A HunOr magyar-orosz párhuzamos korpusz

Szabó Martina Katalin[1], Schmalcz András[2], Nagy T. István[2], Vincze Veronika[3]

[1]Szegedi Tudományegyetem, Magyar Nyelvészeti Tanszék
szabomartinakatalin@gmail.com
[2]Szegedi Tudományegyetem, Informatikai Tanszékcsoport
schmalcz.andras@stud.u-szeged.hu, nistvan@inf.u-szeged.hu
[3] SZTE-MTA Mesterséges Intelligencia Kutatócsoport
vinczev@inf.u-szeged.hu

Kivonat: A jelen dolgozatban a HunOr, egy eleddig hiányzó digitalizált magyar–orosz párhuzamos korpusz létrehozásáról számolunk be. A dolgozat a korpuszépítési munka céljáról, jelenlegi állásáról, az eddigi munka során szerzett tapasztalatokról, a munka folyamatáról és eszközeiről, valamint a HunOr korpusz adatairól igyekszik átfogó képet adni. Az ismertetés során részletesen szólunk azokról az elméleti és gyakorlati jellegű problémákról, amelyek az eddig elvégzett és a jelenleg folyó feldolgozási munkák (mondatra bontás, mondatszintű párhuzamosítás, NE-annotálás) során elméleti vagy gyakorlati szempontból megoldásra váró feladatként léptek fel.

1 Bevezetés

A HunOr korpusz autentikus magyar nyelvű szövegeket, valamint azok orosz fordításait, illetve autentikus orosz nyelvű szövegeket, valamint azok magyar fordításait tartalmazza. A korpusz létrehozásának elsődleges célja, hogy vizsgálati anyagot teremtsünk 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 korpusz mindemellett különféle számítógépes nyelvészeti alkalmazásokhoz, például a gépi fordításhoz is kitűnő segédletet biztosíthat.

2 A HunOr korpusz szöveganyaga

A korpusz feldolgozott szövegállománya jelenleg valamivel több mint 75 000 szövegszót tartalmaz, azonban folyamatos bővítés alatt áll. A korpusz szövegei különböző típusú forrásból (internetes kiadvány, könyvformátum stb.) származnak.

A HunOr a szövegműfajokat illetően három kisebb egységre bontható: szépirodalmi, tudományos, valamint hivatalos alkorpuszra. Hamarosan azonban reményeink

szerint sajtónyelvi, a Russzisztika Központ *Orosz Negyed* című kiadványainak szövegeivel is bővül a korpusz.

A szépirodalmi alkotások közül a korpusz jelenleg a *Kladbiščenskie istorii* című művet tartalmazza, amelynek szerzője a Magyarországon egyelőre csak álnéven, Borisz Akunyinként ismert Grigorij Cshartisvili. A novellákat és esszéket tartalmazó könyv 2005-ben jelent meg. A művet 2008-ban *Temetői történetek* címmel Bagi Ibolya és Sarnyai Csaba ültették magyar nyelvre. A korpuszban található tudományos szövegek a szépirodalomhoz kapcsolódó, orosz forrásnyelvű elemző tanulmányok: Nyikolaj Bergyaev egy hosszabb lélegzetű, 1990-ben, *O „večno-babjom" v russkoj duse* címen publikált művének egy részlete, valamint Vitalij Orlov *Hranitel „nenužnih veščej"* című, 1999-es tanulmánya. A fordításokat 2007-ben Régéczi Ildikó, valamint 2009-ben Józsa György Zoltán készítették. A hivatalos alkorpusz a Magyar Külügyminisztérium honlapján közzétett, *Tények Magyarországról* című kiadvány következő szövegeiből áll: *A magyar kultúra ezer esztendeje; Nemzeti jelképek, nemzeti ünnepek; Magyar Nobel-díjasok egy jobb világért.*

Az alábbi táblázat bemutatja a HunOr jelenlegi feldolgozott állományának összefoglaló adatait:

1. táblázat: A HunOr korpusz adatai.

Szövegtípus	Szövegszavak		Mondatok		Fordítási irány
	orosz	magyar	orosz	magyar	
Szépirodalom	52 798	57 980	3 255	3 313	orosz → magyar
Tudományos	7 014	7 483	360	348	orosz → magyar
Hivatalos	15 924	14 412	710	561	magyar → orosz
Összesen	75 736	79 875	4 325	4 222	

3 A korpusz feldolgozása

A korpusz későbbi hasznosíthatósága érdekében szükségesnek bizonyult a szövegek mondatokra bontása, mondatszintű párhuzamosítása, illetve – ez utóbbival összefüggésben – a szövegek tulajdonnévi annotálása.

3.1 A szövegek mondatokra bontása és mondatszintű párhuzamosítása

A korpusz mondatokra bontása, valamint mondatszintű párhuzamosítása szükségessé tette a mondatnak mint a két művelet alapegységének a pontos meghatározását.

A mondat meghatározásának a feladata korántsem triviális; problematikusak ugyanis az olyan kifejezések, amelyekben a kettősponttal záródó szerzői szavakat egy nagy kezdőbetűvel kezdődő idézet (egyenes beszéd), egy dialógus, egy önálló mondatokból álló felsorolás vagy egy kifejtő magyarázat követi. E szövegtípusok közül az idézés és a dialógus a szépirodalmi, a felsorolás és a kifejtő magyarázat pedig a tudományos és a hivatalos stílusú szövegek gyakori szerkesztésbeli sajátsága. A HunOr korpusz műfaji összetétele okán fontos feladat volt tehát, hogy egységes rendszert

dolgozzunk ki a kettősponttal szerkesztett kifejezések annotálásához. A probléma megoldásának céljából elvégeztük az említett szövegtípusok magyar és orosz helyesírási gyakorlatának összevető vizsgálatát, valamint áttekintettük a vonatkozó orosz és magyar irodalom megjegyzéseit [3, 11, 13, 14]. A tapasztaltak részletes bemutatásától a dolgozat keretei miatt most eltekintünk.

A kettőspont után kis kezdőbetűvel kezdődő kifejezések annotálása nem volt problematikus számunkra, azokat egységesen egy mondatba tartozónak jelöltük az előtte álló, kettősponttal végződő szerzői bevezetővel. A nagy kezdőbetűvel kezdődő, kettőspont után álló idézetek, dialógusok, felsorolások és leírások annotálása azonban már kérdéses volt. A kínálkozó lehetőségek a következők voltak:

a) a kettősponttal záródó kifejezést egy mondatként kezeljük az általa bevezetett mondattal; amennyiben a kettősponttal záródó kifejezést több mondatból álló szövegrész követi, úgy a szerző szavait egy mondatként kezeljük annak első mondatával, majd a többi mondatot önálló mondatokként annotáljuk;

b) a kettősponttal záródó kifejezést, valamint az általa bevezetett, egy vagy több mondatból álló szövegrészt együtt egyetlen mondatként kezeljük;

c) a kettősponttal záródó kifejezést önálló mondatként annotáljuk csakúgy, mint az általa bevezetett mondatot, vagy a több mondatból álló szövegrész minden egyes mondatát.

Vizsgáljuk meg a fenti szegmentálási lehetőségeket az alábbi példán [3] keresztül!

E vizsgálatoknak két formája terjedt el: Az egyik vizsgálati forma az oxitocinterheléses teszt. A méhkontrakciók csökkentik az uterus és az intervillózus tér véráramlását. A másik vizsgálati forma a fizikális terheléses teszt. Fizikai megterhelésre a vázizomzat vérátáramlása fokozódik, többek között a myometrium rovására.

A lehetséges mondatra bontási megoldások tehát a következők:

a) **<S>** E vizsgálatoknak két formája terjedt el: Az egyik vizsgálati forma az oxitocinterheléses teszt. **</S> <S>** A méhkontrakciók csökkentik az uterus és az intervillózus tér véráramlását. **</S> <S>** A másik vizsgálati forma a fizikális terheléses teszt. **</S> <S>** Fizikai megterhelésre a vázizomzat vérátáramlása fokozódik, többek között a myometrium rovására. **</S>**

b) **<S>** E vizsgálatoknak két formája terjedt el: Az egyik vizsgálati forma az oxitocinterheléses teszt. A méhkontrakciók csökkentik az uterus és az intervillózus tér véráramlását. A másik vizsgálati forma a fizikális terheléses teszt. Fizikai megterhelésre a vázizomzat vérátáramlása fokozódik, többek között a myometrium rovására. **</S>**

c) **<S>** E vizsgálatoknak két formája terjedt el: **</S> <S>** Az egyik vizsgálati forma az oxitocinterheléses teszt. **</S> <S>** A méhkontrakciók csökkentik az uterus és az intervillózus tér véráramlását. **</S> <S>** A másik vizsgálati forma a fizikális terheléses teszt. **</S> <S>** Fizikai megterhelésre a vázizomzat vérátáramlása fokozódik, többek között a myometrium rovására. **</S>**

Az (a) és a (b) megoldást támogatja a magyar és az orosz korpuszannotálási gyakorlat [4, 7, 12, 15], amely szerint minden kettőspontot tagmondatok közötti írásjelként annotálnak a készítők. A módszer azonban ellentmondásosnak tűnik, amennyiben szem előtt tartjuk Rozental [13] megjegyzését, miszerint az egyenes beszéd megfelel az önálló mondat szintaktikai kritériumainak, illetve azt, hogy mind a magyar, mind az orosz szerzők [3, 11, 14] különbséget tesznek az önálló mondatokból, valamint a nem önálló mondatokból álló felsorolások között. Amennyiben a korpuszannotálási gyakorlatot követnénk tehát, úgy kettő vagy több, szintaktikai szempontból önálló mondatot egyetlen mondatként jelölnénk be a korpuszban.

Az (a) megoldást támogatja továbbá az orosz helyesírási gyakorlat; az orosz szerzők ugyanis – a magyar gyakorlattal ellentétben [3] – nem ismerik el a kettőspontot mondatvégi írásjelként: a mondatzárók között rendre a pontot, a felkiáltójelet, a kérdőjelet, valamint a három pontot sorolják fel [11, 13, 14]. Amennyiben tehát az orosz helyesírási gyakorlathoz ragaszkodnánk, úgy a pontokat mondatvégi, a kettőspontokat pedig tagmondatok közötti írásjelként kezelnénk, azaz az (a) megoldást alkalmaznánk a korpuszban. Az eljárásmód vitatható volta azonban kiütközni látszik azokban az esetekben, ahol a szerző szavai több mondat vezetnek be. Véleményünk szerint ugyanis semmiféle különbség nem mutatkozik a szerző szavai és az azokat közvetlenül követő mondat, valamint a szerző szavai és az azokat nem közvetlenül követő mondat (vagy mondatok) között, ami alapul szolgálhatna ehhez a sajátos annotálási módhoz.

A (c) megoldást támogatják az (a) és a (b) megoldással szemben tett kritikai észrevételek, ugyanakkor a (c) annotálási mód ellen szól az említetteknek megfelelően a korpuszannotálási gyakorlat, valamint az, hogy az orosz nyelvben nem ismerik el a kettőspont esetleges mondatvégi státusát. Ugyanakkor grammatikáinkban nem találni olyan kritériumot, amely lehetetlenné tenné a kettősponttal végződő mondat feltevését, pl: „[A mondatot] a szerkesztés különféle nyelvtani eszközeinek viszonylagos lezártsága jellemez” [8]; „formai szempontból elsősorban az intonáció egysége, lezártsága jellemzi a magyar mondatot” [6]; „A mondat egy vagy több szóból áll, zárt intonációs szerkezet jellemzi” [2].

Az ismertetett érveket és ellenérveket megfontolva a HunOr korpuszban végül a (c) megoldás alkalmazása mellett döntöttünk. Az általunk választott eljárásmód tehát a következő: azokat a kettőspontokat, amelyek nagy kezdőbetűvel kezdődő, egy vagy több mondatból álló szövegrészt vezetnek be, mondatvégi írásjelekként kezeljük a korpuszban, s a kettősponttal végződő szerzői bevezető utáni mondatot vagy mondatokat önálló egységekként annotáljuk.

Az annotáció az elmondottak alapján tehát szakít a hazai és az orosz korpuszannotálási gyakorlattal. Ugyanakkor, mivel elméleti megfontolásokon alapszik, teoretikus szempontból a többi lehetséges megoldásnál helytállóbbnak tekinthető. Mindemellett érdemes kiemelni azt is, hogy a módszer az egységessége folytán nem teremt kérdéses eseteket, amelynek köszönhetően annak korpuszbeli alkalmazása mind az annotátori döntéshozatal, mind az automatikus munka szempontjából problémamentesen megoldható.

A mondatok párhuzamosításában a fordítási egység hatféle megfeleléstípusát szokás megkülönböztetni [1, 5, 10], a HunOr korpusz építése során azonban egy hetedik típust is detektáltunk ((g)-vel jelölve). A hét megfeleléstípus tehát a következő:

a) 1-1 megfelelés: egy forrásnyelvi mondat egy célnyelvi mondatnak felel meg;
b) 0-1 megfelelés, azaz a beszúrás;
c) 1-0 megfelelés, azaz a kihagyás;
d) 1-N megfelelés, azaz a részekre bontás;
e) N-1 megfelelés, azaz az összevonás;
f) N-M megfelelés, amely a mondathatár eltolódásából fakad;
g) N=M megfelelés, amely a mondatok sorrendjének a cseréjéből fakad: a forrásnyelvi szöveg két, (a) (b) sorrendű mondatának megfelelője a célnyelvű szövegben (b) (a) sorrendben található meg.

A hetedik megfeleléstípust az alábbi, a HunOr korpuszból származó példa szemlélteti:

Dombrovszkij ezt a verset igen szerette.

Kit vulkán edzett jó előre
S a Nemezis kezébe tett:
A bosszú kése vagy szabadság titkos őre,
Bírák bírája bűn és jogtiprás felett!

Лемносский бог тебя сковал
Для рук бессмертной Немезиды,
Свободы тайный страж, карающий кинжал,
Последний судия Позора и Обиды.

Это стихотворение Домбровский очень любил.

3.2 A tulajdonnévi annotálás

Az automatikus párhuzamosítást segítik a szövegben található horgonyelemek, például a számok és tulajdonnevek [9], így a szövegekben két független annotátor bejelölte a tulajdonneveket. Az annotáció során a négy klasszikus tulajdonnévosztályt alkalmaztuk: személy, szervezet, hely és egyéb. Az annotációk közti egyetértési ráta a magyar anyagon 0,8695 és 0,9609, az oroszon pedig 0,7995 és 0,9318 volt (κ-mértékben és mikro F-mértékben megadva). A tulajdonnevek kézi annotálása lehetővé teszi továbbá különféle magyar és orosz tulajdonnév-felismerő rendszerek teljesítményének mérését.

A 2. táblázatból kiderül, hogy a két nyelvben eltérő gyakorisággal fordulnak elő a tulajdonnevek, ami valószínűleg egyrészt nyelvek közti különbségeknek köszönhető: léteznek sajátos, csak az adott nyelvben tulajdonnévnek számító elemek, mint például az orosz *человечество*, melynek magyar megfelelője (*emberiség*) nem számít tulajdonnévnek. Másrészt a fordításnak köszönhetően stilisztikai különbségek is lehetnek a szövegek között: például az egyik nyelvben szereplő tulajdonnév helyett állhat névmás a másik nyelvű szövegben.

2. táblázat: A HunOr korpuszban található tulajdonnevek.

	orosz	**magyar**
Személy	1535	1487
Hely	608	479
Szervezet	137	105
Egyéb	291	224
Összesen	2571	2295

A HunOr korpusz esetében a horgonykeresést illetően több jelentős nyelvi tényezőt kell szem előtt tartanunk: Először is, az általunk feldolgozni kívánt szövegek nem azonos karakterkészletű nyelvekből származnak, hiszen a magyar nyelv a latin, az orosz nyelv a cirill ábécét használja. A tulajdonnevek tehát nem azonos írásmódban fordulnak elő, ami jelentős nehezítő körülmény például egy magyar–angol párhuzamos korpusz létrehozásához képest. További jelentős nehezítő körülmény, hogy az orosz nyelvben az idegen tulajdonneveket nem azok forrásnyelvi betűzése, hanem részben azok kiejtése alapján írják át cirill betűkre, pl. *New York Times* (angol) → *Нью-Йорк Таймс* [Nju Jork Tajms]; *François de la Chaise* (francia) → *Франсуа де ла Шез* [Fransua de la Šez]. E problémákra tehát fokozott figyelmet kell fordítanunk az automatikus párhuzamosítás során.

Ugyanakkor jelentős könnyebbség, hogy a köz- és a tulajdonnevekben a kezdőbetűk nagyságát illetően a két nyelvben nincs alapvető eltérés, illetve, hogy a két nyelv központozási készlete és annak használati sajátságai alapvetően azonosak.

4 A HunOr korpusz hasznosíthatósága

Az elkészült korpuszt a jövőben szeretnénk morfológiai és szintaktikai elemzésnek is alávetni. A morfológiailag és szintaktikailag elemzett párhuzamos korpusz minden bizonnyal kiemelkedő szerepet tölthet majd be a transzferalapú gépi fordítórendszerek fejlesztésében, de többnyelvű információkinyerésben is hasznosítható lesz, ugyanakkor a többszintű annotációnak köszönhetően (morfológia, szintaxis, névelemek) a két részkorpusz a magyar, illetve orosz nyelvű számítógépes nyelvészeti kutatásokat egyaránt ösztönözheti.

Köszönetnyilvánítás

A kutatás – részben – a MASZEKER kódnevű projekt keretében a Nemzeti Fejlesztési Ügynökség, illetve a TÁMOP-4.2.1/B-09/1/KONV-2010-0005 jelű projekt keretében az Európai Unió támogatásával, az Európai Regionális Fejlesztési Alap és az Európai Szociális Alap társfinanszírozásával valósult meg. Szabó Martina Katalin konferencián való részvétele a Szegedi Tudományegyetem Hallgatói Önkormányzata segítségével vált lehetségessé.

Bibliográfia

1. Klaudy K.: A fordítás elmélete és gyakorlata. Angol / francia / német / orosz fordítástechnikai példatárral. Scholastica Kiadó, Budapest (1997)
2. Kugler N.: A mondattan általános kérdései. In: Keszler B. (szerk.): Magyar Grammatika. Nemzeti Tankönyvkiadó, Budapest (2000) 369–393
3. Laczkó K., Mártonfi A.: Helyesírás. Osiris Kiadó, Budapest (2006)
4. Magyar Nemzeti Szövegtár [http://corpus.nytud.hu/mnsz/]
5. Pohl G.: Szövegszinkronizációs módszerek, hibrid bekezdés- és mondatszinkronizációs megoldás. In: Alexin Z., Csendes D. (szerk.): MSzNy 2003 – I. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2003) 254−259
6. Rácz E.: Mondattan. In: Rácz E. (szerk.): A mai magyar nyelv. Nemzeti Tankönyvkiadó, Budapest (1968) 205–458
7. Szeged Korpusz [http://www.inf.u-szeged.hu/projectdirs/hlt/]
8. Tompa J.: A mondat és a mondattan általános kérdései. In: Tompa J. (szerk.): A mai magyar nyelv rendszere. Leíró nyelvtan II. Akadémiai Kiadó, Budapest (1962) 7–22
9. 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
10. Vincze V., Felvégi Zs., R. Tóth K.: Félig kompozicionális szerkezetek a SzegedParalell angol−magyar párhuzamos korpuszban. In: Tanács A., Vincze V. (szerk.): MSzNy 2010 – VII. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2010) 91−101
11. Лопатин, В.В. ред.: Правила русской орфографии и пунктуации. Полный академический справочник. Издательство «Эксмо», Москва (2007)
12. Национальный корпус русского языка [http://www.ruscorpora.ru/]
13. Розенталь, Д.Э.: Русский язык. Пособие для поступающих в вузы. Издание второе, дополненное и переработанное. Московский университет, Москва (1988)
14. Соловьев, Н.В.: Орфографический словарь. Комментарий. Правила. 3-е издание. Издательство «Норинт», Сацкт-Петербург (2000)
15. ХАНКО [http://www.ling.helsinki.fi/projects/hanco/]