HunLearner: a magyar nyelv nyelvtanulói korpusza

Vincze Veronika[1], Zsibrita János[2], Durst Péter[3], Szabó Martina Katalin[4]

[1] MTA-SZTE Mesterséges Intelligencia Kutatócsoport
vinczev@inf.u-szeged.hu
[2] Szegedi Tudományegyetem, Informatikai Tanszékcsoport
zsibrita@inf.u-szeged.hu
[3] Szegedi Tudományegyetem, Hungarológia Központ
durst.peter@gmail.com
[4] Szegedi Tudományegyetem, Magyar Nyelvészeti Tanszék
szabomartinakatalin@gmail.com

Kivonat: Cikkünkben bemutatjuk a HunLearner korpuszt, mely a magyart mint idegen nyelvet tanulók által létrehozott szövegeket tartalmaz. A korpusz tartalmazza a morfológiailag hibás főnevek javított alakjait és a hiba kódját is. A javított alakok kézi annotációja lehetővé tette azt is, hogy megvizsgáljuk a hibák automatikus javításának lehetőségeit. Az eredmények azt mutatják, hogy már egyszerű módszerekkel is jelentősen lehet csökkenteni a hibás szóalakok számát egy nem sztenderd szövegben, ami ígéretesnek mutatkozik a nem sztenderd szövegek automatikus feldolgozására nézve.

1 Bevezetés

A magyar nyelvtechnológia eddig túlnyomórészt sztenderd magyar szövegek elemzésével foglalkozott, azonban számos olyan magyar nyelvű dokumentum létezik, amelynek sajátságai eltérnek a sztenderd nyelvtől. Közéjük tartoznak a webes szövegek, a nyelvjárási szövegek, illetve a magyart idegen nyelvként beszélők, továbbá az agysérültek vagy nyelvi zavarral rendelkezők által létrehozott nyelvi produktumok. Az ilyen jellegű szövegek feldolgozásához egyrészt a meglevő elemzők átalakítása, másrészről pedig annotált korpuszok létrehozása szükséges. Ennek első lépéseként az előadásban egy digitalizált magyar nyelvtanulói korpuszról számolunk be.

Nyelvtanulói korpuszoknak nevezzük azokat a korpuszokat, amelyek egy bizonyos nyelvet idegen nyelvként tanulók írott vagy hangzó szövegeit tartalmazzák (vö. [11]). Létrehozásuk célja, hogy fényt deríthessünk mindazokra a sajátságokra, amelyek a tanulók nyelvezetét (*köztes nyelv*, *interlanguage* [10]) az anyanyelvi beszélőkétől megkülönböztetik (vö. [7]). Mivel a digitalizált nyelvtanulói korpuszok lehetővé teszik a diákok nyelvi produktumainak alapos vizsgálatát, fontos szerepet tölthetnek be a kapcsolódó nyelvészeti kutatásokban, valamint az oktatási anyagok fejlesztésének folyamatában egyaránt. Emellett hathatós segítségül szolgálhatnak a hibakereső rendszerek értékelésében és fejlesztésében, valamint a lexikográfia területén a különböző szótárak, köztük az egynyelvű nyelvtanulói szótárak készítésében is (vö. [3,4,6]). Jelentős gyakorlati hasznuknak köszönhetően a nyelvtanulói korpuszok száma az

elmúlt években jelentősen megnövekedett, legtöbbjük azonban valamely nyugat-európai nyelv köztes nyelvi szövegeit tartalmazza [1]. A magyar nyelv vonatkozásában elmondható, hogy, bár a magyart idegen nyelvként tanulók nyelvi hibái régóta képezik vizsgálat tárgyát, a vonatkozó tanulmányok vizsgálati anyagaként nem digitálisan rögzített anyagokat használtak, és az adatok feldolgozása is manuálisan történt. Emellett a viszonylag kisméretű nyelvi anyagokat (10-20 válaszadó) általában a magyar és valamilyen másik nyelv kontrasztív elemzése alapján elemezték. Tudomásunk szerint ez idáig két olyan magyar nyelvtanulói korpusz készült, amelyet digitális formában dolgoztak fel: a BilingBank kínai–magyar, 11 interjút tartalmazó korpusz, valamint az Indiana Egyetem 14, egyenként 10-15 mondatból álló szöveget tartalmazó korpusza [4]. A HunLearner korpusz újdonsága abban rejlik a korábbiakhoz képest, hogy egyrészt jóval nagyobb méretű, mint az eddigiek, másrészt tartalmazza a morfológiailag hibás főnevek javított alakjait és a hibák kódját is.

2 Elméleti háttér és nemzetközi kitekintés

Bár a viszonylag csekély számú érintett miatt a magyar mint idegen nyelv tanítása soha nem foglalt el kitüntetett helyet a nemzetközi köztudatban, módszertana igen hosszú múltra tekint vissza és kiváló nyelvészek tevékenykedtek ezen a területen. A hazai nyelvészeti vizsgálódások ma is a korszerű nemzetközi kutatásokkal karöltve folynak, a magyar nyelv sajátosságainak figyelembevételével. Így nem hiányoznak az utóbbi évtizedek szakirodalmából a hibaelemzéssel foglalkozó tanulmányok sem, amelyek alapvetően a magyar nyelv tanulása és idegen nyelvként történő használata közben elkövetett hibákat[1] csoportosítják és elemzik.

Az elméleti háttér az utóbbi évtizedekben jelentősen megváltozott, hiszen az anyanyelv és az idegen nyelv részletes kontrasztív elemzésén alapuló, a hibákat előre megjósló és kerülni szándékozó behaviorista szempontú megközelítés helyett mára széles körben ismert és elfogadott fogalom lett a *köztes nyelv* (vö. 1. rész), amely a nyelvtanuló saját nyelvi rendszerére utal. Ebben a folyamatosan változó, szerencsés esetben a célnyelvhez egyre jobban közelítő rendszerben a hétköznapi értelemben vett hibákat a nyelvtanuló saját köztes nyelvének megnyilvánulásaként értelmezzük, amelyek a szabályalkotási folyamatokról tanúskodnak. Ennek megfelelően nem a tanulást akadályozó, zavaró jelenségekként szemléljük őket, hanem a nyelvtanulás folyamatának természetes és szükséges velejárójaként. Az anyanyelvet és a célnyelvet, valamint a köztes nyelv tulajdonságait egyaránt figyelembe vevő hibaelemzés tehát nagy segítséget nyújthat ma is a nyelvtanításban. A tanulói korpuszok számítógépes feldolgozásában a morfológiailag igen gazdag magyar nyelv számos kihívást támaszt, és bár már más finnugor nyelvek tanulói korpuszainak köszönhetően állnak rendelkezésre adatok [9], a hibák javítása és kódolása még ezekben a projektekben sem telje-

[1] A nyelvek tanulásának és elsajátításának vizsgálatakor lényeges feladat a célnyelvi szabályoknak nem megfelelő, rendszerszerű eltérések, azaz a valódi hibák (*error*), valamint a nyelvi szabályok tudásának ellenére, alkalmi jelleggel felbukkanó tévesztések (*mistake*) megkülönböztetése, mivel azonban a jelen tanulmány szempontjából ez a probléma nem releváns, a dolgozatban egységesen a *hiba* terminust használjuk.

sen megoldott. A közelmúlt nemzetközi eredményei inspirálóak: új nyelvtanulói korpuszok építéséből, annotálásából és a hibák kezeléséből álló komplex feladatokat sikerült már megoldani idegen nyelvként ritkábban tanított nyelvek esetében is (l. például a cseh nyelv nyelvtanulói korpuszát [8]). A HunLearner nyelvtanulói korpusz építésével arra törekszünk, hogy e hiányosságot a magyar nyelv vonatkozásában is pótoljuk.

3 A korpusz adatai

A HunLearner korpusz szövegei a Zágrábi Egyetem magyar szakos, horvát anyanyelvű hallgatóitól származnak. A diákok három témában írtak fogalmazást: (1) Nehézségek a magyar nyelv tanulásában; (2) Egy szimpatikus ember; (3) Egy Angliában dolgozó magyar levele a családjának. A fogalmazásokat számítógépen készítették el, amelyre legfeljebb egy óra állt a rendelkezésükre. A munka során szótárt, nyelvkönyvet, illetve internetes forrásokat nem volt szabad használniuk, emellett magyar billentyűzettel kellett dolgozniuk. A tényleges nyelvi anyagon kívül a válaszadókra vonatkozó adatokat is tárolunk, azaz a nyelvtanulók életkorára, nemére, anyanyelvére, egyéb idegen nyelvi ismeretére, a magyar nyelv tanulásával töltött eddigi időtartamra, valamint a célnyelvi országban eltöltött időre vonatkozó információkat. Mindezeket a későbbi elemzésekben szándékozzuk felhasználni. A korpusz főbb adatait az alábbi táblázat foglalja össze.

1. táblázat: A HunLearner korpusz adatai.

	Nehézségek	**Szimpatikus ember**	**Anglia**	**Összesen**
Szövegek száma	18	6	11	35
Mondatszám	559	134	258	951
Tokenszám	10433	1930	3936	16299

Az alábbiakban bemutatunk egy részletet a korpuszból:

Amikor én kisgyerek voltam minden évben apámmal Bosznában utaztam.Ott egy kis faluban megismertem egy öreg embert. A neve Bego volt. Ő nagyon erős volt és bőlscesz is. Amikor három fiatal ember földről nem tudhatott felhozni a fákat ő tudhatta. Egész napon tudhatott nehézek munkákat csinálni, erdőben egyedül fákat levágni,kecskékkelhegyekre sétálni és mindent enekelve és vakáció kivül csinált. Estén a háza előtt ült és gyrekeknek falúból ijedősök meséket elbeszél. Ha én ott is nyartam, minden estén a meséket is hallgattam. Nagyon szép volt ott maradni, mert Bego is tüzet megcsinált. Mindenki szeretti őt. Szomsedeinek mindenben segített és mindig mosolyos volt

4 Morfológiai hibák a korpuszban

A korpuszt a *magyarlanc* elemzővel [15] automatikusan elemeztük, majd az elemző által ismeretlennek minősített szavakat további elemzéseknek vetettük alá. Célunk a morfológiai hibák kategorizálása volt. Első lépésként a *hunspell* helyesírás-ellenőrző [12] segítségével javítottuk a hibásan írt szóalakokat. Azokban az esetekben ahol több lehetőséget is ajánlott a program, kézzel választottuk ki a kontextusba illőt. Ezzel a módszerrel az ismeretlen szavak 60%-ára kaptunk elemzést, a maradék 40% túlnyomó többsége idegen szó vagy tulajdonnév volt. Mivel jelenleg a főnévi hibák javítására koncentrálunk, kiszűrtük a főneveket (a javított szavak 45%-át), majd közülük is kiválasztottuk a morfológiai hibát tartalmazókat (azaz a szegmentálási hibát tartalmazó eseteket figyelmen kívül hagytuk). Így a további vizsgálataink alapját összesen 157 főnévi hibás szóalak képezte, ami a javított szavak közel 40%-át jelentette. A 2. táblázat bemutatja az ismeretlen, illetve a javított szavak korpuszbeli számát és arányát.

2. táblázat: Az ismeretlen, illetve javított szavak száma és aránya a korpuszban.

	Nehézségek	**Anglia**	**Szimpatikus ember**	**Összesen**
Szavak száma	8692	3271	1622	13585
Ismeretlen szavak (aránya)	393 (4,52%)	146 (4,46%)	128 (7,89%)	667 (4,91%)
A helyesírás-ellenőrző által felajánlott javítások	2328	614	679	3621
Az elfogadott javítások (aránya)	237 (60,31%)	110 (75,34%)	50 (39,06%)	397 (59,52)
A javított főnevek (aránya)	100 (42,19%)	58 (52,73%)	24 (48%)	182 (44,84%)
A kiszűrt főnevek (aránya)	80 (33,76%)	56 (50,91%)	21 (42%)	157 (39,55%)

Megjegyezzük, hogy a morfológiai elemző által ismeretlennek minősített szavak aránya jóval nagyobb a *Szimpatikus ember* alkorpuszban, mint a másik kettőben, és ugyanitt az elfogadott javítások aránya is jóval alulmarad a többi alkorpuszhoz képest. Ennek valószínűleg az lehet az oka, hogy a fogalmazások témájából fakadóan számos tulajdonnév, elsődlegesen személy- és helynév szerepel a szövegekben, amelyek elemzésére sem a *magyarlanc*, sem a *hunspell* nem volt képes.

A morfológiai hibák osztályozására egy saját kategóriarendszert és az ennek megfelelő kódrendszert hoztunk létre az általános nyelvtanári tapasztalat, valamint a magyar mint idegen nyelv vonatkozásában készült hibaelemzések alapján [5]. A következőkben az osztályozás részleteit mutatjuk be, példákkal illusztrálva a hibák egyes típusait.

A hibás szóalakoknál először is megvizsgáltuk, hogy a szótő vagy a toldalék-e a hibás (természetesen nem zártuk ki azt az esetet sem, hogy mind a kettő is tartalmazhat hibát egyszerre). A szótőben található hibákat aszerint bontottuk tovább, hogy többalakú tő nem megfelelő alakját tartalmazza-e a szó (pl. **kézem* a *kezem* helyett), illetve egyéb elírást, helyesírási hibát találhatunk benne (pl. **problámát* vs. *problémát*). A

szótő minőségét (helyes, hibás, utóbbi esetben mi a hiba jellege) a hibakódok első pozíciója kódolja.

A toldalékolással kapcsolatos hibákat alapvetően szintén két osztályra bontottuk (a két osztály szintén nem zárja ki egymást). Az első hibaosztály a hasonulással kapcsolatos hibákat foglalja magában, a második pedig a hangrenddel, kötőhangokkal és toldalékallomorfokkal kapcsolatos hibákat tartalmazza. A hibakód második pozíciója jelzi a hasonulási hibákat, a harmadik pozíció pedig a második toldalékolási hibaosztálynak feleltethető meg. A kód negyedik pozíciója azt tartalmazza, hogy egy vagy több morfémából áll-e a toldalék. A hibatípusok összefoglalása az alábbi táblázatban látható, példák segítségével illusztrálva.

3. táblázat: Hibatípusok.

Első pozíció – szótő	**Kód**	**Magyarázat**	**Példa**
	A	helyes	
	B	helyesírási hibát tartalmazó szótő	*problámát*
	C	többalakú tő nem megfelelő alakja	*kézek*
	X	egyéb hiba	
Második pozíció – hasonulás	1	nincs hasonulás és nem is kell	*kézt*, *kezet*
	2	van hasonulás, és jó, de egyéb probléma van a toldalékkal	*cukorram* (=*cukorral*)
	3	van hasonulás, de nem kellene	*hallak* (=*halnak*)
	4	nincs hasonulás, de kellene	*cukorval*
	5	van hasonulás, de hibás	*cukornal* (=*cukorral*)
	X	egyéb hasonulási hiba	
Harmadik pozíció – hangrend, kötőhangok, toldalékok allomorfjai	A	helyes allomorf	
	B	hangrendi hiba	*házben*
	C	rossz kötőhang	*házen* (=*házon*)
	D	fölösleges kötőhang	*söröt*
	E	hiányzó kötőhang	*templomt*
	F	fölösleges j birtokjel	*toldalékja*
	G	hiányzó j birtokjel	*kutyáa*
	H	hangrendi illeszkedés egyalakú toldaléknál	*éjfélker*
	X	egyéb toldalékolási hiba	
Negyedik pozíció – toldalékok száma	0	nincs toldalék	*problém*
	1	egy toldalék	*házben*
	2	egynél több toldalék	*kézemben*

A morfológiai hibák automatikus kódolására kifejlesztettünk egy szabályalapú rendszert, amely a hibás és helyes szóalak összevetése alapján rendeli hozzá a hibakódokat az egyes hibás szóalakokhoz. Az automatikus kódokat a *Nehézségek* alkorpuszon ellenőrizve azt állapítottuk meg, hogy azok minősége megfelel az elvárásoknak, 80 esetből mindössze 2 hibát találtunk.

Az alábbiakban bemutatunk egy mintát az automatikusan kódolt szóalakokból. A korpuszban szereplő alakot követi a javított szóalak, majd a hibakód következik:

viszonyot	*viszonyt*	*A1D1*
hidjai	*hídjai*	*C1A2*
rágozást	*ragozást*	*B1A1*
tanszékon	*tanszéken*	*A1C1*
gyokorlatokon	*gyakorlatokon*	*B1A2*

Az automatikus hibakódolás lehetővé tette az egyes hibatípusok számszerűsítését is. Ezáltal megvalósíthatóvá vált, hogy megállapítsuk a tő- és toldaléktévesztések arányát, illetve a hasonulási és hangrendi problémák arányát. A morfológiai jellegű hibák mellett automatikusan megvizsgáltuk az ékezettévesztéses hibák arányát is, hiszen a korpuszbeli szövegek előzetes tanulmányozása arra engedett következtetni, hogy az ékezetek helyes kitétele gyakori hibaforrás a nyelvtanulók körében. A mért adatokat a 4. táblázat foglalja össze.

4. táblázat: A morfológiai hibák száma a korpuszban.

helyesírási hibát tartalmazó szótő	122
többalakú tő nem megfelelő alakja	12
hangrendi hiba	5
rossz kötőhang	8
fölösleges kötőhang	3
hiányzó kötőhang	1
fölösleges j birtokjel	2
egyéb toldalékolási hiba	8
ékezet	40

Az eredmények szerint a leggyakoribb hibatípus a tőtévesztés (85%) volt, különös tekintettel az ékezetek nem megfelelő használatára (28%). A toldaléktévesztések közül pedig a hibás kötőhang volt a leggyakoribb (29%).

5 Az automatikus hibajavítás lehetőségei

A javított alakok kézi annotációja lehetővé teszi azt is, hogy megvizsgáljuk a hibák automatikus javításának lehetőségeit, így teszteltük néhány egyszerű módszer hatékonyságát a hibák kijavítására. Amennyiben a *hunspell* által javasolt első helyes szóalakot választottuk, akkor 81,86%-os pontosságot értünk el az összes javított szóalakot tekintve, ami az összes ismeretlen szóalak 49%-ának felel meg.

Ezen túl egy másik módszert is alkalmaztunk: megvizsgáltuk, hogy a *hunspell* által javasolt szóalakok közül melyek fordulnak elő a Szeged Treebankben [2], és, amennyiben több javasolt szóalak is szerepelt benne, a leggyakoribbat választottuk. Ez a módszer 83%-os pontosságot eredményezett, azonban csak 318 szó esetében tudtuk

alkalmazni, mivel az adatbázisban előfordultak olyan szóalakok, ahol a javítási javaslatok egyike sem szerepelt a korpuszban, így azokhoz nem tudtunk gyakoriságot hozzárendelni.

A fenti két megoldást végül kombináltuk egymással: első lépésben a leggyakoribb javasolt szóalakot rendeltük a hibás alakhoz, illetve azon szavak esetében, ahol ez nem volt lehetséges, a *hunspell* által javasolt első javított alakkal dolgoztunk. Ez a módszer végül 82,62%-os pontossághoz vezetett.

Eredményeink arra utalnak, hogy már egyszerű módszerekkel is jelentősen, körülbelül felére lehet csökkenteni a hibás szóalakok számát egy nem sztenderd szövegben, ami ígéretesnek mutatkozik a nem sztenderd szövegek automatikus feldolgozására nézve. További javítási lehetőségként a különféle tulajdonnévszótárak beépítése kínálkozik a morfológiai elemzőbe, különös tekintettel a nyelvtanulói korpusz szövegeit létrehozó tanulók nemzetiségére és földrajzi környezetére. A HunLearner esetében például egy horvát személy- és földrajzinév-szótár bizonyulna hasznosnak.

A korpuszban természetesen előfordulhatnak olyan esetek is, amikor a szóalak morfológiailag kifogástalan, azonban szintaktikailag nem illik a mondatba, mert például az ige más vonzatot kíván meg. Az ilyen esetek automatikus felderítése nem valósulhat meg pusztán morfológiai elemzés segítségével, ehelyett a szintaxishoz kell segítségért folyamodni. A korpuszt automatikus függőségi elemzésnek vetettük alá a *magyarlanc* 2.0 [15] függőségi moduljával, majd kinyertük belőle az igei vonzatkereteket. Összesen 953 vonzatkeret szerepel a korpuszban, melyeket összehasonlítottuk a Szeged Dependencia Treebankből [13] kigyűjtött vonzatkeretekkel [14], és amelyek nem szerepeltek benne (306 vonzatkeret, az összes keret 32,11%-a), azokat külön vizsgálat alá vetettük. Tekintve, hogy a magyarban nem kötelező fonológiailag megjeleníteni a névmási vonzatokat, kiszűrtük azokat az igéket, amelyek argumentumszerkezete üres volt, így 278 vonzatkeretet kaptunk (29,17%). Ezek közül 37 esetben az egyik vonzat ismeretlen vagy hibás szóalak szófaji kódot kapott, így a morfológiai elemzés tökéletlensége okán a szintaktikai elemzés sem lehetett kielégítő. Összesen tehát 241 olyan vonzatkeret (25,29%) található a korpuszban, amely további vizsgálatra szorul. Előzetes eredményeink szerint a problémás keretek egy része valóban hibás (pl. az *érdekel* ige részes esetű vonzattal: *nekem nem érdekel*), más esetekben a szintaktikai elemző hibázik, illetve lehetnek olyan vonzatkeretek is, amelyek hibátlanok, pusztán nem fordultak elő a Szeged Dependencia Treebankben, így kerültek ebbe a kategóriába (pl. *felvág vmivel*). A későbbiekben szeretnénk részletesebben is megvizsgálni, hogyan lehet automatikus eszközökkel tovább csökkenteni a hibás vonzatkeretek számát.

6 Összegzés

A cikkben bemutattuk a HunLearner korpuszt, mely a magyart mint idegen nyelvet tanulók által létrehozott szövegeket tartalmaz. A korpusz tartalmazza a morfológiailag hibás főnevek javított alakjait és a hiba kódját is. A javított alakok kézi annotációja lehetővé tette azt is, hogy megvizsgáljuk a hibák automatikus javításának lehetőségeit. Az eredmények azt mutatják, hogy már egyszerű módszerekkel is jelentősen lehet

csökkenteni a hibás szóalakok számát egy nem sztenderd szövegben, ami ígéretesnek mutatkozik a nem sztenderd szövegek automatikus feldolgozására nézve.

A jövőben tervezzük a korpusz további bővítését, továbbá szeretnénk feltérképezni a szintaktikai és szóhasználati hibák automatikus módszerekkel történő javításának lehetőségeit. A korpusz kutatási célokra szabadon elérhető a http://www.inf.u-szeged.hu/rgai/hunlearner oldalon.

Köszönetnyilvánítás

A kutatás a TÁMOP-4.2.2/C-11/1/KONV-2012-0013 jelű futurICT projekt keretében az Európai Unió és az Európai Szociális Alap társfinanszírozásával valósult meg. Vincze Veronikát az A/11/83421 jelű fiatal kutatói ösztöndíj keretében a Deutscher Akademischer Austauschdienst támogatta.

Hivatkozások

1. Centre for English Corpus Linguistics (UCL) [http://www.uclouvain.be/en-cecl-lcWorld.html]
2. Csendes, D., Csirik, J., Gyimóthy, T., Kocsor, A.: The Szeged Treebank. In: Proceedings of the Eighth International Conference on Text, Speech and Dialogue (TSD 2005). Karlovy Vary, Czech Republic 12-16 September, and LNAI series Vol. 3658 (2005) 123-131
3. De Cock, S., Granger, S.: Computer Learner Corpora and Monolingual Learners' Dictionaries: the Perfect Match. Lexicographica, Vol. 20 (2005) 72–86
4. Dickinson, M., Ledbetter, S.: Annotating Errors in a Hungarian Learner Corpus. In: Proceedings of the 8th Language Resources and Evaluation Conference (LREC 2012). Istanbul, Turkey (2012)
5. Durst P.: A magyar mint idegen nyelv elsajátításának vizsgálata – különös tekintettel a főnévi és igei szótövekre, valamint a határozott tárgyas ragozásra. Bölcsészdoktori értekezés. Kézirat. Pécs (2010)
6. Granger, S.: A Bird's-eye View of Computer Learner Corpus Research. In: Granger S., Hung J., Petch-Tyson, S. (eds): Computer Learner Corpora, Second Language Acquisition, and Foreign Language Teaching. Amsterdam & Philadelphia, Benjamins (2002) 3–33
7. Granger, S.: The computer learner corpus: A versatile new source of data for SLA research. In: Granger, S. (ed.): Learner English on Computer. London, Addison Wesley Longman Limited (1998) 3–18
8. Hana, J., Rosen, A., Škodová, S., Štindlová, B.: Error-Tagged Learner Corpus of Czech. In: Proceedings of the Fourth Linguistic Annotation Workshop, ACL 2010. (2010) 11−19
9. Jantunen, J. H.: Kansainvälinen oppijansuomen korpus (ICLFI): typologia,taustamuuttujat ja annotointi [International Corpus of Learner Finnish (ICLFI): typology, variables and annotation]. Lähivõrdlusi. Lähivertailuja Vol. 21 (2011) 86–105
10. Selinker, L.: Interlanguage. IRAL, Vol. 10 (1972) 209–230
11. Szirmai M.: Bevezetés a korpusznyelvészetbe. Budapest, Tinta Kiadó (2005)

12. Trón, V., Németh, L., Halácsy, P., Kornai, A., Gyepesi, Gy., Varga, D.: Hunmorph: open source word analysis. In: Proceedings of ACL (2005)
13. Vincze, V. Szauter, D., Almási, A., Móra, Gy., Alexin, Z., Csirik, J.: Hungarian Dependency Treebank. In: Proceedings of the Seventh Conference on International Language Resources and Evaluation (2010)
14. Vincze, V.: Valency frames in a Hungarian corpus. Kézirat (2012)
15. Zsibrita J., Vincze V., Farkas R.: magyarlanc 2.0: szintaktikai elemzés és felgyorsított szófaji egyértelműsítés. In: Tanács A., Vincze V. (szerk.): IX. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2013) 368-374