Élő vagy élettelen?

Sass Bálint

MTA Nyelvtudományi Intézet és PPKE ITK MMT Doktori Iskola
e-mail: joker@nytud.hu

Kivonat Hogyan lehet megállapítani az igei keretek alanyi pozíciójának élő vagy élettelen voltát? A kidolgozott módszer az igei személyragok eloszlását, valamint az élőre és élettelenre utaló vonatkozó névmások arányát veszi tekintetbe. Az élettelen alanyú keretek 70%-át megtalálja, miközben szinte sosem határoz meg élő alanyú keretet élettelenként. A nyerhető igelistát egy magyar-angol fordítórendszer lexikai erőforrásába építve arra használjuk, hogy a pro-drop magyar mondatok fordításakor a „semmiből" megfelelő testes névmást generáljunk az angol oldalon.

Kulcsszavak: élő, élettelen, gépi fordítás, pro-drop

1. Bevezetés

Hogyan fordítanánk angolra az alábbi két magyar mondatot?

1. *Alszik.*
2. *Elromlott.*

Valószínűleg legtöbben a következő angol megfelelőket tartanák természetesnek, legalábbis abból a szempontból, hogy automatikusan az ige szemantikájának megfelelő élőre illetve élettelenre utaló névmást használnák:

1. *He/she is sleeping.*
2. *It has gone wrong.*

Általánosan fogalmazva arról a kérdésről van tehát szó, hogy a gépi fordítás során mit tehetünk olyan esetekben mikor a forrásnyelv nem specifikál bizonyos tulajdonságokat, jegyeket, a célnyelv viszont ugyanazon a ponton elvárja a tulajdonság egy konkrétan megadott értékét. Az egyik lehetőség, hogy dinamikusan megkíséreljük kitalálni a szövegkörnyezetből az elvárt értéket, a most bemutatandó másik lehetőség pedig az, hogy a lexikonba bekódolt alapértelmezett értékeket használunk. Egyértelmű esetekben ez a módszer hibátlan megoldást ad futásidejű számítási igény nélkül. A javasolt eljárás tehát leegyszerűsítve az lesz, hogy nagyméretű korpuszban mért gyakoriságok alapján megbecsüljük a jegy alapértelmezett értékét, rögzítjük a lexikonban, és ezt az értéket használjuk akkor, ha nincs információnk a jegy aktuális értékéről, esetünkben az alany élő vagy élettelen voltáról.

2. Az élőségi skála jelentősége

Az *élőségi* (vö: *animacy*) skála (vagy élő/élettelen skála) a nyelvi prominenciaviszonyokat meghatározó egyik tényező, sok esetben valamely elem élő illetve élettelen volta szerint választunk két nyelvi forma között [1]. A megértés szempontjából központi szerepe van, lehetővé teszi, hogy a dialógusban követni tudjuk, hogy éppen melyik szereplőről van szó [2]. Univerzálisan kimondható, hogy az egyes szereplők élőségi skálán elfoglalt helye arányos az aktuális esemény befolyásolására való képességükkel [3].

Az élőségi skála a természetesnyelv-feldolgozásban kisebb figyelmet kapott, az alapkérdéssel – főnevek élő illetve élettelen voltának megállapításával – foglalkozó tanulmányok csak az utóbbi időben jelentek meg [4,5]. Éppen a gépi fordítás generálás fázisa az a terület, ahol az élőség fontossága nyilvánvaló [1]. A szemantikai szelekció az igék természetes tulajdonsága, ennek egy esete, hogy bizonyos igék élő ill. élettelen szereplőt várnak el az alanyi pozícióban. A fent felvetett kérdésnek, hogy ti. adott konstrukció adott pozícióját betöltő szóosztályról állapítsuk meg az élőségi értékét, a számítógépes kezelésével nem találkoztam az irodalomban.

Az univerzális *ember* > *állat* > *élettelen* skálán a különböző nyelvek különböző pontokon húznak határvonalakat [3]. A magyar és az angol is az *ember* kategóriát választja el az összes többitől, ennek megfelelően, amikor a továbbiakban élő és élettelen kategóriákról lesz szó, akkor az állatokat nyelvi szempontunk alapján (vö: *ami*-vel és *it*-tel hivatkozunk rájuk) az élettelenek közé soroljuk.

3. A konkrét kérdés

Az angollal ellentétben a magyar pro-drop nyelv, a személyes névmást semleges mondatban nem tesszük ki. Egyes szám harmadik személyben mindkét nyelv elkülöníti az élőre ill. az élettelenre utaló névmást. Probléma akkor merül fel, mikor az egyes szám harmadik személyű magyar mondatban nincs kitéve a névmás, az angol oldalon pedig el kell döntenünk, hogy a „semmiből" élő vagy élettelen testes névmást generáljunk.

Általános megállapítás, hogy az alany hajlamos élő és ágens lenni [3,5]. Ennek tudatában megtehetjük, hogy minden esetben *he/she*-t generálunk (a nemek közötti különbségtétellel jelen dolgozatban nem foglalkozunk). Kiértékeléskor ezt a primitív – azonban meglehetősen jó eredményeket adó – módszert fogjuk baseline-nak tekinteni. Felmerült egy másik baseline módszer lehetősége is, miszerint a tárgyas igék alanya alapértelmezésben élő, a tárgyatlanoké pedig élettelen. Ezt elvetettük, mert a fenti egyszerűbb „mindig élő" baseline rendszeresen jobb eredményt adott.

A fordítórendszer alapértelmezés szerint valóban *he/she*-t generál, így a kidolgozandó módszer felé az az elvárás, hogy lehetőleg soha ne tévedjen abban az irányban, hogy élő helyett élettelent javasol.

4. Módszerek, kiértékelés

4.1. Nyersanyag

A vizsgálatokhoz a Magyar Nemzeti Szövegtár egyvonzatkeretes egységekre bontott változatát [6] használtam. Ezek az egységek egy igét, és a mellette álló bővítményeket tartalmazzák. Így lehetőség van arra, hogy ne csak puszta igékkel, hanem igei keretekkel is dolgozzunk (pl. *tudomásul vesz vmit*, *kiderül vmiről vmi*, *rendben van vmi*), az igék különféle kereteit külön kezeljük. Hiányosság, hogy amikor adott keret megjelenéseit kérdezzük le a korpuszból, akkor csak azt lehet megadni, hogy mely bővítmények szerepeljenek az ige mellett, azt nem lehet meghatározni, hogy mi ne szerepeljen. Következésképpen a *megy* igére vonatkozó lekérdezés az ige bővítményeit különféle variációkban tartalmazni fogja, ezért jóval zajosabb lesz, mint a *nyilvánosságra hoz vmit* keretre vonatkozó.

Az MNSZ gyakoribb igei kereteiből válogattam a mintáimat: konkrétan azok közül a keretek közül, amik 925-nél többször fordulnak elő a Szövegtárban. Mindvégig *type* alapon dolgoztam, azaz egy igei keretet tekintettem egy egységnek, szemben azzal a felfogással, mikor egy adott előfordulás, mondat a vizsgálati egység.

4.2. Előzetes: a *3sz%* módszer

Komlósy megállapítja, hogy bizonyos igék csak egyes szám 3. személyben használatosak, és ezeknek az igéknek „az alanyi vonzata nem jelölhet személyt” [7, 335.o.]. Az 1. és 2. személy tehát élő alanyra utal, sőt valójában mindig élő alanyt jelent, míg a 3. személy jelenthet élőt és élettelent is. (Ennek megfelelően nem véletlen, hogy sok nyelv csak a 3. személyű névmásokban különíti el az élőt és élettelent [3].) Ezen a megfigyelésen alapul a *harmadik-személy% (3sz%)* módszer, mely szerint ha az ige túlnyomó többségében 3. személyben fordul elő, akkor alanya élettelen, különben élő.

1. táblázat. Néhány jellemzően élő ill. élettelen alanyú ige *3sz%*-értéke

ige	*élőség*	*3sz%*-érték
néz	élő	65,4%
alszik	élő	64,0%
megtörténik	élettelen	99,9%
tartalmaz	élettelen	99,9%

Néhány jellemzően élő ill. élettelen alanyú ige manuális vizsgálata (1. táblázat) után az alábbi szabályt állítottam fel:

3sz%-módszer: 3. személy aránya $> 90\% \Rightarrow$ élettelen az alany

Ezt a kiinduló módszert egy 68 véletlenszerűen kiválasztott igei keretből álló kis korpuszon teszteltem, a kereteket előzőleg annotáltam az alany élősége szerint. Az eredményeket a 2. táblázat tartalmazza. A baseline nagyon magas: pusztán azáltal, hogy minden alanyt élőnek veszünk, az igék négyötödét helyes kategóriába soroljuk. A *3sz%* módszer ezt kis mértékben meghaladja, de a teljesítménye nem kielégítő.

2. táblázat. A *3sz%* módszer kiértékelése ($n = 68$). Mértékek: A – megfelelőség (vö: *accuracy*), azaz hogy milyen arányban döntött helyesen a módszer; valamint: P_I – élettelen pontossága, R_I – élettelen fedése, P_A – élő pontossága, R_A – élő fedése.

	A	P_I	R_I	P_A	R_A
3sz%	84%	57%	86%	96%	83%
baseline	79%				

A módszer főleg a kellemetlenebb irányba hibázott, azaz élő helyett élettelennek határozott meg bizonyos alanyokat. A hibák elemzésekor körvonalazódott egy olyan igecsoport, ahol annak ellenére, hogy ezek az igék lényegében kizárólag egyes szám harmadik személyben fordulnak elő, az alany egyértelműen élő (pl. *nyilatkozik*, *vélekedik*, *aláír*, *tárgyal vmiről*). Komlósy fenti állítása tehát ezen az empirikus alapon cáfolhatónak tűnik, a módszert pedig valamilyen módon finomítani szükséges.

4.3. A *k3sz%* módszer

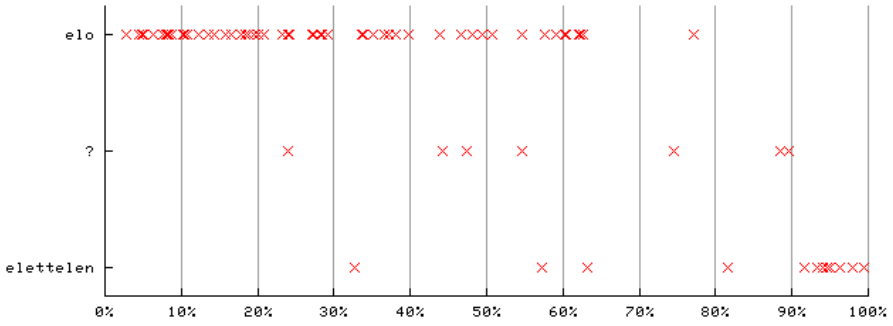
Mint említettük, az 1. és 2. személyű ragozás egyértelműen élő alanyt jelez, a továbbiakban a harmadik személyű mondatokkal foglalkozunk, itt kell megbecsülnünk az élő és élettelen alanyok arányát. Az alapötlet a következő: vannak olyan szópáraink, melyek funkciójukban azonosak, kizárólag abban különböznek, hogy az élő/élettelen jegy beléjük van kódolva: ilyen a speciális *aki/ami* vonatkozó névmás pár. Adott helyen pontosan vagy az egyik vagy a másik szerepel, és hogy melyik, az csakis a referált entitás élőségétől függ.

Ha egy pozíción nagy többségben van az *aki* névmás, akkor valószínűsíthetjük, hogy élő jegyű pozícióról van szó, másként fogalmazva az *aki/ami* arány értékes információval szolgálhat a pozíció élő/élettelen arányáról, annak közelítéseként fogható fel. Megjegyzendő, hogy ezen a ponton hallgatólagosan feltételeztük, hogy élő és élettelen dolgokra ugyanolyan arányban szoktunk vonatkozó névmással hivatkozni. A magyar nyelv sajátosságának megfelelően az *ami*-hoz

hozzá kell vennünk az *amely*-t és a *mely*-t, erre a háromelemű halmazra fogok egyszerűen *ami*-ként hivatkozni, ez fog szemben állni az *aki*-vel.

A *korrigált harmadik-személy% (k3sz%)* módszerben tehát az élettelen alanyok arányának becslését úgy finomítjuk, hogy a 3. személyű alanyok közül csak az *ami* összes alany pozícióban előforduló vonatkozó névmáshoz viszonyított arányának megfelelő számút tekintünk élettelennek, azaz az alábbi mértéket fogjuk alkalmazni:

$$\text{3. személy aránya} \cdot \mathit{ami\%} = \text{3. személy aránya} \cdot \frac{ami}{ami+aki}$$



1. ábra. A *k3sz%* értékek eloszlása a tanulókorpuszon. Minden pont egy igét jelöl. A felső sorban az élő, az alsó sorban az élettelen alanyú igék helyezkednek el. A középső sor azokat az igéket ábrázolja, melyek élő és élettelen alannyal is előfordulnak.

A már említett 68 igei keretet tartalmazó korpuszt tanulókorpuszként használtam fel, és ábrázoltam, hogy milyen a *k3sz%* értékek eloszlása az egyes kategóriákban (1. ábra). Az ábrán egyértelműen elkülönülnek az igék az alany élősége szerint: az élő alanyú igék lényegében 65% alatt, az élettelen alanyú igék lényegében 90% fölött helyezkednek el, a két érték között egy szinte üres sáv van, ahol csak néhány ige található. A 65 és 90%-ot döntési szabályként alkalmazva 5 ige esetén hibáznánk: a *kitesz vmit*, a *feltűnik*, a *kimarad vmiből* a *repül* illetőleg a *megváltoztat vmit* esetében is valójában olyan igei keretekről van szó, melyek természetes módon elképzelhetők élő és élettelen alannyal is. Ennek kapcsán felmerül az annotált korpusz megfelelőségének kérdése.

Ennek a „kézi” tanulási szakasznak a feladata az, hogy a *k3sz%* értékekhez döntési szabályt rendeljünk. Mivel semmiképp nem szeretnénk, hogy élő alanyt élettelenként osztályozzunk, a küszöbértéket magas értéken: 90%-ban állapítottuk meg. A 82% körül lévő élettelen alanyú igei keret outliernek tekinthető, a

küszöbérték leszállítása 80%-ra valószínűleg túltanuláshoz vezetne. A végső szabály tehát a következő:

k3sz%-módszer: 3. személy aránya · *ami%* $> 90\% \Rightarrow$ élettelen az alany

A *tanuló*korpuszon a módszer a 3. táblázatbeli eredményt adja. A módszer jelentősen túllépi a baseline-t, a kívántnak megfelelően csak abban az irányban téved, hogy élettelent néha élőnek mond (azaz a P_I és R_A értékeket 100%-on tartja), emellett az élettelen alanyok nagy részét (71%-át) felismeri. Az előző rész végén említett, lényegében kizárólag egyes szám harmadik személyben előforduló, mégis élő alanyú igéket a módszer helyesen osztályozza.

3. táblázat. A *k3sz%* módszer kiértékelése a *tanuló*korpuszon ($n = 68$). (Mértékeket ld: 2. táblázat)

	A	P_I	R_I	P_A	R_A
k3sz%	94%	100%	71%	93%	100%
baseline	79%				

4.4. A *k3sz%* módszer kiértékelése

Az éles teszteléshez egy nagyobb és megbízhatóbb korpuszt készítettem. Két független annotátor osztályozta a 383 véletlenszerűen kiválasztott igei keretet, a tanulókorpuszhoz hasonlóan három lehetőségből választhattak: az alany élő, az alany élettelen, az adott keret élő és élettelen alannyal egyaránt megfelelő. A 4. táblázat mutatja a különféle annotációk gyakoriságát.

4. táblázat. A tesztelőkorpusz annotációinak gyakorisága. Az annotátorok egyetértése $296/383 = 77\%$ volt.

db	*annotáció*
246	egyértelműen élő
59	élő ↔ mindkettő
18	egyértelműen mindkettő
22	élettelen ↔ mindkettő
32	egyértelműen élettelen
6	élő ↔ élettelen (azaz ellentmondás)

Az egyértelműen élőnek vagy élettelennek megjelölt kereteken lefuttatott tesztelés eredménye a 5. táblázatban látható. Az eredmény hasonló a tanulókorpuszon nyújtott teljesítményhez (vö: 3. táblázat), a baseline itt még magasabb. Egy esetben történt olyan hiba, hogy élő alany helyett élettelen jött ki: a tárgy nélküli *jelent* ige volt ez, a hibát egyértelműen az okozta, hogy a korpuszlekérdezésben az ige élettelen dominanciájú tárgyas formái elfedték a ritkább tárgyatlan változatot (ld: 4.1 rész).

5. táblázat. A *k3sz%* módszer kiértékelése ($n = 278$). (Mértékeket ld: 2. táblázat)

	A	P_I	R_I	P_A	R_A
k3sz%	95%	95%	63%	95%	100%
baseline	88%				

A meg nem talált 12 élettelen alanyú keret a következő: *sért vkit*, *minősül vminek*, *működik vmiben*, *rendben van vmi*, *emelkedik*, *készül vmiben*, *jut vkinek*, *jelentkezik vmiben*, *lesz vmikor*, *kiderül vkiről*, *elpusztul*, *sejtet vmit*. Az első 7 *k3sz%* értéke 80% fölötti, a *működik vmiben* keretet valószínűleg a *közre működik vmiben* élő alanyú keret fedte el. Az másik 5 keret pedig lehet, hogy ténylegesen élő alanyú (pl *lesz vki vmikor vhol*, *elpusztul*).

A megtalált 20 élettelen alanyú keret a következő: *vezet vmihez*, *kezdődik*, *kell vmihez*, *történik vkivel*, *következik vmiből*, *csökken*, *múlik vmin*, *megvalósul*, *létre jön vmi*, *véget ér vmi*, *épül vmire*, *kezdődik vmivel*, *szolgál vmire*, *irányul vmire*, *zajlik*, *keletkezik*, *kialakul vmiben*, *növekedik*, *fennmarad*, *zajlik vmiben*. Ezek valóban kizárólag élettelen alannyal állhatnak.

Gyakorlati célunk az egyértelműen élettelen alannyal járó keretek kiválasztása volt. A magyar-angol fordítórendszerben arra a számos igére is kénytelenek vagyunk meghagyni az alapértelmezett *élő* értéket, amelyek rendesen élő és élettelen alannyal is előfordulnak (pl. *kimarad vmiből*, *feltűnik*, *repül*, *megváltoztat-t*). Ilyen értelemben kettéosztva az igéket az egyik oldalra kerülnek az az egyértelműen élettelen alannyal járók, a másik oldalra pedig az összes többi. Ezzel a felosztással a teljes tesztelőkorpuszon a következő eredményt kaptam (6. táblázat).

A baseline szélsőségesen magas értéke abból adódik, hogy szinte minden igét élő alanyúnak vettünk (kivéve egyedül azt a 32 darabot, amit mind a két annotátor élettelen alanyúnak jelölt). Rosszabbnak tűnő értékeket kaptunk, de mindössze arról van szó, hogy 5 esetben „élő helyett" élettelen alanyt jósolt az osztályozó. A következő igékről van szó: *befolyásol vmit*, *előír vmit*, *sugall vmit*, *tilt vmit*, *erősödik*. Látható, hogy mindegyik természetszerűen járhat élettelen alannyal, ha éppen nem ez a gyakoribb használatuk.

6. táblázat. A *k3sz%* módszer kiértékelése ($n = 383$). (Mértékeket ld: 2. táblázat)

	A	P_I	R_I	P_A	R_A
k3sz%	95%	77%	63%	97%	98%
baseline	92%				

5. Összefoglalás, továbbfejlesztési lehetőségek, alkalmazás

Az ismertetett *k3sz%* módszer alkalmas az élettelen alanyú igei keretek nagy részének kiválasztására, miközben lényegében sosem téved abban az értelemben, hogy élő alanyú igét élettelennek határozna meg.

A módszer kiegészíthető egyéb jegyek vizsgálatával: élő alanyra utal például a felszólítómód használata. Szükséges azonban elválasztani az azonos alakú kötőmódtól, például egyszerűen a *hogy*-gyal kezdődő tagmondatok kiszűrésével. Míg a *megy* ige felszólítómódú alakjainak 75%-a, a *működik*-nek mindössze 10%-a van valódi felszólító tagmondatban.

Kézenfekvő, de jóval bonyolultabb módszer lenne az egyes szám harmadik személyű mondatok alanyi pozícióján megjelenő szavak kimerítő gyűjtése és élő/élettelen kategóriákba sorolása például a WordNet segítségével [4] vagy a szavak élőségének gépi tanulásával [5]. Éppen azt szándékoztam bemutatni, hogy erre nincs szükség, mert a fenti kevesebb erőforrást igénylő módszer is kielégítő eredményt ad.

A módszer minden bizonnyal egyéb nyelvekre is alkalmazható. Az első-második illetve a harmadik személy elkülönítése közvetlenül, az *aki/ami* párnak megfelelő szópárt pedig nyelvspecifikusan kell keresni, angolban a *who/what* megfelelőnek tűnik.

A módszerrel az igék tárgyának ill. egyéb bővítményeinek élőségi értéke is megállapítható. Hasonlóan kezelhető a predikatív melléknév alanya, esetleg birtok birtokosa is, ami magyarban szintén elmaradhat. Az élő alanyok azonosítása esetleg szemantikus taggelés alapját adhatja, amennyiben ez az ágens jó közelítése.

Az *aki/ami* arány mintájára bizonyos esetekben a nemek elkülönítése is megvalósítható: itt két kézzel kialakított szóosztály gyakoriságait lehetne vizsgálni. Illusztrációképpen a *lány,nő/fiú,férfi* arány a *megnősül* esetében 1/20, a *férjhez megy* esetében 108/2. Némely nem ennyire egyértelmű esetben is határozott eltolódás van az egyik nem irányába, a *zokog* esetén a fenti arány 25/9.

A leírt módszerrel megállapított alapértelmezett értékek a MetaMorpho magyar-angol fordítóprogram [8] lexikonjába kerülnek be. A rendszer szabadon elérhető, kipróbálható a `http://www.webforditas.hu` oldalon.

A kutatást a Magyar Tudományos Akadémia *Elnöki kerete* támogatta. Köszönet Munkácsy Dorottyának az annotálás elvégzéséért.

Hivatkozások

1. Zaenen, A., Carletta, J., Garretson, G., Bresnan, J., Koontz-Garboden, A., Nikitina, T., O'Connor, M.C., Wasow, T.: Animacy encoding in English: why and how. In: Proceedings of ACL Workshop on Discourse Annotation, Barcelona (2004)
2. Dahl, Ö.: Animacy and the notion of semantic gender. (1996)
3. Frawley, W.: Linguistic Semantics. Lawrence Erlbaum Associates (1992)
4. Orăsan, C., Evans, R.: Learning to identify animate references. In: Proceedings of ACL Workshop on CoNLL. (2001)
5. Øvrelid, L.: Towards robust animacy classification using morphosyntactic distributional features. In: Proceedings of EACL Student Research Workshop, Trento, Italy (2006)
6. Sass, B.: Igei vonzatkeretek az MNSZ tagmondataiban. In: Alexin Z., Csendes D. (szerk.): IV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2006), Szeged (2006) 15–21
7. Komlósy, A.: Régensek és vonzatok. Strukturális magyar nyelvtan I. Mondattan (1992) 279–529
8. Tihanyi, L., Merényi, C.: A MetaMorpho fordítóprogram projekt 2006-ban. In: Alexin Z., Csendes D. (szerk.): IV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2006), Szeged (2006)