A magyar beteg

Siklósi Borbála[1], Novák Attila[1,2]

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

Kivonat A klinikai szövegek feldolgozása aktív kutatási terület, melynek során az egyik legnagyobb kihívás az ilyen szövegek azon sajátosságainak a kezelése, amelyek tekintetében ezek az általános szövegektől jelentősen eltérnek. Ezek között szerepel többek között a sok szakszó és rövidítés, a szinte csak rövidítésekből és numerikus adatokból álló „mondatok", valamint a jelentős számú helyesírási és központozási hiba, amelyből többek között a mondathatárok felismerésének rendkívül nehéz volta is következik.
Cikkünkben bemutatjuk a rendelkezésünkre álló magyar klinikai korpusz jellemzőit, különös tekintettel az előbb említett tényezőkre, összevetve azt egy általános tartalmú magyar szövegeket tartalmazó korpusszal. A szövegek felszíni tulajdonságai mellett összehasonlításokat végeztünk a leggyakoribb szavak disztribúciós szemantikai viselkedése alapján is, melynek során a jelentésbeli különbségek is kimutathatóak a különböző korpuszok között.

1. Bevezetés

A klinikai dokumentumok olyan szövegek, melyek kórházi körülmények között, mindennapi eseteket dokumentálva a kezelések során jönnek létre. Minőségük tehát nem összehasonlítható az elsősorban angol nyelven szintén aktívan vizsgált orvosi-biológiai szakirodalom nyelvezetével, amelyek többszörös ellenőrzésen keresztülmenve, szigorú nyelvi szabályok betartása mellett keletkeznek [1,2]. A klinikai orvosi szövegek ezzel szemben sietve, minden nyelvi segédeszköz, vagy emberi ellenőrzés nélkül, általában strukturálatlan formában jönnek létre. Jellemző továbbá, hogy keletkezésük során ezeknek a dokumentumoknak a címzettje általában az azt leíró orvos maga, tehát az eredeti célját nem befolyásolja a sajátos nyelvezet, egyedi rövidítések, utalások használata. Ezek a dokumentumok azonban nagyon sok olyan információt és tudást tartalmaznak, amelyeket ezen az elsődleges célon túl, az orvostudomány több területén alkalmazni lehetne. Ehhez arra lenne szükség, hogy a szövegekben leírt tényállásokat olyan formára hozzuk, amely lehetővé teszi ezeknek az információknak a hatékony kinyerését.

Több kísérlet született már a természetes nyelvű szövegek feldolgozásához általánosan használt eszközök orvosi szövegeken való alkalmazására, azonban ezek teljesítménye általában messze elmarad attól a szinttől, amit általános szövegeken elérnek. Ahhoz, hogy a már bevált módszerek, vagy azoknak egy része

adaptálható legyen az orvosi szövegekre, ismernünk kell ez utóbbinak a jellemzőit, illetve az általános szövegektől való főbb eltéréseket.

Ehhez több vizsgálatot végeztünk. A korpusz alapján először a felszíni alakok statisztikai eloszlását, majd ugyanezek egy feldolgozási lépéssel későbbi szintű (szótő, szófaj, névelemek, rövidítések) előfordulását vizsgáltuk, összehasonlítva a kapott mintákat az általános korpuszból kinyert adatokkal. Általános szövegként a Szeged Korpuszt használtuk. Jól elkülöníthetővé váltak a két szövegtípusban jellemzően előforduló nyelvi szerkezetek. Az eredmények elemzése során kimutathatóak azok a szerkezeti bizonytalanságok, amelyek miatt a klinikai szövegek jóval nehezebben értelmezhetőek az általános szövegeknél. Ilyen jellemzők nemcsak a rengeteg szakkifejezés jelenléte, hanem a szövegek gyakran rendkívül pongyola megformálása és az azonos fogalmak jelölésére konkrétan használt írott alakok rendkívüli változatossága is.

Természetesen a lexikai alakok vizsgálata során azok összehasonlítása nem vizsgálható érdemben, hiszen a szakkifejezések előfordulási aránya nyilvánvalóan nagyobb a szakszövegekben. A klinikai dokumentumokra azonban jellemző, hogy az esetleírásoknál, különösen a panaszok felvétele során egészen hétköznapi történetek leírása is szerepel. Ennek a kevert orvosi nyelvnek a statisztikai jellemzői is felismerhetők a korpusz önmagában való vizsgálata során.

Tanulmányunk célja a részletes statisztikai vizsgálatok alapján azon jelenségek bemutatása, amik igazolják az orvosi-klinikai szövegek feldolgozásának nehézségeit, illetve irányadók lehetnek a különböző eszközök fejlesztése során, melyek paraméterei így a specifikus problémákhoz hangolhatóak.

2. Korpuszok

Vizsgálataink során általános nyelvezetű korpuszként a Szeged Korpusz 2-t használtuk. Orvosi korpuszként pedig a rendelkezésünkre álló nyers klinikai dokumentumokat. Ezek 29 különböző osztályról származó kezelési lapok, zárójelentések, egyéb klinikai dokumentumok. A klinikai korpuszon belül külön foglalkoztunk a szemészeti dokumentumokkal, hiszen azok feldolgozottsági állapota a folyamatban lévő kutatásaink miatt sokkal előrébb tart, a már meglévő eszközeink adaptálása a többi osztály dokumentumaira még nem valósult meg. Így a következő három domént vetettük alá összehasonlító vizsgálatainknak: általános szövegek a Szeged Korpusz alapján (SZEG), vegyes orvosi szövegek (MED), illetve szemészeti szövegek (SZEM). A korpuszok méretére vonatkozó részletes adatok az 1. táblázatban találhatóak. A Szeged Korpuszra vonatkozó adatok itt és a továbbiakban is [3]-ból származnak.

A szófajok eloszlását illetően eltérő a két fő domén (SZEG és MED) összetétele. Míg a Szeged Korpuszban a leggyakoribb szófajok közül az első három a főnév, ige, melléknév, addig az orvosi szövegekben a főnevek mellett a melléknevek és a számnevek a leggyakoribbak, míg az igék száma az utóbbi két, közel azonos mennyiségben előforduló szófajhoz képest csak harmadannyiszor szerepel. Jelentős különbség még, hogy az orvosi szövegekben a névelők, kötőszavak és névmások is a rangsor második felében helyezkednek el. Ezek az előfordulási

1. táblázat: A három vizsgált korpusz mérete (tokenek és mondatok száma), az őket jellemző átlagos mondathossz

	tokenszám	mondatszám	átlagos mondathossz
Orvosi korpusz (MED)	7 119 841	734 666	9,69
Szemészeti korpusz (SZEM)	334 546	34 432	9,7
Szeged Korpusz (SZEG)	1 194 348	70 990	16,82

arányok nem is meglepőek, hiszen az orvosi feljegyzések legnagyobb része arról szól, hogy egy állapotot ír le (*valami valamilyen* (FN, MN)), vagy valamilyen vizsgálat eredményét (*valami valamennyi* (FN, SZN)). Az orvosi szövegekben előforduló számnevek túlnyomó része numerikus adat. A részletes szófaji eloszlásokat tartalmazza a 2. táblázat.

2. táblázat: A Szeged Korpusz és az Orvosi korpusz tokenjeinek szófaji eloszlása, illetve rangsora

	FN	MN	SZN	IGE	HAT	NM	DET	NU	KOT
MED	43,02%	13,87%	12,33%	3,88%	2,47%	2,21%	2,12%	1,03%	0,87%
SZEG	21,96%	9,48%	2,46%	9,55%	7,60%	3,85%	9,39%	1,24%	5,58%

	FN	MN	SZN	IGE	HAT	NM	DET	NU	KOT
MED	1	2	3	4	5	6	7	8	9
SZEG	1	3	8	2	5	7	4	9	6

3. Helyesírási különbségek

A klinikai dokumentumok jellegzetessége, hogy gyorsan, utólagos lektorálás, ellenőrzés, illetve automatikus segédeszközök (pl. helyesírás-ellenőrző) nélkül készülnek, ezért a leírás során keletkezett hibák száma igen nagy, valamint sokféle lehet [4]. Így nem csupán a magyar nyelv nehézségeiből eredő problémák jelennek meg, hanem sok olyan hiba is felmerült a szövegekben, melyek a szakterület sajátosságaiból erednek. A legjellemzőbb hibák az alábbiak:

- elgépelés, félreütés, betűcserék,
- központozás hiányosságai (pl. mondathatárok jelöletlensége) és rossz használata (pl. betűközök elhagyása az írásjelek körül, illetve a szavak között),
- nyelvtani hibák,
- mondattöredékek,
- a szakkifejezések latin és magyar helyesírással is, de gyakran a kettő valamilyen keverékeként fordulnak elő a szövegekben (pl. *tensio/tenzio/tensió/tenzió*); külön nehézséget jelent, hogy bár ezeknek a szavaknak a helyesírása

szabályozott, az orvosi szokások rendkívül változatosak, és időnként még a szakértőknek is problémát jelent az ilyen szavak helyességének megítélése,
– szakterületre jellemző és sokszor teljesen ad hoc rövidítések, amelyeknek nagy része nem felel meg a rövidítések írására vonatkozó helyesírási és központozási szabályoknak

A fenti hibajelenségek mindegyikére jellemző továbbá, hogy orvosonként, vagy akár a szövegeket lejegyző asszisztensenként is változóak a jellemző hibák. Így elképzelhető olyan helyzet, hogy egy adott szót az egyik dokumentum esetén javítani kell annak hibás volta miatt, egy másik dokumentumban azonban ugyanaz a szóalak egy sajátos rövidítés, melynek értelmezése nem egyezik meg a csupán elírt szó javításával.

A Szeged Korpusszal összehasonlítva két fő különbséget állapíthatunk meg. Az egyik a rövidítések aránya: míg a Szeged Korpuszban a rövidítések a tokenek 0,08%-át teszik ki, addig az általunk vizsgált anyag 7,15%-a rövidítés [5], tehát a rövidítések gyakorisága két nagyságrenddel nagyobb. Ezt a számítást az orvosi szövegekre egy 15 278 token méretű részkorpusz alapján végeztük. Szintén ebből számítottuk a helyesírási hibákat, amiket kézzel jelöltünk meg az orvosi szövegeknek ebben a részhalmazában. Ezért az ebben előforduló helyesírási hibák típusairól részletesebb statisztikát is tudtunk készíteni, melyet a 3. táblázat tartalmaz. A helyesírási hibák aránya az orvosi korpuszban 8,44%, ezzel szemben a Szeged Korpuszban csupán 0,27%. Ezen belül is az iskolai fogalmazásokat tartalmazó részkorpuszban is mindössze 0,87%, tehát tízszer kevesebb helyesírási hibát ejtettek a Szeged Korpuszban szereplő fogalmazásokat író iskolás tanulók, mint a klinikai szövegeket író orvosok. Az orvosi szövegek esetén ezek a hibák az esetek felében ponthibák (leginkább a pont hiánya a rövidítések végén). Az egybeírási és különírási hibák pedig közel azonos mértékben fordulnak elő, összesen a hibák 10%-át teszik ki. Amellett, hogy a rövidítések végéről gyakran hiányzik a pont, az orvosi szövegekre egyébként is jellemző a központozási hibák magas aránya. Míg a Szeged Korpuszban csak a mondatok 1,04%-a nem végződik pontra (címek), addig az orvosi dokumentumokban ez az arány 48,28%. Hasonló problémák vannak a mondatkezdő nagybetűhasználattal: míg a Szeged Korpuszban csak a mondatok 0,42%-a nem kezdődik nagybetűvel, addig az orvosi korpuszban a mondatok 12,81%-a. Ez teszi a mondatokra bontás látszólag triviális feladatát is rendkívül nehézzé [6].

3. táblázat: Az orvosi szövegek egy részkorpuszában előforduló helyesírási hibák típusai

	hibás	ponthiba	egybeírás	különírás	egyéb
Szeged Korpusz	0,27%	-	-	-	-
Szeged Korpusz – iskolás	0,87%	-	-	-	-
Orvosi korpusz	8,44%	46,55%	5,66%	5,59%	42,2%

4. Szemantikai különbségek

A szófaji és nyelvhasználati különbségek mellett az általános és az orvosi szövegek között gyakran jelentős eltérés mutatkozik meg azoknak a szavaknak a jelentésében is, amelyek mindkét korpuszban előfordulnak, tehát az egyes szavak szemantikája mást fed le a különböző korpuszokban. Ezt a jelenséget a disztribúciós szemantika módszerével vizsgáltuk. 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. [7] 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* és *w'*) közötti hasonlóságot a következő metrikával számoltuk [8] alapján:

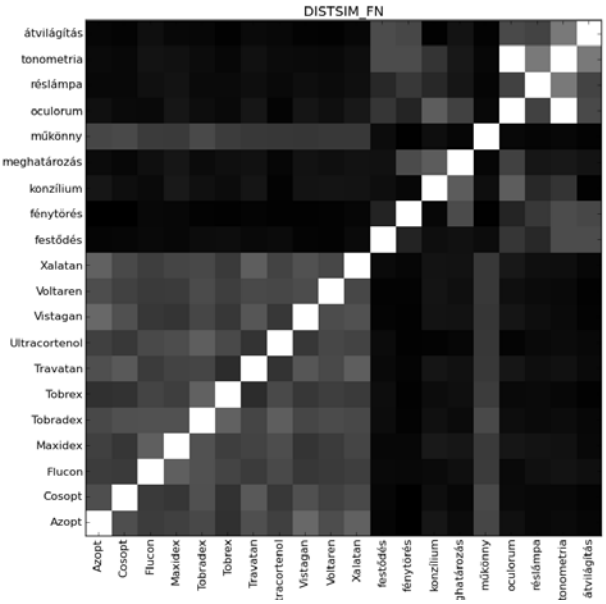
$$\frac{\sum_{(r,w)\in T(w_1)\bigcap T(w_2)}(I(w_1,r,w)+I(w_2,r,w))}{\sum_{(r,w)\in T(w_1)} I(w_1,r,w)+\sum_{(r,w)\in T(w_2)} I(w_2,r,w)},$$

ahol *T(w)* azoknak az *(r,w')* pároknak a halmaza, ahol az *I(w,r,w')* pozitív. Ennek a metrikának a használatával korpuszonként kiszámoltuk a leggyakoribb főnevekre, igékre és melléknevekre a páronkénti disztribúciós hasonlóságukat.

4.1. A szemészeti korpusz disztribúciós szemantikája

A gyakori főnevek vizsgálata során olyan szópárokat kerestünk, melyeknél jól kimutatható a más főnevekhez való viszonyuk. Az 1. ábrán egy ilyen részlet látható. A világosabb mezők jelzik az erősebb szemantikai kapcsolatot az adott két szó között. Az ábrán jól elkülönülő szemantikai terek láthatóak. Például a különböző szemcseppek nevei és a *műkönny* kifejezés egy behatárolható csoportot

alkotnak. Ezek fölé rendelhető a *szemcsepp* fogalom. Hasonlóan, az egyes szemészeti vizsgálatok is egy csoportba kerültek *(átvilágítás, tonometria, réslámpa)*, illetve az ezek által vizsgált jelenségek *(fénytörés, festődés)*.



1. ábra: A szemészeti korpusz leggyakoribb főneveinek hasonlósági mátrixa

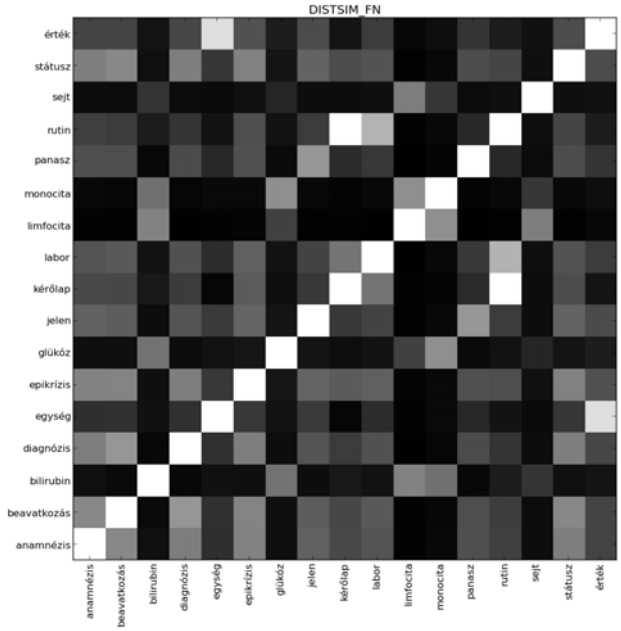
Az igei eloszlásra vonatkozóan is meghatározhatóak a szemantikai együttállások. Így a szemészeti korpusz esetén releváns csoportot alkotnak a *fáj* igéhez tartozó, hozzá hasonló kifejezések: *könnyezik, szúr, viszket, beragad.* Az orvosi korpusz eredményeinél látni fogjuk, hogy a *fáj* igéhez tartozó igék a *köhög, érez, romlik.*

A hasonlóságok kiértékelése során sok esetben nem tudtuk megítélni az egyes szakkifejezések közötti kapcsolat helyességét. A melléknevek esetén olyan hasonlóságokat találtunk, mint a *szélű* és a *határú*, valamint a *bal* és a *jobb* közötti kapcsolat, amelyek mindenképpen helytállóak. Az utóbbi párral kapcsolatban megjegyzendő, hogy a szemészeti szövegekben a *bal szem* és a *jobb szem* vizsgálatai miatt ezek kapcsolata sokkal erősebb és sokkal jobban elkülönülő csoportot alkotnak, mint az *alsó*, vagy *felső* mellékneveké, amik szintén az irányultságot jelzik. Az általános orvosi szövegekben a négy irány már egy csoportot alkot. A szinonímák és antonímák mellett a módszer kollokációkat is kimutat, pl. a *széli* és a *vesszős* hasonló disztribúciója abból adódik, hogy ezek leginkább a *(vaskos) széli vesszős homály* kifejezésben szerepelnek együtt.

4.2. Az orvosi szövegek disztribúciós szemantikája

Az általános orvosi szövegekből álló korpusz tartalma sokszínűbb, mint a szemészeti részkorpusz, ezért a szemantikai csoportok sem annyira kifinomultak,

mint egyetlen szűk domén esetén. Az azonban itt is megállapítható, hogy a létrejött relációk, illetve szemantikai csoportok helytállóak és relevánsak. A 2. ábrán szintén a leggyakoribb, legnagyobb hasonlóságot mutató főnevek szemantikai mátrixa látható. Az ábrán is élesen kiugrik a *limfocita–monocita* szópár, illetve a hozzájuk kapcsolódó *bilirubin*, *glükóz* és *sejt* szavak, melyekkel együtt élesen elhatárolódnak a többi fogalomtól. Hasonlóan jól elhatárolódnak az orvosi feljegyzések egyes részeit jelölő kifejezések: *anamnézis*, *diagnózis*, *epikrízis*, *státusz*. Ezekkel kapcsolatban látszik az, hogy bár a szemészeti dokumentumokban is ugyanezek a részek találhatóak meg, ott nem jelentek meg a leggyakoribb és legerősebb összefüggést mutató csoportok között (természetesen a szemantikai viselkedésük, így a hasonlóságuk ott is fennáll). A vegyes orvosi szövegekben azonban ez a csoport a kevert domén fölött hangsúlyosabb összetartozást mutat.



2. ábra: Az orvosi korpusz leggyakoribb főneveinek hasonlósági mátrixa

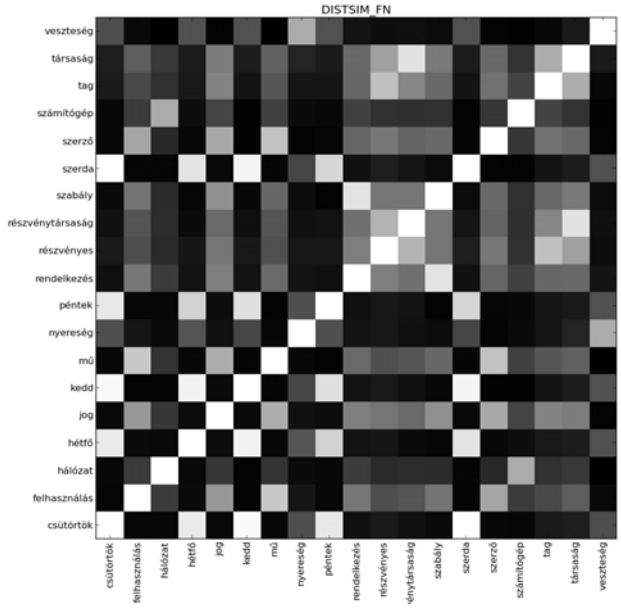
Az orvosi korpuszban vizsgált gyakori igék között olyan csoportok jöttek létre, mint a *mutat*, *igazol*, *látszik*, *ábrázolódik*. Természetesen nem csak az összetartozásnak van jelentősége (igaz ez mindhárom domén esetén mindegyik szófajra vonatkozóan), hanem az elhatárolódásnak is. Így az igék között az *ábrázolódik* és az *elhagy* jó példa arra, hogy ezek szemantikai viselkedése között nincsen hasonlóság.

A melléknevek esetén a már fent említett irányultsági csoportok emelhetőek ki, itt már mind a négy irányra vonatkozóan, illetve megjelenik a szakterületre vonatkozó melléknevek csoportja (*szakápolói*, *neurológiai*, *pszichiátriai*).

4.3. A Szeged Korpusz disztribúciós szemantikája

Az előző két doménhez képest nagy eltérést találunk az általános szövegeket tartalmazó korpuszban.

Bár a Szeged Korpusz a témák sokkal szélesebb körét öleli fel, mint az orvosi korpusz együttesen, vagy különösen mint egy adott orvosi szakterülethez tartozó szövegek, a módszer mégis kiemel a vegyes szöveghalmazból is olyan szemantikai csoportokat, amelyeken belül előforduló szavak erős tematikus összefüggést mutatnak. A főnevekre vonatkozó 3. ábrán jól látszik, hogy egy szemantikai csoportba kerültek a leggyakoribb főnevekre vizsgálva a *részvénytársaság*, *társaság*, *részvényes*, és *tag* kifejezések, illetve a *mű, szerző, felhasználás*, valamint a *jog, rendelkezés* és *szabály* szócsoportok. Nyilvánvalóan a korpuszban további, ebbe a körbe tartozó szavak is megjelennek, azonban az algoritmus igen nagy számításigénye miatt mindegyik esetben csak a leggyakoribb 100 szóra végeztük el a vizsgálatot. Természetesen olyan általánosan gyakori szavakból álló csoportok is felfedezhetőek, mint a hétköznapok nevei, azok igen erős hasonlósága alapján.



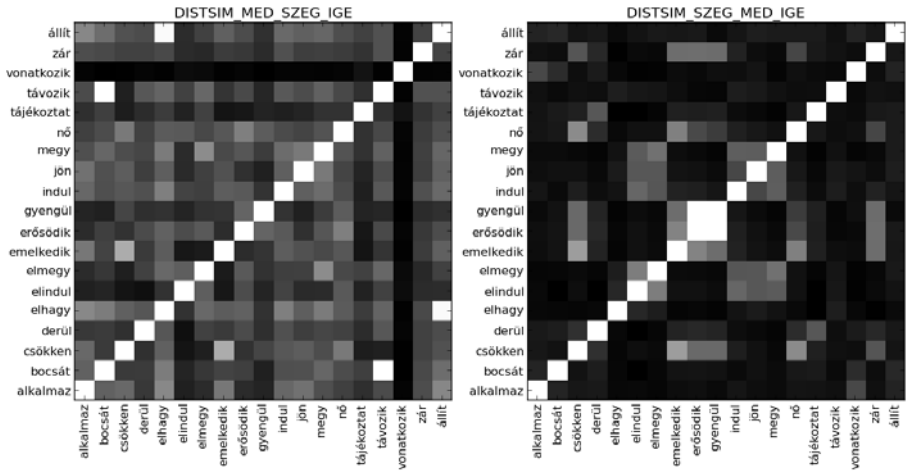
3. ábra: A Szeged Korpusz leggyakoribb főneveinek hasonlósági mátrixa

Az igékre vonatkozóan a létrejött szemantikai csoportok mellett további finomításokat tapasztalhatunk. Míg a *megy–indul*, illetve az *elmegy–elindul* szópárok külön-külön nagyon hasonlóak, addig a két páros egymáshoz való hasonlósága ennél kisebb, amit az igekötős alakok más jellegű viselkedése jól magyaráz.

A melléknevek esetén is kialakultak a fentiekhez hasonló csoportok, azonban ezek az általános mellékneveket tartalmazzák (pl. *tavalyi–idei*).

4.4. A korpuszok összehasonlítása

Amellett, hogy külön-külön megvizsgáltuk az egyes szófajok leggyakoribb példányait, összehasonlító elemzést is végeztünk. Ehhez először létrehoztuk a korpuszok leggyakoribb szavainak metszetét (szófajonként), majd az ebben a listában szereplő szavak disztribúciós hasonlóságát megmértük mindkét, az összehasonlításban részt vevő korpusz esetén. Az összehasonlítások során az általános orvosi és a szemészeti korpuszt vizsgáltuk a Szeged Korpusszal szemben. Általánosságban elmondható, hogy az általános szövegekben kiemelkedő hasonlóságot mutató szavak (például hónapnevek) hasonlósága megmaradt az orvosi szövegek esetén is, azonban a kontraszt az egyes csoportok között kisebbnek bizonyult. A másik általános jelenség, hogy az orvosi szövegekben egyes szavakhoz új jelentéscsoportok jelentek meg, melyek az általános témájú szövegekre nem jellemzőek. A főnevek között ilyen például az *időpont*, ami az orvosi korpuszokban sokkal közelebb kerül például a hét napjainak megnevezéséhez, hiszen a klinikai események során az időpontnak leginkább abban van szerepe, hogy melyik napon esedékes egy vizsgálat.

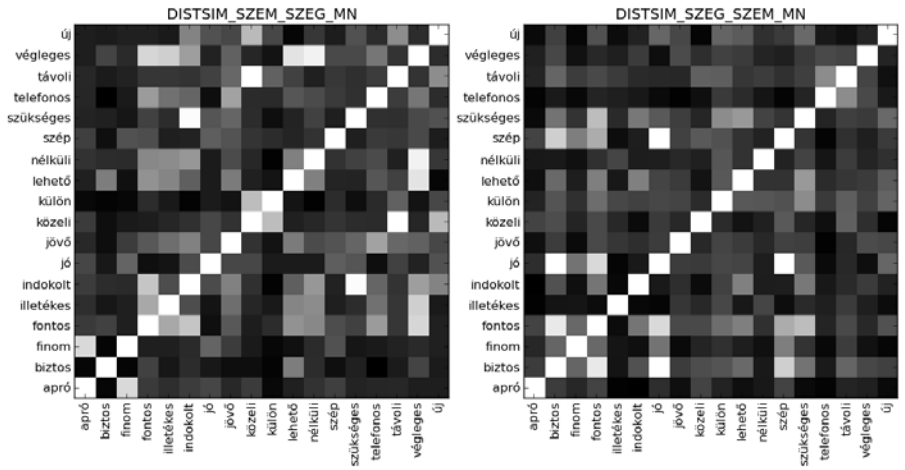


4. ábra: A leggyakoribb igék hasonlósági mátrixai az orvosi korpusz és a Szeged Korpusz alapján

Az igék esetében még jelentősebb különbségeket láthatunk, amit a 4. ábra illusztrál. A *távozik* és a *bocsát* szavak összefüggése az orvosi korpuszban nyilvánvaló (az *otthonába bocsát* kifejezés miatt), ami az ábrán is kiugró értékként jelenik meg, azonban a Szeged Korpuszban vizsgálva ugyanez a két szó egészen távoli. Szintén jól látszik, hogy míg az általános nyelvezetű szövegben az *emelkedik*, *erősödik*, *gyengül*, *csökken*, *nő* szavak egy jelentéskörbe tartoznak, addig az orvosi szövegek esetén ezeknek a viselkedése eltérő. Különbségként látszik még az *állít* és az *elhagy* szavak hasonlósága. Az orvosi korpuszban ezek nagyon hasonló viselkedésűek, mindkettő a gyógyszeres kezelésekkel kapcsolatos *(gyógyszer*

elhagyása, gyógyszer adagolásának beállítása). A Szeged Korpuszban ezek között semmilyen hasonlóság nem látszik.

A melléknevek esetén szintén látszanak olyan különbségek a szemészeti és az általános korpusz között, hogy míg az elsőben a *fontos* és az *indokolt* szavak lettek hasonlóak, amikhez viszont a *jó* egyáltalán nem kapcsolódik, addig az általános szövegekben a *fontos*, a *biztos*, és a *jó* tartoznak egy jelentéskörbe. További jelentésbeli eltolódások láthatóak az 5. ábrán.



5. ábra: A leggyakoribb melléknevek hasonlósági mátrixai a szemészeti korpusz és a Szeged Korpusz alapján

5. Konklúzió

Cikkünkben bemutattuk három korpusz összehasonlítását néhány fő statisztikai jellemzőjük alapján. Látható, hogy a klinikai dokumentumokból álló korpusz szavainak szófaji eloszlása és helyessége jelentősen eltér az általános szövegeket tartalmazó korpusztól, ezért az utóbbiakra széles körben elfogadott megállapítások, illetve az ezekre a megállapításokra alapozott alkalmazások nem feltétlenül érvényesek, nem feltétlenül alkalmazhatóak orvosi szövegek esetén. Mindenképpen szükséges tehát a klinikai szövegek feldolgozására alkalmas eszközök egyedi fejlesztése.

További vizsgálatokat végeztünk az egyes korpuszok disztribúciós szemantikájára vonatkozóan is. Ennek során szintén lelepleződtek az alapvető különbségek, melyek a különböző szövegek közötti tartalmi eltérésből adódnak. Az orvosi szövegeknél, a viszonylag szűk domén miatt, ez a módszer alkalmas lehet disztribúciós tezaurusz építésére magyar nyelvű dokumentumok esetén is, hiszen látható, hogy a kimutatható hasonlósági relációk relevánsak, valódi összefüggéseket jelenítenek meg.

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. 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. Vincze, V.: Domének közti hasonlóságok és különbségek a szófajok és szintaktikai viszonyok eloszlásában. In: IX. Magyar Számítógépes Nyelvészeti Konferencia. (2013) 182–192
4. Siklósi, B., Novák, A., Prószéky, G. Number Lecture Notes in Computer Science 7978. In: Context-Aware Correction of Spelling Errors in Hungarian Medical Documents. Springer Berlin Heidelberg (2013) 248–259
5. Siklósi, B., Novák, A. In: Detection and Expansion of Abbreviations in Hungarian Clinical Notes. Volume 8265 of Lecture Notes in Artificial Intelligence. Springer-Verlag, Heidelberg (2013) 318–328
6. Orosz, G., Novák, A., Prószéky, G. In: Hybrid text segmentation for Hungarian clinical records. Volume 8265 of Lecture Notes in Artificial Intelligence. Springer-Verlag, Heidelberg (2013)
7. 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
8. 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