Gondolatok a gondola-tokról

Morfológiai annotációt javító módszerek tesztelése gold standard korpuszon

K. Molnár Emese[1,2], Dömötör Andrea[1,3]

[1]Digitális Örökség Nemzeti Laboratórium
1088 Budapest Múzeum krt. 6-8.
[2]ELTE BTK Nyelvtudományi Doktori Iskola
1088 Budapest Múzeum krt. 6-8.
[3]PPKE BTK Nyelvtudományi Doktori Iskola
1088 Budapest, Mikszáth Kálmán tér 1.
mesii@student.elte.hu, domotor.andrea@btk.elte.hu

Kivonat Kutatásunk célja az volt, hogy csökkentsük az emberi annotációs munka mennyiségét a gold standard korpusz projektünkben. A cikkben (egy baseline mellett) három egyszerűen kivitelezhető módszert vizsgáltunk meg, amely alkalmas lehet az annotálási hibák megtalálására. A cél elsősorban a magas fedés: olyan módszert keresünk, amely úgy szűkíti le a második annotátor által áttekintendő tokenek körét, hogy a lehető legtöbb hibát lefedje. A három módszer eredményeit összegezve azt az eredményt kaptuk, hogy a tokenek 31,65%-át újraannotálva a szövegben előforduló annotálási hibák arányát 2%-ra csökkenthetjük, melynek közel fele csak a részletes (HuMor) elemzést érinti. Ez a munkaterhelésben lényeges csökkentést jelent, miközben magas minőséget is biztosít.
Kulcsszavak: annotáció, morfológia, korpusz, gold standard

1. Bevezetés

Jelenleg magyar nyelvre kevés megbízható minőségű elemzett korpusz áll rendelkezésre, ezért egy új, függőségi elemzést is tartalmazó korpusz (treebank) hiánypótló munka lenne az egész magyar NLP-közösség és a korpuszfelhasználó nyelvészek számára is. A kézzel készült, ellenőrzött annotáció előállítása azonban rendkívül idő- és erőforrásigényes. Kutatásunk ezért azt vizsgálja, hogyan lehetséges minőségi annotációt előállítani kevesebb humán erőforrás felhasználásával.

A klasszikus gold standard módszertan szerint minden szöveget 2 annotátornak kell annotálni, majd egy harmadiknak felülvizsgálni az eltéréseket. Tekintve azonban, hogy a mai gépi elemző eszközök eleve elég magas pontossággal működnek, úgy gondoljuk, hogy nincs szükség ilyen mennyiségű emberi annotációra.

A cikkben (egy baseline mellett) három egyszerűen kivitelezhető módszert vizsgáltunk meg, amely alkalmas lehet az annotálási hibák megtalálására. A cél elsősorban a magas fedés: olyan módszert keresünk, amely úgy szűkíti le a

második annotátor által áttekintendő tokenek körét, hogy a lehető legtöbb hibát lefedje.

Mivel gold standardről beszélünk, az első körös emberi annotációt nem akarjuk megspórolni. Abból indulunk ki tehát, hogy egy annotátor végigannotálja a szöveget, és az ő munkájában keresünk hibajelölteket.

2. Motiváció

Kutatásunk motivációját egy gold standard korpuszépítési projekt adta. Ennek célja egy olyan többszintű nyelvi annotációt tartalmazó általános referenciakorpusz létrehozása, amely adatvezérelt kutatások adatbázisaként, valamint természetes nyelvet feldolgozó gépi tanulásos algoritmusok tanítóanyagaként egyaránt felhasználható. Terveink szerint a korpusz morfológiai, szintaktikai (függőségi), tagmondat- és NP-annotációt is tartalmazni fog. A munka első (jelenlegi) fázisában a morfológiai annotálást végezzük.

A folyamatosan bővülő korpusz jelenleg 350 ezer tokennyi szöveget tartalmaz egy annotátor által (morfológiailag) annotálva. Ennek nagy része blogokból, valamint tudományos-ismeretterjesztő és kulturális weboldalakról származik, illetve megtalálható még a korpuszban néhány 20. századi irodalmi szöveg is. A későbbiekben további műfajok, például sajtó-, hivatalos és szakszövegek felvételét is tervezzük.

Az annotáció munkafolyamata Vadász (2020) és Simon és Vadász (2021) módszertanát követi. Az annotáláshoz az emMorph (Váradi és mtsai, 2018) kódkészletét használtuk. A szövegeket emtsv-vel (Indig és mtsai, 2019) előelemeztük, az annotátorok az emMorph modul kimenetét javították. Korpuszunk morfológiai annotációja tartalmazza a szótőt, a szófajcímkét és a részletes morfológiai elemzést (Novák, 2014) is, az annotátoroknak ezek mindegyikét javítaniuk kellett. A kézi annotáláshoz egy saját fejlesztésű Java alapú felületet használtunk. Ezen az annotátornak lehetősége van a tokenizálás és a morfológiai elemzés javítására is. A tokenizálás javítása során az annotátor egy tokent összevonhat az előtte és az utána lévő tokennel, szétszedhet egy tokent tetszőleges helyen, illetve ha szükséges, át is írhatja a tokent.

A program felületén előhívható egy szó összes, az e-magyar által megadott lehetséges elemzése, amelyek egymás alatt, listaszerűen sorolódnak fel ("Elemzés választása" menüpont). A lehetséges elemzések megjelenítései tartalmazzák a lemmát, valamint a szónak az e-magyar által megadott részletes és egyszerű morfológiai elemzését is. Az e-magyar egyértelműsítő modulja által helyesnek tartott elemzés zöld pipával jelenik meg. Amennyiben nem értünk egyet az e-magyar által helyesként megadott elemzéssel, akkor a listában egy másik lehetséges elemzésre kattintva megváltoztathatjuk azt. Ha az e-magyar által megadott lehetséges elemzések közül egyiket sem tartjuk helyesnek, akkor megadható egy teljesen új elemzés is egy erre szolgáló üres mező kitöltésével. Az annotáló felület bármikor képes az e-magyartól új elemzést lekérni egy-egy tokenre, ha a tokenizálás javítása során a token vagy a mondat megváltozik. A programnak ez a

funkciója jelentősen leegyszerűsíti a munkafolyamatot, hiszen így egy lépésben lehet elvégezni a tokenizálást és a morfológiai elemzést.

Az annotálás gyorsításának érdekében a Szeged Treebank alapján létrehoztunk egy olyan listát, amely az egyértelmű szavakat tartalmazza, tehát azokat, amelyekhez minden esetben csak egyféle elemzés tartozhat. A listát nyelvészek ellenőrizték. A listán szereplő szavakat az előfeldolgozás után az annotáló felület már ellenőrzött tokenként jeleníti meg, így az annotátornak nincs vele munkája[1]. A lista jelenleg 31881 szót tartalmaz, ami általában lefedi az annotálandó szövegek egyharmadát.

A korpusz eredetileg XML formátumú, de a hibakereső kísérletekhez készítettünk egy egyszerűsített TSV formátumot is. Terveink szerint a későbbiekben (a függőségi elemzés bevezetése után) a korpuszt CoNLL-U+ formátumban is közzétesszük.

3. Vizsgált módszerek

A vizsgálandó módszerek kiválasztásakor Ide–Pustejovsky nyelvi annotációs kézikönyvének vonatkozó fejezete (Dickinson és Tufis, 2017) volt a kiindulópontunk. Ez a fejezet számos, az annotáció minőségének javítását célzó módszert foglal össze. Külön alfejezetekben tárgyalja azokat a módszereket, amelyek kész korpuszokra alkalmazhatók jól, és azokat, amelyek folyamatban lévő annotálás esetén is használhatók az inkonzisztens annotációk (és esetleg a séma hibáinak) szűrésére.

Kutatásunkban egy intuitív baseline mellett három, a kézikönyv által ihletett módszer hatékonyságát teszteltük. Mindegyik módszer lényege az, hogy hibajelöltek keresésével csökkentsük a második annotátor által átnézendő tokenek számát úgy, hogy a hibák minél nagyobb részét lefedjük.

3.1. Baseline: többértelmű és egyedi szavak mint hibajelöltek

Baseline módszerünk abból a gondolatból indul ki, hogy azok a szavak, amelyek többször, következetesen ugyanazzal az annotációval szerepelnek a korpuszban, nagy valószínűséggel nem hibásak. Ennek megfelelően hibajelöltnek tekintünk minden olyan tokent, amely a már elkészült korpuszban többféle elemzéssel is előfordul. Ezen kívül hozzá kell adni a hibajelöltekhez minden hapaxot is, hiszen ezeknél nem tudjuk mérni az annotáció következetességét. Így várhatóan sok hibajelöltünk lesz ugyan, de bízhatunk a magas fedésben. A módszer előnye továbbá, hogy mindhárom elemzési szintre (egyszerű elemzés, részletes elemzés, szótövesítés) használható.

3.2. Gépi és emberi annotáció összevetése

A kézikönyv több olyan kísérletet is említ, ahol a szerzők egy automatikus elemző eszköz kimenetével vetették össze az emberi annotációt arra alapozva, hogy a

[1] Kivéve ha a szó egy cím vagy tulajdonnév része, l. 4.2.

gépi eszköz alapvetően konzisztens viselkedésre van betanítva, így az általa adott elemzésektől való eltérés alkalmas lehet az inkonzisztenciák detektálására.

Az egyik ehhez kapcsolódó kísérletünkben az annotálási munkafolyamatban kiindulásnak használt emtsv eredeti elemzéseit vetettük össze az annotátorok által javított változattal abban a reményben, hogy a hibák nagy része azokból az esetekből fog származni, ahol az annotátor módosított a gépi előelemzésen.

Mivel az emtsv nem képes a részletes elemzések egyértelműsítésére (erről bővebben l. 4.3. pont), így ez a módszer csak az egyszerű elemzésekre és a szótövekre alkalmazható.

3.3. Annotáció összevetése másik elemző kimenetével

Az előző módszer továbbfejlesztéseként a kézi annotációkat összehasonlítottuk egy másik, az annotációs munkafolyamattól független elemző kimenetével is. Ehhez az emtsv-be integrált UDPipe és Stanza modulok, illetve a HuSpaCy (Orosz és mtsai, 2022) jött szóba. Ezek mindegyike az UDv2 kódkészletét használja, ezért az összehasonlíthatóság kedvéért a kézi annotációkat az emtsv emmorph2ud2 moduljával konvertáltuk az UDv2 címkekészletre.

Az összehasonlítás során azok a tokenek lettek hibajelöltek, ahol a kézi elemzés és az elemző kimenete eltérő volt. Külön gyűjtöttük a szótő, és külön a morfológiai annotációk eltéréseit. Utóbbi esetben mind a szófajcímke, mind a morfológiai jegyek eltéréseit figyelembe vettük, hiszen ezek együttesen feleltethetők meg az emMorph egyszerű elemzéseinek.

A UDPipe és a Stanza esetén váratlanul nagyszámú eltérést tapasztaltunk. Ennek az volt az oka, hogy az elemzők kimenetében használt kódkészlet nem volt teljesen kompatibilis az emmorph2ud2 által konvertált címkékkel. (Az elemzők jellemzően többféle morfológiai jegyet jelenítettek meg.) Emiatt az eredményeknél csak a HuSpaCy-vel kapott eredményeket ismertetjük.

3.4. Érvénytelen bigram módszer

Květoň és Oliva (2002) módszerének lényege az, hogy érvénytelen szófajcímkebigramokat keres a javítandó korpuszban. Ehhez egy kisebb, validált korpuszra van szükség, amelyből az érvényes bigramok kinyerhetők. Esetünkben a 4.1. pontban leírt tesztkorpusz felelt meg erre a célra.

A tesztkorpusz eredeti annotált és javított (gold standard) változatát is 10 részre osztottuk, és minden iterációban a gold standard 9 részéből készítettük el az érvényes bigramok listáját (csak az egyszerű elemzést figyelembe véve), és a fennmaradó egy rész eredeti annotált változatából kigyűjtöttük a hibajelölteket. A hibajelöltek ebben az esetben azok a tokenek voltak, amelyek egy érvénytelen (azaz az érvényes bigramok listáján nem szereplő) bigram részei voltak.

4. Eredmények

4.1. Tesztkorpusz

Tesztkorpusznak 6 szöveget választottunk ki a korpuszunk eddig elkészült részéből. Ezek összesen 14147 tokent tartalmaznak. A válogatásnál szempont volt, hogy minél több annotátortól kerüljön be szöveg a tesztkorpuszba, illetve a kiválaszott szövegek között legyenek régebben és újabban annotáltak is. Ez utóbbi szempontra azért volt szükség, mert a több mint egy éve tartó annotálási munkafolyamat során előfordultak változtatások az annotálási útmutatóban (ezeket a következő pontban fejtjük ki részletesebben).

A kiválasztott szövegeket a klasszikus gold standard módszer szerint még egy annotátor annotálta az első annotátortól függetlenül. A két annotáció eltéréseit részletesen megvizsgáltuk, és közös megegyezéssel döntöttünk a végleges annotációkról. Az így létrejött gold standardet tekintjük referenciának a hibakereső módszerek kiértékelésénél. A megtalálandó hibák listáját az első annotátorok annotációi alapján állítottuk össze.

4.2. Hibatípusok a tesztkorpuszban

A hibákat egyaránt vizsgáltuk a lemmák, az összetett és az egyszerűsített elemzések esetében is. A különböző elemzési lehetőségek hibái részben átfedik ugyan egymást, azonban a részletes kiértékelés és az egyes hibatípusok pontos azonosítása indokolttá tette, hogy minden csoportot külön is vizsgáljunk. Összesen 1261 hibát azonosítottunk az első annotátorok és a gold standard változatok összevetésével. Ebből 196 esetben a lemma volt hibás, 762 esetben az összetett elemzésben, míg 303 esetben az egyszerűsített elemzésben volt hiba.

Az egyes elemzések hibáit kategorizáltuk és altípusokba osztottuk. Az elemzési szintek hibatípusainak összefoglalását példákkal és gyakoriságértékekkel az 1–3. táblázatok tartalmazzák.

A lemmáknál a hibás szótövesítést (`lem`), a kis- illetve nagybetűk tévesztését (`lett`), az elütéseket (`typ`), a tokenizálási hibákat (`tok`), a szövegben előforduló karakterhibákat (`cerr`), valamint az annotálás közben történt sémaváltozások nyomán kialakuló hibákat (`schem`) különböztettük meg. A hibás szótövesítés esetén az annotáror hibásan határozta meg a szó szerkezetét, ezért nem megfelelő szótő került az elemzésbe. A kis- és nagybetűk tévesztése főként a tulajdonnevek esetén fordult elő, ugyanis az emMorph bizonyos esetekben következetesen kisbetűvel jelölt egyes tulajdonneveket, amelyeket az annotátorok nem minden esetben javítottak. Elütésként határoztuk meg, ha egy-egy betű volt téves a szótövön belül. A tokenizálási hiba esetén a szövegben előforduló gépelési vagy helyesírási hibákat az annotátor nem javította, így ez hibát eredményezett a szótőben. A szövegben megjelentek olyan karakterhibák is, amelyek az írásjeleket és a számjegyeket érintették, ezeket az annotátorok nem minden esetben javították. A sémaváltozás pedig azokat az esetek jelöli, ahol az annotációs séma (útmutató) időközben megváltozott, ezért az első annotátor változata eltér

a gold standardtől. Ez főként a tulajdonnevek és az elváló igekötős igék annotálásában jelent változást. A tulajdonnevek esetében a nyelvi egység státusz (Tolcsvai Nagy, 2008) jelölése érdekében az annotálás során úgy határoztunk, hogy csak az utolsó tagon jelöljük az esetragot, a megelőző tagokat eset nélküli főnévnek vagy melléknévnek címkézzük, tehát például `Pázmány[/N] Péter[/N] Katolikus[/Adj] Egyetemen[/N][Supe]` formában. Ezen belül a címeket speciális főnévi csoportoknak tekintjük, ezért a címek utolsó szavát minden esetben főnévi (`[/N]`) címkével jelöljük a megfelelő esetraggal, a cím többi szava `[None]` (korábban `[/X]`), azaz nem elemzett címkét kap, így például `Jojo[None] + nyuszi[/N] + ban[Ine]` formában elemezzük. Ha az igekötő elválik az igéjétől, az igét, Pethő és mtsai (2022) javaslata alapján igekötős igeként elemezzük. Az igekötő megjelenik az ige lemmájában és zéró morféma formájában a részletes elemzésben is: `kinéz | ki[/Prev]=[] + néz[/V] + [Prs.NDef.3Sg]`.

LEMMA			
Hibakód	**Helytelen**	**Helyes**	**Gyakoriság**
lem	*szó*	*szóval*	70
lett	*nap*	*Nap*	15
typ	*Millenniumi*	*Milleniumi*	5
tok	*nem-megújuló*	*nem*	3
cerr	*15_000*	*15000*	7
schem	*Csillagok*	*[None]*	96
		Összesen:	**196**

1. táblázat. A szótövek hibatípusainak összefoglalása

Az összetett elemzés esetében a helytelen címkekiosztásból származó hibás elemzéseket (`err`), az egyes kifejezések túlelemzéséből (`oana`) vagy alulelemzésből (`uana`) származó hibákat, a szövegben hibásan szereplő tokenek nyomán létrejövő szóhibákat (`word`), hiányzó elemzéseket (`miss`), az annotáció szintaxisában megjelenő hibákat (`syn`), egyéb egyedi hibákat (`x`), az emMorph sémától eltérő esetekben ejtett hibákat (`emM`) és a sémaváltozás folytán kialakuló hibákat (`schem`) azonosítottunk. A hibás elemzésnél az annotátor rosszul határozta meg a szófajt vagy szószerkezet valamelyik elemét, ezért helytelen az összetett elemzés. Azokban az esetekben, amelyekben a szófaj alapvetően megfelelően lett meghatározva, azonban például az összetételek vagy a képzett szavak esetén tovább vagy nem elég szóelemre bontott az annotáror, túlelemzésről vagy alulelemzésről beszélhetünk. Bizonyos szavaknál az is előfordult, hogy az elemző nem kínált fel összetett elemzést, az annotátorok pedig nem pótolták ezeket. Továbbá a megfelelően meghatározott morfológiai elemzések ellenére hibát okozott az is, ha az annotátor az annotáció szintaxisában rontott. Főként az írásjeleket – kötőjeleket és gondolatjeleket – érintő hibákat egyéb egyedi hibáknak tekintettük. Az annotáció során a megváltozott séma következtében felmerülő eltéréseket sémaváltozásként szintén külön kategóriaként kezeltük. Az annotálás kezdetétől két esetben egyértelműsítettük az emMorph sémáját: egyrészt a létigék töveinek meghatározásában, másrészt a számok annotálásában. A létigék esetén minden

esetben szigorúan morfológiai alapon határozzuk meg a szótövet, tehát a *le-* kezdetű létigéket mindig a *lesz* szótőre vezetjük vissza, sohasem a *van*-ra. A többjegyű számok esetén az emMorph bizonyos esetekben egy számjegyként kezeli a számokat, más esetekben viszont külön karakterekre bontja őket. Az egységes és következetes sémakialakítás miatt mi az előbbi megoldást választottuk, tehát a többjegyű számjegyeket egyben elemeztük, nem bontottuk karakterekre és egy címkét kaptak.

RÉSZLETES ELEMZÉS			
Hibakód	**Helytelen**	**Helyes**	**Gyakoriság**
err	mód[/N] + on[Supe]	módon[/Adv]	217
oana	hossz[/N] + ú[_Adjz:Ú/Adj] + [Nom]	hosszú[/Adj] + [Nom]	167
uana	ökoszisztéma[/N] + [Nom]	öko[/CmpdPfx] + szisztéma[/N] + [Nom]	117
word	Malayam[/Adj|nat] + ból[Ela]	Malayalam[/Adj|nat] + ból[Ela]	13
miss		android[/N] + [Nom]	33
syn	Tejút=Tejut + at[Acc]	Tejút[/N]=Tejut + at[Acc]	14
x	IndiaPass[/N] + -[Hyph:Hyph] + t[Acc]	IndiaPass[/N]=Indiapass- + t[Acc]	12
emM	van[/V]=le + gyen[Sbjv.NDef.3Sg]	lesz[/V]=le + gyen[Sbjv.NDef.3Sg]	28
schem	ír[/V] + ni[Inf]	le[/Prev]=[] + ír[/V] + ni[Inf]	161
		Összesen:	**762**

2. táblázat. A részletes elemzések hibatípusainak összefoglalása

Az egyszerűsített elemzésnél különbséget tettünk a szófajtévesztésből származó (`lerr`), a csak a morfológiai jegyekben jelentkező hibák (`morph`), a szövegben hibásan szereplő tokenek nyomán létrejövő szóhibák (`word`), az annotáció szintaxisában mutatkozó hibák (`syn`) és sémaváltozásból adódó hibák (`schem`) között. A szófajtévesztésnél az annotátor alapvetően rossz szófajba sorolta be az adott szót. Hibás morfológiai jegyek esetén azonban a szófaj megfelelő volt, viszont egyéb morfológiai jegyek, mint például a főneveknél az eset vagy az igéknél a szám és személy meghatározása volt hibás. Emellett az egyszerűsített elemzésnél szintén előfordultak olyan hibák, amelyek abból adódtak, hogy az annotátorok nem javították a szövegben szereplő helyesírási hibákat. Továbbá itt is megjelentek a címkék szintaxisában történő tévesztések. Az annotáció közben változó séma az egyszerűsített elemzésben is okozott eltéréseket az első annotátor és a gold standard változat között.

EGYSZERŰ ELEMZÉS			
Hibakód	**Helytelen**	**Helyes**	**Gyakoriság**
lerr	*elhunyt [/V][Pst.NDef.3Sg]*	*elhunyt [/N][Nom]*	150
morph	*sejlik [/V][Prs.NDef.3Sg]*	*sejlik [/V][Prs.Def.3Pl]*	40
word	*rossztanulókat [/N][Pl][Acc]*	*rossz [/Adj][Acc]*	14
syn	*Keralából /N][Ela]*	*Keralából [/N][Ela]*	2
schem	*Fanny [/X]*	*Fanny [None]*	97
		Összesen:	**303**

3. táblázat. Az egyszerű elemzések hibatípusainak összefoglalása

4.3. Az emtsv és az annotátorok teljesítményének kiértékelése

Az emtsv által adott elemzéseket összevetettük a tesztkorpuszunkkal. Az egyszerű elemzésekben 989 (6,99%), a szótövekben pedig 711 (5,03%) hibát találtunk. A részletes elemzések helyességét nem értékeltük ki, hiszen ezekhez nem tartozik egyértelműsítő modul, és mivel sok szóalaknál azonos tő és azonos egyszerű elemzés mellett is lehetséges többféle részletes elemzés (1. ábra), így az emtsv eszközeivel lehetetlen ezek közül választani.

1. ábra: Példa konkurrens részletes elemzésekre

Az annotátorok teljesítményét aszerint értékeltük ki, hogy mennyit találtak meg és javítottak ki az emtsv hibáiból (azaz az eltéréseket vetettük össze az emtsv hibalistájával). Az eredményeket a 4. táblázat tartalmazza. A pontosságértékek szerint a módosítások nagyjából 80-90%-a sikeres javítást jelentett. (Egyébként minden módosított token hibás volt az emtsv elemzésében, csak az annotátorok által megadott javított verziók sem voltak feltétlenül helyesek.) A fedéseredmények pedig azt mutatják, hogy az emtsv hibáinak bő egyharmadával egyetértettek az annotátorok, vagy ha meg is találták őket, nem jól javították ki. A kézi annotáció így összességében 2-3%-ot javított az emtsv eredeti elemzésein.

	Pontosság	**Fedés**
Lemma	81,50%	59,48%
Egyszerű elemzés	91,23%	62,95%

4. táblázat. Az annotátorok eredményei az emtsv hibáinak megtalálásában

Mindez megerősíti azt a kiinduló kérdésfelvetésünket, miszerint a dupla annotáció és az eltérések áttekintése nem túlságosan pazarló módszer-e egy olyan annotálási munkafolyamatban, ahol már egy eleve jól teljesítő gépi elemző eszköz kimenetéből indulunk ki.

4.4. Az annotációt javító módszerek eredményei

Bár a vizsgált módszereknek nem mindegyike alkalmas minden annotációs szintünk ellenőrzésére, a különböző szintű hibák átfedéseinek nagy száma miatt

mindegyik módszert kiértékeltük mindegyik szinten. Az eredményeket elemzési szintenként elkülönítve az 5–7. táblázatok tartalmazzák. Összességében elmondható, hogy a legjobb fedéseredményeket minden esetben a baseline módszerrel kaptuk, de a többi módszerhez képest jóval nagyobb számú átnézendő hibajelölt árán.

	Hibajelöltek száma	**Megtalált hibák**	**Pontosság**	**Fedés**
Baseline	8796	247	2,81%	83,17%
e-magyar	535	99	18,50%	33,33%
HuSpaCy	2714	147	5,42%	49,49%
Érvénytelen bigram	967	130	13,44%	43,77%

5. táblázat. A hibakereső módszerek eredményei az egyszerű elemzésekre

Az egyszerű elemzéseknél a legjobb eredményt a HuSpaCy érte el: az emberi annotációk és a HuSpaCy kimenetének összevetésével a korpusz 20%-ának áttekintése elegendő a hibák közel felének megtalálásához.

A részletes elemzésekre értelemszerűen rosszabb eredményeket kaptunk, hiszen a baseline-on kívül egyik módszer sem alkalmas ennek az elemzési szintnek az ellenőrzésére. Az itt kapott találatok így annak köszönhetőek, hogy ahol az egyszerű elemzés hibás, ott általában a részletes is.

	Hibajelöltek száma	**Megtalált hibák**	**Pontosság**	**Fedés**
Baseline	8284	678	8,18%	88,98%
e-magyar	1476	295	19,99%	38,71%
HuSpaCy	2714	258	9,51%	33,86%
Érvénytelen bigram	967	159	16,44%	20,81%

6. táblázat. A hibakereső módszerek eredményei a részletes elemzésekre

A szótövek gyenge eredményei azonban meglepőek, úgy tűnik a szótőhibák detektálása (önmagában) különösen nehéz feladat. Ebben is a HuSpaCy-vel értük el a legjobb fedést.

	Hibajelöltek száma	**Megtalált hibák**	**Pontosság**	**Fedés**
Baseline	5099	160	3,14%	81,63%
e-magyar	365	32	8,77%	16,33%
HuSpaCy	801	65	8,11%	33,16%
Érvénytelen bigram	967	42	4,34%	21,43%

7. táblázat. A hibakereső módszerek eredményei a szótövekre

Megvizsgáltuk azt is, hogy az egyes módszerek milyen típusú hibákat találnak meg a legsikeresebben. Ezek elemzési szintek szerinti összefoglalását a 8–10. táblázatok tartalmazzák.

	e-magyar	**HuSpaCy**	**Bigram**
Szófajtévesztés	38,00%	48,00%	48,00%
Morfológiai jegyek	20,00%	55,0%	70,00%
Szintaxis	100%	50,0%	100%
Sémaváltozás	29,90%	50,52%	25,77%
Szóhiba	21,43%	21,43%	21,43%

8. táblázat. Az egyszerű elemzések megtalált hibáinak aránya hibatípusok szerint

Az egyszerű elemzéseknél a szintaktikai hibák megtalálása mind az e-magyar, mind az érvénytelen bigram módszerekkel könnyűnek bizonyult, ami nem is meglepő, hiszen szintaktikai hiba csak akkor fordulhat elő az annotációban, ha azt az annotátor kézzel írta be (tehát ezek értelemszerűen el fognak térni az eredeti e-magyar elemzéstől). Az érvénytelen bigram módszer emellett kiemelkedően teljesít a morfológiai jegyek hibáinak detektálásában.

	e-magyar	**HuSpaCy**	**Bigram**
Szintaxis	100%	71,43%	71,43%
Túlelemzés	11,38%	15,57%	7,19%
Alulelemzés	82,05%	23,93%	14,53%
Hibás elemzés	34,10%	41,47%	34,10%
Sémaváltozás	45,96%	35,40%	15,53%
Eltérés az emMorphtól	3,57%	71,43%	42,86%
Szóhiba	38,46%	23,08%	23,08%
Hiányzó elemzés	6,06%	48,48%	39,39%
Egyéb hiba	83,33%	66,67%	16,67%

9. táblázat. A részletes elemzések megtalált hibáinak aránya hibatípusok szerint

A részletes elemzéseknél szintén a szintaktikai hibák megtalálása volt a legkönnyebb. Ezen kívül az e-magyartól való eltérések nagy arányban jeleztek alulelemzést és egyéb hibát, a HuSpaCy segítségével pedig viszonylag jól megtalálhatók az emMorph sémától való eltérések (ennek okát nem vizsgáltuk meg részletesen).

A szótöveknél a kis- és nagybetűk hibái nagy arányban megtalálhatók a HuSpaCy segítégével. Ezen kívül az e-magyar eltérésekkel értünk el viszonylag jó eredményt az elütések detektálásában.

Látható tehát, hogy az egyes hibakereső módszerek különböző hibatípusok megtalálásában sikeresek, ezért érdemes őket kombinálni. Elvégeztük a három módszer (a baseline-t a nagyszámú hibajelölt miatt kihagytuk) összes hibajelölt-

	e-magyar	HuSpaCy	Bigram
Elütés	60,0%	40,0%	20,0%
Karakterhiba	14,29%	42,86%	0%
Hibás szótő	31,43%	55,07%	38,57%
Kis-/nagybetű	0%	86,67%	13,33%
Sémaváltozás	6,25%	8,33%	12,50%
Tokenizálás	0%	0%	0%

10. táblázat. A szótövek megtalált hibáinak aránya hibatípusok szerint

jének[2] összevetését az összes hibával, ezúttal már nem annotációs szintenként, hanem tokenenként. Ez utóbbi összevonást azért is érdemes megtenni, mert láthattuk, hogy a szótőhibák specifikus keresése nem volt sikeres, ám mivel a szótőhibák gyakran járnak együtt a morfológiai elemzés(ek) hibájával, így az elemzések hibajelöltjei egyúttal kiadhatják a szótőhibákat is. Másrészről, az annotálási munka során az egyes annotációs szinteket együtt ellenőrizzük, így a hibajelöltek kiválasztásánál az az elsődleges feladat, hogy határozzuk meg azokat a tokeneket, amelyek potenciálisan hibás annotáció(ka)t tartalmaznak.

Az eredmények összesítését a 11. táblázat tartalmazza. Az annotált szövegekben összesen 777 token volt hibás (a korpusz 5,49%-a). A hibakereső módszerek együttesen 4478 hibajelöltet adtak ki, és a hibák 63,71%-át találták meg. Ez összességében azt jelenti, hogy az egy annotátor által annotált korpusz 31,65%-ának átnézésével a hibás tokenek arányát 2%-ra tudjuk csökkenteni.

Hibajelöltek		**Összes hiba**		**Megtalált hibák**		**Megmaradt hibák**	
Száma	**Aránya**	**Száma**	**Aránya**	**Száma**	**Fedés**	**Száma**	**Aránya**
4478	31,65%	777	5,49%	495	63,71%	282	2,0%

11. táblázat. A vizsgált módszerek eredményeinek összesítése

Az előbbi eredményeket tovább árnyalja, ha megnézzük, hogy miyen típusú hibák maradtak a korpuszban a hibakeresések összesítése[3] után (12. táblázat). Ebből látható, hogy a megmaradt hibák közel fele túl- vagy alulelemzés, ami sok esetben meglehetősen szubjektív megítélésű, és a többi elemzési szintet (szótő, egyszerű elemzés) nem érinti. Az annotálási útmutató változásaiból fakadó hibák (schem) szintén nem jelentenek nagy problémát, hiszen ezek csak a szöve-

[2] A hibajelöltek között megtartottuk azokat a szavakat is, amelyek szerepelnek a 2. pontban említett egyértelmű szavak listáján, sok esetben ugyanis éppen a címekben és egyéb tulajdonnevekben hibáztak az annotátorok, mely esetekre a lista kivételesen nem érvényes.

[3] Több hiba esetén elsősorban az egyszerű elemzés, másodsorban a szótő, és harmadsorban a részletes elemzés hibáit vettük figyelembe. Így például az `oana` típusba sorolt tokenek annotációja biztosan nem tartalmaz más hibát, míg a `morph` típusba soroltaknál több hiba is lehetséges.

gek egy részét érintik, és mivel a sémaváltozások ismertek, ezekre az eltérésekre valószínűleg érdemes lehet specifikusan keresni.

Hibakód	**Megmaradt hibák száma**
oana	118
schem	69
err	37
lerr	30
morph	6
lem	6
uana	4
miss	4
word	3
emM	3
typ	1
cerr	1

12. táblázat. A megmaradt hibák típus szerinti eloszlása

5. Összegzés

Kutatásunk célja az volt, hogy csökkentsük az emberi annotációs munka mennyiségét a gold standard korpusz projektünkben. Három hibakereső módszert teszteltünk részletesen. Ebből kettő az emberi és a gépi annotáció összevetésén alapul. Az annotált tesztkorpuszunkat összehasonlítottuk az előfeldolgozásra is használt e-magyar, és az annotálási munkafolyamattól teljesen független HuSpacy kimenetével is. A harmadik vizsgált módszer, az érvénytelen bigram módszer más megközelítést alkalmaz. Ennek lényege, hogy egy kisebb validált korpuszból kinyerjük a lehetséges szófajcímke-bigramokat, és az annotációk ezzel nem kompatibilis bigramjait alkotó tokeneket tekintjük hibajelöltnek.

Bár önmagában egyik módszer sem adott kiemelkedő eredményt, a megtalált hibák részletesebb vizsgálatából kiderült, hogy a különböző módszerek jellemzően különböző típusú hibákat képesek megtalálni. A három módszer eredményeit összegezve azt az eredményt kaptuk, hogy a tokenek 31,65%-át újraannotálva a szövegben előforduló annotálási hibák arányát 2%-ra csökkenthetjük. Ez a munkaterhelésben lényeges csökkentést jelent, miközben magas minőséget is biztosít.

Korpuszunk és a kutatáshoz használt annotált szövegek megtalálhatók a githubon[4]. További terveink között szerepel specializált módszerek kidolgozása a gyakori hibatípusok megtalálására és javítására.

[4] Gold standard korpusz projekt: https://github.com/ELTE-DH/gold-standard
A cikkben ismertetett kutatás: https://github.com/ELTE-DH/gold-standard-eval

Hivatkozások

Dickinson, M., Tufis, D.: Iterative enhancement. In: Ide, N., Pustejovsky, J. (szerk.) Handbook of Linguistic Annotation, pp. 257–276. Springer (2017)

Indig, B., Sass, B., Simon, E., Mittelholcz, I., Kundráth, P., Vadász, N.: emtsv – Egy formátum mind felett. In: Berend, G., Gosztolya, G., Vincze, V. (szerk.) XV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2019). pp. 235–247. Szegedi Tudományegyetem Informatikai Tanszékcsoport, Szeged (2019)

Květoň, P., Oliva, K.: Achieving an almost correct pos-tagged corpus. In: Sojka, P., Kopeček, I., Pala, K. (szerk.) Text, Speech and Dialogue. pp. 19–26. Springer Berlin Heidelberg, Berlin, Heidelberg (2002)

Novák, A.: A New Form of Humor – Mapping Constraint-Based Computational Morphologies to a Finite-State Representation. In: Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (szerk.) Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). pp. 1068–1073. European Language Resources Association (ELRA), Reykjavik, Iceland (May 2014), aCL Anthology Identifier: L14-1207

Orosz, G., Szántó, Z., Berkecz, P., Szabó, G., Farkas, R.: HuSpaCy: an industrial-strength Hungarian natural language processing toolkit (2022)

Pethő, G., Sass, B., Kalivoda, A., Simon, L., Lipp, V.: Igekötő-kapcsolás. In: Berend, G., Gosztolya, G., Vincze, V. (szerk.) MSZNY 2022, XVIII. Magyar Számítógépes Nyelvészeti Konferencia. pp. 77–91. Szegedi Tudományegyetem, Informatikai Intézet, Szeged (2022)

Simon, E., Vadász, N.: Introducing nytk-nerkor, A gold standard hungarian named entity annotated corpus. In: Ekstein, K., Pártl, F., Konopík, M. (szerk.) Text, Speech, and Dialogue - 24th International Conference, TSD 2021, Olomouc, Czech Republic, September 6-9, 2021, Proceedings. Lecture Notes in Computer Science, vol. 12848, pp. 222–234. Springer (2021)

Tolcsvai Nagy, G.: A tulajdonnév jelentése. In: Bölcskei, A., N. Császi, I. (szerk.) Név és valóság. A VI. Magyar Névtudományi Konferencia előadásai. pp. 30–41. Károli Gáspár Református Egyetem BTK Magyar Nyelvtudományi Tanszéke, Budapest (2008)

Vadász, N.: KorKorpusz: kézzel annotált, többrétegű pilotkorpusz építése. In: Berend, G., Gosztolya, G., Vincze, V. (szerk.) XVI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2020). pp. 141–154. Szegedi Tudományegyetem, TTIK, Informatikai Intézet, Szeged (2020)

Váradi, T., Simon, E., Sass, B., Mittelholcz, I., Novák, A., Indig, B., Farkas, R., Vincze, V.: E-magyar – A Digital Language Processing System. In: chair), N.C.C., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., Tokunaga, T. (szerk.) Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan (May 7-12, 2018 2018)