Német-magyar nyelvtanulói korpusz (Dulko)

Kappel Péter[1], Modrián-Horváth Bernadett[1],
Andreas Nolda[2], Vargáné Drewnowska Ewa[1]

[1] Szegedi Tudományegyetem, Germán Filológiai Intézet
kappelp@lit.u-szeged.hu,
{bernadett.modrianhorvath,ewa5drewa}@gmail.com
[2] Berlin-Brandenburgische Akademie der Wissenschaften, Jägerstraße 22/23,
10117 Berlin, Germany
andreas@nolda.org

Kivonat: Cikkünkben bemutatjuk a Dulko korpuszt, amely magyar anyanyelvű, a németet mint idegen nyelvet tanuló egyetemisták által létrehozott szövegeket tartalmaz. A német-magyar nyelvtanulói korpusz várhatóan 2020-tól szabadon hozzáférhető és (több nyelvtanulói korpuszhoz hasonlóan) az ANNIS keresőrendszerrel (Krause és Zeldes, 2016) online kutatható lesz. A nyelvtanulói szövegek többszintű annotációja a szóalakokon kívül többek között lemmán, szófajokon, metaadatokon (pl. a célnyelv tanulásával töltött időtartam) és hibakategóriákon alapuló lekérdezéseket is lehetővé tesz. A korpusz építése során mind tartalmi, mind korpusztechnológiai szempontból számos innovatív elemet alkalmazunk. Kiemelendő egyrészt az explicit hibajelölés és -kategorizálás, másrészt egy olyan nyílt forráskódú szoftver kifejlesztése, amely (többek között beépített lemmatizálóval és szófaji egyértelműsítővel) megkönnyíti a német nyelvű szövegek annotációját, és a nyelvtanulói korpuszokkal szemben támasztott elvárásokhoz igazodva lehetővé teszi a nyelvi adatok többszintű, transzparens elemzését.

Kulcsszavak: nyelvtanulói korpusz, korpuszépítés, hibaannotáció

1 DULKO - egy új német-magyar nyelvtanulói korpusz szükségességéről

A Dulko korpusz (*D*eutsch-*u*ngarisches *L*erner*ko*rpus) része egy nemzetközi projektnek, amelyben három terület kap központi szerepet: a korpusztechnológia, a nyelvészet és a nyelvdidaktika.[1] A Dulkót a projekt futamideje, azaz három év alatt, a Szegedi Tudományegyetemen tanuló germanisztika szakos hallgatók által írt esszék és fordítások alapján hozzuk létre és tesszük szabadon hozzáférhetővé (CLARIN PUB+BY+SA+PRIV). A Dulko projekt elsődleges célját a germanista hallgatók írásbeli szövegalkotásában jelentkező nyelvi hibák[2] empirikus vizsgálata képezi. A kor-

[1] A részletekhez az említett nemzetközi projektről vö. http://arts.u-szeged.hu/kutatas-tudomany/dulko

[2] Itt fontosnak tartjuk előre leszögezni, hogy germanista hallgatók által vétett nyelvi hibákat eltéréseknek tekintjük a nyelvtanuló köztes nyelvének rendszere és a célnyelv rendszere kö-

puszba a nyelvi hibákkal kapcsolatos adatokat, pl. az egyes hibatípusokat is integráljuk, ezzel a tipikus lexikai, nyelvtani és helyesírási hibák elektronikusan is kereshetővé válnak.

Ezzel a célkitűzéssel a Dulko a nyelvelsajátítás-kutatást és korpusznyelvészetet kapcsolja össze. Korpusznyelvészeti szempontból nézve feltehető a kérdés, hogyan lehet nyelvi eltéréseket egy tanulói korpuszban jól áttekinthetően és nyomon követhető módon a célhipotézisek felállítása által (vö. Reznicek és mtsai, 2013) hibaként interpretálni, kategorizálni és többdimenziós szófaji annotáció és lemmatizáció segítségével a mai elvárások szerint kereshetővé tenni. A projekt céljának megvalósítására csoportunk egyik munkatársa egy olyan eljárást dolgozott ki, amely lehetővé teszi a hallgatók szövegeinek elektronikus feldolgozását nyelvtanulói korpusz formájában (vö. 2-4. fejezet).

Ezen korpusz létrehozása mellett három egymással összefüggő érv szól. Mindenekelőtt figyelemre méltó a korpusz sajátossága és egyedülállósága nyelvészeti és nyelvdidaktikai szempontból. A Dulko szövegei szerzőinek anyanyelve és az általuk megfogalmazott szövegek nyelve nyelvtipológiailag lényegesen különböznek egymástól, ebben rejlik a Dulko korpusz sajátossága pl. a rokon nyelveken alapuló tanulói korpuszokkal szemben. Az indogermán német és a finnugor magyar nyelv nem állnak genetikai rokonságban egymással. Morfoszintaktikai szempontból nézve a német egy leginkább fuzionáló-analitikus-izoláló nyelv, míg a magyar elsősorban agglutináló nyelvnek számít. Ebből számos különbség következik a két nyelv között a fonetikai, fonológiai, morfológiai, szintaktikai, lexikológiai, frazeológiai, pragmatikai és szövegnyelvészeti szinten.[3] Ezek a nyelvtipológián alapuló kontrasztok sok potenciális hibaforrást[4] képezhetnek a magyar anyanyelvű németül tanulók számára. Számos hiba vezethető vissza a német és a magyar mondatok eltérő információs szerkezeti felépítésére vagy olyan különbségekre, mint a nyelvtani nem megléte a németben és hiánya a magyarban (lásd pl. 1. ábra). A Dulko egyik központi tartalmi célkitűzése az annotált nyelvi hibák leírása és visszavezetése a két nyelv között fennálló nyelvtipológiai különbségekre.[5] A korpusz másik ehhez kapcsolódó célja, hogy megfelelő adatokat szolgáltasson a tanulói nyelv (Lernersprache) kutatásához.[6] A tanulói nyelv egy még mindig kevéssé feltárt, nagyon releváns kutatási területnek számít. Hazai és nemzetközi viszonylatban a legtöbb publikáció a kezdő szintű tanulói nyelvváltozat kutatásával foglalkozik, a német-magyar nyelvtanulói korpuszunk sajátossága tehát

zött. A 'köztes nyelv' (másként: 'interimnyelv') – a nyelvtanulás folyamata során kialakult sajátos nyelvrendszer, amely úgy a tanuló anyanyelve, mint az általa elsajátítandó a célnyelv jellemzőit tartalmazza. Ezen kívül a köztes nyelv rendszerében más jellemzők is találhatók, amelyek sem az anyanyelvben, sem a célnyelvben nem lépnek fel. (vö. Selinker, 1972; Fekete 2016).

[3] Részletes nyelvtipológiai leírásokhoz a német és a magyar nyelv között vö. Brdar-Szabó, 2010b; Gunkel és mtsai, 2017; Pilarský, 2018.

[4] A cikk korlátozott terjedelme miatt itt nem térünk ki más releváns faktorokra, amelyek szintén hibákhoz vezethetnek a német nyelvű szövegek fogalmazásánál.

[5] Ezzel együtt e célkitűzés a német és a magyar nyelv közötti esetleges hasonlóságokat természetesen nem hagyja figyelmen kívül. A fő elv itt az, hogy a hibaelemzés a célnyelv és a tanulók anyanyelvének összevetésén alapuljon.

[6] A nyelvi hiba, a hibaelemzés és a nyelvek közötti kontraszt fontosságához a nyelvelsajátításban és nyelvtanításban vö. Brdar-Szabó, 2010a; Fekete, 2008, 2016.

többek között abban rejlik, hogy kontrollált körülmények között, haladó nyelvtanulóktól gyűjtött autentikus adatokból áll.[7]

Második érvként az olyan tananyagok hiánya jelölhető meg a német mint idegen nyelv tanításában, melyek a következő, a jelenlegi Nemzeti Alaptantervben (2018) megnevezett tanulói kompetenciák fejlesztését segítenék: „Az anyanyelv és az idegen nyelv különbségének felismerése, ennek megfogalmazása a diák saját szavaival"; „Az anyanyelvi és az idegen nyelvi ismeretek összevetése, az egyes jelenségek egyre pontosabb megnevezése"; „Az anyanyelvhez és az idegen nyelvhez kötődő sajátosságok összevetése az általános nyelvészeti ismeretek felhasználásával."[8] Ezeknek az elvárt kompetenciáknak a magyar iskolákban a német nyelv tanításához használt tankönyvek alig felelnek meg, mivel egyáltalán nem, vagy csak nagyon ritkán kerül bennük szóba a német és a magyar nyelv összevetése. Erről meggyőződhettünk, amikor 2016-ban a projektünk elindítása előtt egy felmérést végeztünk a szegedi gimnáziumi tanárok között. Hasonló eredményekre jutott Fekete (2016), aki a „Schulbus", „Das Deutschmobil", „Start! Neu" és „Unterwegs" c. tankönyveket abból a szempontból vizsgálta meg, hogy a nyelvtan ismertetése mennyire alapul a nyelvi különbségek figyelembe vételén. A német és magyar nyelv összevetésének teljes hiánya, vagy a nyelvi különbségek csak nagyon következetlen figyelembevétele ezekben a tankönyvekben nyomós okot szolgáltat arra, hogy kiegészítő, rendszeresen használható és progresszíven felépített tananyag készüljön. Itt a Dulko korpusz potenciális forrásként és alapként szolgálhat az új tananyagok kialakításához.

A Dulko tudomásunk szerint az egyetlen, az ANNIS keresőrendszerrel (Krause és Zeldes, 2016) online kutatható német-magyar nyelvtanulói korpusz (a részletekhez l. 4. és 5. fejezet). Ennek köszönhetően az érdeklődő szakemberek számára könnyen és széles körben hozzáférhetővé válik. Itt elsősorban német nyelvtanárokra és nyelvi kontraszttal, nyelvtipológiával foglalkozó nyelvészekre gondolunk.[9]

Harmadik, záró érvként megemlítendő, hogy a Dulkónak a korpusztechnológia területén belül is fontos szerep jut: A tanulmány elején megnevezett nemzetközi projektnek, amelynek a Dulko részét képezi, többek között az a célja, hogy olyan eljárásokat dolgozzunk ki, melyek lehetővé teszik, hogy nyelvi tulajdonságokat szövegkorpuszok alapján egy elemzést támogató szoftver segítségével hasonlíthassunk össze. Ennek előfeltétele az összehasonlítható német és magyar nyelvű korpuszok fejlesztése. Az összehasonlítható korpuszok kialakításánál a DeReKo (Deutsches Referenzkorpus, a német nyelv reprezentatív korpusza, vö. Institut für Deutsche Sprache, 2004ff.) és az MNSZ (Magyar Nemzeti Szövegtár, vö. Váradi, 2002) korpuszokra támaszkodunk. A két korpusz technológiai harmonizációját a mannheimi IDS által

[7] Juhász (1970) az interferencia problémáival foglalkozó monográfiája nem tanulói korpuszon, hanem kísérleti alapon gyűjtött, kevésbé autentikus adatokon alapul. A Falko alkorpuszokban csak nagyon kevés adat van magyar anyanyelvű tanulókról (vö. Reznicek és mtsai, 2012). Fekete (2016) longitudinális elemzése egy 90 írott szövegből álló korpuszon alapul. A szövegek magyar gimnazistáktól származnak. Ezzel szemben a Dulko korpusz esetében haladó tanulói nyelvváltozatról van szó Walter és Grommes (2008) értelmében. Ezen kívül Fekete korpusza online nem elérhető.

[8] Vö. NAT (2018) „Anyanyelvi kultúra, ismeretek az anyanyelvről" c. 5. fejezete 37. o.

[9] Az ANNIS-keresést támogató formátum segítségével a Dulkóban levő annotátumok integrálhatók lesznek a Falko korpuszba (a részletes technikai leíráshoz vö. 3.1 fejezet). Ez is lényegesen növelheti a Dulko korpusz nyilvános jellegét és hozzáférhetőségét.

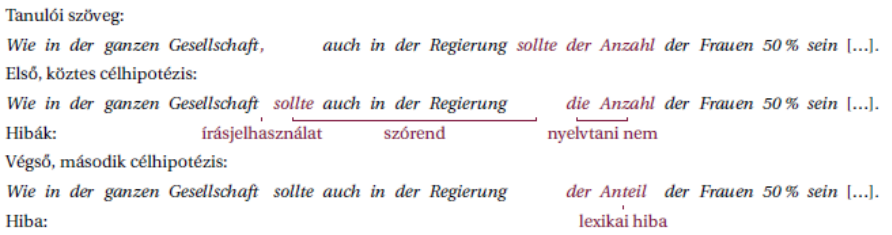
fejlesztett KorAP rendszerrel (Korpusanalyseplattform der nächsten Generation, a következő generáció korpuszelemzési plattformja)[10] kívánjuk megoldani. A Dulko korpuszt nyelvi hibák annotációjának továbbfejlesztése mellett az újgenerációs KorAP rendszerébe is be szeretnénk ágyazni.

2 Az annotációs eljárás alapelvei

A Dulko korpusz adatgyűjtés és -kezelés tekintetében a Falko korpuszon alapszik. A Falko egy nyelvtanulói szövegkorpusz, amelyet 2005 óta a berlini Humboldt Egyetemen fejlesztenek (vö. Reznicek és mtsai, 2012). A Falkótól azonban abban különbözik a Dulko, hogy itt a nyelvi hibák explicit többdimenziós (többszintes) annotációját is elvégezzük. Így a Dulko a következő releváns pontokban tér el a Falkótól (vö. Hirschmann és Nolda, 2019; Nolda, 2019):

1. A Dulko-féle annotációs eljárásban a célhipotézisek mennyisége tetszés szerint adható meg.
2. A Dulko-eljárásban a hibák és ezek területei explicit módon annotálhatók a hibakategóriák segítségével különböző nyelvi szinteken.
3. A Dulko-eljárásban minden célhipotézishez hozzá lehet rendelni hibakategóriát bármely nyelvi szinten.

A célhipotézisek segítségével lépésenként közelíthető a tanulói szöveg a hibáktól megtisztított célnyelvi megfelelőjéhez. Az alábbi szövegrészletben a következő eltéréseket tekinthetjük hibának (vö. 1. ábra):



1. ábra: Egy esszészöveg részletének hibaelemzése, 2017/2018. I. félév, SZTE

- A tanulói szövegben kitett vessző helyesírási hibának számít.
- Szórendi hiba van a *sollte* igealak esetében.
- A *der Anzahl* főnévi csoport *der* névelőjén jelzett nyelvtani neme hibás.
- A *der Anzahl* főnévi csoport *Anzahl* lexikai egysége nem illik a kontextushoz, lexikai hibaként értékelhető: az *50 %* nem számot (*Anzahl*), hanem arányt (*Anteil*) fejez ki.

[10] http://www1.ids-mannheim.de/kl/projekte/korap/

Amennyiben nem adnánk meg az első, köztes célhipotézist, a *der Anzahl* hibás nyelvtani német nem lehetne explicit formában láthatóvá tenni, mivel az *Anzahl* főnév nyelvtani neme megegyezik az *Anteil* főnévével.

Ahogy a fenti példa mutatja, a köztes célhipotézisek azokat a hibákat teszik láthatóvá, amelyek a végső hipotézisnél az átfedések miatt láthatatlanná válnak. A köztes célhipotézishez igény szerint alternatív hipotézisek is megadhatók (vö. Nolda, 2019).

Köztes célhipotézisekre akkor is szükség lehet, amikor például morfológiai vagy lexikai és szórendi hiba együtt fordul elő, ugyanazon a nyelvi szinten két vagy több hiba található, illetve ha más okból nem oldható meg egy lépésben a korrektúra.

3 Az EXMARaLDA (Dulko) annotációs szoftver

Projektünk koncepciója a Falko korpuszcsalád nyomán, arra nagymértékben építve alakult ki, így annotációs eljárásunk kiindulópontja is a Falkóban használt konvenció volt. Ezt Andreas Nolda munkatársunk Hagen Hirschmann közreműködésével továbbfejlesztette és az EXMARaLDA-Partitur-Editor programra[11] implementálta. 2018-ban az „Innovatív eljárás idegennyelv-tanulók nyelvi adatainak annotációjára nyelvtanulói korpuszokban: Koncepció, modellálás, programozás” munkájáért a Szegedi Tudományegyetem innovációs díját kapta a műszaki tudományok területén.

A Dulko (akárcsak a Falko) annotációs eljárása automatikus és manuális elemeket is magába foglal. Automatikus elsősorban a tokenizálás, a mondathatárok szerinti, a szófaj szerinti (pos) annotáció, a lemmatizáció, valamint a tanulói szöveg és a célhipotézis, ill. az egyes kumulatív célhipotézisek közti eltérések annotációja. A manuális annotáció főként a tanulói szöveg egyes mondataihoz rendelt célhipotézisekre, valamint az eltérések különböző nyelvi szinteken való explicit hibaannotációjára irányul.

Az alábbiakban röviden bemutatjuk az annotációs program működési mechanizmusát, valamint a főbb transzformációs műveleteket.

3.1 Az annotációs program működési háttere: felhasznált szoftverek és kereshetőség

A program az annotációs lépéseket az EXMARaLDA-Partitur-Editor (vö. Schmidt, 2004) segítségével végzi, ennek módosításaként jött létre a tanulói adatokat annotáló EXMARaLDA (Dulko). Ez egy nyílt forráskódú szoftver, mely a Bitbucket projekt hosting platformon keresztül ingyenesen hozzáférhető (a licenc GPL, 2. verziójú).[12] Hasonló programról nincsen tudomásunk. A program futtatható Linux, MacOS és Windows operációs rendszerekkel is.

Az EXMARaLDA (Dulko) projektspecifikus transzformációs szcenáriókat tartalmaz, amelyek mindegyike egy, a share-jegyzékben tárolt XSLT-stylesheethez kapcsolódik. A tokenizálás után, melyet az EXMARaLDA-Partitur-Editor végez, történik a mondathatárok szerinti annotáció, a pos-tagging és a lemmatizálás a TreeTagger (Schmid, 1997) programon belüli alkalmazásával. A TreeTagger a német nyelv szófa-

[11] https://exmaralda.org/de

[12] https://bitbucket.org/nolda/exmaralda-dulko/downloads/

ji meghatározásánál sztenderdnek számító STTS-tagsetet (Schiller és mtsai, 1997) használja. Ezt követi a célhipotézisek, ill. a célhipotézisek és a tanulói szöveg közötti eltérések regisztrálása. Az EXMARaLDA (Dulko) projektspecifikus transzformációi közé tartozik t. k. a tanulói produktum célhipotézisbe másolása, ill. – kumulatív célhipotézisek esetén – a célhipotézisek következő szintű célhipotézisbe való másolása, a hibaannotálás célhipotézisenként négy szintjének kialakítása, valamint ide kapcsolódik a hibatageket tartalmazó XML-dokumentum, az annotációs panel is.

Az annotáció során nyert XML-dokumentumot egy további EXMARaLDA (Dulko)-transzformáció segítségével az ANNIS keresővel (Krause és Zeldes, 2016) kompatibilis formátumba lehet hozni, illetve html-formátumba is lehet konvertálni. Az előbbi az internetes (vagy helyi hálózatokon belüli) ANNIS-keresést[13] támogatja, ezáltal az annotátumok integrálhatók lesznek a Falko korpuszba. A html-verzió egy mondatonként tördelt változatot tartalmaz, amely közvetlen olvasásra a leginkább alkalmas. Céljaink között szerepel a mannheimi Institut für Deutsche Sprache KorAP rendszerével való kompatibilitás kialakítása is (vö. 1.).

3.2 Transzformációk az EXMARaLDA (Dulko) szoftverben

A projektspecifikus transzformációk elsősorban a következőket tartalmazzák:

1. Metaadatok átvitele a Dulko-template-ből: az egész projektre érvényes adatok importálása, és a hallgatóra vonatkozó, valamint a szöveggel kapcsolatos metaadatok beviteléhez alkalmas sablonok átvétele. Az adatközlők anonimitásának megőrzése érdekében a hallgatói kódokból md5-kódokat generálunk és ezeket a program az annotáció minden sorához hozzárendeli.
2. Word-(szóalak-)sor (korábban: tok-sor) generálása: a bemásolt hallgatói szöveg tokenizálása, ill. ennek aktualizálása.
3. A hallgatói szöveghez (word-sor) kapcsolódó mondathatárok szerinti és pos-annotáció, valamint lemmatizáció és ezek aktualizálása.
4. Orig-sor, layout-sor és graph-sor: a program új (17.0-tól) verziójában lehetőség van a hallgatók által véghez vitt javítások (kihúzás, betoldás, javítás stb.) és az eredeti szöveg sortöréseinek és bekezdéseinek jelölésére. Erre a célra a word-sor adatai másolódnak az orig-, layout, ill. graph-sorokba, ezeket lehet manuálisan megváltoztatni.
5. Különbségek felismerése a word- és az orig-/layout-/graph-sorok között, valamint ezek aktualizálása.
6. Trans-sor (fordítás) hozzáadása: fordított szövegek esetén mondatonként kerül rögzítésre a fordítás alapját képező forrásnyelvi feladatszöveg.
7. Célhipotézis-sor és a hozzá kapcsolódó hibasorok létrehozása, ill. aktualizálása. A célhipotézisnél szintén a word-sorban található hallgatói szöveg kerül átmásolásra, melyet manuálisan lehet módosítani a célnyelvi normának megfelelően. Lehetőség van második, ill. további célhipotézisek hozzáadására is. A kapcsolódó hibasorok négy nyelvi szintnek megfelelő sorokat rendelnek minden célhipotézishez (helyes-

[13] https://corpus-tools.org/annis/

írás, morfológia, szintaxis és szemantika), melyekbe az annotációs panelben definiált hibakategóriák szerint lehet hibatageket beszúrni.
8. A célhipotézishez (vagy második, ill. további célhipotézishez) kapcsolódó mondathatárok szerinti és pos-annotáció, valamint lemmatizáció, valamint ezek aktualizálása.
9. Különbségek felismerése a word- és a célhipotézis-sorok között, ill. kumulatív célhipotézisek esetén az egyes célhipotézisek között (betoldás, törlés, mozgatás, egyesítés, hasítás) és ezek aktualizálása.
10. „Tisztító" transzformációk: a mondatszakaszok rendezésére, ill. a már nem használt időpontok[14] törlésére vonatkozó műveletek.
11. Konvertálás: html-, ill. ANNIS-kompatibilis verzió létrehozása.

3.3 Az annotációs eljárás szemléltetése

Az annotáció főbb lépéseinek szemléltetésére álljon itt néhány képernyőkép, melyeket egy fordításszöveg egy mondatának annotálása során készítettünk. Az elsőn (2. ábra) a tokenizált tanulói szöveg látszik, a következőn a mondathatár és szófaj szerint taggelt, lemmatizált változat.

[word]	Ich	konnte	vieles	besuchen	ohne	dass	,	ich	bei	den	lokalen	Menschen	bemerkt	worden	wäre	.

2. ábra Tokenizált hallgatói szöveg az EXMARaLDA (Dulko) programban

[word]	Ich	konnte	vieles	besuchen	ohne	dass	,	ich	bei	den	lokalen	Menschen	bemerkt	worden	wäre	.
[S]	s1															
[pos]	PPER	VMFIN	PIS	VVINF	APPR	KOUS	$,	PPER	APPR	ART	ADJA	NN	VVPP	VAPP	VAFIN	$.
[lemma]	ich	können	viele	besuchen	ohne	dass	,	ich	bei	die	lokal	Mensch	bemerken	werden	sein	.

3. ábra Tokenizált, mondathatár és szófaj szerint annotált és lemmatizált hallgatói szöveg az EXMARaLDA (Dulko) programban

A 4. ábrán a program beszúrta a fordítási sort, melybe manuálisan bekerült a kiinduló nyelvi szöveg. Ezután a célhipotézis előállításához automatikusan bemásolásra kerül a word-sor (a tokenizált tanulói szöveg), valamint ekkor jelennek meg a hibataggelésre szolgáló sorok is (5. ábra). Ezeket manuálisan kell megváltoztatni, majd a különbségek felismerésére szolgáló transzformáció segítségével megjeleníteni a tanulói szöveg és a célhipotézis közötti eltéréseket (6. ábra).

Szükség esetén – például, ha egy nyelvi-annotációs szinten több hiba fordul elő, vagy egyidejűleg szórendi és más hiba is előfordul – második, illetve további célhipotézisek előállítására is sor kerülhet, melynek másolása, módosítása és annotálása az első célhipotézisével analóg módon történik; a végeredményt a 7. ábra mutatja. A két célhipotézist jelen esetben kumulatívan kell értelmezni, tehát a hallgatói szövegtől a lehető legkevésbé eltérő, de a célnyelvi normának már megfelelő változat itt a 2. célhipotézisben olvasható.

[14] Mivel az EXMARaLDA egy eredetileg beszélt nyelvi korpuszok annotációjára készült program, itt az egyes tokenek időpontokhoz vannak rendelve.

[word]	Ich	konnte	vieles	besuchen	ohne	dass	,	ich	bei	den	lokalen	Menschen	bemerkt	worden	wäre	.
[S]	s1															
[pos]	PPER	VMFIN	PIS	VVINF	APPR	KOUS	$,	PPER	APPR	ART	ADJA	NN	VVPP	VAPP	VAFIN	$.
[lemma]	ich	können	viele	besuchen	ohne	dass	,	ich	bei	die	lokal	Mensch	bemerken	werden	sein	.
[trans]	Sok mindent megnézhettem anélkül, hogy a helyieknek feltűntem volna.															

4. ábra Fordítási sor hozzáfűzése az EXMARaLDA (Dulko) programban

[word]	Ich	konnte	vieles	besuchen	ohne	dass	,	ich	bei	den	lokalen	Menschen	bemerkt	worden	wäre	.
[S]	s1															
[pos]	PPER	VMFIN	PIS	VVINF	APPR	KOUS	$,	PPER	APPR	ART	ADJA	NN	VVPP	VAPP	VAFIN	$.
[lemma]	ich	können	viele	besuchen	ohne	dass	,	ich	bei	die	lokal	Mensch	bemerken	werden	sein	.
[trans]	Sok mindent megnézhettem anélkül, hogy a helyieknek feltűntem volna.															
[ZH]	Ich	konnte	vieles	besuchen	ohne	dass	,	ich	bei	den	lokalen	Menschen	bemerkt	worden	wäre	.
[FehlerOrth]																
[FehlerMorph]																
[FehlerSyn]																
[FehlerLex]																
[FehlerSem]																

5. ábra Célhipotézis automatikus előállítása az EXMARaLDA (Dulko) programban

4 A korpuszépítés folyamata

A tanulói korpusz építésében hét projekttagunk, valamint pályázat útján kiválasztott hallgatói segéderők vesznek részt. A korpusz építése a mintavétel megtervezésétől az adatgyűjtésen át az annotációs folyamat végrehajtásáig, illetve a korpusz publikálásáig tart. Az adatok feldolgozásával párhuzamosan az első időszakban a hibatagset optimalizálása is fontos feladatot jelentett projektcsoportunk számára, hiszen egyrészt néhány hibatípus csak nagyobb mennyiségű szöveg kiértékelése után fordult elő, másrészt bizonyos hibatagek összevonhatónak bizonyultak az annotációs munka során.

4.1 Adatgyűjtés, metaadatok gyűjtése

Az adatgyűjtés félévente történik intézetünkben kontrollált körülmények között, részben tanóra, részben vizsga keretében. A feladat (esszéírás vagy fordítás) elvégzéséhez a hallgatók semmilyen segédeszközt (szótárat, internetet stb.) nem vehetnek igénybe, a munka elkészítése – ellentétben a Falko szövegeinek legnagyobb részével – kézírással történik. A mintavétel megtervezése során törekszünk a longitudinális vizsgálatok lehetővé tételére, tehát ugyanazon hallgatói csoportoktól igyekszünk több egymást követő félévben is adatokat gyűjteni.

[word]	Ich	konnte	vieles	besuchen		ohne	dass	,	ich	bei	den	lokalen	Menschen	bemerkt	worden	wäre	.
[S]	s1																
[pos]	PPER	VMFIN	PIS	VVINF		APPR	KOUS	$,	PPER	APPR	ART	ADJA	NN	VVPP	VAPP	VAFIN	$.
[lemma]	ich	können	viele	besuchen		ohne	dass	,	ich	bei	die	lokal	Mensch	bemerken	werden	sein	.
[trans]	Sok mindent megnézhettem anélkül, hogy a helyieknek feltűntem volna.																
[ZH]	Ich	konnte	vieles	besuchen	,	ohne	dass		ich	von	den	lokalen	Menschen	bemerkt	worden	wäre	.
[ZHDiff]					MOVT			MOVS		CHA							
[ZHS]	s1																
[ZHpos]	PPER	VMFIN	PIS	VVINF	$,	APPR	KOUS		PPER	APPR	ART	ADJA	NN	VVPP	VAPP	VAFIN	$.
[ZHlemma]	ich	können	viele	besuchen	,	ohne	dass		ich	von	die	lokal	Mensch	bemerken	werden	sein	.
[FehlerOrth]					ZS			ZS									
[FehlerMorph]																	
[FehlerSyn]										ValV							
[FehlerLex]																	
[FehlerSem]																	

6. ábra Manuálisan módosított, hibataggelt célhipotézis a hallgatói szövegtől való eltérések jelölésével az EXMARaLDA (Dulko) programban

[word]	Ich	konnte	vieles	besuchen		ohne	dass	,	ich	bei	den	lokalen	Menschen	bemerkt	worden	wäre	.
[S]	s1																
[pos]	PPER	VMFIN	PIS	VVINF		APPR	KOUS	$,	PPER	APPR	ART	ADJA	NN	VVPP	VAPP	VAFIN	$.
[lemma]	ich	können	viele	besuchen		ohne	dass	,	ich	bei	die	lokal	Mensch	bemerken	werden	sein	.
[trans]	Sok mindent megnézhettem anélkül, hogy a helyieknek feltűntem volna.																
[ZH]	Ich	konnte	vieles	besuchen	,	ohne	dass		ich	von	den	lokalen	Menschen	bemerkt	worden	wäre	.
[ZHDiff]					MOVT			MOVS		CHA							
[ZHS]	s1																
[ZHpos]	PPER	VMFIN	PIS	VVINF	$,	APPR	KOUS		PPER	APPR	ART	ADJA	NN	VVPP	VAPP	VAFIN	$.
[ZHlemma]	ich	können	viele	besuchen	,	ohne	dass		ich	von	die	lokal	Mensch	bemerken	werden	sein	.
[FehlerOrth]					ZS			ZS									
[FehlerMorph]																	
[FehlerSyn]										ValV							
[FehlerLex]																	
[FehlerSem]																	
[ZH]	Ich	konnte	vieles	besuchen	,	ohne	dass		ich	von	den	Einheimischen		bemerkt	worden	wäre	.
[ZHDiff]												MERGE					
[ZHS]	s1																
[ZHpos]	PPER	VMFIN	PIS	VVINF	$,	APPR	KOUS		PPER	APPR	ART	NN		VVPP	VAPP	VAFIN	$.
[ZHlemma]	ich	können	viele	besuchen	,	ohne	dass		ich	von	die	Einheimische		bemerken	werden	sein	.
[FehlerOrth]																	
[FehlerMorph]																	
[FehlerSyn]																	
[FehlerLex]												Lex					

7. ábra Tokenizált hallgatói szöveg az EXMARaLDA (Dulko) programban

A szövegeken kívül egyúttal a hallgatókra vonatkozó metaadatok gyűjtésére is sor kerül, a Granger és Paquot (2017) által, tanulói korpuszokra kifejlesztett standard alapján. Ez a hallgatók korán és nemén kívül egyrészt a hallgatók nyelvtudásának felmérését, másrészt az ún. nyelvi biográfia felvázolását foglalja magába. Előbbi az e célra kifejlesztett tesztek segítségével sorolja a hallgatói teljesítményeket a Közös Európai Referenciakeretben meghatározott szintekre, ezek közül a B2-es, ill. a C1+ (C1 és afölötti) szintet elérő hallgatók szövegei kerülnek csak annotálásra. A nyelvi biográfia a hallgatók által beszélt nyelveket, ezek sorrendjére, az elsajátítás módjára, ill. a célnyelvi környezetben töltött időre vonatkozó információkat tartalmazza.

4.2 Az annotálási folyamat gyakorlati szempontból és hozzáférhetőség

A gyűjtött anyagok papíralapú és elektronikus archiválását, a kézírásos szövegek begépelését és az annotációt a hallgatói segéderők végzik a projekt mentoráló tagjainak felügyeletével és segítségével. Ez – az annotációt megelőző bevezetésen túl – a szövegek, ill. annotátumok ellenőrzését, a felmerülő kérdések megválaszolását és állandó visszacsatolást jelent. A célhipotéziseket anyanyelvi projekttagjaink lektorálják.

A hibatagek alkalmazásával kapcsolatos problémás eseteket projektmegbeszélések keretében oldjuk meg. Az elmúlt két évben így számos tag revideálásra, ill. megszüntetésre került, például a sok esetben nehezen elhatárolható „kongruencia a determinatívfrázisban" hibakategória, amely a németben gyakori szinkretizmusjelenségek miatt gyakran egybeesett más (melléknév)ragozási hibákkal. Más hibakategóriákat a ritka előfordulás miatt összevontunk, például a különböző (rémakiemelő ill. tagadó) partikulák szórendi hibáit egy egységes tag használatával tesszük kereshetővé. Új hibakategóriaként jelent meg például legutóbb a kötött bővítményeken kívül megjelenő szemantikai viszonyok hibás grammatikai kifejezését kódoló „SemRel" tag.

Az annotátumok feldolgozásának utolsó lépése az ANNIS-kompatibilis verzióba való konvertálás, mely által a szövegek az ANNIS által nyújtott keresési lehetőségek számára hozzáférhetővé válnak (l. 5. pont).

Az annotáció következő szakaszában az új fejlesztésű orig-, graph- és layout-sorok, tehát a szövegeket író megnyilatkozók általi változtatások, ill. az eredeti hallgatói szövegek tördelésének integrálása történik meg az első változatban publikált annotátumokba.

Jelenleg mintegy 63 szöveg (kb. huszonkétezer token) annotációja készült el, melyek a Dulko korpuszának 1.0 verzióját (Dulko-v1.0) fogják képezni. A Falko esszékorpusz (FalkoEssayL2v2.4) méretéhez képest (248 szöveg, 144.619 token) a Dulko mérete talán kicsinek tűnhet, viszont összehasonlítva a negyvennégy különböző anyanyelvvel rendelkezőktől gyűjtött szövegekből álló részkorpuszokkal a Dulko mérete tekintélyesnek mondható.[15] Az ANNIS-kompatibilis verziók egyelőre csak a helyi gépeken installált ANNIS programon keresztül hozzáférhetők, internetes publikálás az év végéig várható.

5 Alkalmazási lehetőségek

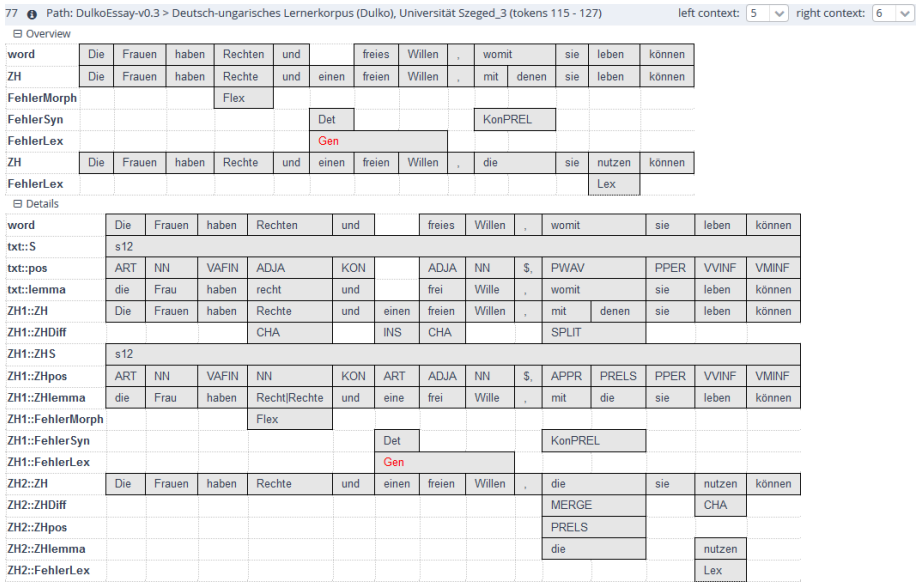
A Dulko annotációs módszerét a Szegedi Tudományegyetem mellett Németországban több egyetemen (Gießen, Lipcse, Marburg és Potsdam), valamint a Genti Egyetemen és Kínában (Hangcsou, Zhejiang University) is használják. Külön örvendetes, hogy a vietnami nyelvtanulói korpusz (az ún. „*Viet*namesisches *L*ernerkorpus", Vielko), amely két vietnami Egyetem (HANU, Hanoi University és a szintén Hanoiban működ-

[15] Egy példával szemléltetve, a Falko esszék szerzői között a negyedik legmagasabb aránnyal képviseltetik magukat a francia anyanyelvűek (vö. Reznicek és mtsai, 2012), az általuk írt szövegek mennyiségénél (17 szöveg, 10.756 token, vö. https://korpling.german.hu-berlin.de/falko-suche/) a Dulko esszéket tartalmazó részkorpusza (34 szöveg, 12.283 token) nagyobbnak bizonyul.

dő ULIS, University of Languages & International Studies) és két németországi egyetem (Lipcse és Gießen) kooperációjában fog létrejönni, ugyancsak a Dulko módszerével készül. A berlini Humboldt-Egyetemen, ahol a Falko korpuszt építették, szintén használják már az EXMARaLDA (Dulko) annotációs szoftvert.

A Dulko a fent említett, közeljövőben építendő korpuszokkal is összevethető lesz, viszont már jelenleg is számos alkalmazási lehetőséget rejt. A következőkben azt kívánjuk szemléltetni, hogy a korpusz milyen kérdésfelvetések tisztázásához járulhat hozzá, egyrészt önmagában, másrészt egyéb korpuszokkal való összehasonlításban.

A fent vázolt annotációs eljárásnak köszönhetően a korábbiakhoz képest egyszerűsödik a különböző hibatípusok elemzése. Ugyan a Falko nyelvtanulói korpuszokban[16] is kereshetők hibatípusok, viszont ez több esetben csak úgy oldható meg, hogy a nagy mennyiségű találati listákból manuálisan választjuk szét a valóban a keresett jelenséghez tartozó, helyesen felismert találatokat (true positives) a hibás, hamis pozitív találatoktól (false positives). A Dulkóban a hibatagek használatával egyszerűen lehívhatók a találatok az egyes hibatípusokhoz, például a Falkóban nehezen kereshető nyelvtani nem tévesztése hibatípusnál a „Gen” hiba-tag segítségével (vö. 8. ábra).



8. ábra Példa az ANNIS keresőfelületéről a „Gen” hibataggel való lekérdezésre kapott találatra

A lekérdezést a metaadatok segítségével is korlátozhatjuk. Így könnyebbé válik különböző tényezők, például a nyelvtanulók nyelvi szintje és az egyes hibatípusok aránya közötti összefüggések vizsgálata. A fenti hibakategória részkorpuszok szerinti megoszlását szemléltetve elmondható, hogy a szintfelmérő nyelvi teszteken a Közös Európai Referenciakeret alapján B2-es nyelvi szinten álló hallgatók a C1-es nyelvi szinten állókhoz képest sokkal több esetben tévesztik el a nyelvtani nemet. Az esszé

[16] https://korpling.german.hu-berlin.de/falko-suche/

részkorpusz aktuális változata (DulkoEssay-v0.3, össz. 12.283 token, 679 mondat) alapján a B2-es részkorpuszban 9871 tokenre (549 mondatra) jut 99 hiba, a C1-es részkorpusz esetén 18 hiba található 2412 token (130 mondat) mennyiségű szövegben.[17] A B2-es részkorpuszban tehát minden 100. tokenre jut egy ilyen hiba, míg a C1-es részkorpuszban csak minden 134.-re. Ugyanígy a célnyelv tanulásával töltött eddigi időtartam és egyéb más metaadatok is könnyen kombinálhatók a hibatagekkel. Így egyszerűbben vizsgálható az a kérdés is, hogy az egyes nyelvi szinteken található hibák mennyiben korrelálnak egymással és mennyiben függnek egyéb tényezőktől, pl. másik tanult idegen nyelv (angol) hatásától.

A magyar anyanyelvű nyelvtanulók esszéit más anyanyelvű nyelvtanulók esszéivel összevetve arról is képet kaphatunk, milyen szerkezetek fordulnak gyakrabban elő az előbbiekben.[18] A Falko nyelvtanulói korpusz esszé részkorpusza (falkoEssayL2v2.4) például mindössze öt találatot tartalmaz 248 szövegben (144.619 token, 6484 mondat) a *solch* (’az a’, ’olyan’) + főnév után álló vonatkozói mellékmondatokra, míg a jóval kisebb Dulko esszékorpuszban (DulkoEssay-v0.3) 12.283 tokenre jut három ilyen szerkezet. A magyar anyanyelvű nyelvtanulók esszéiben tehát minden 4049. tokenre jut egy ilyen szerkezet, míg a Falko korpuszban ez a szám 28.924, azaz 7-szer magasabb. A Falko korpusz német anyanyelvűek által írt esszéiben (falkoEssayL1v2.3) pedig a megfelelő keresőparancs nulla találathoz vezet.[19] Az amúgy a német anyanyelvűek szövegeiben is (bizonyos esetekben) használatos szerkezet tehát a magyar anyanyelvűek szövegeiben sokkal gyakrabban fordul elő (overuse). Ezzel kapcsolatban az egyik lényeges kérdés az interferencia jelensége, azaz hogy milyen esetekben befolyásolják magyar anyanyelvi szerkezetek a célnyelv használata során választott szerkezeteket. Az ilyen egyértelmű eredmények empirikus bizonyítékkal szolgálnak az interferencia hatására. Úgy véljük, a nyelvi hibák elemzésén túl a magyar nyelvtanulók köztes nyelvéről az ilyen kvantitatív elemzések (overuse és underuse) is sokat elárulnak.

Különösen fontos, hogy a nyelvtanítás során tisztában legyünk azzal, milyen jellegű hibák jellemzők a magyar anyanyelvű nyelvtanulók nyelvhasználatára. Ilyen például az *auch* (’is’) fókuszpartikula (rémakiemelő partikula) hibás használata. A magyar anyanyelvű nyelvtanulók annotált esszészövegeiben ez minden hatodik esetben, azaz 93-ból 15 nyolc esetben, – szintén a magyar nyelv hatására – nem a fókusz előtt, hanem azután áll. Az ilyen szórendi hiba a Falko esszékorpuszában (falkoEssayL2v2.4) több mint háromszor ritkább (853 esetből kevesebb mint 45 szórendi hiba tartozik ide, mivel a lemma="auch" & ZH2Diff="MOVS" & #1_=_#2 keresőparancs hamis pozitív találatokat is eredményez). További empirikus vizsgálatokkal a

[17] A keresőparancs a B2-es részkorpusz esetén FehlerLex="Gen" & meta::learner_level_CEFR_conversion="B2", a C1-es részkorpusz esetén pedig FehlerLex="Gen" & meta::learner_level_CEFR_conversion="C1".

[18] A Falko korpusz mellett a Kobalt-DaF projekt (vö. Zinsmeister és mtsai, 2012) nyelvi adataival való összehasonlítás is lehetséges, melyek a Falko korpuszban alulreprezentált kínai vagy svéd anyanyelvű nyelvtanulók nyelvi produktumaival való összehasonlításhoz kínálnak megfelelő alapot.

[19] A Falko esszékorpuszok esetében a lemmára, szófajokra, írásjelre, valamint ezek sorrendjére vonatkozó keresőparancs: lemma="solch" & pos="NN" & lemma="," & pos="PRELS" & #1.#2 & #2.#3 & #3.#4. A Dulko esszékorpusz esetében ez csak annyiban tér el, hogy az újabb lemmatizáló a „solche” szótári alakkal dolgozik a régebbi „solch” helyett.

magyar anyanyelvű nyelvtanulók nyelvhasználatára jellemző hibák azonosítása azért is fontos, mert bizonyos hibák csak nehezen leküzdhetők, ezek sok esetben fosszilizálódhatnak, a haladó szintű nyelvtanulók nyelvhasználatában is megmaradhatnak. Ha ezeknek az általános és középiskolai nyelvtanítás során nem szentelnek kellő figyelmet, a hallgatóknak az egyetemi nyelvoktatás során kell a helyes célnyelvi használatot elsajátítaniuk. A nyelvtanulói korpuszok elemzése tehát az egyetemi nyelvtanítás optimalizálásához is hozzájárulhat. A még haladó szintű nyelvtanulók által is elkövetett (de reményeink szerint az egyetemi évek végére leküzdött) tipikus hibák ismerete az általános és középiskolai nyelvtanárok számára is fontos, hogy diákjaikat minél eredményesebben tudják hozzásegíteni a célnyelv magas szintű használatához.

6 Összegzés

A cikkben bemutattuk a Dulko német-magyar nyelvtanulói korpuszt, amelyben magyar anyanyelvű, németül tanuló germanisztika szakos hallgatók nyelvi adatait annotáljuk és tesszük elektronikusan kutathatóvá. Az annotáció során a Falko korpusz módszerét követjük (pl. szófajok és lemmák automatikus annotációja, metaadatok, célhipotézisek), viszont számos újdonságot vezettünk be (pl. hibatagek használata, fordításszövegek annotációja, a kézirat önjavításainak jelölése). Az EXMARaLDA programra épülő nyílt forráskódú program megkönnyíti a nyers szövegtől az annotált szövegig tartó folyamat technikai megvalósítását és az annotátumok más korpuszokkal való kompatibilitását. A programunkat már jelenleg is több korpuszprojektben használják, és örömmel várjuk a további együttműködéseket.

A Dulko reményeink szerint nemcsak a szabadon kereshető, magyar anyanyelvű németül tanulók nyelvi adataival járul hozzá a nyelvtanulók köztes nyelvének vizsgálatához, hanem a nyílt forráskódú programmal és a javított annotációs eljárással más nyelvtanulói korpuszok építéséhez is mintát és hathatós segítséget nyújt.

Hivatkozások

Brdar-Szabó, R.: Nutzen und Grenzen der kontrastiven Analyse für Deutsch als Fremd- und Zweitsprache. In: Krumm, H.-J., Fandrych, C., Hufeisen, B., Riemer, C. (szerk.) Deutsch als Fremd- und Zweitsprache. Ein internationales Handbuch (Sprach- und Kommunikationswissenschaft 352). pp. 518–531. de Gruyter, Berlin, New York (2010a)

Brdar-Szabó, R.: Kontrastive Analyse Ungarisch-Deutsch. In: Krumm, H.-J., Fandrych, C., Hufeisen, B., Riemer, C. (szerk.) Deutsch als Fremd- und Zweitsprache. Ein internationales Handbuch (Sprach- und Kommunikationswissenschaft 352). pp. 732–737. de Gruyter, Berlin, New York (2010b)

Fekete, O.: Forschungsmethodologische Aspekte zur Kasusverwendung bei ungarischen DaF-Lernenden. In: Böttger, L., Masát, A. (szerk.) Jahrbuch der ungarischen Germanistik 2008. pp. 163–183. Gondolat, Budapest, Bonn (2009)

Fekete, O.: Komplexität und Grammatikalität in der Lernersprache : eine Längsschnittstudie zur Entwicklung von Deutschkenntnissen ungarischer Muttersprachler. Waxmann, Münster, New York (2016)

Granger, S., Paquot, M.: Core metadata for learner corpora. Draft 1.0. Kézirat. Louvain-la-Neuve (2017)

Gunkel, L., Murelli, A., Schlotthauer, S., Wiese, B., Zifonun, G.: Grammatik des Deutschen im europäischen Vergleich. Das Nominal. Unter Mitarbeit von C. Günther und U. Hoberg. 2 Bände. (Schriften des IDS 14) de Gruyter, Berlin, Boston (2017)

Hirschmann, H., Nolda, A.: Dulko – auf dem Weg zu einem deutsch-ungarischen Lernerkorpus. In: Eichinger, L., Plewnia, A. (szerk.) Neues vom heutigen Deutsch: Empirisch – methodisch – theoretisch (Institut für Deutsche Sprache, Jahrbuch 2018). pp. 339–342. de Gruyter, Berlin (2019)

Institut für Deutsche Sprache: Deutsches Referenzkorpus – DeReKo. Archiv der Korpora geschriebener Gegenwartssprache. Institut für Deutsche Sprache, Mannheim (2004ff.) [http://www1.ids-mannheim.de/kl/projekte/korpora]

Juhász, J.: Probleme der Interferenz. Akadémiai Kiadó, Budapest, München (1970)

Krause, T., Zeldes, A.: ANNIS3: A new architecture for generic corpus query and visualization. Digital Scholarship in the Humanities 31, 118–139 (2016) [http://dsh.oxfordjournals.org/content/31/1/118]

NAT - Nemzeti alaptanterv - Hatály: 2018.I.1. - Magyar joganyagok - a Nemzeti alaptanterv kiadásáról, bevezetéséről és alkalmazásáról. [https://net.jogtar.hu/getpdf?docid=a1200110.kor&targetdate=&printTitle=]

Nolda, A.: Annotation von Lernerdaten mit EXMARaLDA (Dulko). Kézirat. Berlin (2019) [https://andreas.nolda.org/publications/nolda_2019_annotation_lernerdaten.pdf]

Pilarský, J. (szerk.): Deutsch-ungarische kontrastive Grammatik. 2. kiadás. Egyetemi Kiadó, Debrecen (2018)

Reznicek, M., Lüdeling, A., Krummes, C., Schwantuschke, F., Walter, M., Schmidt, K., Hirschmann, H., Andreas, T.: Das Falko-Handbuch. Korpusaufbau und Annotationen Version 2.01. Humboldt-Universität zu Berlin, Institut für deutsche Sprache und Linguistik – Korpuslinguistik, Berlin (2012) [https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/falko/FalkoHandbuchV2/]

Reznicek, M. Lüdeling, A., Hirschmann, H.: Competing target hypotheses in the Falko corpus: A flexible multi-layer corpus architecture. In: Díaz-Negrillo, A., Ballier, N., Thompson, P. (szerk.) Automatic treatment and analysis of learner corpus data (Studies in Corpus Linguistics 59). pp. 101–123. John Benjamins, Amsterdam (2013)

Schiller, A., Teufel, S., Stöckert, C., Thielen, C.: Guidelines für das Tagging deutscher Textcorpora mit STTS (kleines und großes Tagset). Kézirat. Universität Stuttgart, Institut für maschinelle Sprachverarbeitung und Universität Tübingen, Seminar für Sprachwissenschaft (1999) [http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf]

Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: Jones, D. B., Somers, H. L. (szerk.) New Methods in Language Processing, pp. 154–164. Routledge, London (1997)

Schmidt, T.: EXMARaLDA – ein Modellierungs- und Visualisierungsverfahren für die computergestützte Transkription gesprochener Sprache. In: Buchberger, E. (szerk.) Proceedings of Konvens 2004, Schriftenreihe der Österreichischen Gesellschaft für Artificial Intelligence 5, Österreichische Gesellschaft für Artificial Intelligence, Wien (2004) [https://www.exmaralda.org/files/Konvens_Paper.pdf]

Selinker, L.: Interlanguage. International Review of Applied Linguistics. Language Teaching 10, 209–231 (1972)

Váradi, T.: The Hungarian National Corpus. In: Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002). pp. 385–389. European Language Resources Association, Las Palmas de Gran Canaria (2002)

Walter, M., Grommes, P.: Fortgeschrittene Lernervarietäten. Korpuslinguistik und Zweitspracherwerbsforschung. Niemeyer, Tübingen (2008)

Zinsmeister, H., Reznicek, M., Brede, J. R., Rosén, C., Skiba, D.: Das Wissenschaftliche Netzwerk „Kobalt-DaF“. Korpusbasierte Analyse von Lernertexten für Deutsch als Fremdsprache. Zeitschrift für Germanistische Linguistik 40(3), 457–458 (2012)