Egy magyar nyelvű kérdezőrendszer

Novák Attila[1,2], Laki László János[1,2], Novák Borbála[1,2], Dömötör Andrea[2,3], Ligeti-Nagy Noémi[2,3], Kalivoda Ágnes[2,3]

[1]Pázmány Péter Katolikus Egyetem, Információs Technológiai és Bionikai Kar
[2]MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport
Budapest, Práter u. 50/a.
[3]Pázmány Péter Katolikus Egyetem, Bölcsészet- és Társadalomtudományi Kar
2087 Piliscsaba, Egyetem u. 1.
{vezetéknév.keresztnév}@itk.ppke.hu

Kivonat Cikkünkben egy folyamatban lévő kutatásról számolunk be, amelynek keretében olyan korpuszannotációt hozunk létre, amely alkalmas a feldolgozott szöveggel kapcsolatban releváns kérdéseket megfogalmazni képes elemzőrendszer betanítására. A cikk terjedelmi korlátai által biztosított határok között röviden bemutatjuk a kutatás célkitűzéseit, a kiindulásul használt magyar UD korpusz javításával, a tematikus vonzatkeret-lexikon létrehozásával, a szabad határozók osztályozásával és a vonzatkeretek korpusz-előfordulásokra való illesztésével kapcsolatos eddigi erőfeszítéseinket.

1. Bevezetés

Az utóbbi években a korábbiakat meghaladó színvonalú eredményeket nyújtó módszernek bizonyult a neurális mélytanuló hálózatokon alapuló olyan ún. end-to-end rendszerek alkalmazása, amelyek semmilyen grammatikai elemzést nem tartalmaznak, ezért kétségek merültek fel azzal kapcsolatban, hogy van-e értelme egyáltalán grammatikai elemzéssel foglalkozni. Ugyanakkor az end-to-end rendszerek betanítása rendszerint hatalmas mennyiségű tanítóanyagot igényel, amely a legtöbb nyelven nem áll rendelkezésre. Ezért azt gondoljuk, hogy továbbra is lehet értelme egy grammatikai elemzést előállító rendszernek, amennyiben az elemzés eredménye közvetlenül felhasználható olyan feladatok végrehajtásához, amely a hétköznapi felhasználók számára is relevanciával bír.

Nem lehetünk elégedettek azonban egy olyan elemzéssel, amely olyan teljesen absztrakt kategóriákkal dolgozik, amelyeket nem lehet egyértelműen olyan fogalmakra lefordítani, ami hétköznapi emberek számára is érthető módon összefüggésbe hozható azzal, hogy mit jelent az adott szöveg. A szövegértés lényeges eleme, hogy képesek vagyunk értelmes kérdéseket feltenni az adott szöveggel kapcsolatban, és ez a képességünk szorosan összefügg azzal, hogy képesek vagyunk kérdésekre válaszolni is. Olyan elemzőrendszer létrehozását tűztük ki tehát célul, amely ténylegesen alkalmas arra, hogy releváns kérdéseket tegyen fel azzal a szöveggel kapcsolatban, amit feldolgoz. Ehhez számtalan olyan distinkcióra van

szükség, amiknek az eddigi elemzőrendszerekben nem láttuk nyomát. Jelen cikk ennek a munkálatnak az első fázisát mutatja be, amelyben célunk egy olyan annotált korpusz létrehozása, ahol az annotáció tartalmazza mindazokat a jegyeket, amik az adott szöveggel kapcsolatos kérdések generálásához szükségesek.

2. A hagyományos elemzés hiányosságai

Mivel olyan rendszer létrehozása a célunk, amely értelmes kérdéseket tud feltenni, ezért úgy döntöttünk, hogy az annotációban használt megkülönböztetések létjogosultságát alapvetően az határozza meg, hogy az adott konstrukcióval kapcsolatban milyen kérdéseket lehet föltenni. A **névszói csoportokra** vonatkozó kérdéseknél például alapvető a *ki?/mi?* megkülönböztetés, ezért a rendszernek pontosan meg kell tudnia különböztetni a személyeket a dolgoktól. Ugyanakkor a csoportokra vagy szervezetekre attól függően kérdezünk *ki?*-vel vagy *mi?*-vel, hogy milyen szerepet töltenek be az adott mondatban. Egy bank például nyelvileg személyként viselkedik, ha számlalevelet küld, de dologként, ha felszámolják. Az állítmányként használt névszói csoportokkal kapcsolatos kérdések generálásához pedig egy még ennél is jóval részletesebb osztályozásra van szükség. A *Lajos orvos* mondattal kapcsolatban a *Lajos ki?* kérdés nem túl kifinomult, a *Lajosnak mi a foglalkozása?* jóval pontosabban kérdez rá arra, ami a mondatban az állítás. A fogalmak foglalkozásként, állatként, eszközként, viselkedésként, stb. való osztályozása a névszói csoportok nem predikatív előfordulásaival kapcsolatban is jóval specifikusabb kérdések megfogalmazására ad lehetőséget: pl. *Milyen állatot láttál a kertben?* szemben a *Mit láttál a kertben?* kérdéssel. Különösen lényeges ez a koordinált frázisok esetében, ahol az egyik koordinált összetevőre csak akkor tudunk a kérdezett számára is azonosítható módon rákérdezni, ha a kérdés eléggé specifikus.

A **határozókkal** kapcsolatos kérdések megfogalmazásához is nagyságrendekkel részletesebb osztályozásra van szükség még a legminimálisabb szinten is, mint amivel a létező hagyományos elemzőrendszerek szolgálni tudnak. Az inesszívusz ragos szóalakok például rengeteg különböző funkciót tölthetnek be, és így különböző kérdés tartozik hozzájuk:

- *szeptemberben*: mikor?,
- *Londonban*: hol?,
- *fájdalmában*: mitől?,
- *magában (bízik)*: kiben?,
- *hármasban*: hányan?,
- *elemében (van)*: erre nem kérdezünk,
- stb.

Az **állítmánnyal kapcsolatos kérdések** megfogalmazása nemcsak a névszói állítmányok, hanem az igék esetében is olyan ismereteket igényel, amelyekkel a létező grammatikai leírások nem tudnak szolgálni. Hogy hogyan kérdezzünk az állítmányra annak egy adott vonzatát horgonyként használva, az attól függ, hogy az adott vonzat milyen tematikus szerepet tölt be az igei vonzatkeretben. A *Mit*

csinált Jancsi Ferivel? adekvát kérdés, ha *Jancsi* ágens és *Feri* páciens. Ugyanebben a helyzetben a *Mi történt Ferivel?* és a *Mit csinált Jancsi?* ugyanígy helyes kérdés.

A vonzatkeretek argumentumhelyeinek tematikus osztályozására szükség van az **oblikvuszi vonzatok és a szemantikailag tartalmas viszonyok** megkülönböztetéséhez is. Például: *bízik valamiben* szemben azzal, hogy *van valahol.*

Szükség van ugyanakkor a félig kompozicionális, illetve **idiomatikus szerkezetek** kompozicionális szerkezetektől való megkülönböztetésére is. Vicc lesz belőle, ha az előbbiekre kérdezünk:

- *Mit hozott Édesapám?*
- *Döntést.*

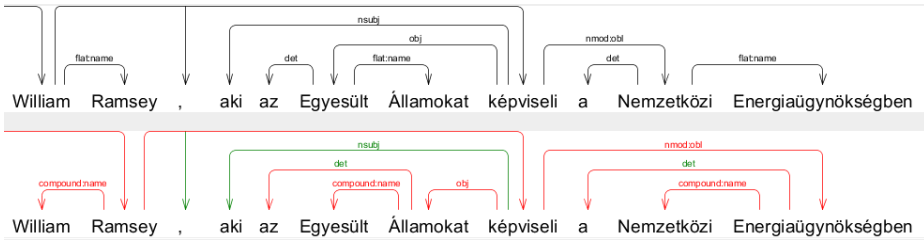
3. A korpusz

Kiindulási anyagként a Universal Dependencies (UD) korpusz [1] 1800 mondatból (42000 token) álló magyar alkorpuszát választottuk, hogy nemzetközi szinten is értelmezhető kontextusba helyezzük az általunk javasolt annotációs sémát. Az UD korpusz nagyjából egységes elvek és kategóriák felhasználásával sok nyelv szövegeire tartalmaz morfoszintaktikai és szintaktikai függőségi elemzést. Eredeti tervünk az volt, hogy a magyar UD korpuszban szereplő annotációt pusztán kiegészítjük, illetve finomítjuk a kérdések megfogalmazásához szükséges információkkal. Kiderült azonban, hogy a magyar alkorpuszban szereplő annotáció sok szempontból nem felel meg az érvényes UD specifikációnak, illetve sok véletlenszerű annotációs hibát tartalmaz, ezért a feladat része lett ezeknek a hibáknak a javítása.

Az UD 2.0 specifikációja[1] szerint a **több szavas kifejezések** belső szerkezetét `flat, fixed` vagy `compound` függőségi viszonyok alkalmazásával kell leírni. A `fixed` viszonyt kizárólag a teljesen megkövült funkciószó-szerű több szavas kifejezések leírására használják. A `compound` viszonyt kell használni azoknak a szerkezeteknek a leírására, amelyeknek van feje. Számos nyelvben, például az angolban, a több szavas neveket általában lapos endocentrikus szerkezeteknek tekintik, ezért a `flat` viszony használatát javasolják ezeknek a neveknek a leírására. Az UD 2.0 annotációs specifikációja azonban kategorikusan kizárja ennek a típusú elemzésnek a használatát azokban az esetekben, amikor a névnek szabályos szintaktikai szerkezete van (pl. címek, illetve az intézménynevek nagy része), ahol a szokásos szintaktikai viszonyok használatát írja elő, illetve az endocentrikus szerkezetű nevek esetében, ahol a `compound` viszonyt, illetve ennek valamelyik alváltozatát kell használni. A magyar névszói szerkezetek mindig endocentrikus szerkezetek, amelyek rendszerint jobb fejűek, ezért a nem szabályos szerkezetű és kompozicionális jelentésű nevek esetében a magyarban mindig a `compound` viszonyt kell használni. Ez biztosítja például, hogy a mindig a szerkezet fején megjelenő esetragok közvetlenül elérhetőek legyenek. Ezért a feldolgozás egyik lépéseként a korpuszban hibásan `flat` szerkezetűnek annotált több szavas

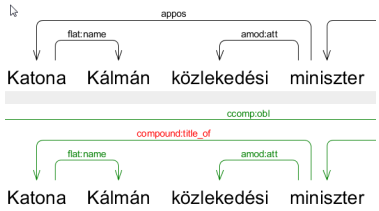
[1] `http://universaldependencies.org/guidelines.html`

neveket automatikusan `compound` szerkezetekké konvertáltuk. Egyelőre elmaradt a teljesen szabályos szerkezetű nevek konverziója, hiszen ezeket kézzel kellene kiválogatni és újraannotálni (1. ábra).



1. ábra. A nevek annotációjának javítása

A tévesen jobb fejű appozitív szerkezetként annotált *Katona Kálmán közlekedési minisztert*-típusú szerkezetekben[2] az UD 2.0 specifikációval kompatibilis módon `compound:title_of` viszonyt vettünk fel a név és a foglalkozás/funkció között (2. ábra).



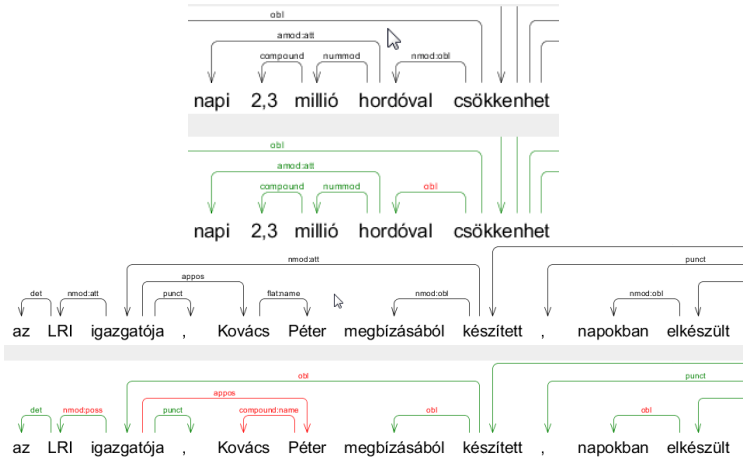
2. ábra. Név és foglalkozás javítása

Az alanyon, tárgyon és részeshatározón kívüli **névszói vonzatok** jelölésére az UD 2.0 specifikáció az `obl` relációt írja elő akkor is, ha a fej nem ige. Ez a korpuszban sokszor igei fejek esetén sem így szerepelt. Igei és igenévi fejek esetében tudtuk automatikusan javítani ezeket a annotációkat – amennyire lehetett (3. ábra).

Az **igekötős igék** lemmája nem tartalmazta az igekötőt azokban az esetekben, ahol az ige és az igekötő nem volt egybeírva. A vonzatok tematikus szerepeit tartalmazó lexikonban szereplő annotáció korpuszra vetítéséhez szükséges volt, hogy az igekötő része legyen ezekben az esetekben is a lemmának. Ezért ezt a hibát is kijavítottuk.

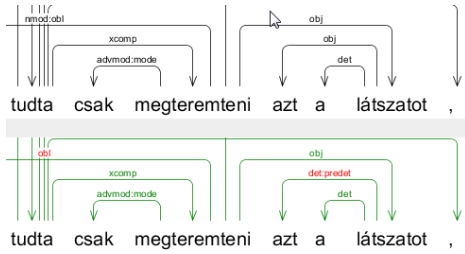
Az *azt a kutyát*-típusú **egyeztetett predeterminánst** tartalmazó szerkezetekben a mutató névmás sokszor tévesen ugyanazzal a címkével volt a névszói

[2] Az appozitív szerkezetekben esetegyeztetés van a két elem között, itt erről nincs szó.



3. ábra. Az `obl` reláció javítása igei és igenévi fejeknél – a második esetben az *igazgatója* szó rossz fejhez volt kötve, így az annotáció továbbra is hibás maradt

csoport fejéhez csatolva, mint amilyen funkciót a teljes NP betölt. Ezeket és az összes ilyen predeterminánsz címkéjét `det:predet` címkére cseréltük (4. ábra).

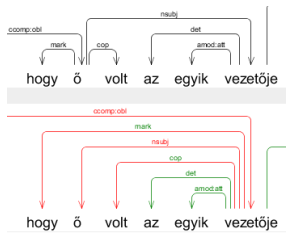


4. ábra. Hibásan annotált mutató névmás javítása

A **birtokos szerkezetekben** a birtokos annotációját `nmod:att`-ról `nmod:poss`-ra javítottuk (l. a 3. ábrán alul).

A **névutókat** egységesen `case` viszonnyal kapcsoltuk a névszói csoport fejéhez.

A harmadik személyű **névszói állítmányt** tartalmazó tagmondatok annotációjában az alany és az állítmány sok esetben meg volt cserélve, mert a fókuszt összetévesztették az állítmánnyal. A korábbiakban leírt javításokat programozottan végeztük. Ezeket a szerkezeteket azonban kénytelenek voltunk félig manuális módszerrel javítani: kézzel jelöltük meg azokat a mondatokat, ahol aztán az alany és állítmány annotációját programozottan javítottuk (5. ábra).



5. ábra. Felcserélt alany és állítmány javítása

4. Vonzatkeret-adatbázis

A magyar UD korpuszban szereplő összes ige és igenév tövét kigyűjtöttük, és a [2] cikkben leírt elemzett korpuszból épített szóbeágyazási modellben szereplő vektorreprezentációjuk alapján klasztereztük a [3,4] cikkekben leírt módon. A hasonló disztribúciójú (és vonzatkeretű) igék így egy-egy klaszterben gyűltek össze. A listát kiegészítettük minden egyes igéhez a *Magyar igei szerkezetek: a leggyakoribb vonzatok és szókapcsolatok szótára* magyar vonzatkeretszótárban [5] szereplő az adott igéhez tartozó leírással. Ezt a kiinduló reprezentációt ihletforrásként használva kézzel készítettük el az egyes igék lehetséges vonzatkereteinek leírását, amelyben az egyes vonzatok tematikus szerepe, formai jegyei (esetrag, névutó, birtokos végződés, stb.), esetleges opcionalitása, és a rájuk vonatkozó esetleges lexikai/szemantikai megszorítások szerepelnek.

Az igei vonzatkeretek leírásánál a fő szempont az volt, hogy minél több olyan információt adjunk meg, amelyek segítségével a lehető legjobb, legpontosabb kérdések tehetők fel. Éppen ezért a vonzatkeret-leírásokban használt tematikusszerep-készlet, bár azokból indul ki, legfőképpen abban követi az általánosan ismert tematikusszerep-hierarchiákat, hogy részleteiben éppen úgy különbözik azoktól, mint azok egymástól. Az igék leírása igyekszik minden lehetséges jelentést (vonzatkeretet) lefedni. Az, hogy a hasonló jelentésű és vonzatkeretű igék eleve összegyűjtve szerepeltek az adatbázisban, lehetővé tette, hogy több ige közös vonzatkeretét csak egyszer kelljen megadni, és az egy csoportba tartozó igék automatikusan öröklik az így megadott vonzatkereteket. Emellett természetesen az egyes igéknek egyéb csak rájuk jellemző vonzatkeretei is lehetnek, amelyek hozzáadódnak az igecsoportra jellemző vonzatkeretekhez.

Az igéhez tartozó vonzatokat és opcionális bővítményeket szerepek szerint vagy lexikálisan adtuk meg, minden esetben a szükséges esetragokkal vagy névutókkal kiegészítve. A szerepek meghatározása aszerint történt, hogy milyen kérdés tehető fel az adott mondatrészre, illetve a mondatrésszel az igére. Például az ágens kérdése a *mit csinál?*, a páciensé pedig a *mi történik vele?*.

Bizonyos szerepek egyúttal egyfajta szemantikai kategóriát is jelölnek, ilyen például a kontent (`CONT`), amely valamilyen kifejthető tartalomra, információra utal, vagy a cselekvést - elsősorban főnévi igenevet - jelölő `ACT`. Azok a vonzatok, amelyeknek nincs meghatározott tematikus szerepe, nem igazán lehet őket

horgonyként használva az állítmányra kérdezni, a semlegesnek tekinthető téma (TH) szerepet kapták.

Az idiomatikus vagy félig kompozicionális igei szerkezetek vonzatait nem szerep szerint, hanem lexikálisan, a szó vagy lexikális kategória megadásával jelöltük. Ahol indokolt volt, ezek a szerkezetek - önálló egységként értelmezve őket - külön vonzatkeret-leírást kaptak. Így például a *sor kerül* leírását nem a *kerül* igénél adtuk meg, hanem a kifejezéshez mint külön tételhez rendeltünk saját vonzatkeretet.

Az igék és igei szerkezetek vonzataihoz rendelt tematikus szerepeket az 1. táblázat foglalja össze.

A táblázatban felsoroltakon kívül külön jelet kaptak a mozgó szereplők, így például a mozgó ágens jele az AGMV lett. A leírásoknál alapvetően abból indultunk ki, hogy egy igéhez nem tartozhat több azonos szerepű vonzat, ahol erre mégis szükség volt, ott a co- prefixszel jelöltük a társszereplőt, így például a *sétál valakivel* jelölése AG_coAG-vAl.

Az előzőek szerint leírt vonzatkeretek speciális szemantikai besorolást is kaphattak, melyek segítségével a kérdések tovább finomíthatók. Az ehhez felhasznált kategóriák a következők:

- biotünet (pl.: *izzad*)
- érzékelés (pl.: *lát*)
- érzelem (pl.: *örül*)
- feltétel (pl.: *múlik valami valamin*)
- hang (pl.: *zeng*)
- helyzet (pl.: *szorít az idő*)
- kezdet (pl. *megalakul*)
- kognitív (pl.: *egyetért*)
- kommunikáció (pl. *értesít*)
- matematikai (pl. *összead*)
- nemverbális kommunikáció (pl. *int*)
- önjáró (a mozgáshoz nem használ eszközt, pl. *lép*)
- pénzügyi (pl. *utal*)
- pusztítás (pl. *szabotál*)
- pusztulás (pl. *kiszárad*)
- természeti (pl. *esik az eső*)
- transzformáció (pl. *felgyorsul*)
- viselkedés (pl. *kikezd valakivel*)
- viszony (pl. *támogat*)

Végül, a vonzatkeretekhez tartozik egy polaritásérték is, ami azt jelzi, hogy az adott esemény a páciensre vagy experiensre nézve pozitív, negatív vagy semleges.

A 6. ábrán a *sodródik, hull, zuhan, esik* igék leírása látható a vonzatkeret-adatbázisban. A részlet elején szereplő PATMV és PATMV_PATH keret, illetve a @.-tal jelölt semleges polaritás mindegyik igére vonatkozik, az egyes igéknél +-szal jelölt keretek ezekhez adódnak hozzá. A leírásokban szereplő kerek zárójelek az opcionalitást, a szögletes zárójelek pedig a valamilyen szemantikai kategóriát meghatározó példák felsorolását tartalmazzák.

```
PATMV PATMV_PATH
@.
sodródik[IGE] +CHAR_ár-vAl +PAT_TH-bA
hull[IGE] +AG_térd-rA_(CHAR~előtt) +hó +PAT~[haj|könny]-A +PAT@-pusztulás
zuhan[IGE] +EXP_álom-bA@.biotünet
esik[IGE] +[eső|hó]@.nature +szó_CONT-rŐl@.komm +PAT_[áldozat|fogoly]-Ul_TH-nAk
  +AGPAT_[késedelem|hiba|túlzás]-bA_(TH-bAn/-vAl~kapcsolatban/-t~illetően) +CHAR_tartomány-bA
  +csorba_PAT~[jóhír|hírnév|becsület|...]-A-n +PAT_fogság-bA +EXP_pánik-bA_(ST-tŐl)@-érzelem
  +PAT_has-rA_(CAU-tŐl) +választás_CHAR-rA +PAT-nAk_baj-A +EXP-nAk_nehéz-A-rA_ST
  +AGPAT_gondolkodó-bA_(TH-rŐl/-vAl~kapcsolatban/-t~illetően) +EXP_[kísértés|révület]-bA_(ST-tŐl)
  +szégyen_PAT-vAl +PAT_teher-bA_(TH-tŐl)
```

6. ábra. Részlet a vonzatkeret-adatbázisból

Jel	Név	Kérdés az igére	Példa
AG	ágens	Mit csinál AG?	**Feri** felmászott a fára.
CHAR	jellemzett	Mi jellemző CHAR-ra?	A **szaktudás** előnyt jelent.
ATTR	attribútum	–	A szaktudás **előnyt** jelent.
EXP	experiens	Mit érez/érzékel EXP?	**Feri** szereti Julit. **Feri** meglátott egy fecskét.
PAT	páciens	Mi történt PAT-tal?	Feri megcsókolta **Julit**.
PATDST	páciens-célpont	Mi történt PATDST-vel? Hova került PAT?	A gyerek a **falra** kente a főzeléket.
TH	téma	–	Feri a **megérzéseire** hagyatkozik.
ST	stimulus	Milyen érzést kelt ST (EXP-ben)? Milyen hatást vált ki ST (EXP-ben)?	Feri szereti **Julit**. Feri megijedt az **árnyékától**.
CONT	információtartalom	–	Feri ismertette a **tervet** Lajossal.
REC	recipiens	–	Feri ismertette a tervet **Lajossal**. Juli kapott egy **levelet**.
RES	eredmény	Honnan lett RES?	Feri hajtogatott egy **repülőt**.
INS	eszköz	Mire használta AG INS-t?	Feri **rollerrel** jár dolgozni.
CAU	okozó	Mit okozott CAU? Mi lett CAU következménye?	Feri **baleset miatt** késett.
MOT	cél	–	Feri **mérnöknek** tanul.
LOC	hely	Mi történt LOC-ban/-n...?	Feri megcsókolta Julit a **moziban**.
SRC	forrás, kiindulópont	–	Feri kijött a **szobából**. Feri megkérdezte **Lajostól** az állást. Juli kapott egy levelet **Feritől**.
DST	célpont	–	Feri bement a **szobába**.
HOW	mód	–	Feri **ügyesen** felmászott a fára.
ASPECT	tekintet	–	Feri nem áll rosszul **anyagilag**.
ACT	cselekvés	–	Feri rollerrel jár **dolgozni**.

1. táblázat. A vonzatkeretek leírásához használt tematikus szerepek

A vonzatkeret-adatbázis a cikk írásakor 1574 ige 5394 különböző vonzatkeretét tartalmazza valamennyi vonzat tematikus szerepével együtt. Bár az opcionális vonzatokat tartalmazó keretek (pl. `olvas AG_(HOW)_(PAT-t)_(REC-nAk)_(TH-rŐl)_(LOC-bAn)`) a gyakorlatban számtalan látszólag különböző szerkezetként jelennek meg, az előbbi számot úgy kaptuk, hogy az opcionális vonzatokat és az esetleges tematikusszerep-variánsokat tartalmazó kereteket egy keretnek számoltuk.

5. A szabad határozók szerepének azonosítása

Fontos feladat a mondatban hagyományosan „szabad határozóként” emlegetett esetragos névszók szerepének pontosabb meghatározása is. Ha ugyanis az esetragok felől közelítjük meg a kérdést, első közelítésben azt mondhatnánk, hogy az inesszívuszi esetragot magán viselő névszó valamilyen helyviszonyt jelöl, és a *Hol?* kérdésre válaszol. A *Hol diplomázott Fanni?* kérdésre azonban vicc az a válasz, hogy *Álmában.* Nyilvánvaló, hogy az irányhármasságot kifejező, *Hol?*, *Hová?* és *Honnan?* kérdésre válaszoló 3-3-3 esetrag (inesszívuszi *-bAn*, adesszívuszi *-nÁl*, szuperesszívuszi *-On*; illatívuszi *-bA*, allatívuszi *-hOz* és szublatívuszi *-rA*; illetve az elatívuszi *-bÓl*, ablatívuszi *-tÓl* és delatívuszi *-rÓl*) nem minden esetben a hely, a forrás vagy a cél megjelölésére szolgál. Ezért a szótő kategóriájának és az esetragnak a kombinációjával határoztuk meg az egyes szóalakok szerepét.

A feladat megfogalmazható úgy is, hogy határozókat csoportosítunk: vannak természetesen helyhatározók, mint a *sarkon*, vagy a *bankban*, vannak időhatározók, mint a *télen*, *decemberben*. De persze találkozunk időtartam-határozókkal is, mint az *Öt hónapra béreltük a lakást.* mondatban a *hónapra*. Összesen 31 főkategóriát állapítottunk meg, amelyek közül némelyik több alkategóriára osztható. Alkategóriákkal együtt 51 csoportba osztottuk a korpuszban található, helyhatározói esetraggal szabad bővítményi státuszban álló szótöveket. Az alkategóriák szemléltetésére a valóban helyhatározást szolgáló, *loc* kategóriába sorolt töveket hozzuk.

kategória	példa	bAn	nÁl	On
loc all	*szekrény*	hol	hol	hol
loc ade	*Microsoft*	miben	hol	min
loc ine	*állam*	hol	minél	min
loc sup	*címoldal*	miben	minél	hol
loc ine-sup	*könyv*	hol	minél	hol
loc city-ine	*Altenkirchen*	hol	hol	melyik városon
loc city-sup	*Kaposvár*	melyik városban	hol	hol
loc country	*Afganisztán*	hol	hol	melyik országon

A táblázat azt mutatja, hogy az adott főkategória (jelen esetben a *loc*) adott alkategóriájába (*all*, *ine*, *city-sup* stb.) tartozó szótövek adott esetrag (*-bAn*, *-nÁl*, *-On*) esetén milyen kérdést vonnak maguk után - azaz pontosan milyen szerepük van az adott mondatban. Az irányhármasság körébe tartozó esetragos határozók osztályozásával kapcsolatos eredményeinkről részletesebben is beszámolunk a jelen kötetben megjelent másik tanulmányunkban [6].

6. Félig kompozicionális szerkezetek automatikus azonosítása

Az idiomatikus és félig kompozicionális szerkezetek azonosításakor is azt a célt tartottuk szem előtt, hogy egy kifejezés az arra vonatkozó releváns kérdés megfogalmazása szempontjából hogyan viselkedik. A fent említett *döntést hoz* esetén nem jó kérdés a *Mit hoz?*, a *szóba hoz* esetében a *Hova/mibe hoz?*.

Az ilyen kifejezések összegyűjtésére saját algoritmust dolgoztunk ki. Ehhez először egy 644,5 millió token méretű angol-magyar párhuzamos korpusz [7] 7-gramjaira vonatkozó szómegfeleltetési (alignment) modellt hoztunk létre fast align programmal [8] úgy, hogy minden szót egy vagy két token reprezentált mind a magyar, mind az angol oldalon: a szótő a fő szófajcímkével és az esetleges egyéb morfoszintaktikai címkék. A párhuzamos korpuszból így kinyert frázispárokból azokat vettük figyelembe, amelyeknél mind az angol, mind a magyar oldalon pontosan egy ige szerepelt. Ezekből a frázispárokból minden magyar igéhez összegyűjtöttük az összes olyan főnevet a magyar oldalról, ami az angol oldalon szereplő, a magyar igéhez kötött igéhez volt kötve. Például a *döntést hoz* kifejezés esetén a vizsgált ige a *hoz*, és ha az angol oldalon a *decide* ige szerepel, akkor a *döntést* főnév szintén ehhez az igéhez van hozzárendelve, hiszen az angol oldalon nem szerepel külön szóként. Ezzel szemben például a *táskát hoz* esetén az angol oldalon a *bring* és a *bag* is szerepel, ezek megfelelően vannak hozzárendelve a magyar megfelelőikhez. Végül az egyes magyar igékhez összegyűjtött főnevek listáját gyakoriságuk és az adott igéhez tartozó homogenitás alapján normalizáltuk és sorba rendeztük. Az így kapott lista végét levágtuk (ahol már csak olyan kifejezések gyűltek össze, amik jelentése kompozicionális). Az algoritmus kiértékeléséhez a Szeged Korpuszból és a SzegedParalell korpuszból készült félig kompozicionális igei szerkezeteket tartalmazó listát [9] használtuk, illetve a saját algoritmusunk által nem azonosított, de ezen a listán szereplő és a kérdezőrendszer szempontjából valóban releváns kifejezéseket is felvettük a vonzatkeret-lexikonunkba kiegészítve azt a vonzatkeret kompozicionális elemeivel, illetve azok tematikus szerepeivel. Az idiomatikus és félig kompozicionális igei szerkezetek párhuzamos korpusz felhasználásával történő azonosításával kapcsolatos eredményeinkről a jelen kötetben megjelent másik tanulmányunkban [10] számolunk be részletesebben.

7. A vonzatkeretek korpuszbeli előfordulásokra való illesztése

A vonzatkereteket az UD korpuszbeli igeelőfordulásokra illesztő algoritmus első lépésben beolvassa és szintaktikailag ellenőrzi a vonzatkeret-leírásokat tartalmazó forrásfájlokat, és az öröklődési mechanizmust alkalmazva előállítja az egyes igék teljes vonzatkeret-leírását az igecsoporthoz tartozó vonzatkeretek és a csak az adott igére jellemző leírás összeolvasztásával.

A vonzatkeret-leírásokban szereplő explicit, illetve az egyes tematikus szerepek által implikált implicit formai megszorításokat (ragok, névutók, stb.) a ma-

gyar UD korpuszban használt morfológiai és szintaktikai annotációban szereplő jegyegyüttesekre fordítjuk le, és ezek felhasználásával illesztjük a vonzatkereteket az egyes igékhez a korpuszban. A hely (`LOC`), végpont (`DST`) és kiindulópont (`SRC`) szerepű kifejezések az irányhármasságra jellemző ragokat, névutókat és névmásokat tartalmazó névszói csoportokra, illetve a megfelelő határozószókra illeszkednek. Számos ige vonzatkeretében szerepel az útvonal (`PATH`) tematikus szerep, amely a végpont, a kiindulópont és érintett hely (`VIA`) szerepek tetszőleges kombinációjával helyettesíthető. A vonzatkeretlistában a könnyebb olvashatóság érdekében a ragok a mögöttes fonológiai alakjukban szerepelnek. Az illesztőalgoritmus ezeket a leírásokat alakítja át az UD korpuszban szereplő morfoszintaktikai jegyleírások formalizmusára.

Tekintettel a magyar pro drop jellegére, a hiányzó alanyokat és tárgyakat a megfelelő helyen implicit névmásokkal helyettesítjük, ha a vonzatkeret tartalmaz ilyen vonzatot és az adott tagmondatban nem jelenik meg testes alany, illetve tárgy. Az infinitívusz és az igenevek vonzatkereteit az adott igenévtípusra jellemző transzformációval hozzuk létre az alapige vonzatkereteiből.

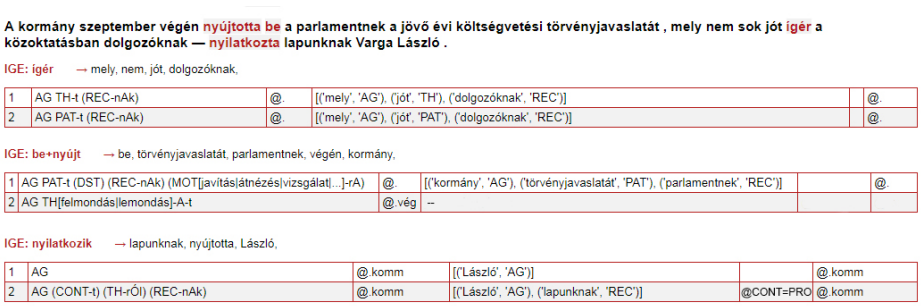
A félig kompozicionális szerkezetek egy része olyan formailag birtokos alakokat tartalmaz, amelyeknél nem a kifejezés fejét alkotó birtokjeles szóalak kapja a tényleges tematikus szerepet, hanem annak a birtokosa. Például: *a szomszédjának a nyakára küldte az adóhatóságot.* Ezeket a szerkezeteket a névutós szerkezetekhez hasonló alakúvá alakítjuk és a tényleges vonzat (*szomszédja*) lesz a módosított szerkezetben a vonzatként szereplő szerkezet feje. Ehhez már közvetlenül hozzárendelhető a megfelelő tematikus szerep.

Számos vonzatkeretben (az ige egy konkrét jelentése esetében) szemantikailag kötött típusú valamelyik argumentum. Például: `felkel [égitest]`, `átvesz [lábbeli|ruha]-A-t`. Az ilyen keretek illesztésénél a [11]-ben leírt módon morfológiailag elemzett korpuszból és lexikai szemantikai erőforrás felhasználásával épített szóbeágyazás alapú „Dologfelismerő” modellt használjuk. Ez a modell a szavakhoz lexikai szemantikai címkéket rendel. Ha az adott argumentum feje rendelkezik a vonzatkeretben meghatározott címkével, akkor a vonzatkeret illeszkedik. Például *felkel a nap*, *átveszi a tornacipőjét.*

A 7. ábrán egy minta látható arra, hogy egy adott mondat igéire milyen vonzatkeretek szerepeltek az adatbázisban, és ezek hogyan illeszkednek az adott mondatra.

8. Konklúzió

Cikkünkben egy olyan folyamatban lévő kutatásról számoltunk be, amelynek keretében létrehozott korpuszannotáció alkalmas a feldolgozott szöveggel kapcsolatban releváns kérdéseket megfogalmazni képes elemzőrendszer betanítására. A továbbiakban a lehetséges vonzatkeret-illeszkedések rangsorolása, a szabad határozók szerepének azonosítására szolgáló erőforrás rendszerbe illesztése, és ezek felhasználásával a kézi ellenőrzés alapjául szolgáló annotáció előállítása a célunk.



A kormány szeptember végén nyújtotta be a parlamentnek a jövő évi költségvetési törvényjavaslatát , mely nem sok jót igér a közoktatásban dolgozóknak — nyilatkozta lapunknak Varga László .

IGE: igér → mely, nem, jót, dolgozóknak,

1	AG TH-t (REC-nAk)	@.	[('mely', 'AG'), ('jót', 'TH'), ('dolgozóknak', 'REC')]	@.
2	AG PAT-t (REC-nAk)	@.	[('mely', 'AG'), ('jót', 'PAT'), ('dolgozóknak', 'REC')]	@.

IGE: be+nyújt → be, törvényjavaslatát, parlamentnek, végén, kormány,

1	AG PAT-t (DST) (REC-nAk) (MOT[javítás|átnézés|vizsgálat|...]-rA)	@.	[('kormány', 'AG'), ('törvényjavaslatát', 'PAT'), ('parlamentnek', 'REC')]		@.
2	AG TH[felmondás|lemondás]-A-t	@.vég	--		

IGE: nyilatkozik → lapunknak, nyújtotta, László,

1	AG	@.komm	[('László', 'AG')]		@.komm
2	AG (CONT-t) (TH-rÓl) (REC-nAk)	@.komm	[('László', 'AG'), ('lapunknak', 'REC')]	@CONT=PRO	@.komm

7. ábra. Példa a vonzatok tematikus szerepeinek illesztésére a vonzatkeret-adatbázisból

Köszönetnyilvánítás

Jelen kutatás az FK 125217 és a PD 125216 számú projekt keretében az FK 17 és a PD 17 pályázati program finanszírozásában a Nemzeti Kutatási Fejlesztési és Innovációs Alap által biztosított támogatással és az Emberi Erőforrások Minisztériuma ÚNKP-18–3-III-PPKE-26 kódszámú Új Nemzeti Kiválóság Programjának támogatásával valósult meg. Szeretnénk köszönetet mondani Fegyó Kingának és Bognár Ivettnek az igei vonzatkeretek és a vonzatok tematikus szerepeinek leírásában végzett munkájukért.

Hivatkozások

1. Nivre, J., de Marneffe, M.C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C.D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., Zeman, D.: Universal dependencies v1: A multilingual treebank collection. In Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., eds.: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Paris, France, European Language Resources Association (ELRA) (2016)
2. Novák, A., Novák, B.: Pos, ana and lem: Word embeddings built from annotated corpora perform better. In Gelbukh, A., ed.: Computational Linguistics and Intelligent Text Processing: 17th International Conference, CICLing 2018, Hanoi, Vietnam, Springer International Publishing, Cham. (2018)
3. Siklósi, B.: Using embedding models for lexical categorization in morphologically rich languages. In Gelbukh, A., ed.: Computational Linguistics and Intelligent Text Processing: 17th International Conference, CICLing 2016, Konya, Turkey, Springer International Publishing, Cham. (2016)
4. Siklósi, B., Novák, A.: Közeli rokonunk, az autó. XII. Magyar Számítógépes Nyelvészeti Konferencia (2016)
5. Sass, B., Váradi, T., Pajzs, J., Kiss, M.: Magyar igei szerkezetek: a leggyakoribb vonzatok és szókapcsolatok szótára. A magyar nyelv kézikönyvei. Tinta Könyvkiadó (2010)

6. Ligeti-Nagy, N., Novák, A.: Hol ugat a kutya? Örömében. helyhatározói esetragos névszók pontosabb annotációja. In: XV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2019), Szeged, SZTE (2019)
7. Lison, P., Tiedemann, J.: Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., eds.: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Paris, France, European Language Resources Association (ELRA) (2016)
8. Dyer, C., Chahuneau, V., Smith, N.A.: A simple, fast, and effective reparameterization of ibm model 2. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics (2013) 644–648
9. Vincze, V.: Semi-Compositional Noun + Verb Constructions : Theoretical Questions and Computational Linguistic Analyses. PhD thesis, University of Szeged (2011)
10. Novák, A., Laki, L.J., Novák, B.: Mit hozott édesapám? döntést – idiomatikus és félig kompozicionális magyar igei szerkezetek azonosítása párhuzamos korpuszból. In: XV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2019), Szeged, SZTE (2019)
11. Novák, A., Novák, B.: Cross-Lingual Generation and Evaluation of a Wide-Coverage Lexical Semantic Resource. In Calzolari, N., 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., eds.: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, European Language Resources Association (ELRA) (2018)