SzegedKoref: kézzel annotált magyar nyelvű koreferenciakorpusz

Vincze Veronika[1,2], Hegedűs Klára[3], Farkas Richárd[1]

[1]Szegedi Tudományegyetem, Informatikai Tanszékcsoport
Szeged Árpád tér 2., e-mail: {vinczev,rfarkas}@inf.u-szeged.hu
[2]MTA-SZTE Mesterséges Intelligencia Kutatócsoport
[3]Szegedi Tudományegyetem, Pszichológiai Intézet
e-mail: klarahegedus92@gmail.com

Kivonat Ebben a munkában bemutatjuk a SzegedKoref nevű, magyar nyelvű, teljes egészében kézzel annotált koreferenciakorpuszunkat, amely nagy méretének köszönhetően a későbbiekben alkalmas lehet különféle koreferenciafeloldó algoritmusok tanítására és kiértékelésére is. Ismertetjük a korpusz felépítését, az annotációs elveket, majd statisztikai adatokat közlünk az annotált nyelvi jelenségekről.
Kulcsszavak: koreferencia, anafora, korpusz

1. Bevezetés

A természetes nyelven írt szövegekben általában megfigyelhető, hogy a szerzők szóhasználatukban változatosságra törekednek, kerülik az ismétléseket, többféle kifejezést használnak ugyanarra az entitásra. A nyelvek mind lexikai, mind grammatikai síkon is lehetőséget biztosítanak erre a változatosságra. A grammatikai eszközök közé sorolhatjuk a koreferenciajelenségeket, azaz az anaforát és a kataforát. Anaforának nevezzük azt a jelenséget, amikor a szövegen belül egy elem visszautal egy másik elemre (antecedensre) oly módon, hogy a két elem azonos entitást jelöl, azaz koreferensek. Ennek ellentéte a – jóval ritkábban előforduló – katafora, amikor egy elem a szövegen belül előreutal egy később előforduló elemre. Koreferenciaviszonyokat leggyakrabban névmások, határozószók és bizonyos főnevek (például személyek esetén általában nemet vagy rangot jelölő főnevek) fejeznek ki. Lexikai szinten pedig többnyire szinonimák alkalmazása segíti a szóhasználatbeli változatosság elérését.

A szövegek jelentésének megértéséhez szükséges annak ismerete, hogy a szövegbeli egyedek a világ mely egyedeire referálnak, illetve melyek azok a szövegbeli egyedek, amelyek azonos egyedre utalnak a világban. A számítógépes nyelvészetben egyrészt a normalizálás feladata a szövegbeli egyedek egységes formára hozása (például az *OTP*, *Országos Takarékpénztár* és *OTP Bank* kifejezések egymáshoz rendelése), másrészt pedig a koreferenciafeloldás segítségével lehetséges meghatározni az azonos egyedre utaló szövegrészeket. Míg a normalizálás általában névelemekre alkalmazott eljárás, így azzal a sajátsággal rendelkezik, hogy a szövegben előforduló minden egyes *OTP Bank* kifejezés az *Országos*

Takarékpénztárra utal, addig a koreferenciafeloldás nem csak tulajdonnevekre alkalmazható, mivel a szövegbeli összes antecedens azonosítása fontos részfeladat nyelvtechnológiai célalkalmazások (főleg az információkinyerés) számára, továbbá ugyanannak az anaforikus elemnek akár mondatról mondatra is változhat az antecedense (például személyes névmások esetében).

Ebben a munkában bemutatjuk a SzegedKoref nevű, magyar nyelvű, teljes egészében kézzel annotált koreferenciakorpuszunkat, amely nagy méretének köszönhetően a későbbiekben alkalmas lehet különféle koreferenciafeloldó algoritmusok tanítására és kiértékelésére is. Cikkünkben ismertetjük a korpusz felépítését, az annotációs elveket, majd statisztikai adatokat közlünk az annotált nyelvi jelenségekről.

2. Kapcsolódó irodalom

A világ számos nyelvére létezik koreferenciára annotált korpusz, például az OntoNotes adatbázis [1, 2] angol, kínai és arab nyelvre tartalmaz koreferenciaannotációt. Ez a adatbázis szolgált a CoNLL-2011 [3] és CoNLL-2012 [4] versenyek alapjául, ahol a feladat automatikus koreferenciafeloldás volt.

Francia és német nyelvre beszélt nyelvi korpuszokban, a DIRNDL és ANCOR_Centre korpuszokban találhatunk koreferenciaannotációt [5,6]. Japán nyelven a NAIST Text korpusz tartalmaz koreferenciajelölést, a predikátum-argumentum viszonyok jelölése mellett [7]. Lengyel nyelvre is készült nagyméretű, koreferenciaannotált korpusz [8, 9], emellett holland [10] és cseh [11] nyelvekre is elérhetők korpuszok.

Magyar nyelvre is készült már egy kisméretű, kézzel annotált koreferenciakorpusz [12]. Jelen cikkben egy nagyméretű, teljes egészében kézzel annotált magyar nyelvű koreferenciakorpusz elkészítését ismertetjük, mely a későbbiekben méreténél fogva alkalmas lehet gépi tanuláson alapuló koreferenciafeloldó rendszerek tanítására és kiértékelésére is, ami a manapság legelterjedtebb eljárás koreferenciaviszonyok azonosítására (vö. [4]).

Morfológiailag gazdag nyelvek esetében – mint amilyen a magyar is – a koreferenciaviszonyok jelölése nehézségekbe ütközhet bizonyos nyelvi jelenségek kapcsán. Többek között a fonológiailag meg nem jelenő személyes névmások kezelése igényel különös figyelmet, vö. [13] a lengyel nyelvben tapasztalt nehézségekről. Emellett az utalószavak és mellékmondatok kapcsolatának jelölésére is külön figyelmet kell fordítani. Cikkünkben erről a két jelenségről is szót ejtünk.

3. A korpusz

Az annotálás alapjául a Szeged Korpuszt [14] választottuk, újabb kézi annotációs réteggel bővítve a szövegeket. Mivel koreferenciaviszonyokat hosszabb, összefüggő szövegekben érdemes vizsgálni, a Szeged Korpuszon belül is ki kellett választani, mely alkorpuszokban hasznos bejelölni a koreferenciakapcsolatokat. A gazdasági rövidhíreket tartalmazó alkorpuszban a hírek pusztán 1-2 mondatból állnak, így úgy döntöttünk, ezen az alkorpuszon nem végezzük el az annotálást.

Az annotálási munkálatok jelenleg is folyamatban vannak. 2014 novemberéig 181 dokumentum (újsághír, illetve iskolai fogalmazás) annotációja készült el, azonban ez a szám folyamatosan nő. A teljes koreferenciakorpusz tehát – az eddig elkészült anyagokon túl – további újságcikkeket, regényeket, jogi és számítástechnikai szövegeket, valamint iskolai fogalmazásokat fog tartalmazni.

4. Annotációs elvek

Az annotáció során összekötjük az antecedenseket és a velük koreferens elemeket. Jelöljük a névmási, főnévi, határozószói és igei anaforákat is, ahogy a következő példák is mutatják:

- Névmási anafora
  - Személyes névmás: ***Mari*** *észrevette Józsit, de a fiú nem látta* ***őt****.*
  - Mutató névmás: *Megvettem* ***a labdát****, de* ***az*** *hamarosan kidurrant.*
  - Kölcsönös névmás: ***Józsi és Mari*** *látta* ***egymást****.*
  - Visszaható névmás: ***Józsi*** *látta* ***magát*** *a tükörben.*
  - Vonatkozó névmás: *Ismertem* ***a lányt****,* ***aki*** *épp átjött az úton.*
  - Birtokos névmás: ***Józsi*** *nem tudta eldönteni, melyik labda az* ***övé****.*
  - Zéró névmás: ***A tanárok*** *látták előre* ***a konfliktust****, de (****ők****) nem tudták megakadályozni (****azt****).*
    Józsi *bejött a szobába. A(z* ***ő****) kutyája követte.*
- Főnévi anafora (NP)
  - Ismétlés: *Józsi este találkozott* ***a lánnyal****.* ***A lány*** *piros ruhát viselt.*
  - Variáns: ***Pálffy János gróf*** *személyében magyar főparancsnokot neveztek ki a császári sereg élére.* ***Pálffy*** *tárgyalásokat kezdett Károlyi Sándor báróval.*
  - Szinonima: *Józsi kapott* ***egy biciklit****. Másnap* ***az új kerékpárral*** *jött munkába.*
  - Hipernima: *Az udvaron volt* ***egy kutya****.* ***Az állat*** *keservesen ugatott.*
  - Hiponima: *Az udvaron volt* ***egy kutya****.* ***Szegény uszkár*** *meg volt kötve.*
  - Meronima: *Jól játszott* ***a csapat****,* ***a kapus*** *különösen kiemelkedett a mezőnyből.*
  - Holonima: *Defektes lett* ***a jobb első kerék****, így* ***az autónak*** *ki kellett állnia a versenyből.*
  - Epitheton: ***Józsi*** *nem tudott bejutni, mert* ***a szerencsétlen*** *otthon hagyta a kulcsot.*
  - Appozíció: *Pálffy tárgyalásokat kezdett* ***Rákóczi megbízottjával****,* ***Károlyi Sándor báróval****.*
- Határozói anafora:
  - Mutató határozószó: *Elindultunk* ***a hotelba****, a többiekkel* ***ott*** *találkozunk.*
  - Vonatkozó határozószó: *Hol jársz* ***itt****,* ***ahol*** *a madár se jár?*
- Igei anafora: *Juli* ***elénekelt*** *tegnap* ***egy dalt****, ma pedig Józsi is* ***így tett****.*
- Anafora képzett alakokkal: *Józsi mindig* ***énekel*** *a fürdőben.* ***Az éneklés*** *nagyon zavarja a többi lakót.*
  Józsi mindig ***énekel*** *a fürdőben. Az* ***éneklő*** *férfi nagyon zavarja a többi lakót.*

Az annotáció során a fenti fő kategóriákat jelöltük a szövegben. Főnévi anafora esetében jelöltük az altípust is, a névmási és határozói anaforák esetében azonban a szavak morfológiai elemzéséből kiderül, hogy melyik altípusról van szó, ezek külön jelölését tehát mellőztük.

Magyar nyelvű szövegekben az anaforák bejelölését nehezítik az ún. zéró névmások. Az alanyi és tárgyas igeragozás különbségének megléte folytán nem szükséges kitenni a tárgyi névmásokat, illetve az alanyt jelző személyes névmás kitétele sem kötelező, sőt birtokos szerkezetben is elmarad(hat) a személyes névmási birtokos. A koreferenciaviszonyok szempontjából ez annyit tesz, hogy az anaforikus elem látható formában nincs jelen a mondatban, csak zéró névmás (pro) formájában, így azokat az annotáció megkezdése előtt be kellett illeszteni a szövegbe. Egy példa:

Látta a kertjében. → ***proSUBJ*** *látta* ***proOBJ*** *a* ***proPOSS*** *kertjében.*

A zéró anaforikus névmások beszúrása a szövegbe automatikusan, morfológiai és szintaktikai megkötésekre épülő nyelvészeti szabályok alapján történt.

Külön figyelmet fordítottunk arra is, hogy a mellékmondatokra vonatkozó utalószavak is össze legyenek kötve az adott mellékmondattal, akár teljes alakban, akár zéró névmás formájában jelennek meg. Így tehát az alábbi példák mindegyikében jelöltük a névmás és a mellékmondat kapcsolatát:

Mondtam ***proOBJ****, hogy mindjárt itt a karácsony.*
Azt *mondtam, hogy mindjárt itt a karácsony.*

Az alábbiakban közlünk egy példát az annotált szövegre, indexekkel jelölve az összetartozó elemeket, illetve külön szerepeltetve az anaforikus láncokat.

Az úton **[sok ismerőssel]**$_i$ találkoztunk, **[akik]**$_i$ újságolták **[proOBJ]**$_j$ nekünk, hogy **[milyen jó a hangulat a majálison]**$_j$. Amikor leérkeztünk, már nagy volt a nyüzsgés, finom illatok szálltak a levegőben, és folytak **[a koncert]**$_k$ előkészületei, ugyanis – ha még nem írtam **[proOBJ]**$_l$ volna – **[a [Bestiák]**$_m$ **énekeltek aznap nekünk]**$_l$. Én ugyan nem nagyon szeretem **[ezt az együttest]**$_m$, de **[miattuk]**$_m$ nem hagyhattam ki **[ezt az eseményt]**$_k$. Amíg **[a koncert]**$_k$ nem kezdődött el, addig édességet ettünk a haverjaimmal, és hülyéskedtünk. Aztán egyszer csak **[sipító hangot]**$_n$ hallottunk. **[Azt]**$_o$ hittük, hogy **[a Bestiák**$_m$ **egyik énekese [az]**$_n$**]**$_o$, de **[proSUBJ]**$_p$ kiderült, hogy **[csak egy mikrofon hibásodott meg]**$_p$. Rövid várakozás után végül elkezdődött **[a koncert]**$_k$. **[A hangulat]**$_q$ a **[proPOSS]**$_q$ tetőfokára hágott, **[mindenki]**$_r$ tombolt, és együtt **[proSUBJ]**$_r$ énekelt **[a lányokkal]**$_m$. Több visszatapsolás és ráadás-dal után véget ért **[a koncert]**$_k$.

Anaforikus láncok:

sok ismerőssel – akik
proOBJ – milyen jó a hangulat a majálison

a koncert – ezt az eseményt – a koncert –a koncert – a koncert
proOBJ – a Bestiák énekeltek aznap nekünk
Bestiák – ezt az együttest – miattuk – Bestiák – a lányokkal
sipító hangot – az
azt – a Bestiák egyik énekese az
proSUBJ – csak egy mikrofon hibásodott meg
a hangulat – proPOSS
mindenki – proSUBJ

5. Statisztikai adatok

A korpusz jelenleg 309 mondatot és 9782 tokent tartalmaz az újsághírekből, illetve 3712 mondatot és 45981 tokent az iskolai fogalmazásokból, összesen 4021 mondat és 55763 token szerepel tehát a korpusz 2014 novemberi változatában. Ezekben összesen 2456 anaforikus lánc található (2191 az iskolai fogalmazásokban, 265 pedig az újsághírekben). Az anafora típusa szerinti (százalékos) eloszlást az 1. táblázat mutatja.

1. táblázat. Az anaforatípusok eloszlása.

Anafora	**Fogalmazás**	**%**	**Újsághír**	**%**	**Összesen**	**%**
névmási	1531	33,51	320	39,22	1851	34,37
ismétlés	1176	25,74	86	10,54	1262	23,44
szinonima	329	7,20	252	30,88	581	10,79
hipernímia	445	9,74	0	0,00	445	8,26
holonímia	350	7,66	34	4,17	384	7,13
epitheton	17	0,37	23	2,82	40	0,74
appozíció	117	2,56	70	8,58	187	3,47
határozói	339	7,42	1	0,12	340	6,31
igei	5	0,11	0	0,00	5	0,09
képzés	76	1,66	30	3,68	106	1,97
egyéb	184	4,03	0	0,00	184	3,42
Összesen	4569	100	816	100	5385	100

A táblázatból látszik, hogy a névmási anafora és az ismétlés a leggyakoribb anaforatípusok, e két kategória együttesen lefedi az adatok mintegy felét. Így tehát az automatikus koreferenciafeloldó rendszereknek e kategóriákra fokozott figyelmet kell fordítaniuk.

A 2. táblázat azt is elárulja, hogy a szövegekben számos zéró névmás szerepel anaforikus lánc részeként, sőt a névmási anaforák jelentős részében (mintegy kétharmadában) zéró névmás szerepel. Így a magyar nyelvű koreferenciafeloldó algoritmusoknak ezeknek a kezelésére is célszerű felkészülniük.

2. táblázat. Az anaforikus zéró névmások eloszlása.

Zéró névmás	**Fogalmazás**	**Újsághír**	**Összesen**
proSUBJ	594	119	713
proOBJ	181	9	190
proPOSS	212	128	340
Összesen	987	256	1243

6. Alkalmazási lehetőségek

A koreferenciaviszonyokra annotált korpusz, illetve a rá épülő automatikus koreferenciafeloldó rendszer felhasználási lehetőségei számos területre terjednek ki. A koreferenciaviszonyok információkinyerő rendszerek számára is hasznosak, hiszen például egy adott cégről szóló információkat nemcsak a cég nevére keresve lehet így megtalálni, hanem a cégre anaforikusan utaló elemek kikeresésével is többletinformációkra lehet szert tenni.

Fordítóprogramok is hasznosíthatják a bejelölt koreferenciakapcsolatokat, hiszen például míg a magyarban nincsenek nyelvtani nemek, addig számos nyelvben léteznek. Ha egy magyar névmás össze van kapcsolva antecedensével, ennek segítségével meg lehet határozni, hogy az idegen nyelven hímnemű, nőnemű vagy semlegesnemű névmás felel-e meg neki.

7. Összegzés

Ebben a munkában bemutattuk a SzegedKoref korpuszt, melyben kézzel megjelöltük a koreferenciaviszonyokat. Példákon keresztül ismertettük az annotálás alapelveit, illetve statisztikai adatokat közöltünk az elkészült anyagról. A jövőben szeretnénk a korpuszt bővíteni, illetőleg az annotált anyagra építve egy automatikus koreferenciafeloldó rendszert létrehozni.

Az annotált korpuszt kutatási és oktatási célokra ingyenesen elérhetővé tesszük.

Köszönetnyilvánítás

Szeretnénk megköszönni Miháltz Mártonnak, Anders Björkelundnak és Arndt Riesnernek az annotációs elvek kialakításában nyújtott önzetlen segítségüket.

A jelen kutatás a futurICT.hu nevű, TÁMOP-4.2.2.C-11/1/KONV-2012-0013 azonosítószámú projekt keretében az Európai Unió támogatásával és az Európai Szociális Alap társfinanszírozásával valósult meg.

Hivatkozások

1. Weischedel, R., Hovy, E., Marcus, M., Palmer, M., Belvin, R., Pradan, S., Ramshaw, L., Xue, N.: OntoNotes: A Large Training Corpus for Enhanced Processing. In: Handbook of Natural Language Processing and Machine Translation. (2011)

2. Pradhan, S.S., Ramshaw, L., Weischedel, R.M., MacBride, J., Micciulla, L.: Unrestricted Coreference: Identifying Entities and Events in OntoNotes. In: ICSC, IEEE Computer Society (2007) 446–453
3. Pradhan, S., Ramshaw, L., Marcus, M., Palmer, M., Weischedel, R., Xue, N.: CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes. In: Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task. CONLL Shared Task '11, Stroudsburg, PA, USA, Association for Computational Linguistics (2011) 1–27
4. Pradhan, S., Moschitti, A., Xue, N., Uryupina, O., Zhang, Y.: CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In: Proceedings of the Sixteenth Conference on Computational Natural Language Learning (CoNLL 2012), Jeju, Korea (2012)
5. Muzerelle, J., Lefeuvre, A., Schang, E., Antoine, J.Y., Pelletier, A., Maurel, D., Eshkol, I., Villaneau, J.: ANCOR_Centre, a large free spoken French coreference corpus: description of the resource and reliability measures. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S., eds.: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, European Language Resources Association (ELRA) (2014) 843–847 ACL Anthology Identifier: L14-1169.
6. Björkelund, A., Eckart, K., Riester, A., Schauffler, N., Schweitzer, K.: The Extended DIRNDL Corpus as a Resource for Coreference and Bridging Resolution. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S., eds.: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, European Language Resources Association (ELRA) (2014) 3222–3228 ACL Anthology Identifier: L14-1683.
7. Iida, R., Komachi, M., Inui, K., Matsumoto, Y.: Annotating a Japanese Text Corpus with Predicate-Argument and Coreference Relations. In: Proceedings of the Linguistic Annotation Workshop, Prague, Czech Republic, Association for Computational Linguistics (2007) 132–139
8. Ogrodniczuk, M., Kopeć, M., Savary, A.: Polish Coreference Corpus in Numbers. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S., eds.: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, European Language Resources Association (ELRA) (2014) 3234–3238 ACL Anthology Identifier: L14-1066.
9. Ogrodniczuk, M., Głowińska, K., Kopeć, M., Savary, A., Zawisławska, M.: Polish coreference corpus. In: 6th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics. Volume 3., Wydawnictwo Poznańskie (2013) 494–498
10. Hendrickx, I., Bouma, G., Coppens, F., Daelemans, W., Hoste, V., Kloosterman, G., Mineur, A.M., Vloet, J.V.D., Verschelde, J.L.: A Coreference Corpus and Resolution System for Dutch. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Tapias, D., eds.: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco, European Language Resources Association (ELRA) (2008) http://www.lrec-conf.org/proceedings/lrec2008/.
11. Nedoluzhko, A., Mírovskỳ, J., Ocelák, R., Pergler, J.: Extended coreferential relations and bridging anaphora in the Prague Dependency Treebank. In: Proceedings

of the 7th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2009), Goa, India. (2009) 1–16
12. Miháltz, M.: Tudásalapú koreferencia- és birtokosviszony-feloldás magyar szövegekben. Általános Nyelvészeti Tanulmányok **XXIV** (2012) 151–166
13. Ogrodniczuk, M., Głowińska, K., Kopeć, M., Savary, A., Zawisławska, M.: Interesting Linguistic Features in Coreference Annotation of an Inflectional Language. In Sun, M., Zhang, M., Lin, D., Wang, H., eds.: Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. Volume 8202 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2013) 97–108
14. Csendes, D., Csirik, J., Gyimóthy, T., Kocsor, A.: The Szeged TreeBank. In Matousek, V., Mautner, P., Pavelka, T., eds.: Proceedings of the 8th International Conference on Text, Speech and Dialogue, TSD 2005. Lecture Notes in Computer Science, Berlin / Heidelberg, Springer (2005) 123–132