot kigyűjtöttünk ($1 < n < 10$), mindegyikre meghatároztuk a hozzá tartozó c-value-t, ami az adott n-gram kifejezés voltára utaló mérőszám. Ez az érték négy komponens alapján határozható meg:

- a kifejezésjelölt gyakorisága;
- annak gyakorisága, hogy hányszor fordul elő hosszabb kifejezés részeként;
- az ilyen, hosszabb kifejezések száma; és
- a kifejezés hossza.

Ezeket a statisztikákat a korpusz alapján lehet meghatározni. A c-value számítását végző algoritmus részletei [16]-ban találhatóak meg. A 2. táblázatban látható néhány többszavas kifejezés, amit egy dokumentumból nyertünk ki, a hozzájuk tartozó c-value értékkel.

2. táblázat. Egy dokumentumból kinyert többszavas kifejezések a hozzájuk tartozó c-value értékkel

Kifejezés	c-value
bal szem	2431.708
ép papilla	1172.0
tiszta törőközeg	373.0
békés elülső szegmentum	160.08
hátsó polus	47.5
tompa sérülés	12.0

3.3. Disztribúciós szemantikai modellek

A releváns kifejezések csoportosításához szükség van egy hasonlósági metrikára is, ami két kifejezés jelentésbeli távolságát határozza meg. Erre szintén olyan nem felügyelt módszert alkalmaztunk, amely a hasonlóságokat nem egy külső erőforrás, ontológia alapján határozza meg, hanem a kifejezések korpuszbeli előfordulásai, az adott korpuszban való használatuk alapján.

A disztribúciós szemantika lényege, hogy a szemantikailag hasonló szavak hasonló környezetben fordulnak elő. Tehát két szó jelentésének hasonlósága meghatározható a környezetük hasonlósága alapján. A szavak környezetét olyan jellemzőhalmazokkal reprezentáltuk, ahol minden jellemző egy relációból (r) és az adott reláció által meghatározott szóból (w') áll. Ezek a relációk más alkalmazásokban általában függőségi relációk, azonban a klinikai szövegekre ilyen elemzés a zajos mivoltuk miatt nem végezhető el kellően jó eredménnyel. Carrol et al. [17] szintén klinikai szövegekre alkalmazva csupán a vizsgált szó meghatározott méretű környezetében előforduló szavak lexikai alakjának felhasználásával építettek ilyen szemantikai modellt. Mivel a mi esetünkben a morfológiai elemzés is rendelkezésre állt, ezért a következő jellemzőket vettük figyelembe:

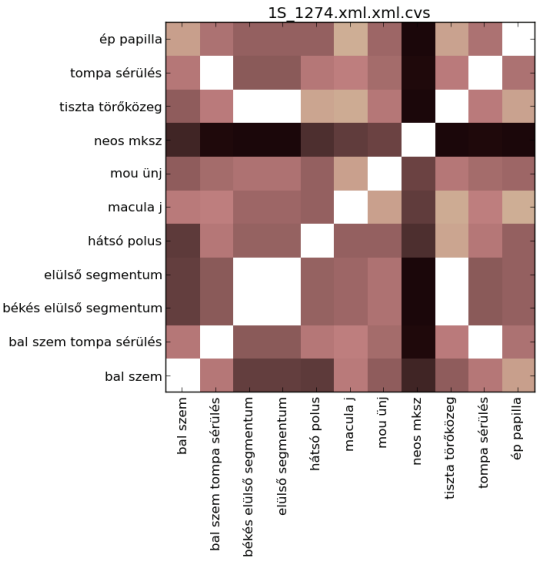
- prev_1: a szót megelőző szó lemmája
- prev_w: a szó előtt 2-4 távolságon belül eső szavak lemmái
- next_1: a rákövetkező szó lemmája
- next_w: a szó után 2-4 távolságon belül eső szavak lemmái
- pos: a szó szófaja
- prev_pos: a szót megelőző szó szófaja
- next_pos: a szót követő szó szófaja

Minden egyes jellemzőhöz meghatároztuk a korpuszbeli gyakoriságát. Ezekből a gyakoriságokból határozható meg a (w,r,w') hármas információtartalma

(I(w,r,w')) maximum likelihood becsléssel. Ezután a két szó (w_1 és w_2) közötti hasonlóságot a következő metrikával számoltuk [18] alapján:

$$SIM(w_1, w_2) = \frac{\sum_{(r,w)\in T(w_1)\bigcap T(w_2)} (I(w_1, r, w) + I(w_2, r, w))}{\sum_{(r,w)\in T(w_1)} I(w_1, r, w) + \sum_{(r,w)\in T(w_2)} I(w_2, r, w)}$$

ahol T(w) azoknak az (r,w') pároknak a halmaza, ahol az I(w,r,w') pozitív.



1. ábra. Egy dokumentumhoz tartozó kifejezések páronkénti hasonlóságát megjelenítő hőtérkép

Ezzel a metrikával bármely két kifejezés közötti disztribúciós hasonlóság meghatározható. Bár elvileg két bármilyen típusú szóra alkalmazhatjuk, egy ige és egy főnév összehasonlításának gyakorlati szempontból nem sok haszna lenne. Ezért a metrika többszavas kifejezésekre való alkalmazásából fakadó komplexitás elkerülése érdekében a többszavas kifejezéseket összevonva és az [FN][MT] címkével ellátva vettük figyelembe. Mivel minden kifejezésjelölt, az előző alfejezetben leírtaknak megfelelően, főnévi csoport, ezért ez a hasonlóság megfelel a két kifejezés közötti hasonlóságnak. A 1. ábra egy hőtérképen jeleníti meg az egy dokumentumhoz tartozó kifejezések páronkénti hasonlóságát. Minél világosabb egy négyzet, a hozzá tartozó kifejezések annál hasonlóbbak. Látható, hogy a *"tiszta törőközeg"* és a *"békés elülső segmentum"* hasonló viselkedést mutatnak, míg például a *"neos mksz"* az aktuális dokumentumból kinyert kifejezések közül egyikhez sem igazán hasonlít.

3.4. Fogalmi klaszterek és mintázatok

A szavak és kifejezések páronkénti hasonlóságából kiindulva fogalmi hierarchiát határozhatunk meg. Ehhez a leggyakoribb kifejezések és szavak csoportján agglomeratív klaszterezést hajtottunk végre. Bár adná magát, hogy a klaszterezés során szükséges távolságmetrikának a fent definiált disztribúciós hasonlóságot használjuk, ez önmagában nem bizonyult elégségesnek, illetve a klaszterezési algoritmusok rugalmassága is csökkent volna, ha csupán ezt a metrikát használjuk. Ezért az egyes kifejezéseket a többi kifejezéshez való hasonlóságukból álló jellemzővektorokkal ábrázoltuk. Így az egy kifejezéshez tartozó $c(w)$ vektor c_i eleme $SIM(w, w_i)$. Az egyes kifejezésekhez így létrehozott jellemzővektorokat klaszteresztük, ahol a klaszterek távolságát Ward ([19]) módszere alapján határoztuk meg. Ennek köszönhetően a kapott dendrogram alsó szintjein tömör, egymáshoz közel álló kifejezésekből álló csoportok jöttek létre.

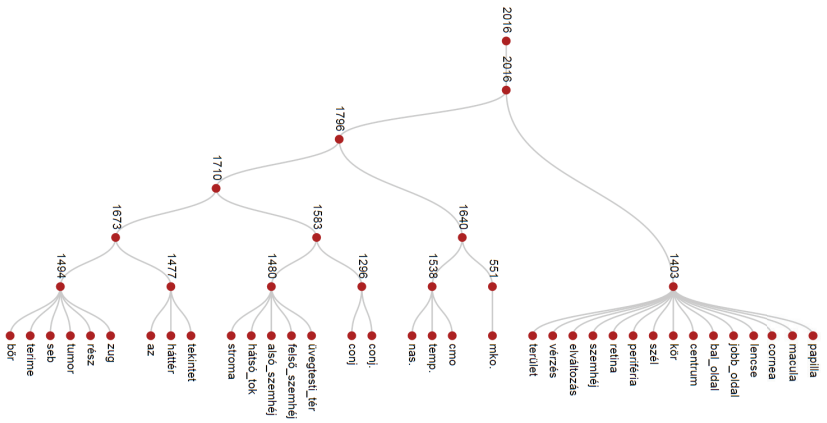
Ezeket egy, a dendrogramról szabad szemmel is jól leolvasható küszöbérték alatt összevontuk. Így a 3. táblázatban példaként felsoroltakhoz hasonló csoportok jöttek létre.

3. táblázat. Klaszterezés által létrejött levél csomópontok

ID:1403	ID:1636	ID:1549	ID:1551	ID:2045
papilla	hely	tbl	folyamat	ép papilla
macula	kh	medrol	kivizsgálás	halvány papilla
cornea	kötőhártya	üveg	érték	jó színű papilla
lencse	szaru	szemcsepp	idegentest	szűk ér
jobb oldal	conjunctiva	gyógyszer	gyulladás	ép macula
bal oldal	szemrés		retinaleválás	fénytelen macula
centrum	szempilla		látásromlás	kör fekvő retina
kör	pilla			fekvő retina
szél	könnypont			rb.
periféria				tiszta törőközeg
retina				bes
szemhéj				békés elülső szegmentum
elváltozás				békés es
vérzés				
terület				

A létrejött csoportok vagy rokonértelmű kifejezéseket, vagy szemantikai szempontból azonos szerepet betöltő kifejezéseket (pl. hét napjai, gyógyszernevek, stb.) tartalmaznak, akár rövidített alakokkal együtt (pl. *"bes"*, *"békés elülső szegmens"*, *"békés elülső szegmentum"*, *"békés es"*). Ezek mellett létrejöttek azonban olyan absztrakt csoportok is, amelyek az orvosi eljárás egyes fázisaihoz kapcsolódó kifejezések csoportjai (pl. az *"éhgyomor"* az időpontokhoz kapcsolódó kifejezésekkel került egy klaszterbe, illetve a *"strab"* és a *"párhuzamos szem"* kifejezések is az orvosi jelentőségük miatt kerültek egy csoportba). Ezeket a levél

csomópontokat összevonva, a magasabb szintű hierarchia természetesen megmarad, illetve a létrejött fa bármilyen küszöbérték mentén tovább vágható. Ezt kihasználva olyan részfákat vágtunk ki a teljes hierarchiából, amelyek a korábban kinyert kifejezéscsoportok közül az egymáshoz közel állókat fogja össze, megjelenítve azok hierarchiáját is. Egy ilyen részfa látható a 2. ábrán.



2. ábra. A teljes hierarchiából kivágott részfa, a leveleken összevont kifejezéscsoportokkal

Az orvosi szövegekben található kifejezések ilyen módon létrejött csoportosítása és rendszerezése önmagában is felhasználható lehet egy, az orvosi gyakorlatban használatos kifejezésekből álló ontológia kiindulásaként. Mivel azonban minden csoportot (illetve a létrejött hierarchia minden csomópontját) egy egyedi azonosítóval láttunk el, ezért ezek az eredeti szövegekbe visszahelyettesíthetők. Így egy tetszőleges szintű, de az eredetinél absztraktabb reprezentációját hoztuk létre az egyes dokumentumoknak. Ebből a reprezentációból könnyen azonosíthatók a dokumentumokban gyakran előforduló mintázatok, függetlenül attól, hogy a konkrét alakja egy-egy állításnak mennyire ritka, vagy gyakori kifejezést tartalmaz. A 4. táblázatban látható néhány mondat ilyen módon behelyettesített változata. A behelyettesítés következtében a mondatok nem csak egyszerűbbé válnak, hanem gyakran ismétlődő mintázatok is kiemelhetőek lesznek. A példában is megjelenő *"1889 1706 1706"* minta arról szól, hogy valamilyen állapotában valamelyik szemről milyen adatokat jegyeztek fel. Ennek a mintának a néhány leggyakoribb megjelenési formája a szövegekben: *"st. o. u."*, *"st. o. s."*, *"st. o. d."*, *"moct o. d."*, *rl. o. u."*, *"rl. o. sin."*, *"status o. s."*, *"távozáskor o. d."*, *"b-scan o. d."*, stb. Jellemző továbbá erre a mintára, hogy sor elején jelenik meg. A példamondatokban látható az is, hogy hol jelennek meg a 2. ábrán látható *2016*-os azonosítóval ellátott részfa egyes kifejezései, amik a mondatban leírt állítás helyét jelölik meg.

Bár vannak esetek, amikor az ugyanazzal az azonosítóval ellátott kifejezések nem mind tekinthetőek egy csoportba sorolhatónak, a közöttük fennálló kapcsolat hierarchiáját a részfákból meg lehet határozni, illetve a csoportokat meghatározó küszöbérték megfelelő hangolásával állítható a klaszterezés finomsága. Tapasztalataink szerint nem érdemes nagyon szűk csoportokat definiálni, hiszen sok esetben az egymással lazább kapcsolatban álló kifejezések is jól illeszthetőek az eredményül kapott mintázatokba. Jó példa erre a *NUMx* (a NUM valamilyen számot jelöl) és a *th.*, illetve ezek változatainak csoportja, ahol az egy gyógyszer adagolására vonatkozó meghatározás tulajdonképpen egy terápia. Ezzel szemben a *NUMán* sokkal inkább a vizsgálatokhoz, időpontokhoz besorolt kifejezés. Ha a csoportosításnál használt küszöbértéket alacsonyabbra állítjuk, akkor az ilyen tágabb asszociációkat elveszítjük. Az egyensúly megtalálása további kutatás témája, melyhez orvosszakértők bevonására is szükség van.

4. táblázat. Néhány példamondat, ahol a szavakat és kifejezéseket a hozzájuk tartozó klaszterazonosítóval helyettesítettük

1518 1706 1706 : ***2016*** tiszta üti *2007* , *2045* , szemfenék-szerte *2007* , *1956* , a macula_kemény_exsudatum , *2007* . *fu. o. u* : ***mko.*** tiszta üti *tér* , *ép_papilla* , szemfenék-szerte *ma-k* , *pontszerű_vérzés* , a macula_kemény_exsudatum , *oedema* .
2071 1706 1706 : ***2016*** felett és nasalisan szivárgó , ischaemiás_terület , kis neovasc._burjánzás , *2049* *flag o. d.* : ***macula*** felett és nasalisan szivárgó , ischaemiás_terület , kis neovasc._burjánzás , *macula_oedema*
2016 sima , csillogó , *2007* és a *1789* tiszta . ***cornea*** sima , csillogó , *állománya* és a *hátlapja* tiszta .
2016 tizsta . *1706* friss *1884* nem látható . ***lencse*** tizsta . *funduson* friss *kóros* nem látható .
2016 tiszta , *1789* tiszta , *1789* tiszta , *1789* békés , *1789* jól reagál . ***stroma*** tiszta , *hátlap* tiszta , *csarnok* tiszta , *iris* békés , *pupilla* jól reagál .
2016 nem vizsgálható erős_fénykerülés miatt . ***periféria*** nem vizsgálható erős_fénykerülés miatt .
1889 1706 1706 : *1998* , halvány_conjunctiva , ***2016*** epithelialis pontszerű_kiemelkedő_szürkésfehér_laesi_(j»b, balon csak NUM-NUM laesio) , a cornea_mély_rész épek ,transparens , *1812* mély , tiszta , *1789* békés , *1812* tág , kerek , centrális , *2007* jól reagál . *st. o. u* : *ép_védőszerv* , halvány_conjunctiva , ***corneán*** epithelialis pontszerű_kiemelkedő_szürkésfehér_laesi_(j»b, balon csak NUM-NUM laesio) , a cornea_mély_rész épek , transparens , *csarnok_kp* mély , tiszta , *iris* békés , *pupilla_kp* tág , kerek , centrális , *fényre* jól reagál .

4. Konklúzió

A klinikai dokumentumok mind a tartalmuk, mind a nyelvhasználatuk miatt egy doménspecifikus alnyelvet reprezentálnak. Ezeknek a szövegeknek a fő tulajdonsága a nagy mennyiségű zaj jelenléte, ami a helyesírási hibákból, a rövidítésekből és a hiányos szintaktikai szerkezetekből adódik. A szövegekben található információk kinyerését tovább nehezíti, hogy az olyan kis nyelvekre, mint a magyar, nincsenek kész lexikai erőforrások, amiket más nyelvek esetén a szövegekben szereplő kifejezések és a közöttük fennálló kapcsolatok azonosítására gyakran használnak. Ezért az ilyen lexikonok előállítása orvosszakértői feladat. A nyers dokumentumok kezdetleges előfeldolgozással történő átalakítása azonban jelentősen megkönnyítheti és hatékonyabbá teheti ezt a munkát. Cikkünkben olyan korpuszalapú módszereket mutattunk be, amik jól teljesítenek rövidítések automatikus feloldására, többszavas kifejezések felismerésére és ezek hasonlóságának meghatározására. A dokumentumoknak egy olyan félig strukturált reprezentációja jött létre ezeknek a moduloknak az alkalmazásával, ami az emberi feldolgozást támogatja, hatékonyabbá teszi. Továbbá a hasonlósági metrikák által definiált szorosabb vagy lazább kapcsolatok egy relációs tezaurusz építése során annak kezdeti állapotát képző információk is lehetnek.

Hivatkozások

1. Sager, N., Lyman, M., Bucknall, C., Nhan, N., Tick, L.J.: Natural language processing and the representation of clinical data. Journal of the American Medical Informatics Association **1**(2) (1994)
2. Meystre, S., Savova, G., Kipper-Schuler, K., Hurdle, J.: Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform **35** (2008) 128–44
3. Siklósi, B., Orosz, Gy., 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.
4. 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) 248–259
5. Orosz, Gy., Novák, A., Prószéky, G.: Lessons learned from tagging clinical Hungarian. International Journal of Computational Linguistics and Applications **5** (2014)
6. Harris, Z.S.: The structure of science information. J. of Biomedical Informatics **35**(4) (2002) 215–221
7. Friedman, C., Kra, P., Rzhetsky, A.: Two biomedical sublanguages: a description based on the theories of Zellig Harris. Journal of Biomedical Informatics **35**(4) (2002) 222–235
8. Kate, R.J.: Unsupervised grammar induction of clinical report sublanguage. J. Biomedical Semantics **3**(S-3) (2012) S4

9. Orosz, Gy., Novák, A., Prószéky, G. In: Hybrid text segmentation for Hungarian clinical records. Volume 8265 of Lecture Notes in Artificial Intelligence. Springer-Verlag, Heidelberg (2013)
10. Siklósi, B., Novák, A.: A magyar beteg. In: X. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Szegedi Tudományegyetem, Informatikai Tanszékcsoport (2014) 188–198
11. Siklósi, B., Novák, A., Prószéky, G.: Context-aware correction of spelling errors in Hungarian medical documents. Computer Speech and Language (2014)
12. Navigli, R.: A quick tour of word sense disambiguation, induction and related approaches. In: Proceedings of the 38th Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM). (2012) 115–129
13. Nasiruddin, M.: A state of the art of word sense induction: A way towards word sense disambiguation for under-resourced languages. CoRR **abs/1310.1425** (2013)
14. Siklósi, B., Novák, A., Prószéky, G.: Resolving abbreviations in clinical texts without pre-existing structured resources. In: Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing, LREC 2014. (2014)
15. Siklósi, B., Novák, A.: Detection and Expansion of Abbreviations in Hungarian Clinical Notes. In: MICAI 2013: 12th Mexican International Conference on Artificial Intelligence. Volume 8265 of Lecture Notes in Artificial Intelligence., Heidelberg, Springer-Verlag (2013) 318–328
16. Frantzi, K., Ananiadou, S., Mima, H.: Automatic recognition of multi-word terms:. the C-value/NC-value method. International Journal on Digital Libraries **3**(2) (2000) 115–130
17. Carroll, J., Koeling, R., Puri, S.: Lexical acquisition for clinical text mining using distributional similarity. In: Proceedings of the 13th international conference on Computational Linguistics and Intelligent Text Processing - Volume Part II. CICLing'12, Berlin, Heidelberg, Springer-Verlag (2012) 232–246
18. Lin, D.: Automatic retrieval and clustering of similar words. In: Proceedings of the 17th international conference on Computational linguistics - Volume 2. COLING '98, Stroudsburg, PA, USA, Association for Computational Linguistics (1998) 768–774
19. Ward, J.H.: Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association **58**(301) (1963) 236–244