Jelentés az első digitalizált magyar–orosz párhuzamos korpusz (HunOr) építéséről

Problémák és lehetőségek[1]

munka során szerzett tapasztalatokról, valamint a jövőbeli feladatokról igyekszem átfogó képet adni. Az ismertetés során részletesen kitérek mindazokra az elméleti és gyakorlati jellegű problémákra, amelyek a korpuszépítési munkát akadályozzák, illetőleg lassítják.

2. A HunOr korpusz létrehozásának a célja. – Párhuzamos korpuszoknak (parallel corpus) nevezzük azokat a két- vagy többnyelvű korpuszokat, amelyek forrásnyelvi szövegeket és azok fordításait tartalmazzák (vö. KÁROLY 2003: 19). A HunOr korpusz autentikus magyar nyelvű szövegeket, valamint azok orosz fordításait, illetve autentikus orosz nyelvű szövegeket, valamint azok magyar fordításait tartalmazza.

A korpusz létrehozására elsősorban azért törekszem, hogy vizsgálati anyagot teremtsek a magyar–orosz, illetve az orosz–magyar fordításkutatás számára. Ugyanakkor, mivel a korpusz nem csupán fordított, hanem autentikus szövegeket is tartalmaz mindkét nyelven, számos egyéb tudományterület kérdéskörébe tartozó nyelvészeti probléma számítógéppel támogatott vizsgálatát is lehetővé fogja tenni. A továbbiakban a korpusz fordításkutatásbeli hasznáról szólok részletesebben.

A fordításkutatás alapvető célja, hogy feltérképezze és rendszerbe foglalja a fordítás során alkalmazott műveleteket, majd ezek alapján hathatós segítséget nyújtson a fordítások készítőinek a lehető legmegfelelőbb fordítási megoldások kiválasztásához (vö. KLAUDY 2007: 13–5). Ahhoz, hogy a fordítási megoldásokról pontos képet kaphassunk, jelentős mennyiségű, különböző műfajba tartozó szöveg átvizsgálására lenne szükség, azonban egy ilyen vállalkozás manuálisan szinte kivitelezhetetlen feladat. Éppen ezért számos, a fordításkutatás területén tevékenykedő szerző hangoztatja a digitalizált párhuzamos korpuszokra való jelentős igényt (vö. KLAUDY 2001: 150–1, SZABÓMIHÁLY 2003: 64, ДОБРОВОЛЬСКИЙ et. al. 2005: 271–3, HORVÁTH 2008: 40). KLAUDY (2001: 150–1) véleménye szerint „a korpuszalapú megközelítés segítségével megalapozottan lehetne fordítási stratégiákat javasolni”,

[1] Köszönettel tartozom témavezetőimnek, dr. Bibok Károlynak és dr. Forgács Tamásnak értékes javaslataikért és támogatásukért, valamint kutatótársaimnak, Vincze Veronikának, Nagy T. Istvánnak és Schmalcz Andrásnak az előrevivő közös munkáért. Ugyancsak hálával tartozom a dolgozat lektorainak a hasznos megjegyzéseikért, valamint dr. Horváth Lászlónak a dolgozat végső változatának az elkészítéséhez nyújtott önzetlen segítségéért.

illetve az átváltási műveleteknek a műfajonkénti módosulása is feltárható lenne a korpusznyelvészet segítségével, kvantitatív módszerrel. A párhuzamos korpuszoknak a fordítás- és az idegennyelv-oktatásban való jelentős gyakorlati hasznát hangsúlyozza KOHN JÁNOS (1999). A szerző (i. m. 67–9) úgy látja, hogy az elektronikus párhuzamos szöveg lehet az az eszköz, amely mind a fordításkutatásban, mind az oktatásmódszertan területén – az összehasonlító szövegelemzésnek köszönhetően – új dimenziókat nyithat.

Jelentős gyakorlati hasznukból kifolyólag egyre bővül a párhuzamos szövegkorpuszok száma, s közöttük immáron hazai projektek keretében készült korpuszokat is találunk (Hunglish Korpusz; SzegedParalell korpusz).

Digitalizált magyar–orosz párhuzamos korpusz információim szerint eddig sem magyarországi, sem külföldi projekt keretében nem készült. Munkám keretében arra törekszem, hogy e hiányzó korpusz létrehozásával lehetőséget teremtsek a magyar és az orosz nyelvű, fordított és nem fordított szövegek nyelvi sajátságainak számítógéppel támogatott, összevető vizsgálatára. Ennek megfelelően a korpuszt majdan az érdeklődők számára szabadon hozzáférhetővé szándékozom tenni.

3. A HunOr korpusz építői. – Mielőtt rátérnék a HunOr korpusz bemutatására, említést kívánok tenni egy a munkát érintő üdvözlendő változásról.

Időközben a korpusz építéséhez a Magyar Tudományos Akadémia Mesterséges Intelligencia Kutatócsoportjának, valamint a Szegedi Tudományegyetem Informatikai Tanszékcsoportjának néhány munkatársa is csatlakozott az egyes programozási feladatok, valamint a munka során felmerülő magyar és orosz nyelvészeti problémák megoldása érdekében.

Magam elsősorban a magyar és az orosz nyelvet érintő elméleti kérdések vizsgálatával, valamint a korpuszban feldolgozható szövegek gyűjtésével és többsíkú manuális feldolgozásával foglalkozom. A dolgozat további részében a HunOr korpusz rövid bemutatása után ezeket a feladatokat ismertetem részletesen.

4. A HunOr korpusz szöveganyaga. – A korpusz teljes szövegállománya jelenleg valamivel több mint 130 000 szövegszós, azonban folyamatos bővítés alatt áll. A korpusz szövegei különböző típusú forrásokból származnak, többet közülük nyomtatott változatból digitalizálnom kellett.

A korpusz szövegállományának bővítése kapcsán érdemes megemlíteni, hogy a megfelelő szövegek megtalálása gyakran jelentős problémát okoz, ahogyan a szövegek felhasználására a jogtulajdonosok engedélyének a megszerzése is. Mindezeknek a következménye az a műfajonkénti és fordítási irányonkénti kiegyenlítetlenség, ami a HunOr szövegállományát jelen pillanatban jellemzi.

A HunOr a szövegműfajokat illetően jelenleg három kisebb egységre bontható: szépirodalmi, tudományos, valamint hivatalos alkorpuszra. Hamarosan azonban sajtónyelvi, a Ruszisztikai Központ Orosz Negyed című kiadványának szövegeivel bővítem a korpuszt.

Az 1. táblázat bemutatja a HunOr jelenlegi teljes szövegállományának az összefoglaló adatait:

1. táblázat
A HunOr Korpusz összefoglaló adatai

Szövegtípus	A szövegszavak száma		A fordítási irány
	orosz	**Magyar**	
1. Szépirodalmi	43 889	48 358	orosz → magyar
2. Tudományos	5 649	6 193	orosz → magyar
3. Hivatalos	13 744	12 515	magyar → orosz

A szépirodalmi alkotások közül a korpusz jelenleg a Kladbisenskie istorii című könyvet tartalmazza, amelynek szerzője a Magyarországon egyelőre csak álnéven, Borisz Akunyinként ismert Grigorij Cshartisvili. A könyv novellákat, valamint esszéket tartalmaz. A korpuszban található tudományos szövegek a szépirodalomhoz kapcsolódó, orosz forrásnyelvű elemző tanulmányok. A hivatalos alkorpusz a Magyar Külügyminisztérium honlapján közzétett, Tények Magyarországról című kiadvány egyes szövegeiből áll. (A korpusz jelenlegi szövegállományának adatairól részletesebben lásd SZABÓ et. al. 2011: 342.)

Terveink között szerepel, hogy a szöveggyűjtés folyamán egy-egy forrásnyelvi szöveghez több fordítást is felvegyünk a korpuszba, lehetőséget teremtve ezzel azok összevető vizsgálatára.

5. A szövegek feldolgozása. – A szövegek feldolgozásának a tervezett menete a következő: a szövegek előkészítését a szövegek mondatokra, valamint fordítási egységekre bontása, majd névelem-annotálása követi, végül a munkát a morfológiai és a szintaktikai annotálás zárja.

A) Az eddig elvégzett és a jelenleg folyó munka. – A korpuszba jelenleg tartozó összes szöveg előkészítése már megtörtént, tehát a szövegeket az elemzéshez szükséges egyszerűszöveg-formátumúra alakítottam, illetve a munka során a szövegekbe került hibákat manuálisan korrigáltam. Jelenleg a szövegek mondatokra és fordítási egységekre bontásán dolgozom.

Párhuzamos korpusz építésekor rendkívül fontos lépés a fordítási egységek megadása, másképpen szinkronizálás vagy mondatszintű párhuzamosítás; ez teszi lehetővé ugyanis a szövegekben való párhuzamos keresést, hiszen a segítségével együtt láthatjuk a forrás- és a célnyelvi mondatokat. A „mondatszintű párhuzamosítás" terminus kissé megtévesztő, hiszen a mondatszinten párhuzamosított korpusz a gyakorlatban – nevének látszólag ellentmondva – nem mondatokat, hanem fordítási egységeket feleltet meg egymással. VÁRADI TAMÁS (2001) a jelenséget a következőképpen magyarázza: azok az egységek, „amelyek a forrásnyelvben egyetlen mondatot jelentenek, a célnyelvben esetleg többet is. Így tehát a célnyelvi egységek adott esetben túlnyúlnak egy mondat határán" (VÁRADI 2001: 269). Bár VÁRADI megállapítása kétségtelenül igaz, valójában a fordítási egységben található megfelelésnek a szakirodalom (vö. KLAUDY 1997: 172–86, POHL 2003: 256, VINCZE et. al.

2010: 95–6) alapján hat típusát különböztethetjük meg; sőt magam a munka során egy hetedik típust (g) is detektáltam. A hét megfeleléstípus tehát a következő:

a) 1-1 megfelelés: egy forrásnyelvi mondatnak egy célnyelvi mondat felel meg;

b) 0-1 megfelelés, azaz a beszúrás: a célnyelvi mondat tartalmi megfelelője nem szerepel a forrásnyelvi szövegben;

c) 1-0 megfelelés, azaz a kihagyás: a forrásnyelvi mondatnak nincs tartalmi megfelelője a célnyelvi szövegben;

d) 1-N megfelelés (ahol $N \geq 2$), azaz a részekre bontás: a forrásnyelvi mondatnak egynél több célnyelvi mondat felel meg;

e) N-1 megfelelés (ahol $N \geq 2$), azaz az összevonás: egynél több forrásnyelvi mondatnak egy célnyelvi mondat felel meg;

f) N-M megfelelés (ahol N és $M \geq 2$, és N=M lehetséges), amely a mondathatár eltolódásából fakad: több forrásnyelvi mondatnak több célnyelvi mondat felel meg;

g) N=M megfelelés (ahol N és M = 2), amely a mondatok sorrendjének cseréjéből fakad: a forrásnyelvi szöveg két, (a) (b) sorrendű mondatának megfelelője a célnyelvű szövegben (b) (a) sorrendben található meg. Ennek következtében, bár az egységben a két nyelv mondatainak száma azonos, azokat egyetlen egységként kell kezelnünk a mondatok fordított sorrendje okán.

A párhuzamosság jelentősége miatt fontos lenne tehát, hogy a HunOr teljes szövegállományát szinkronizáljuk. Mivel azonban nincs tudomásunk olyan algoritmusról, amelyik a magyar és az orosz nyelvű szövegek szinkronizálását automatikusan elvégezné, a csoport arra vállalkozott, hogy létrehozza ezt az egyedülálló feldolgozó eszközt, illetve egy már meglévő, nyelvfüggetlen szinkronizáló programot fejleszt erre a célra tovább. Az algoritmus fejlesztéséhez, valamint a működésének teszteléséhez nagy mennyiségű referenciaszövegre van szükség. Magam jelenleg ennek létrehozásán dolgozom.

A referenciaszövegek három altípusra oszlanak: szinkronizált, mondatra bontott, valamint névelem-annotált szövegekre. A szinkronizált szövegek a HunOr jelenlegi szövegállományának egynegyed részét teszik ki, a mondatra bontás a teljes állományon folyik. A szinkronizált szövegek jelentősége abban áll, hogy segítségükkel mérhetővé válik a szinkronizáló program hatékonysága a manuálisan végzett munka eredményességének viszonylatában. A szinkronizált szövegek a mondatra bontott szövegekkel együtt pedig azért hasznosak, mert valamelyest képet adnak a korábban bemutatott megfeleléstípusok szövegműfajonkénti gyakorisági megoszlásáról. Ennek segítségével például amennyiben egy adott műfajú szövegben az 1-1-megfelelés a leggyakoribb, ott az automatikus munka során nagyban támaszkodhatunk majd a mondatvégi írásjelekre.

A névelem-annotálás a mondatra bontáshoz hasonlóan a HunOr teljes szövegállományát érinti. A művelet lényege a tulajdonnevek és a különféle azonosítók kigyűjtése és osztályokba sorolása (vö. VINCZE–FARKAS 2009). Ennek megfelelően tehát nem csupán a nyelvészetben is tulajdonnévnek tekintett elemek, hanem más

szövegbeli entitások azonosítása (pl. e-mail címek, rendszámok stb.) is beletartozik a számítógépes névelem-felismerés feladatkörébe.

A névelem-annotáció jelentősége abban áll, hogy a névelemek úgynevezett horgonyokként szolgálhatnak az algoritmus számára az egyes összetartozó szövegrészek megtalálásához. A munka keretében a személy-, a hely- és a szervezetneveket látom el különböző jelölésekkel, illetve az e három kategóriába be nem sorolható névelemek egy negyedik típusú jelölést kapnak.

A tulajdonnevek annotálása számos elméleti problémát vet fel. Mindenekelőtt nehézkes lehet már annak a meghatározása is, hogy mi számít egyáltalán tulajdonnévnek. Példaként említsük meg azokat a köznév és tulajdonnév határmezsgyéjén mozgó szavakat, amelyeknek az átmenetisége a köznévvé, illetve a tulajdonnévvé válás folyamatából fakad, például *Internet – internet* (vö. VINCZE–FARKAS 2009). A helyzetet tovább bonyolítja, hogy a tulajdonnév mibenlétének megítélése nyelvenként eltérő lehet. Így például az angolban a napok és a hónapok nevei, valamint a népnevek a tulajdonnevekre jellemzően nagy kezdőbetűvel írandók, a magyarban azonban egyértelműen köznévként kezelendők (vö. FARKAS 2007: 178–9, VINCZE–FARKAS 2009). (A tulajdonnevek meghatározásának itt említett és további problémáiról – például a kis és a nagy kezdőbetű kérdése, a tulajdonnevek terjedelmének, valamint a névtartozékok osztályozásának a problémája stb. – részletesen lásd FARKAS 2007, VINCZE–FARKAS 2009.)

A tulajdonnévi horgonyokat sikerrel alkalmazzák a különböző nyelvű szövegek automatikus párhuzamosításában (vö. POHL 2003), mivel a programok hatékonyan támaszkodnak az egymással megegyező nyelvi elemekre. Fontos szem előtt tartanunk azonban, hogy a forrásnyelvi tulajdonnevet a fordítók számos esetben jelentősen átalakítják, akár köszóval helyettesítik vagy kihagyják (vö. FARKAS 2007: 174), illetve természetesen az az eset is lehetséges, hogy a célnyelvi szövegben jelenik meg egy tulajdonnév ott, ahol a forrásnyelvi szövegben az nem szerepelt (például egy személyes névmás személynévvel történő fordításakor). Ezek a fordítói műveletek magától értetődően korlátozzák a tulajdonnevekre mint horgonyokra való támaszkodás lehetőségét (a tulajdonnevek fordítási problémáiról részletesen lásd VERMES 2005, HORVÁTH 2008).

A HunOr korpuszt illetően a tulajdonnevekre való támaszkodást jelentősen megnehezíti az a körülmény, hogy a feldolgozni kívánt szövegek nem azonos karakterkészletű nyelvekből származnak, hiszen a magyar nyelv a latin, az orosz nyelv a cirill ábécét használja. A tulajdonnevek tehát a HunOr korpusz különböző nyelvű szövegeiben nem azonos alakban fordulnak elő, aminek köszönhetően a két nyelv összetartozó elemeinek automatikus felismerése jóval nehezebb feladat, mint például ugyanez egy magyar–angol párhuzamos korpusz esetében. A helyzetet még bonyolultabbá teszi, hogy az orosz nyelvben az idegen tulajdonneveket nem azok forrásnyelvi betűzése, hanem részben azok kiejtése alapján írják át cirill betűkre, ahogyan azt az alábbi, a HunOr korpuszból származó példa is mutatja: *New York Times → Нью-Йорк Таймс* [Nju-Jork Tajmsz]. Mindezen túlmenően át kell hidalnunk azt a problémát is, hogy mind a magyar, mind az orosz nyelvben változnak a főnévi végződések a nyelvek ragozási sajátságai folytán, például m. *Anna – Annával*; or. *Anna – sz Annoj*. Ugyanakkor jelentős könnyebbség, hogy a köz- és a tulaj-

donnevekben a kezdőbetűk kis vagy nagy voltát illetően a két nyelvben nincs alapvető eltérés (vö. például a német nyelvvel, amelyben a nagy kezdőbetű a főnevek sajátsága attól függetlenül, hogy azok tulajdonnevek vagy köznevek-e), illetve, hogy a két nyelv központozási készlete és annak használati sajátságai alapvetően azonosak; ez ugyanis jelentős segítségül szolgálhat, különösen az 1-1 típusú megfelelések azonosításában.

B) A tervezett szintaktikai annotálás elméleti problémái. – A korpusz feldolgozásának utolsó tervezett lépése a szintaktikai annotálás. Mivel a korpuszok szintaktikai annotálása általánosságban véve is számos teoretikus problémát vet fel, már a munka ezen szakaszában folytatok vonatkozó kutatásokat.

A korpusz szintaktikai annotálását érintő problémákat alapvetően az okozza, hogy bármely elméleti alapot választjuk is meg a feladat végrehajtásához, a hevesen vitatott kérdések esetében elkerülhetetlenül felmerül az annotáció elméletfüggőségének a veszélye. Az elméletfüggőség azonban egy kutatóeszköz létrehozásakor elviekben szigorúan tilos; kívánatos ugyanis, hogy a korpuszokból kinyert információk ne egy bizonyos elméletet, hanem a nyelvi valóságot tükrözzék. A problémát még összetettebbé teszi, hogy esetünkben egy párhuzamos korpuszt kívánunk szintaktikailag annotálni. Ilyenkor ugyanis még kontrasztív nyelvészeti szempontokat is figyelembe kell venni a különböző nyelvű szövegekben végzett annotálás elméleti hátterének az azonosságához. A dolgozat keretei miatt most csupán két problémáról ejtek röviden szót.

1. Az orosz számneves frázisok kezelési módja. – A korpusz orosz nyelvű szövegeinek szintaktikai annotálását érintő egyik kardinális probléma a számneves frázisok kezelésének a módja. Az orosz számneves kifejezések viselkedését három tényező befolyásolja: egyrészt, hogy milyen számnév vesz részt a szerkezetben, másrészt, hogy a frázisban szereplő főnév [+élő] jegyű-e, harmadrészt, hogy a teljes számneves kifejezés milyen külső szintaktikai esetű pozícióban áll. Mindezek következtében a frázis hol homogén, hol heterogén szintaktikai viselkedést mutat az esetadási sajátságoktól függően, a 2. táblázatban bemutatottaknak megfelelően:

2. táblázat
Az orosz számneves frázisok morfoszintaktikai viselkedése

<table>
<tr><th rowspan="2">A frázisban szereplő számnév</th><th colspan="4">Morfoszintaktikai viselkedés</th></tr>
<tr><th colspan="3">Nem oblikuszi esetű pozícióban</th><th>Oblikuszi esetű pozícióban</th></tr>
<tr><td>ogyin (’egy’)</td><td colspan="4">Homogén</td></tr>
<tr><td>alacsony (vagy: paucal) számnevek: dva (’kettő’), tri (’három’), csetire (’négy’)</td><td></td><td>NOM.</td><td>AKK.</td><td>Homogén</td></tr>
<tr><td></td><td>[-élő]</td><td colspan="2">Heterogén</td></tr>
<tr><td></td><td>[+élő]</td><td>Heterogén</td><td>Homogén</td></tr>
<tr><td>magas számnevek: pjaty (’öt’) – dvadcaty (’húsz’)</td><td colspan="3">Heterogén</td><td>Homogén</td></tr>
<tr><td>tiszjacsa (’ezer’), million (’millió’), milliard (’milliárd’)</td><td colspan="4">Heterogén</td></tr>
</table>

A táblázatban bemutatottaknak megfelelően az *ogyin* sohasem ad genitívuszt az őt követő főnévnek, hanem vele nemben és számban mindig egyeztetve van. Ezzel ellentétben a *tiszjacsa*, a *million*, valamint a *milliard* genitívuszi esetadása kivétel nélküli, a főnévvel sohasem történik egyeztetés sem a nem oblikuszi, sem az oblikuszi esetű pozícióban. A skála említett végpontjai közé eső számnevek az ismertetett szintaktikai viselkedési módok olyan sajátos keverékét mutatják, amelyeket külső és belső tényezők egyaránt befolyásolnak. (A jelen dolgozat keretei nem teszik lehetővé, hogy az orosz számneves kifejezések szintaktikai viselkedését részleteiben, példák segítségével is bemutassam. Erre vonatkozóan FRANKS leírását – lásd 1994: 32–4 – ajánlom az érdeklődő olvasó figyelmébe.)

Az orosz számneves kifejezések szintaktikai szempontból igen bonyolult rendszere jelentős mennyiségű dolgozat tárgyát képezi az orosz nyelvről szóló szintaktikai szakirodalomban. Számos szerző tett már kísérletet arra, hogy valamiképpen elszámoljon e frázisok kategoriális státusával (vö. PESETSKY 1982, BABBY 1987, NEIDLE 1988, CORBETT 1993, FRANKS 1994, 1995, GIUSTI–LEKO 2001, ГРАЩЕНКОВ 2002, RAPPAPORT 2002, 2003, BAILYN 2003). Az élénk, napjainkban is folyó elméleti vita tétje az, hogy eldöntsük: mi az orosz számneves kifejezés feje, hogyan épül fel a számneves frázis szintaktikai struktúrája. Az orosz számneves frázisok szintaxisát érintő alapvető kérdések tehát a következők:

a) Lehetséges-e az, hogy – a fentebb ismertetett sajátságok alapján – például az *ogyin rubl* (’egy rubel’) és a *million rublej* (’[egy] millió rubel’) frázisoknak azonos kategoriális státust tulajdonítsunk annak ellenére, hogy szintaktikai sajátságaik jelentősen különböznek egymástól?

b) Hogyan számoljunk el a két „véglet” közé eső, szintaktikai sajátságaikat tekintve külső és belső tényezőktől egyaránt függő számneves kifejezésekkel?

Az alacsony és a magas számnevekkel alkotott frázisok szintaktikai státusát illetően a szakirodalomban semmiféle egységesség nincs. Egyes kutatók amellett érvelnek, hogy minden esetben – még az *ogyin* esetében is – a számnevet kell a frázis fejének tekinteni (vö. BABBY 1987). Más szerzők tipológiai vizsgálatokra hivatkozva a főnév kizárólagos fej volta mellett érvelnek úgy, hogy ez alól még a *tiszjacsa*, a *million* és a *milliard* számnevekkel alkotott frázisokat sem tekintik kivételnek (vö. ГРАЩЕНКОВ 2002). Azok a szerzők, akik szakítani kívánnak az egységes kezelési móddal, a megoldások széles skáláját vonultatják fel (vö. PESETSKY 1982, FRANKS 1994, BAILYN 2003).

A HunOr korpusz annotálását illetően a probléma részbeni megoldását jelentheti, ha az *ogyin*, valamint a *tiszjacsa*, a *million* és a *milliard* számnévvel alkotott frázisok kezelésében RAPPAPORT (2003: 2) alapján a következőképpen járunk el: az *ogyin* számnévvel alkotott frázist annak minden szintaktikai helyzetben mutatkozó tiszta melléknévi viselkedésére hivatkozva melléknévként kezeljük, s a frázis fejének a főnevet tekintjük. Ennek megfelelően a kifejezést NP-ként jelöljük a korpuszban. Ugyanakkor a *tiszjacsa*, a *million* és a *milliard* számnevekkel alkotott frázisokat azok önálló egyes és többes számú ragozási paradigmája, illetve kivétel nélküli esetadási sajátsága miatt a frázis fejének tekintjük, s a vele alkotott kifejezést QP-ként annotáljuk. Ez azonban csupán a probléma egy szegmensének a lehetséges megoldása: arra a kérdésre továbbra sem adtunk választ, hogy hogyan járjunk el az e két „véglet" közé eső számnevekkel alkotott frázisok esetében.

A probléma megoldásához az sem visz közelebb, ha figyelembe vesszük az orosz nyelv talán legreprezentatívabbnak tekinthető korpuszában, a Nacionalnij Korpusz Russzkovo Jazikaban alkalmazott vonatkozó kezelési módot. A korpuszban a számneves frázisokat a következőképpen annotálják a készítők: az *ogyin*nal, valamint az alacsony és a magas számnevekkel alkotott frázisokban mindig a főnév tölti be a fej szerepét, függetlenül a frázis esetadási sajátságaitól. Ugyanakkor a *tiszjacsa*, a *million* és a *milliard* számnevekkel álló kifejezésekben rendre a számnevek elemeztetnek fejként. Véleményem szerint megkérdőjelezhető a korpuszban alkalmazott megoldás helytálló volta, hiszen, ha az esetadási sajátságok okán különböző szintaktikai szerkezetet tesznek fel a korpusz építői az *ogyin*, valamint a *tiszjacsa*, a *million* és a *milliard* számnevekkel alkotott frázisok esetében, mire hivatkozva döntenek az alacsony és a magas számnevekkel álló kifejezések esetében az NP feltevésének a javára?

A bemutatott nyelvi sajátságok, az egymással vitatkozó elméletek, valamint az orosz nyelv nemzeti korpuszában alkalmazott, véleményem szerint teoretikus szempontból megkérdőjelezhető eljárásmód okán a HunOr korpusz szintaktikai annotálásához további alapos vizsgálatokat tartok feltétlenül szükségesnek.

2. A szintaktikai hiányok annotálásának problematikája. – Talán a legösszetettebb, a legkörültekintőbb vizsgálatot igénylő szintaktikai probléma a szintaktikai hiányok, azaz a fonetika szintjén nem realizálódó elemek annotálásának a kérdése. A szintaktikai hiányok korpuszbeli annotálását ugyanis számos jelentős probléma nehezíti, az elmélet és a gyakorlat szintjén egyaránt.

Annotálásuk elméleti vonatkozású nehézsége az, hogy jobbára nincs egységesen elfogadott feltevés a problémakör egyetlen részletét illetően sem: a szintaktikai hiányok mind az orosz, mind a magyar szintaktikai szakirodalomban egymással élesen vitázó elméletek tárgyát képezik. A problémát – a korábban említetteknek megfelelően – még összetettebbé teszi, hogy párhuzamos korpuszban kívánjuk végrehajtani az annotálást. A két nyelvre vonatkozó szintaxiselméleteket tehát még közös nevezőre is kellene valahogyan hozni.

A szintaktikai hiányok annotálásának gyakorlati nehézsége abban áll, hogy az automatikus módszerrel végzett szövegfeldolgozás során egyelőre nem megoldott ezeknek az elemeknek az algoritmussal való biztonságos felismertetése (vö. КОПОТЕВ 2007: 2–3). Ma már az egyre nagyobb méretű korpuszokra való igény miatt az annotálást teljes egészében manuálisan végrehajtani szinte lehetetlen, ezért a korpuszszövegek feldolgozása automatikus vagy félig automatikus módszerekkel történik. Gépi elemzés azonban csak akkor lehetséges, ha – egyszerűen fogalmazva – van mit elemezni; azaz, ha az elemeztetni kívánt nyelvi kifejezés jelen van a struktúrában.

Mivel a szintaktikai hiányok kezelése jelentős elméleti és gyakorlati problémát okoz a számítógépes szövegfeldolgozás számára, a korpuszok vonatkozó szintaktikai annotációjában jelentős hiányosság mutatkozik. Figyelemre méltó például, hogy a magyar nyelv legreprezentatívabbnak tekinthető korpuszában, a Magyar Nemzeti Szövegtárban nem találni a szintaktikai hiányok feldolgozását célzó annotációt. Emellett meglepően csekély azoknak a tanulmányoknak a száma, amelyek kísérletet tennének a probléma korpusznyelvészeti megközelítésére. Az említett két tényező erősen problematikus, hiszen ennek köszönhetően egyrészt egy, az elméleti nyelvészetben élénken vizsgált kérdéskör, az alkalmazott nyelvészet terén nem kap elegendő figyelmet, másrészt kétségkívül fontos lenne, hogy a szintaktikai hiányok problémáját korpuszok segítségével is vizsgálni lehessen. Amennyiben ugyanis egy adott korpuszban nem annotálnak szintaktikai hiányokat a készítők, úgy azok lekérdezésére csupán az úgynevezett közvetett lekérdezési módszer segítségével nyílik lehetőség. Ebben az esetben a kutató a listázni kívánt szintaktikai hiányok helyett azok valamely feltételezett szintaktikai környezetére keres rá. A közvetett módszer azonban közel sem tökéletes megoldás. Mindenekelőtt számos lekérdezés sok, az aktuális kutatás szempontjából irreleváns elemet is tartalmazhat. Emellett a keresés ideje igen hosszúra nyúlhat. Végül nem szabad megfeledkeznünk arról sem, hogy a szintaktikai hiányok egyes típusainak megtalálása ezzel a módszerrel egyáltalán nem is lehetséges (vö. КОПОТЕВ 2007: 3).

6. Összegzés. – Dolgozatomban a tudomásom szerinti első digitalizált magyar–orosz párhuzamos korpusz, a HunOr jelenleg folyó építéséről kívántam számot adni. A célom az volt, hogy átfogó képet adjak a korpuszépítés folyó és jövőbeli munkálatairól, illetve ezzel összefüggésben mindazokról az elméleti és gyakorlati jellegű problémákról, amelyek a korpuszépítési munkát akadályozzák, illetőleg lassítják.

Amint arra igyekeztem rávilágítani, a HunOr korpusz létrehozásához számtalan elméleti és gyakorlati jellegű probléma vár még megoldásra. Ugyanakkor bízom

abban, hogy a munka eredményeképpen sikerül létrehozni egy olyan egyedülálló vizsgálati eszközt, amely hathatós segítségül szolgálhat mind a magyar–orosz, mind az orosz–magyar fordításkutatás és -oktatás számára.

Források

Hunglish Korpusz. URL: http://mokk.bme.hu/resources/hunglishcorpus.

Magyar Nemzeti Szövegtár. URL: http://corpus.nytud.hu/mnsz.

Nacionalnij Korpusz Russzkovo Jazika (Национальный корпус русского языка). URL: http://www.ruscorpora.ru.

SzegedParalell korpusz. URL: http://www.inf.u-szeged.hu/rgai/corpus_paralell.

A hivatkozott irodalom

BABBY, LEONARD H. 1987. Case, Prequantifiers, and Discontinuous Agreement in Russian. Natural Language and Linguistic Theory 5: 91–138.

BAILYN, JOHN FREDERICK 2003. The Case of Q. In: ARNAUDOVA, OLGA ed. Formal Approaches to Slavic Linguistics 12. Ann Arbor, Michigan Slavic Publications, Michigan. 1–36.

CORBETT, GREVILLE G. 1993. The head of Russian numeral expressions. In: CORBETT, GREVILLE G. – FRASER, NORMAN M. – MCGLASMAN, SCOTT eds. Heads in grammatical theory. Cambridge University Press, Cambridge. 11–35.

FARKAS TAMÁS 2007. A tulajdonnevek fordíthatóságáról és napjaink fordítási hibáiról, közszók és tulajdonnevek példáján. Névtani Értesítő 29: 167–88.

FRANKS, STEVEN 1994. The Functional Structure of Russian Numeral Phrases. In: TOMAN, JINDŘICH ed. Formal Approaches to Slavic Linguistics 2. Ann Arbor, Michigan Slavic Publications, Michigan. 31–76.

FRANKS, STEVEN 1995. Parameters of Slavic Morphosyntax. Oxford University Press, Oxford.

GIUSTI, GIULIANA – LEKO, NEDZAD 2001. The categorial Status of Quantity Expressions. In: ZYBATOW, GERHILD – JUNGHANNS, UWE – MEHLHORN, GRIT – SZUCSICH, LUKA szerk. Current Issues in Formal Slavic Linguistics. Peter Lang, Frankfurt am Main. 96–105.

HORVÁTH PÉTER IVÁN 2008. Személynevek a szakfordításban. Névtani Értesítő 30: 35–40.

KÁROLY KRISZTINA 2003. Korpusznyelvészet és fordításkutatás. Fordítástudomány 5/2: 18−26.

KLAUDY KINGA 1997. A fordítás elmélete és gyakorlata. Angol / francia / német / orosz fordítástechnikai példatárral. Scholastica Kiadó, Budapest.

KLAUDY KINGA 2001. Mit tehet a fordítástudomány a magyar nyelv „korszerűsítéséért”? Magyar Nyelvőr 145–52.

KLAUDY KINGA 2007. Nyelv és fordítás. Válogatott fordítástudományi tanulmányok. Tinta Könyvkiadó, Budapest.

KOHN JÁNOS 1999: Párhuzamos szövegek számítógéppel segített elemzése a fordításoktatásban (1. rész). Fordítástudomány 1/1: 67–78.

NEIDLE, CAROL 1988. The Role of Case in Russian Syntax. Kluwer Academic Publishers, Dordrecht.

PESETSKY, DAVID 1982. Paths and Categories. Doctoral dissertation. MIT, Cambridge, Massachusetts.

POHL GÁBOR 2003. Szövegszinkronizációs módszerek, hibrid bekezdés- és mondatszinkronizációs megoldás. In: ALEXIN ZOLTÁN – CSENDES DÓRA szerk. Magyar Számítógépes Nyelvészeti Konferencia. Egyetemi Nyomda, Szeged. 254−9.

RAPPAPORT, GILBERT C. 2002. Numeral Phrases in Russian: A Minimalist approach. Journal of Slavic Linguistics 10: 329–42.

RAPPAPORT, GILBERT C. 2003. The Grammatical Role of Animacy in a Formal Model of Slavic Morphology. In: MAGUIRE, ROBERT A. – TIMBERLAKE, ALAN eds. American Contributions to the Thirteenth International Congress of Slavists. Bloomington, Slavica. 149–66.

SZABÓ MARTINA KATALIN – SCHMALCZ ANDRÁS – NAGY T. ISTVÁN – VINCZE VERONIKA 2011. A HunOr magyar–orosz párhuzamos korpusz. In: TANÁCS ATTILA – VINCZE VERONIKA szerk. VIII. Magyar Számítógépes Nyelvészeti Konferencia. JATEPress, Szeged. 341–7.

SZABÓMIHÁLY GIZELLA 2003. A szlovákiai magyar szakfordítások minőségének javításáról és az objektív fordításkritika megteremtésének feltételeiről. Fórum Társadalomtudományi Szemle 4: 55−68.

VÁRADI TAMÁS 2001. Kontrasztív szemantikai kutatások párhuzamos korpusz segítségével. In: GECSŐ TAMÁS szerk. Kontrasztív szemantikai kutatások. Tinta Könyvkiadó, Budapest. 268–76.

VERMES ALBERT PÉTER 2005. Proper names in translation: A relevance-theoretic analysis. Kossuth Egyetemi Kiadó, Debrecen.

VINCZE VERONIKA − FARKAS RICHÁRD 2009. Tulajdonnevek a számítógépes nyelvészetben. Kézirat.

VINCZE VERONIKA − FELVÉGI ZSUZSANNA − R. TÓTH KRISZTINA 2010. Félig kompozicionális szerkezetek a SzegedParalell angol–magyar párhuzamos korpuszban. In: TANÁCS ATTILA – VINCZE VERONIKA szerk. VII. Magyar Számítógépes Nyelvészeti Konferencia. Egyetemi Nyomda, Szeged. 91−101.

ГРАЩЕНКОВ, ПАВЕЛ ВАЛЕРЬЕВИЧ 2002. Родительный падеж при русских числительных: типологическое решение одной «сугубо внутренней» проблемы. Вопросы Языкознания 3: 74–119.

ДОБРОВОЛЬСКИЙ, ДМИТРИЙ ОЛЕГОВИЧ − КРЕТОВ, АЛЕКСЕЙ АЛЕКСАНДРОВИЧ − ШАРОВ, СЕРГЕЙ АЛЕКСАНДРОВИЧ 2005. Корпус параллельных текстов. Архитектура и возможности использования. In: Национальный корпус русского языка: 2003–2005. Результаты и перспективы. Индрик, Москва. 263–96.

КОПОТЕВ, МИХАИЛ ВЯЧЕСЛАВОВИЧ 2007. Разметка синтаксической неполноты в корпусе. Компьютерная лингвистика и интеллектуальные технологии. In: ИОМДИН, ЛЕОНИД Л. –ЛАУФЕР, НАТАЛИЯ И. – НАРИНЬЯНИ, АЛЕКСАНДР С. – СЕЛЕГЕЙ, ВЛАДИМИР П. ред. Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог 2007». РГГУ, Москва. 307–9.

SZABÓ MARTINA KATALIN (szabomartinakatalin@gmail.com)

Difficulties and Possibilities of Creating the First Hungarian–Russian Parallel (HunOr) Corpus

This paper is a progress report on a project aimed at creating the first Hungarian–Russian parallel corpus. I give an overall picture of the goal and the present state of the project, the observations taken in the course of the foregoing work, the tools and planned steps of the corpus's process, as well as the data of the HunOr corpus. In the course of the paper I give details of the practical and theoretical difficulties interfering with the project.

SZABÓ, MARTINA KATALIN