# NORMÁTÓL ELTÉRŐ (ZAJOS) SZÖVEGEKET ÉRTELMEZŐ ALGORITMUSOK KIDOLGOZÁSA ÉS ALKALMAZÁSUK MAGYAR NYELVŰ KLINIKAI DOKUMENTUMOKRA

Doktori értekezés tézisei

Siklósi Borbála


Roska Tamás Műszaki és Természettudományi Doktori Iskola

Pázmány Péter Katolikus Egyetem, Információs Technológiai és Bionikai Kar

Témavezető:

Dr. Prószéky Gábor

2015

## Normától eltérő (zajos) szövegeket értelmező algoritmusok kidolgozása és alkalmazásuk magyar nyelvű klinikai dokumentumokra

A legtöbb kórházban az orvosi feljegyzések tárolása csupán archiválás, illetve az egyes esetek dokumentálása céljából történik. Az így felhalmozódott adattömegek felhasználása jelenleg csupán az egyes betegek kórtörténetének visszakeresésére korlátozódik. A nyelvtechnológia, a számítógépes ontológiák és a statisztikai szövegfeldolgozó algoritmusok lehetővé teszik a folyó szövegekben rejlő összefüggések, rejtett struktúrák felfedését, a feljegyzésekben található információhalmaz elérését, abból tudás kinyerését.

Angol nyelvterületen az ilyen irányú kutatások előrébb járnak, azonban alkalmazhatóságuk a magyar nyelv sajátosságai miatt sokszor nem egyértelmű, továbbá számos olyan nyelvi erőforrás, ami az angol nyelvre hozzáférhető, magyarra nem létezik. Az orvosi dokumentumok feldolgozása során nem csak a magyar nyelv nyelvtani sajátosságait kell figyelembe venni, hanem az orvosi szövegekre különösen jellemző, olykor hiányos szintaktikai szerkezeteket, rövidítéseket, idegen kifejezéseket is kezelni kell.

Ezen tapasztalatok alapján fogalmazódott meg az igény, hogy a magyar nyelvű klinikai dokumentumok feldolgozását a más nyelveken már létező alkalmazások adaptálása, továbbfejlesztése és alkalmazhatóvá tétele révén aktívan kutatott területté tegyük, tekintettel a kutatás várható hasznára.

# 1

# Bevezetés

Az orvosi szövegek feldolgozása a nyelvtechnológia egyik egyre népszerűbb feladata. Angol nyelvre több olyan módszer létezik, ami orvosbiológiai dokumentumokból bizonyos információk kinyerésére alkalmazható. Így ezek az információk hozzáférhetővé válnak a kutatók és az orvosok számára. Mivel azonban ezek az információk nem univerzálisak, hanem – az orvosi protokollok, szokások és a helyi közösségek sajátosságai miatt – országonként vagy akár területenként is eltérőek lehetnek, ezért nem elegendő az angol (vagy más) nyelven kinyert információk átvétele. Ezek kinyeréséhez az adott nyelven létrejött szövegek feldolgozása szükséges.

Meystre et al. (2008) rávilágít arra, hogy fontos különbséget tenni a klinikai és az orvosbbiológiai szövegek között. A klinikai dokumentumok kórházakban jönnek létre, a mindennapi esetek dokumentálása céljából. Ezek minősége messze elmarad az orvosbiológiai cikkek minőségétől, melyekkel szintén sok nyelvtechnológiai kutatás foglalkozik. Az elsősorban angol nyelven írt orvosbiológiai cikkek a tudományos folyóiratokban, könyvekben, kötetekben megjelenő tanulmányok. Ezek nyelvezete követi a hivatalos normákat és helyesírási szabályokat (Sager et al., 1994; Meystre et al., 2008). Ezzel szemben a klinikai dokumentumok olyan strukturálatlan szövegek, amelyek gyakran sietve, a helyesírás ellenőrzése nélkül jönnek létre. Így ezek a szövegek, amik eleve a latin, az angol és a helyi nyelv (esetemben a magyar) sajátos keverékét használják, tele vannak elírásokkal és egyedi szóalakokkal (Siklósi et al., 2012, 2013). Jellemző rájuk továbbá a rövidítések igen nagy aránya, amiket sok esetben szintén minden norma figyelmen kívül hagyásával alkalmaznak a dokumentumok lejegyzői. Gyakran előfordul, hogy teljes állítások, akár mondatok, is csupán rövidített szóalakokból állnak (Barrows et al., 2000).

Jellemző továbbá, hogy keletkezésük során ezeknek a klinikai dokumentumoknak a címzettje általában az azt leíró orvos maga, tehát az eredeti célját nem befolyásolja a sajátos nyelvezet, egyedi rövidítések, utalások használata. Az egyes betegek kórtörténetének rögzítése mellett, számos egyéb célra is felhasználhatóak az ezekben a kórlapokban implicit tárolt információk. Ezek kinyerésének előfeltétele azonban egy, az állításokat és tényállásokat hatékony tároló reprezentáció létrehozása.

A klinikai dokumentumok feldolgozása, azok tartalmi és nyelvi sajátosságai miatt tehát komoly kihívást jelent. **Jelen dolgozatban bemutatott kutatói célkitűzésem a magyar nyelvű klinikai dokumentumok előfeldolgozására irányuló módszerek kidolgozása volt, melyek lehetővé teszik egy olyan reprezentáció létrehozását, melyek alapján a szövegek mögött rejlő implicit összefüggések is kinyerhetővé válnak.** Minden szövegfeldolgozó algoritmus teljesítménye függ az általa feldolgozott szöveg eredeti minőségétől. Helyes és releváns információ csak akkor nyerhető ki, ha létezik. Célom

az volt, hogy a klinikai kórlapok zajos és torz állapotából egy olyan alak jöjjön létre, ami alkalmas további feldolgozási szintek bemeneteként való felhasználásra és az orvosok által megfogalmazni szándékozott eredeti tartalom visszaállítására.

Két fő megközelítés létezik a klinikai dokumentumok kezelésére (Meystre et al., 2008). Az első a szabály alapú algoritmusok alkalmalzása. Ezek általában kicsi, domain-specifikus módszerek, amiknek a létrehozása igencsak idő-, és erőforrásigényes, viszont precíz eredmények várhatók tőlük. A másik lehetőség a statisztikai algoritmusok alkalmazása. Ezek a módszerek egyre elterjedtebbek, azonban hátrányuk, hogy nagy méretű tanítóanyag szükséges ahhoz, hogy kielégítő eredményt nyújtsanak. A klinikai kórlapok esetén ennek létrehozása nem csak mennyiségi, de etikai kérdést is jelent. Továbbá, felügyelt tanulási algoritmusok esetén kézi annotációra is szükség van, ami a kórlapok feldolgozása esetén mind nyelvészeti, mind orvosi kompetenciát igényel. Szükség lehet még a feldolgozás során külső, domain-specifikus lexikai erőforrásokra is (ontológiák, tezauruszok), amik szintén kézi ellenőrzéssel készülnek. Így, bár a statisztikai algoritmusok betanítása egyszerűbb, az olyan nyelvek esetén, mint a magyar, ahol kevés külső erőforrás áll rendelkezésre, illetve az adatmennyiség is kevés, több emberi munkát igényel ez a megközelítés is.

A meglévő eszközök adaptálásának támogatására és strukturált erőforrások létrehozásának előkészítése érdekében magyar nyelvű szemészeti kórlapokból álló korpuszt vizsgáltam. Kutatásaim során elsősorban statisztikai módszereket alkalaztam, melyek során a nyers szöveget használtam. A létrehozott modulok együttes, illetve tetszőleges kombinációjú, alkalmazása, illetve a dokumentumokból kinyert félig-strukturált információk lehetővé teszik a korpusz hatékony felhasználását.

# 2

# Az új tudományos eredmények összefoglalása

A klinikai kórlapok minősége sokkal rosszabb, mint bármilyen általános nyelvezetű szövegé. Jelen kutatás célja ezeknek a zajos szövegeknek egy normalizált reprezentációját létrehozó algoritmusok kidolgozása, melyek a további feldolgozási lépések számára megfelelő inputot állítanak elő az eredeti szövegekből. Az alkalmazott módszerek elsősorban statisztikai algoritmusok, melyek (külső erőforrások hiánya és a dokumentumlétrehozás szabványainak figyelmen kívül hagyása miatt) magában a korpuszban lévő információt használják fel az egyes előfeldolgozási lépések megvalósításához.

## 2.1 Reprezentációs séma

Annotált korpuszok tárolására bevett szokás az XML struktúra használata. A nyers klinikai dokumentumok szerkezetét két szinten definiáltam. Először az egységek meghatározása egy egyszerű szabály alapú mintaillesztő eljárással történt, mely a rekordok szemmel is látható tagolására épül. Így a folyó szövegekben meglévő formázási elemeket transzformáltam a szerkezetet meghatározó jellemzőkké. Az így kapott struktúra jól elkülöníti a dokumentumok egyes részeit, azonban korántsem elegendő ahhoz, hogy a szöveges részek önállóan kezelhetőek legyenek. Ezért a már elkülöníthető, valóban szöveges információt tartalmazó egységek kifinomultabb reprezentációját valósítottam meg. Így minden mondat minden szava külön egység, ahol a szavakhoz, mondatokhoz tatozó további kiegészítő információk tárolására is lehetőség van úgy, hogy ezek száma rugalmasan bővíthető, illetve minden egység egyedileg azonosítható. Kiegészítő információk a kiindulási sémában például az eredeti szóalak mellett a javított szóalak, a szótő, a szófaj, névelem típusa (ha része annak), rövidítés-e, és ha igen, akkor annak feloldása.

Ezen a ponton a szöveges tartalom, ami eredetileg több különböző tartalmi szekcióból áll, egy *content* címke alatt szerepelt. Ezek a részek azonban több alcím alá besorolható részt tartalmaznak, mint pl. *fejléc*, *diagnózisok*, *beavatkozások*, *státusz*, *műtét*, *anamnézis*, stb. Ezek tartalma is jelentősen eltérő, ezért a feldolgozás során is érdemes külön kezelni a különböző szekcióba tartozó állításokat. Ezért megvalósítottam egy olyan módszert, melynek segítségével az egyes állítások külön-külön kategóriákba kerültek besorolásra. Ehhez először a dokumentumok formázása alapján kinyerhető címkéket definiáltam. A második lépésben az ezekkel a címkékkel ellátott sorokat használtam tanítóanyagként a címkézetlen sorok

kategorizálását megvalósító tf-idf súlyozott koszinusz hasonlóságmetrikát használó algoritmus során.

**1. TÉZIS:**

*Definiáltam egy rugalmas reprezentációs sémát magyar nyelvű klinikai dokumentumok tárolására és megvalósítottam egy algoritmust, ami az eredeti dokumentumokat átalakítja ennek a sémának megfelelően.*

Kapcsolódó publikációk: 1, 4, 10, 16, 17

## 2.2 Automatikus helyesírás-javítás

Mivel ezek a szövegek egyrészt mindenféle kontroll (pl. helyesírás-ellenőrző) alkalmazása nélkül készültek, másrészt az adott szövegtípusban nagyon magas arányban fordulnak elő a köznapi nyelvhasználattól idegen szóalakok: latin szavak, rengeteg rövidítés, gyógyszernevek, amelyeknek a helyesírására vonatkozó normákkal a szövegek íróinak nagy része nyilvánvalóan nincs tisztában, az ilyen szövegeknek a javítása nem könnyű feladat. A megvizsgált klinikai szövegekben jellemzően jelen vannak a hivatalos normától eltérő használatból fakadó, de következetesen elkövetett hibák, a véletlen melléütések, a következetlen szóhasználat, illetve az olyan többértelmű elírások, melyek helyességének megítélése még orvosi szakértelemmel sem egyértelmű (pl. elírt rövidítések). Emellett jellemző még az általános helyesírás-ellenőrzés során is felmerülő további probléma is: önmagukban helyes, de az adott környezetben téves szóalakok is előfordulnak.

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 feladat másik nehézségét az jelentette, hogy egyáltalán nem állt rendelkezésünkre nagyméretű, helyesen írt klinikai korpusz, ami alapján elő tudtunk volna állítani a javításhoz használható hibamodelleket.

Megvalósítottam egy rendszert, ami a helyesírási hibák felismerése és javítása során a szövegkörnyezetet is figyelembe veszi. A rendszer a statisztikai gépi fordítás (SMT) módszereit, és egy szóalapú javaslatgeneráló által javasolt jelöltlistát használ. A rendszer paraméterezhetőségének köszönhetően, az aktuális feladatra könnyen optimalizálható, az eredmény pedig az orvosi szövegeknek egy (feltételezett) normához az eredetinél jóval közelebb álló formája.

### 2.2.1 A szóalapú javaslatgeneráló rendszer

Először egy szóalapú rendszert implementáltam, ami egyedülálló szavakhoz generál javításjelölteket különböző szólisták, gyakorisági adatok és egy lineáris rangsolorás alapján.

A tokenizálás után egy stopword-lista és egy rövidítéslista alapján kiszűrtük azokat a szavakat, amelyekre nem hajtunk végre javítást. A többi szóalak mindegyikéhez létrejön egy javaslathalmaz, mely az egy Levenshtein távolságra lévő szóalakokat, illetve a morfológia (Prószéky and Kis, 1999; Novák, 2003) által generált lehetséges javaslatokat rangsorolva tartalmazza.

A rangsorolás alapját a korpuszból és általános, illetve orvosi szótárakból épített modellek és a morfológia által együttesen meghatározott tényező képezi. Mivel minden szóalakhoz létrejönnek javaslatok, nem csak azokra, amiket a morfológia rossznak ítél, ezért azt az információt, hogy az eredeti alakot a morfológia elfogadja-e, a javaslatok rangsorolásánál kell figyelembe venni. A rangsorolás végén a lehetőségek közül az első öt javaslat a lehetséges javítások halmaza. Amennyiben az első és a második helyezett között elég nagy különbség van, akkor az első javaslat automatikusan elfogadható helyes javításnak. A korpusz sajátos jellegének figyelembe vétele miatt - az előzetes feltételezésnek megfelelően - a meglévő korpuszra épülő modellek magasabb súllyal való figyelembe vétele, a morfológiával kiegészítve hozta a legjobb eredményt. A javaslatok sorrendjéről elmondható, hogy amikor nem az első eredmény tartalmazza a helyes alakot (az esetek kb. fele), akkor az első 5 javaslatban az esetek 99,12%-ban fellelhető a helyes szóalak.

### 2.2.2 Statisztikai gépi fordítás alkalmazása hibajavításra

Fontos azonban, hogy a rendszer képes legyen a generált javaslatok közül a valóban helyeset automatikusan kiválasztani. A legjobb javítás kiválasztásához kevésnek bizonyult a kizárólag morfológiára és különböző szóstatisztikákra épülő rangsorolás. Az automatikus javítás pontosságának növeléséhez szükséges az egyes szavakhoz tartozó szövegkörnyezet figyelembevétele is.

E két követelmény alkalmazására a statisztikai gépi fordítás területén széles körben alkalmazott Moses keretrendszert (Koehn et al., 2007) alkalmaztam. A fordítás során forrásnyelvnek az eredeti hibás szöveget tekintem, míg a célnyelv ennek javított formája. Ebben az esetben a rendszer bemenete a hibás mondat: $E = e_1, e_2 \dots e_k$, melynek megfelelő javított mondat a $C = c_1, c_2 \dots c_k$ a várt kimenet. A helyesírás-javító rendszer zajos csatornamodellként tehát úgy fogalmazható meg, hogy az eredeti üzenet a helyes mondat, ami helyett a csatornán átért jel a zajos, azaz hibás mondat. Így a javítás az a $\hat{C}$ mondat lesz, melyre a

$$\hat{C} = argmax P(C|E) = argmax \frac{P(E|C)P(C)}{P(E)} \tag{2.1}$$

feltételes valószínűség a maximális. Mivel P(H) értéke állandó, ezért a nevező elhagyható, így a számlálóban lévő szorzat a fordítási- és nyelvmodellből számított statisztika alapján számítható.

Ezeket a modelleket a hagyományos statisztikai gépifordító-rendszerek esetén a forrás- és célnyelvű párhuzamos korpuszból számolt valószínűségek képezik. Ilyen korpusz azonban

az orvosi szövegek esetében nem áll rendelkezésre, ezért a fordítási modellt a korábban létrehozott rendszer rangsorolásához használt számítási értékek valószínűségekké konvertálása képezi. A szövegkörnyezet figyelembevétele érdekében pedig a SRILM eszköz (Stolcke et al., 2011) segítségével létrehozott nyelvmodell módosítja a "fordítás" során kapott eredményeket. A gépi fordításon alapuló, szövegkörnyezetet is figyelembe vevő rendszer teljesítménye felülmúlja a szóalapú rendszer teljesítményét, így alkalmasabbnak bizonyult a helyesírási hibák automatikus javítására.

**2. TÉZIS:**

*Létrehoztam egy módszert, a sok idegen kifejezést és rövidítést tartalmazó, egyedi nyelvhasználattal írt magyar nyelvű klinikai szövegekben nagy gyakorisággal előforduló helyesírási hibák automatikus javítására. Megmutattam, hogy a statisztikai gépi fordítás módszerei, a szövegkörnyezetért felelős nyelvmodellel alkalmazva, jól teljesítenek ebben a feladatban.*

Kapcsolódó publikációk: 1, 2, 6, 16, 17

## 2.3 Rövidítések felismerése és feloldása

A klinikai dokumentumokban előforduló rövidítések önmagukban sokszor többértelműek, nem csak a jelentésük, hanem a dokumentumokban előforduló változatos alakjuk miatt is (például: *o.sin./o sin/o.s./os/OS*, stb.). Többértelműség fakad továbbá az egytagú rövidítések során azok számtalan feloldási lehetőségéből is (például: *o./p./f.*, stb.). A reguláris kifejezések segítségével megfogalmazott szabályok alapján felismert rövidítéseket tehát azok környezetében vizsgáltam. A szabályok megfogalmazása során figyelembe kellett venni a rövidítések gyakori normától eltérő használatát. A végződést jelölő pont általában hiányzik, a kis- és nagybetűk használata változatos, az összetett kifejezések rövidítésére több minta is található. Egyetlen rövidítésnek tekintettem azokat a többtagú láncolatokat, amelyeket nem tör meg egyetlen teljes szóalak sem. Ezekben az egységekben az önmagukban szinte értelmezhetetlen, sokszorosan többértelmű (tipikusan egybetűs) rövidítések feloldási lehetőségeinek köre jelentősen csökken. A feloldás során célom az volt, hogy az így kapott több tagú rövidítéssorozatok optimális felbontását megtalálva, annak minél több tagja illeszthető legyen a felhasznált lexikonokban található kifejezésekre, illetve a továbbfejlesztett algoritmus során magára a korpuszra. A feloldás során három módszert vizsgáltam.

Az első esetben csak a korpuszban kerestem az adott rövidítés lehetséges feloldásait, illetve annak az összes lehetséges felosztására illeszthető olyan kifejezéseket, amelyekben a rövidített tagok kifejtett alakja szerepel. Ebben az esetben nem támaszkodtam semmilyen külső erőforrásra.

A második esetben kétféle külső lexikont használtam a korpuszban való keresés helyett. A magyar nyelven elérhető orvosi szótárak és a BNO-kódrendszer alapján az általam

vizsgált szemészeti részek kiválasztásával épült szólistákat, illetve egy szakértő bevonásával készített korpuszspecifikus listát használtam. A megtalált rövidítésláncolatok összes lehetséges felbontásából generált reguláris kifejezéseket illesztve ezekre a lexikonokra kerestem meg a lehetséges feloldásokat, melyeket a lefedettség és az egyes fragmentumok hossza alapján számított pontszám alapján rangsoroltam. A legmagasabb pontszámot elért javaslatot fogadtam el feloldásként.

A harmadik esetben a két módszert együttesen alkalmaztam: a felismert rövidítésekre először a korpuszban kerestem illeszkedő mintát, ezután alkalmaztam a lexikont, ami a fennmaradó részek feloldására szolgált csupán. A rangsoroláshoz használt pontszámok számítása során a korpusz-gyakoriságot is figyelembe vettem az egyes rövidítéseknél. Így egyrészt nőtt a rendszer pontossága az in-domain feloldások biztosításával. Másrészt pedig a minőség sokkal kisebb romlása mellett csökkenthető volt a kézzel épített lexikon szerepe, ami a más szakterületekre való alkalmazhatóság szempontjából fontos lehet.

A kiértékelés során a teljes (több szavas) rövidítések minőségére, illetve a tokenszintű feloldás minőségére nézve mértem az egyes módszerek teljesítményét. Eredményeim azt mutatják, hogy az előre definiált lexikonok és a korpusz kombinált használatán alapuló feloldás csökkentett méretű kézzel készült lexikon használatakor is jól teljesít, míg a korpusz bevonása nélkül a saját lexikon csökkentése radikális teljesítménycsökkenéssel jár.

Megmutattam, hogy a rövidítések felismerése és feloldása során a korpusz mint elsődleges forrás önmagában ugyan nem elégséges a feladat megoldására, de a szakterületnek megfelelő feloldások biztosítása miatt jelentősen javítja a rövidítésfeloldó rendszer minőségét. Ez a rövidítések feloldása során 96.5%-os f-mértéket, az egytagú rövidítések feloldása során 80.88%-os, míg a többtagú rövidítéssorozatok esetén 88.05%-os f-mértéket jelent.

**3. TÉZIS:**

*Megvalósítottam egy algoritmust magyar nyelvű klinikai szövegekben található rövidítések felismerésére és feloldására, melynek során a kézzel készített lexikon mellett nagyobb hangsúllyal vettem figyelembe magát az orvosi korpuszt.*

**3.a TÉZIS:**

*Megmutattam, hogy a többértelmű rövidítések rövidítéssorozatok tagjaként könnyebben értelmezhetőek, továbbá egy szűk szövegkörnyezet figyelembe vétele is hozzájárul a feloldás pontosságához.*

**3.b TÉZIS:**

*Megmutattam, hogy a rövidítések feloldása során egy domain-specifikus lexikon felhasználása szükséges, azonban ez kezelhető méretű és ennek összetétele a korpusz alapján meghatározható.*

Kapcsolódó publikációk: 1, 7, 12, 13, 14

## 2.4 A klinikai dokumentomok félig-struktúrált reprezentációja

A klinikai dokumentumok nyelvezete egy alnyelvet képez, mind a tartalmát, mind a nyelvhasználatot illetően. Ezeknek a szövegeknek a zajos volta és a magyar nyelvű lexikai erőforrások hiánya megnehezíti a más nyelvekre bevált módszerek alkalmazását a szövegre jellemző releváns kifejezések azonosítására. A használható ontológiák és fogalmi rendszerek létrehozása általában szakértők által, kézzel történik, de automatikus módszerekkel segíthető ez a munka. Megmutattam, hogy a korpuszra épülő statisztikai módszerek jól teljesítenek többszavas kifejezések felismerése és disztribúciós modellek építése során. Ezeket az orvosi korpuszra alkalmazva, egy olyan kezdeti fogalmi rendszert hoztam létre, mely a további emberi elemzést megkönnyíti, illetve a dokumentumok normalizálásában felhasználható.

### 2.4.1 Többszavas kifejezések kinyerése

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 *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 Frantzi et al. (2000) foglalja össze, majd az egymásba ágyazott kifejezések problémájára is megoldást nyújtó c-value módszert ismerteti.

Ennek a c-value algoritmusnak egy módosított változatát használtam. Először egy nyelvi szűrőt alkalmaztam 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. Miután az összes, a fenti mintára illeszkedő n-gramot kigyűjtöttem ($1 < n < 10$), mindegyikre meghatároztam 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 Frantzi et al. (2000)-ben találhatóak meg.

**4. TÉZIS:**

*Megmutattam, hogy az egyszerű nyelvi szűrővel kiegészített korpuszalapú módszerek jól alkalmalzhatóak többszavas kifejezések kinyerésére magyar nyelvű klinikai szövegekben.*

Kapcsolódó publikációk: 1, 4, 5, 11

### 2.4.2 A klinikai korpusz disztribúciós viselkedése

A releváns kifejezések csoportosításához szükség van egy hasonlósági metrikára is, ami két kifejezés jelentésbeli távolságát határozza meg. Erre szintén olyan nem felügyelt módszert alkalmaztam, 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. Carroll et al. (2012) szintén klinikai szövegekre alkalmazva csupán a vizsgált szó meghatározott méretű környezetében előforduló szavak lexikai alakjának felhasználásával építettek ilyen szemantikai modellt.

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 számoltam Lin (1998) 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.

Ezzel a metrikával bármely két kifejezés hasonlósága meghatározható. A módszer alkalmazhatóságának vizsgálata érdekében azonban egyelőre csak főnevekre és többszavas (nominális) kifejezésekre alkalmaztam.

Az eredmények rávilágítanak arra, hogy a kifejezések közötti hasonlósági értékek tükrözik azok valós kapcsolatát. Ezért ezt a hasonlósági metrikát felhasználtam egy fogalmi hierarchia építésére. Ezt agglomeratív klaszterezési algoritmussal valósítottam meg, ahol az egyes kifejezéseket az összes többi kifejezéstől való távolságukból álló vektor reprezentálta. Az eredményül kapott rendszerezés egyrészt felhasználható egy orvosi ontológia kiindulásaként. Másrészt, mivel a klaszterezés során minden csomópont külön egyedi azonosítót kapott, ezért ezeket az eredeti szövegbe behelyettesítve, a dokumentumok egy magasabb szintű, absztrakt reprezentációja jön létre.

**5. TÉZIS:**

*Megvalósítottam a magyar nyelvű klinikai korpusz disztribúciós szemantikai modelljét. Az ebből a modellből levezetett hasonlósági metrika alapján létrehoztam a korpuszban szereplő releváns kifejezések hierarchikus rendszerét.*

**5.a TÉZIS:**

*A megvalósított módszer felhasználásával automatikusan létrehoztam az orvosi (szemészeti) kifejezések egy rendszerét, melynek kifinomultsága a paraméterek változtatásával könnyen állítható. A létrejött fogalmi rendszer egy szakértők által, kézzel létrehozott ontológia alapja lehet.*

**5.b TÉZIS:**

*Megmutattam, hogy a fogalmi rendszerezés alkalmas az eredeti dokumentumok absztrakt szintre emelésére, valamint releváns minták felfedezésére a fogalmi hierarchia csomópontjaihoz rendelt egyedi azonosítók felhasználásával.*

Kapcsolódó publikációk: 1, 4, 5, 11, 14

# 3

# A szerző publikációi

## Folyóiratpublikációk

1 **Borbála Siklósi**, Attila Novák, Gábor Prószéky (2016): Context-aware correction of spelling errors in Hungarian medical documents, *Computer Speech & Language*, Vol.35, pp. 219-233, ISSN 0885-2308.

2 **Borbála Siklósi**, Attila Novák, György Orosz, Gábor Prószéky (2014): Processing noisy texts in Hungarian: a showcase from the clinical domain, *Jedlik Laboratories Reports*, Vol. II, no.3, pp. 5-62, ISSN 2064-3942

3 László János Laki, Attila Novák, **Borbála Siklósi**, György Orosz (2013): Syntax-based reordering in phrase-based English-Hungarian statistical machine translation. *International Journal of Computational Linguistics and Applications*, Vol. 4 no. 2. pp. 63–78, ISSN 0976-0962.

## Könyvfejezetek

4 **Borbála Siklósi** (2015): Clustering Relevant Terms and Identifying Types of Statements in Clinical Records, In: A. Gelbukh (Ed.), *Lecture Notes in Computer Science Volume 9042: Computational Linguistics and Intelligent Text Processing* Springer International Publishing, Berlin Heidelberg. Part II pp. 619–630. ISBN 978-3-319-18116-5.

5 **Borbála Siklósi**, Attila Novák (2014): Identifying and Clustering Relevant Terms in Clinical Records Using Unsupervised Methods. In: Besacier, L.; Dediu, A.-H. and Martín-Vide, C. (Eds.), *Lecture Notes in Computer Science Volume 8791: Statistical Language and Speech Processing* Springer International Publishing, Berlin Heidelberg. pp. 233-243 ISBN 978-3-319-11396-8.

6 **Borbála Siklósi**, Attila Novák, Gábor Prószéky (2013): Context-Aware Correction of Spelling Errors in Hungarian Medical Documents. In: Adrian-Horia Dediu, Carlos Martín-Vide, Ruslan Mitkov, Bianca Truthe (Eds.), *Lecture Notes in Computer Science Volume 7978: Statistical Language and Speech Processing* Springer, Berlin Heidelberg. pp. 248–259 ISBN 978-3-642-39592-5.

7 **Borbála Siklósi**, Attila Novák (2013): Detection and Expansion of Abbreviations in Hungarian Clinical Notes. In: F. Castro, A. Gelbukh, M.G. Mendoza (Eds.): *Lecture Notes in Computer Science, Vol. 8265: Advances in Artificial Intelligence and Its Applications.* Springer, Berlin Heidelberg. pp. 318–328. ISBN 978-3-642-45114-0

8 György Orosz, László János Laki, Attila Novák, **Borbála Siklósi** (2013): Improved Hungarian Morphological Disambiguation with Tagger Combination. In: Habernal, Ivan; Matousek, Vaclav (Eds.) *Lecture Notes in Computer Science, Vol. 8082: Text, Speech, and Dialogue* Pilsen, Czech Republic. Springer, Berlin Heidelberg. pp. 280–287. ISBN: 978-3-642-40584-6.

# Konferenciapublikációk

9 Novák Attila, **Siklósi Borbála** (2015): Automatic Diacritics Restoration for Hungarian. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal: Association for Computational Linguistics. pp. 2286–91.

10 **Siklósi Borbála**, Novák Attila (2015): Restoring the intended structure of Hungarian ophthalmology documents. BioNLP Workshop at the $53^{rd}$ Annual Meeting of the Association for Computational Linguistics, ACL 2015. Beijing, China, July 26-31, 2015

11 **Siklósi Borbála**, Novák Attila (2015): Nem felügyelt módszerek alkalmazása releváns kifejezések azonosítására és csoportosítására klinikai dokumentumokban. In: Tanács Attila, Varga Viktor, Vincze Veronika (szerk.) *XI. Magyar Számítógépes Nyelvészeti Konferencia.* Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szeged. pp. 237-248

12 **Borbála Siklósi**, Attila Novák, Gábor Prószéky (2014): Resolving Abbreviations in Clinical Texts Without Pre-existing Structured Resources. In: *Proceedings of the Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing (BioTxtM 2014).* Reykjavík

13 **Siklósi Borbála**, Novák Attila (2014): Rec. et exp. aut. Abbr. mnyelv. KLIN. szöv-ben – rövidítések automatikus felismerése és feloldása magyar nyelvű klinikai szövegekben. In: Tanács Attila, Varga Viktor, Vincze Veronika (szerk.) *X. Magyar Számítógépes Nyelvészeti Konferencia.* Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szeged. pp. 167–176. ISBN 978-963-306-246-3

14 **Siklósi Borbála**, Novák Attila (2014): A magyar beteg. In: Tanács Attila, Varga Viktor, Vincze Veronika (szerk.) *X. Magyar Számítógépes Nyelvészeti Konferencia.* Szegedi Tudományegyetem, Informatikai Tanszékcsoport, Szeged. pp. 188–198. ISBN 978-963-306-246-3

15 **Siklósi Borbála**, Novák Attila, Prószéky Gábor (2013): Helyesírási hibák automatikus javítása orvosi szövegekben a szövegkörnyezet figyelembevételével. In: Tanács Attila; Vincze Veronika (szerk.) *A IX. Magyar Számítógépes Nyelvészeti Konferencia előadásai.* SZTE, Szeged. pp. 148–158 ISBN 978-963-306-189-3

16 **Borbála Siklósi**, György Orosz, Attila Novák, Gábor Prószéky (2012): Automatic structuring and correction suggestion system for Hungarian clinical records. In: *LREC-2012: SALTMIL-AfLaT Workshop on "Language technology for normalisation of less-resourced languages".* Istanbul, Turkey, 2012. pp. 29–34

17 **Siklósi Borbála**, Orosz György, Novák Attila (2011): Magyar nyelvű klinikai dokumentumok előfeldolgozása. In: Tanács Attila, Varga Viktor, Vincze Veronika (szerk.) *VIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2011).* Szegedi Tudományegyetem, pp. 143–340

18 Laki László, Novák Attila, **Siklósi Borbála** (2013): Hunglish mondattan – átrendezésalapú angol-magyar statisztikai gépifordító-rendszer. In: Tanács Attila; Vincze Veronika (szerk.) *A IX. Magyar Számítógépes Nyelvészeti Konferencia előadásai.* SZTE, Szeged. pp. 71–82 ISBN 978-963-306-189-3

19 György Orosz, László János Laki, Attila Novák, **Borbála Siklósi** (2013): Combining Language-Independent Part-of-Speech Tagging Tools. In: J. P. Leal, R. Rocha, and A. Simoes (Eds.), *2nd Symposium on Languages, Applications and Technologies.* Porto: Schloss Dagstuhl–Leibniz-Zentrum für Informatik. pp. 249–257 ISBN 978-3-939897-52-1

20 László János Laki, Attila Novák, **Borbála Siklósi** (2013): English-to-Hungarian Morpheme-based Statistical Machine Translation System with Reordering Rules. In: Marta R. Costa-jussa, Reinhard Rapp, Patrik Lambert, Kurt Eberle, Rafael E. Banchs, Bogdan Babych (Eds.) *Proceedings of the Second Workshop on Hybrid Approaches to Machine Translation (HyTra).* Association for Computational Linguistics. pp. 42–50

# Hivatkozások


Barrows, J. R., Busuioc, M., and Friedman, C. (2000). Limited parsing of notational text visit notes: ad-hoc vs. NLP approaches. *Proceedings of the AMIA Annual Symposium*, pages 51–55.

Carroll, J., Koeling, R., and Puri, S. (2012). Lexical acquisition for clinical text mining using distributional similarity. In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing*, volume 7182 of *Lecture Notes in Computer Science*, pages 232–246. Springer Berlin Heidelberg, Berlin, Heidelberg.

Frantzi, K., Ananiadou, S., and Mima, H. (2000). Automatic recognition of multi-word terms:. the C-value/NC-value method. *International Journal on Digital Libraries*, 3(2):115–130.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, pages 177–180, Prague. Association for Computational Linguistics.

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on Computational linguistics - Volume 2*, COLING '98, pages 768–774, Stroudsburg, PA, USA. Association for Computational Linguistics.

Meystre, S., Savova, G., Kipper-Schuler, K., and Hurdle, J. (2008). Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform*, 35:128–44.

Novák, A. (2003). Milyen a jó Humor? [What is good Humor like?]. In *I. Magyar Számítógépes Nyelvészeti Konferencia*, pages 138–144, Szeged. SZTE.

Prószéky, G. and Kis, B. (1999). 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, pages 261–268, College Park, Maryland. Association for Computational Linguistics.

Sager, N., Lyman, M., Bucknall, C., Nhan, N., and Tick, L. J. (1994). Natural Language Processing and the Representation of Clinical Data. *Journal of the American Medical Informatics Association*, 1(2).

Siklósi, B., Novák, A., and Prószéky, G. (2013). Context-Aware Correction of Spelling Errors in Hungarian Medical Documents. In Dediu, A.-H., Martín-Vide, C., Mitkov, R., and Truthe, B., editors, *Statistical Language and Speech Processing*, volume 7978 of *Lecture Notes in Computer Science*, pages 248–259. Springer Berlin Heidelberg.

Siklósi, B., Orosz, G., Novák, A., and Prószéky, G. (2012). Automatic structuring and correction suggestion system for Hungarian clinical records. In De Pauw, G., De Schryver, G.-M., Forcada, M., M. Tyers, F., and Waiganjo Wagacha, P., editors, *8th SaLTMiL Workshop on Creation and use of basic lexical resources for less-resourced languages*, pages 29–34, Istanbul, Turkey.

Stolcke, A., Zheng, J., Wang, W., and Abrash, V. (2011). SRILM at Sixteen: Update and Outlook. In *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, Waikoloa, Hawaii.