Félig kompozicionális szerkezetek a SzegedParalell angol–magyar párhuzamos korpuszban

Vincze Veronika[1], Felvégi Zsuzsanna[2], R. Tóth Krisztina[3]

[1] Szegedi Tudományegyetem, Informatikai Tanszékcsoport
vinczev@inf.u-szeged.hu
[2] Szegedi Tudományegyetem, Nyelvtudományi Doktori Iskola,
Angol Alkalmazott Nyelvészet PhD Program
felvegi@gyakg.u-szeged.hu
[3] MTA-SZTE Oktatáselméleti Kutatócsoport
tothkr@inf.u-szeged.hu

Kivonat: A természetes nyelvi feldolgozásban az egyik legnehezebb problémát a többszavas kifejezések azonosítása és megfelelő kezelése jelenti. Ezt megkönnyítendő, a SzegedParalell angol–magyar párhuzamos korpusz egy részében kézzel bejelöltük a félig kompozicionális szerkezeteket. A szerkezeteket mindkét nyelven annotáltuk, lehetővé téve ezáltal az angol és magyar szerkezetek automatikus párosítását. Az annotált korpusz jól használható tanuló adatbázisként mind egynyelvű, mind többnyelvű alkalmazásokban, de a kontrasztív nyelvészetben, stilisztikában, illetve a nyelvoktatásban is hasznosítható.

1 Bevezetés

A természetes nyelvi feldolgozásban, különösen a gépi fordítás és fordítástámogatás területén az egyik legnehezebb problémát a többszavas kifejezések megfelelő kezelése jelenti. A többszavas kifejezések sikeres kezelésének első lépése, hogy felismerjük, többszavas kifejezéssel van dolgunk, hiszen például a többszavas kifejezések szintaktikai felépítése hasonlít más, produktív szerkezetek felépítéséhez (*veri az ördög a feleségét – veri a szomszéd a kutyáját*). Ennek automatikus eldöntése igen nehéz feladat, a feladatot tovább nehezíti, ha a felismerés mellett különböző nyelven írt többszavas kifejezések megfeleltetését tűzzük ki célul. A probléma nehézségét mutatja, hogy egy többszavas kifejezés idegen nyelvű megfelelője a legritkább esetben szó szerinti fordítása az eredetinek. Például a félig kompozicionális szerkezetek esetén – melyek olyan, főnévből és igéből álló többszavas kifejezések, ahol a szemantikai fej a főnév, míg az ige pusztán csak a szerkezet igeiségéért felelős –, csak a kifejezés egy részét fordíthatjuk szó szerint (a főnévi tagot: *szerződést köt – make a contract*), az idiómák azonban a bennük szereplő szavak szintjén semmiképpen sem feleltethetők meg egymásnak (*ez nem az én asztalom – it's not my cup of tea*), ezért problémát jelent(h)e(t)nek a különféle számítógépes alkalmazások számára. Az azonosításukra képes algoritmusok fejlesztéséhez és teszteléséhez annotált korpuszokra van szükség.

2 A többszavas kifejezések

E fejezetben bemutatjuk a többszavas kifejezések főbb tulajdonságait, amelyek megnehezítik a többszavas kifejezések automatikus felismerését. Példáink az angol és a magyar nyelvből származnak, ezzel is kiemelve a probléma általánosságát (azaz nyelvektől való függetlenségét).

2.1 A többszavas kifejezések idioszinkratikus tulajdonságai

A többszavas kifejezések olyan lexikai egységek, melyek több szóból állnak, és szintaktikai, szemantikai, pragmatikai vagy statisztikai szempontból idioszinkratikus sajátosságokat mutatnak [1, 4, 8]. A lexikálisan idioszinkratikus kifejezések részei nem helyettesíthetők más, azonos (vagy hasonló) értelmű szavakkal anélkül, hogy elvesztenék eredeti jelentésüket [5, 7]. Például a *fűbe harap* idiómában nem cserélhetjük ki a *fű* szót *pázsit*ra (*pázsitba harap*) a jelentés megőrzése mellett. Hasonlóan az angolban: a *kick the pail* 'felrúgja a vödröt' csakis szó szerint értelmezhető, míg a 'meghal' jelentést csak a *kick the bucket* idióma képes hordozni, noha a *pail* és *bucket* szavak önmagukban szinonimák.

A szintaktikailag idioszinkratikus kifejezések szintaktikai tulajdonságai nem következnek a részek szintaktikai tulajdonságaiból. A magyarban például a *kerek perec* kifejezés határozóként viselkedik, noha szigorúan véve egy melléknév és egy főnév kapcsolatáról van szó. Az *all of a sudden* angol többszavas kifejezés szintén határozószói szerepet tölt be, azonban egy névmás, egy prepozíció, egy névelő és egy melléknév alkotja.

A többszavas kifejezések jelentése többnyire nem (teljesen) kompozicionális, azaz nem számítható ki pusztán részeinek jelentésére és azok kapcsolódási módjára támaszkodva (szemantikai idioszinkrázia). Tipikus példák az idiómák: a fenti *fűbe harap* kifejezés 'meghal' jelentése semmiképpen sem számítható ki a *fű*, a *harap* és a *–bA* toldalék jelentéséből. A kompozicionalitás hiányára az is rámutat, hogy a fenti kifejezés helyes angol változata a *kick the bucket*, amely szavak szintjén semmiképpen sem jó fordítása az eredeti kifejezésnek, viszont jelentés szintjén tökéletesen megfelel egymásnak a két kifejezés.

A pragmatikailag idioszinkratikus kifejezések többnyire egy adott helyzetben vagy adott körülmények között használatosak: például a *Jó étvágyat!* mondat jellemzően étkezés előtt hangzik el, a *How do you do?* kifejezés pedig bemutatkozáskor használatos.

A statisztikai idioszinkrázia azt takarja, hogy a többszavas kifejezés tagjai statisztikailag szignifikánsan nagy valószínűséggel együttesen fordulnak elő, ezeket nevezik kollokációknak [8]. Megjegyezzük azonban, hogy a kollokációkra nem (feltétlenül) jellemző a szemantikai vagy szintaktikai sajátos viselkedés. Néhány példa: a *fekete-fehér* melléknév *fehér-fekete* sorrendben is ugyanazzal a jelentéssel rendelkezne, azonban jelentősen többször fordul elő *fekete-fehér* alakban, mint fordítva, illetve az angol *pepper and salt* 'bors-só' jelentése egyenértékű a *salt and pepper* 'só-bors' szókapcsolatéval, mégis utóbbi számít a bevett kifejezésnek.

Természetesen nem minden egyes idioszinkratikus tulajdonság érvényes minden többszavas kifejezésre: léteznek olyan többszavas kifejezések, melyek például szintaktikailag szabályosan viselkednek, azonban jelentésük nem kompozicionális (ilyen a legtöbb idióma).

2.2 A többszavas kifejezések szintaktikai viselkedése

A többszavas kifejezések szintaxisuk szerint lehetnek kötöttek, félig kötöttek, illetve produktívak [8, 11]. A kötött kifejezések nem mutatnak szintaktikai változatosságot: mindig ugyanabban a formában fordulnak elő [5]. Ilyenek például a közmondások (l. alább). Az automatikus felismerést tovább nehezíti, hogy a félig kötött kifejezések bizonyos fokig módosíthatók: az igék például ragozhatók az idiómákon belül, az összetett főnevek pedig többes száma tehetők. A legnehezebb problémát a szintaktikailag produktív többszavas kifejezések jelentik – a félig kötött kifejezéseknél említett módosítások mellett –, melyek szabadabban változtathatók: tagjaik módosíthatók (például egy jelző módosíthatja a félig kompozicionális szerkezetek főnévi tagját), sőt a szerkezet tagjai nem is feltétlenül szerepelnek egymás mellett a mondatban.

2.3 A többszavas kifejezések típusai

A többszavas kifejezések több csoportba sorolhatók egyrészt az őket alkotó szavak szófaja alapján, másrészt szintaktikai viselkedésük alapján. Az alábbiakban a magyarra és angolra egyaránt jellemző, gyakran előforduló többszavas kifejezések, azaz az összetett szavak, idiómák, közmondások és a félig kompozicionális szerkezetek azon tulajdonságait ismertetjük röviden, amelyek megnehezítik azok automatikus cél- és forrásnyelvi megfeleltetését.

Az összetett szavak olyan lexikai egységek, melyek két vagy több önállóan is létező szóból állnak. Helyesírásukat tekintve léteznek egybeírt összetett szavak (*iskolaigazgató*), kötőjellel írt összetett szavak (*időjárás-jelentés*), illetve szóközt tartalmazó összetett szavak (jellemzően az angol nyelvben, például *power plant*, míg a magyar helyesírási gyakorlat az összetett szavak határterületéhez tartozónak minősíti a különírt állandósult szókapcsolatokat, például *kútba esés*). Természetesen az összetett szavak nem csak főnevek lehetnek, léteznek összetett melléknevek (*red haired*, *nagyotmondó*), összetett határozószók (*above all*, *csakazértis*), összetett prepozíciók (*in front of*) és összetett kötőszavak (*in order that, nehogy*) is. A legtöbb összetett szó szintaktikailag szabályos viselkedést mutat, de pontos jelentésük – azaz az összetétel tagjai közti viszony jellege – még azonos szintaktikai felépítés esetén is változhat: a *repceolaj* repcéből készül, de a *babaolaj* babák számára.

Az idiómák jelentése nem határozható meg részeinek jelentéséből [6, 8], noha szintaktikai viselkedésük általában véve szabályos, szemantikájuk teljességgel megjósolhatatlan: a *fekete bárány* 'a megszokottól eltérően viselkedő (nemkívánatos) személy' jelentésének semmi köze sem az állathoz, sem a fekete színhez. Az idiómák általában mutatnak morfológiai változatosságot (például a bennük szereplő ige ragozható).

A közmondások a legtöbb ember által igaznak tartott állításokat fejeznek ki, leginkább egy teljes mondat formájában (*Ki korán kel, aranyat lel*). Emiatt általában ugyanabban a formában fordulnak elő, szemben az idiómákkal.

A félig kompozicionális főnév + ige szerkezetekben (pl.: *tanácsot ad, döntést hoz, virágba borul*) a kifejezés szemantikai tartalmát nagyrészt a főnév hordozza, ugyanakkor az ige vállal főszerepet a szerkezet szintaxisának kialakításában [14]. Mivel jelentésük nem teljesen kompozicionális, a szerkezet elemeinek egyenkénti lefordítása nem (vagy csak nagyon ritkán) eredményezi a szerkezet idegen nyelvű megfelelőjét. Emellett a félig kompozicionális szerkezetek (*választ kap*) szintaktikailag hasonló felépítéssel bírnak, mint más, produktív (kompozicionális) szerkezetek (*pulóvert kap*), illetve idiómák (*vérszemet kap*) [2], így azonosításuk nem valósulhat meg pusztán szintaktikai mintákat figyelembe véve. Végül, mivel a szerkezet szintaktikai és szemantikai feje nem azonos, a szerkezet nyelvi elemzésekor célszerű a főnevet és az igét egy komplex egységként kezelni – az angol vonzatos igékhez (phrasal verbs) hasonlóan.

A fenti okokból kifolyólag a többszavas kifejezések kezelése különleges figyelmet érdemel a természetes nyelvi alkalmazásokban. Ennek első lépéseként azonosítani kell a többszavas kifejezéseket, mely célhoz különféle algoritmusok fejlesztése segíthet hozzá. Ebben sikeresen hasznosítható egy kézzel annotált tanító adatbázis: korpuszunk építésekor ezt a szempontot tartottuk szem előtt.

3 A SzegedParalell párhuzamos korpusz

A SzegedParalell korpusz építésére két, nemzetközi viszonylatban is alkalmazott alternatíva merült fel: 1) a korpusz építéséhez már meglévő, egynyelvű annotált korpuszt használnak fel, annak szövegeit lefordítják a célnyelvre, majd a lefordított szövegeket is feldolgozzák; 2) a korpusz építésekor két nyelven elérhető szövegeket keresnek, majd ezeket a nyers szövegeket mindkét nyelven annotálják. A SzegedParalell építése során az utóbbi megoldást választottuk, mert a fordítás hosszadalmasabb és költségesebb procedúra, mint a kétnyelvű szövegek gyűjtése. A korpusz felépítését tekintve az alábbi témákból tartalmaz szövegeket.

1. táblázat: A SzegedParalell korpusz felépítése.

Témakör	ME (db)
tankönyvi mondatok	3.496
Európai Unióról szóló szövegek	1.518
kétnyelvű magazinok	5.320
irodalmi és történelmi alkotások	88.716
egyéb	695
összesen	***99.745***

Megjegyzés: Mondatszinkronizációs egységek (ME), l. lentebb.

Az első korpuszrész többnyire Dévainé Angeli Mariann *Angol nyelvtani gyakorlatok* és Dohár Péter *Kis angol nyelvtan* című könyvének különálló párhuzamos mondataiból áll, melyek az angol nyelvtan sajátosságait hivatottak reprezentálni. A nyelvtankönyvi mondatok mellett autentikus szövegeket is beépítettünk a párhuzamos korpuszba, így biztosítva az egyensúlyt a mesterkélt és a természetes nyelvi szerkezetek között. Az autentikus szövegek elsősorban kétnyelvű magazinokból, interneten található általános nyelvezetű, hétköznapi, gazdasági és jogi témájú szövegeket tartalmaznak, és változatos szókinccsel rendelkeznek. A kétnyelvű magazinok alkorpuszát a *Horizon Magazin*, illetve a *Resource Ingatlan Info* c. újságok alkotják. A Horizon Magazin (a Malév fedélzeti magazinja) sokféle, hétköznapi témát ölel fel, mint kultúra, utazás, interjúk hírességekkel, nevezetesebb városok bemutatása, a Resource Ingatlan Info pedig elsősorban területfejlesztésről, építőipari beruházásokról, logisztikáról, és az ezekhez kapcsolódó témakörökről közöl cikkeket. A korpusz tartalmaz továbbá Európai Unióról[1] szóló rövid ismeretterjesztő cikkeket (pl.: az EU történetéről, zászlójáról, himnuszáról, pénzneméről stb).

Az irodalmi szövegeket tartalmazó korpuszrész a Hunglish korpusz [3] irodalmi műveinek egy részét tartalmazza. Elsősorban modern angol irodalmi műveket építettünk be a korpuszunkba. Továbbá a Magyar Elektronikus Könyvtár weboldalán elérhető kétnyelvű szövegeket (történelmi és irodalmi művek) emeltünk be a korpuszba.

Az egyéb kategóriában rövid terjedelmű szövegek szerepelnek, melyek korábbi internetes gyűjtések eredményei: rövid beszámoló kulturális eseményről, tudományos feltárások és receptek.

A szövegek párhuzamosítása során először a szövegeket normalizáltuk, majd ellenőriztük a fordítás helyességét, szükség esetén javítottuk azt. Az egyes magyar és angol nyelvi szövegeket külön fájlokban mentettük. A fájlok szinkronizálása után a parallel szövegek *bekezdésszintű* összerendelését automatikusan végeztük, mert szakfordítói tapasztalatokra alapozva a forrásnyelvi és a célnyelvi szövegek egyenlő számú bekezdést tartalmaznak, és ezek sorrendje nem felcserélhető.

A *mondatillesztés* alapaspektusa az a fordítási tény, hogy a fordítási egységek nem nyúlhatnak át bekezdések határán. A mondatok illesztése a bekezdés-összerendeléssel szemben nem egy kölcsönösen egyértelmű reláció, melynek okai:

(1) a bekezdésben szereplő mondatok sorrendje felcserélhető, illetve
(2) egy forrásnyelvi mondat a célnyelven több mondatnak is megfelelhet, így a célnyelvi egységek adott esetben túlnyúlhatnak egy mondat határán.

A párhuzamos szövegek feldolgozása során az alábbi összerendelési lehetőségek (mondatszinkronizációs típusok) fordultak elő:

- 1:1 egy kiindulási nyelvi mondatnak egy célnyelvi megfelelője van (*egyezés*)
- 1:0 a kiindulási nyelvi mondatnak nincs tartalmi megfelelője a célnyelven (*kihagyás*)
- 0:1 a célnyelvi mondat nem szerepel a forrásnyelvi szövegben (*betoldás*)
- 1:N egy kiindulási nyelvi mondatnak több célnyelvi megfelelője van (*szétbontás*)
- N:1 több kiindulási nyelvi mondatnak egy célnyelvi mondat felel meg (*összevonás*)

[1] A http://europa.eu.int weboldalról és a Wikipédia weboldaláról.

- N:M több kiindulási nyelvi mondatnak több célnyelvi mondat felel meg (ez többnyire szépirodalmi művekben fordul elő)

Ezeket az egymásnak megfeleltetett egységeket mondatszinkronizációs egységeknek nevezzük, melyek segítségével feltérképezzük a magyar és angol nyelvi félig kompozicionális szerkezetek előfordulási gyakoriságát, illetve az angol és magyar nyelvű kifejezéseinek automatikus megfeleltetése során felmerülő problémákat.

4 A korpusz annotálása

A többszavas kifejezések természetes nyelvi feldolgozását megkönnyítendő a SzegedParalell angol–magyar párhuzamos korpuszban [10] kézzel bejelöltük a félig kompozicionális szerkezeteket (rövidítve: FX) [13].

4.1 Az annotált korpusz

A korpusz szövegei közül elsődlegesen az újságcikkeket, Európai Unióról szóló szövegeket és a tankönyvi mondatokat annotáltuk, merthogy a félig kompozicionális szerkezetek nagyszámú előfordulása a sajtónyelvben és a gazdasági-politikai doménben várható [14], illetve a nyelvkönyvekben fordításra szánt mondatok valószínűleg nagy arányban tartalmaznak többszavas kifejezéseket (azaz olyan nyelvi elemeket, amelyeknek a fordítása nem szó szerinti). Az összehasonlítás végett azonban néhány irodalmi szövegben is jelöltük a félig kompozicionális szerkezeteket. Így egyrészt a különféle témájú szövegekből nyert adatok segítségével kvantitatívan alá tudjuk támasztani (vagy meg tudjuk cáfolni) a fenti feltételezéseket, másrészt pedig a nyelvek és domének közti összevetésből olyan minőségi jellegű kérdések megválaszolására is sor kerülhet, hogy miként fordítja a szépirodalmi fordítás vagy egy gazdasági-jogi témájú szöveg a félig kompozicionális szerkezeteket (azaz a félig kompozicionális szerkezetnek szintén szerkezet felel-e meg a másik nyelvben, avagy szabadabban fordítják).

Az annotált korpusz méretét a következő táblázat mutatja:

2. táblázat: Az annotált korpusz felépítése és mérete.

Téma	Szövegek száma (db)	ME (db)	FX - magyar	FX - angol
EU	30	1518	295 (19,4%)	227 (15%)
Újságcikkek	151	5320	477 (9%)	400 (7,5%)
Tankönyvi mondatok	7	3496	85 (2,4%)	131 (3,7%)
Irodalmi szövegek	3	3232	247 (7,6%)	336 (10,4%)
Egyéb	5	695	8 (1,2%)	6 (0,8%)
Összesen	196	14261	1112 (7,8%)	1100 (7,71%)

A korpuszban levő angol és magyar félig kompozicionális szerkezetek száma megközelítőleg ugyanaz, emiatt a mondatszinkronizációs egységek (ME) közel ugyanakkora hányada tartalmaz félig kompozicionális szerkezetet (l. százalékos értékek). Ez azonban nem jelenti azt, hogy minden egyes félig kompozicionális szerkezetnek megvan a másik nyelvű megfelelője – más szóval, vannak olyan szerkezetek, amelyek csak a magyar, illetve csak az angol korpuszban szerepelnek.

4.2 Annotációs elvek

A félig kompozicionális szerkezeteket angol és magyar nyelven egyaránt annotáltuk. A szövegeken három annotátor dolgozott egységes annotációs útmutató alapján egy nyelvész szakértő irányításával. Ugyanazon szöveg forrás- és célnyelvi változatában ugyanaz a személy jelölte be a félig kompozicionális szerkezeteket.

Mivel a félig kompozicionális szerkezetek szintaktikailag produktívak (vö. 2.2), többféle formában is előfordulhattak a szövegekben. Hasonlóan a Szeged Korpusz korábbi annotálási elveihez [13], itt is a következő altípusokba soroltuk a szerkezeteket az annotáció során:

Főnév + ige kombinációja (VERB): *bejelentést tesz, igénybe vesz, take a look, pay a visit*

Igenevek (PART): *gondot viselő, kézbe véve, photos taken, taking part*

Főnévi változat (NOM): *szerződéskötés, bérbe vétel, service provider, decision-maker*

Különálló szerkezet (SPLIT): *előadást fog tartani, kapott tőlük engedélyt, contest is held, effort you make*

A szerkezetek altípusainak megoszlása a következő táblázatban látható:

3. táblázat: A félig kompozicionális szerkezetek altípusainak megoszlása.

	VERB		**PART**		**NOM**		**SPLIT**	
	angol	magyar	angol	magyar	angol	magyar	angol	magyar
EU	132	158	30	76	24	32	41	29
Újságcikkek	281	330	48	86	16	23	55	38
Tankönyvi mondatok	106	62	4	10	13	3	8	10
Irodalmi szövegek	222	196	13	11	5	0	96	40
Egyéb	4	7	1	0	0	0	1	1
Összesen	*745*	*753*	*96*	*183*	*58*	*58*	*201*	*118*

Míg az igei és a főnévi alakok száma nagyrészt megegyezik a két nyelvben, addig a melléknévi igenevek és a különálló szerkezetek száma jelentősen eltér egymástól. Ez a különbség valószínűleg nyelvtani okokra vezethető vissza: a SPLIT kategóriába sorolt elemek nagy része az angolban például passzív szerkezetet alkot, ahol is a szerkezet főnévi komponense alanyi funkciót tölt be, ezáltal nem szomszédos az igével. A PART kategória esetében pedig előfordul, hogy míg a magyarban a főnévi komponensnek előmódosítója van, mely megköveteli az igei komponens melléknévi igenév formájában való jelenlétét is, addig az angolban utómódosítót találunk, amely elől elmaradhat az igenév, például:

*az emberi jogokba **vetett** hit*
a belief in human rights

Természetesen további részletes vizsgálatok más tendenciákra is fényt deríthetnek e különbségek elemzésében.

5 A kétnyelvű szerkezetek párosítása

A kétnyelvű annotáció (l. alábbi példa) lehetővé teszi az angol és magyar szerkezetek automatikus párosítását, mivel a mondat szinten párhuzamosított, annotált korpuszban egyszerűen megtalálható az adott kifejezés másik nyelvű megfelelője, ha a mondatszinkronizációs egységek egy-egy félig kompozicionális szerkezetet tartalmaznak:

*A mulatozások fő időszaka a 15-16. századra tehető, amikor ezek a bálok **fontos szerepet játszottak** a párválasztásban.*
*Such revelry can be claimed to have reached its height in the 15th and 16th centuries, when the dances **played an important role** for those in search of a good match.*

A szerkezetek automatikus szinkronizálása természetesen külön kézi ellenőrzést igényel, amennyiben az adott mondaton belül több félig kompozicionális szerkezet is előfordul. Sass Bálint [9] beszámol egy igei szerkezetek párhuzamos korpuszból való kinyerésére szolgáló eljárásról, mely egy korábbi, igéket és azok bővítményeit kinyerő algoritmusra épül. A módszer lényege, hogy a tagmondatok igéit egymás mellé rendelve egy komplex ige jön létre, melyhez a bővítményeket halmazként rendeljük hozzá, felcímkézve őket aszerint, hogy melyik nyelvű részkorpuszból származnak. Az így kapott reprezentációból az eredeti algoritmus segítségével lehet kigyűjteni az egyes nyelvekre jellemző igei szerkezeteket. A módszer előnye, hogy nemcsak a szerkezet-szerkezet párokat képes megtalálni, hanem azokat az eseteket is, amikor a szerkezetnek ige felel meg. A módszer nyelvfüggetlen, tehát korpuszunkra is alkalmazható, a kézi annotációnak köszönhetően pedig a kiértékelés is egyszerűbb.

6 Eredmények

A különböző doméneket tekintve elmondhatjuk, hogy valóban a gazdasági-jogi témájú, Európai Unióról szóló szövegekben fordul elő arányukban a legtöbb félig kompozicionális szerkezet. Ezzel ellentétben a tankönyvi példamondatok elenyésző részében találhatunk félig kompozicionális szerkezetet, vagyis vélhetően nem annyira a lexikai, mint inkább a nyelvtani szempontok domináltak a mondatsorok összeállításakor. További érdekesség, hogy az angol irodalmi szövegek jelentősen nagyobb arányban tartalmaznak félig kompozicionális szerkezetet, mint magyar megfelelőik, ez különösen Swift *Gulliver utazásai* c. regényére igaz (a regény mondatainak 16%-a tartalmaz félig kompozicionális szerkezetet, az angol nyelvű szövegek közül ez a legmagasabb arány). Mivel a regény 1726-ban jelent meg[2], a korai XVIII. századi angol nyelvállapotot tükrözi, azonban – kellő számú adat híján – elhamarkodott lenne arra a következtetésre jutni, hogy a korabeli angol nyelvben jóval több félig kompozicionális szerkezet szerepel, mint a mai angolban: ennek alátámasztásához további – nyelvtörténeti – vizsgálatokra van szükség.

Ha a szövegek témája szerint vizsgáljuk a félig kompozicionális szerkezetek megfeleltetését, azt találjuk, hogy az újságcikkekben és az EU-ról szóló szövegekben nagy arányú a megfelelés a szerkezetek között (azaz egy magyar félig kompozicionális szerkezet angol párja is nagy valószínűséggel félig kompozicionális szerkezet), azonban a szépirodalmi szövegre ez nem áll. Egyrészt az angol irodalmi szövegekben szám szerint jóval több a félig kompozicionális szerkezet, mint a magyar szövegekben (a Mark Twain-regény kivételével, ahol kiegyenlített a számuk), másrészt igen gyakori az a jelenség, hogy a félig kompozicionális szerkezetnek a másik nyelvi megfelelője nem szerkezet (sőt nem is mindig ige), például:

[…] during which time, the emperor ***gave orders*** *to have a bed prepared for me.*
[…] ez idő folyamán, a császár ***parancsára****, fekvőhelyet készítettek nekem.*

Mindebből az következik, hogy a szépirodalmi szövegek kevésbé használhatók mint tanító, illetve tesztadatbázis a szerkezetek automatikus szinkronizálásához, mint például a gazdasági-jogi jellegű szövegek vagy újságcikkek.

Az angol és a magyar nyelvű félig kompozicionális szerkezetek összehasonlításával megállapítható, hogy számos forrásnyelvi félig kompozicionális szerkezetnek szintén egy célnyelvi szerkezet felel meg, ezáltal lehetővé válik az automatikus párosítás (l. fent). Vannak azonban olyan esetek is, amikor az egyik nyelvben félig kompozicionális szerkezetet találunk, a másik nyelv viszont igét alkalmaz:

I don't usually moan or ***make any special requests****.*
Nem vagyok nyűgös, ***nincsenek extra kívánságaim****.*

[2] A SzegedParalellben Vajdafy Ernő 1906-os fordítása szerepel, amely azonban tudatosan archaizáló nyelvezetű, így a keletkezésük között eltelt közel két évszázad ellenére is összemérhetőnek ítéltük meg a két szöveget.

Ennek speciális esete az, amikor jellemzően a szerkezet főnévi komponensével azonos tőből származó, azonos jelentésű igei variáns [12] szerepel a célnyelven:

*It **decided** to welcome 10 more countries to join the EU on 1 May 2004.*
*A Tanács **meghozta a döntést** arról, hogy 2004. május 1-jén 10 új államot vesznek fel az Unió tagállamai sorába.*

Egy másik érdekesség, hogy az angol passzív szerkezetek magyar megfelelői időnként a *kerül* igét tartalmazó félig kompozicionális szerkezetek:

*The song "Auld Lang Syne" **was** partially written by Robert Burns and **published** after his death in 1796.*
*A híres „Auld Lang Syne" („Régóta már") című dalt részben Robert Burns írta, és halála után, 1796-ban **került kiadásra**.*

E nyelvek közti különbségek elemzése mind a gépi fordítás, mind a fordítástudomány számára haszonnal bírhat.

7 A korpusz felhasználhatósága

Az annotált korpusz jól használható tanuló adatbázisként más mind egynyelvű, mind többnyelvű alkalmazásokban (például többnyelvű információ-visszakeresés), de a kontrasztív nyelvészetben, stilisztikában, illetve a nyelvoktatásban is hasznosítható.

Az adatbázis oktatási és kutatási célokra ingyenesen elérhető a Creative Commons licenc alatt a www.inf.u-szeged.hu/rgai/nlp címen.

8 Összegzés

Cikkünkben bemutattuk a SzegedParalell angol-magyar párhuzamos korpusznak félig kompozicionális szerkezetekre annotált verzióját. Az elkészült adatbázis mintegy 1100 szerkezetet tartalmaz mind angol, mind magyar nyelven (noha a forrásnyelvi többszavas kifejezés célnyelvi megfelelője nem minden esetben többszavas kifejezés). Az annotált korpusz hasznosítható különféle, a többszavas kifejezések automatikus felismerésére készített algoritmusok tanításában és kiértékelésében, ezenkívül a gépi fordítás és a többnyelvű információ-visszakeresés számára is haszonnal bírhat, de nyelvészek is sikeresen építhetik be kutatásaikba az itt elért eredményeket.

Köszönetnyilvánítás

Szeretnénk köszönetet mondani a korpusz annotátorainak áldozatos munkájukért.

A kutatást – részben – a MASZEKER kódnevű projekt keretében az NKTH támogatta.

Bibliográfia

1. Calzolari, N., Fillmore, C., Grishman, R., Ide, N., Lenci, A., MacLeod, C., Zampolli, A.: Towards best practice for multiword expressions in computational lexicons. In: Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002). Las Palmas (2002) 1934–1940
2. Fazly, A., Stevenson, S.: Distinguishing Subtypes of Multiword Expressions Using Linguistically-Motivated Statistical Measures. In: Proceedings of the Workshop on A Broader Perspective on Multiword Expressions. Association for Computational Lingustics (2007) 9-16
3. Halácsy P., Kornai A., Németh L., Sass B., Varga D., Váradi T., Vonyó A.: A hunglish korpusz és szótár. In: Alexin Z., Csendes D. (szerk.): MSzNy 2005 – III. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2005) 134–142
4. Kim, S.N.: Statistical Modeling of Multiword Expressions. PhD thesis, University of Melbourne (2008)
5. Manning, C., Schütze, H.: Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA (1999)
6. Nunberg, G., Sag, I. A., Wasow, T.: Idioms. Language, Vol. 70 (1994) 491–538
7. Oravecz, Cs., Nagy, V. Varasdi, K.: Lexical idiosyncrasy in MWE extraction. In: Proceedings from the Corpus Linguistics Conference Series, Vol. 1, No. 1. Birmingham (2005)
8. Sag, I.A., Baldwin, T., Bond, F., Copestake, A., Flickinger, D.: Multiword Expressions: A Pain in the Neck for NLP. In: Gelbukh, A. (ed.) Proceedings of Conference on Intelligent Text Processing and Computational Linguistics 2002. Mexico City (2002)
9. Sass B.: Párhuzamos igei szerkezetek közvetlen kinyerése párhuzamos korpuszból. In: Tanács A., Vincze V. (szerk.): VII. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2010) 102–110
10. Tóth, K., Farkas, R., Kocsor, A.: Hybrid algorithm for sentence alignment of Hungarian-English parallel corpora. Acta Cybernetica Vol. 18, No. 3 (2008) 463–478
11. Váradi T.: Többszavas kifejezések kezelése MT szótárban. In: Alexin Z., Csendes D. (szerk.): MSzNy 2005 – III. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2005) 233–244
12. Vincze V.: Angol–magyar főnév + ige szerkezetek és igei párjaik. In: Váradi T. (szerk.): II. Alkalmazott Nyelvészeti Doktorandusz Konferencia. MTA Nyelvtudományi Intézet, Budapest (2009) 113–123
13. Vincze V.: Félig kompozicionális szerkezetek a Szeged Korpuszban. In: Tanács A., Szauter D., Vincze V. (szerk.): VI. Magyar Számítógépes Nyelvészeti Konferencia (MSzNy 2009). Szegedi Tudományegyetem, Szeged (2009) 390–393
14. Vincze, V., Csirik, J.: Hungarian Corpus of Light Verb Constructions. In: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010). Coling 2010 Organizing Committee, Beijing, China (2010) 1110–1118