Nem lexikalizált fogalmak a Magyar WordNetben

Vincze Veronika, Almási Attila

Szegedi Tudományegyetem, Informatikai Tanszékcsoport
Szeged, Árpád tér 2.
vinczev@inf.u-szeged.hu, vizipal@gmail.com

A Magyar WordNet (HuWN) építése során az annotátorok viszonylag nagy számú olyan fogalommal találkoztak, melyeknek nem volt megfelelőjük a magyar nyelvben. E dolgozatban bemutatjuk a HuWN-be bevezetett nem lexikalizált synsetek két (non-lex és t non-lex) típusát, megvizsgáljuk a non-lex jelenség hátterét, statisztikákat is közlünk, a két wordnetből vett példákkal rávilágítunk bizonyos problémákra, majd megoldásokra is javaslatot teszünk többszavas kifejezések kezelésének kérdését is körüljárjuk és egy esetleges jövőbeli HuWN revízió non-lex irányú felülvizsgálatát is javasoljuk.

1 Bevezetés

A wordnetek olyan lexikai adatbázisok, amelyek jelentésük alapján klaszterekbe rendeződnek és különféle szemantikus és lexikai relációk segítségével kapcsolódnak össze egy konceptuális hierarchiába (lexikai ontológiába). Eredetileg azért alkották meg ezeket, hogy bemutassák, hogyan szerveződnek a nyelvi ismeretek az emberi elmében [6].

A wordnetek méretüket tekintve ugyan eltéréseket mutatnak, de ezeket – különösen a Princeton WordNetet (PWN) – tekintik egy adott nyelv legnagyobb nyelvi információt tartalmazó adatbázisainak.

A wordnetek létrehozásánál a többnyelvűség is fontos szempont: az építők rendszerint a PWN-hez igazítják új adatbázisaikat, így azokat olyan – mind egy-, mind pedig többnyelvű – alkalmazásokban lehet felhasználni a számítógépes nyelvészetben mint pl. a jelentés-egyértelműsítés, a gépileg támogatott fordítás, dokumentumklaszterezés stb.

Azonban két nyelv sosem fedi egymást teljesen sem a konceptuális, sem pedig lexikai szinten. Dolgozatunkban fogalmak megfeleltetése szempontjából vetjük össze a magyar és angol wordnetet, ismertetjük a felmerült problémákat és megoldási javaslatokat is teszünk. Először röviden bemutatjuk a magyar és angol wordnetet, majd példákkal világítjuk meg a nem lexikalizált (*non-lex*) és technikailag nem lexikalizált (*t non-lex*) synseteket. Ezt követően arra teszünk javaslatot, hogy hogyan kerülhetjük el a *non-lex* címke alkalmazását, végül pedig rámutatunk arra, hogy noha ideális esetben egy, a nyelv konceptuális hierarchiáját ábrázoló wordnetnek nem kellene non-lex elemeket tartalmaznia, mégis hasznosnak bizonyulhatnak olyan kutatási területek számára, mint a pszicholingvisztika, néprajz és kontrasztív nyelvészet.

2 Wordnetek a nagyvilágban

Az első wordnetet a Princeton Egyetemen hozták létre angol nyelvre. A ’90-es évek óta folyamatosan fejlesztik és mostanra a legnagyobb angol nyelven hozzáférhető lexikai adatbázissá vált, mely könnyen illeszthető különféle számítógépes alkalmazásokhoz. A Princeton WordNet 3.0 hozzávetőleg 155 000 szót és mintegy 117 000 synsetet tartalmaz.

Azóta egyéb wordneteket is létrehoztak, így pl. a EuroWordNetet, holland, olasz, spanyol, német, francia, cseh és észt nyelvekre [2]; a BalkaNetet, az EuroWordNet kiterjesztéseként bolgár, görög, török, szerb és román nyelvekre [9,10]. Ezeken kívül wordneteket fejlesztettek még arab, horvát, kínai, dán, szlovén, lengyel, orosz, perzsa, hindi, tulu, dravida, tamil, telegu, szanszkrit, bodo, asszámi és filippínó nyelvekre [3,8].

A Magyar WordNetet (HuWN) a Magyar Tudományos Akadémia Nyelvtudományi Intézete, a Szegedi Tudományegyetem Informatikai Tanszékcsoportja és a MorphoLogic Kft. Fejlesztette ki egy hároméves projekt keretében [1,5]. A HuWN jelenleg több mint 40 000 synsetet tartalmaz, melyből 2 000 synset a gazdasági, 650 synset pedig a jogi szakontológia részét képezi.

A HuWN alapjául a Princeton WordNet 2.0 szolgált, pontosabban a BalkaNet Concept Setbe (BCS) tartozó synsetek lettek kiválogatva és magyarra fordítva. A wordnet készítői ezt követően szerkesztették, javították és kiterjesztették őket szinonimákkal a VisDic szerkesztőprogram segítségével. Később a fogalmak körét koncentrikusan terjesztették ki, azaz a már meglévő synsetek „utódait” synsetjelöltekként kezelték. A végső döntést, arról, hogy felvegyék őket vagy sem, több tényező is befolyásolta, mint pl. a fogalom gyakorisága vagy jelenléte más wordnetekben [5].

3 Nem lexikalizált synsetek

A munka kezdetén a magyar wordnet fejlesztői az úgynevezett expand[1] módszer mellett döntöttek. Ez azt vonta maga után, hogy a HuWN a PWN hierarchiáját örökölte. A HuWN főnévi és melléknévi része a következő módszer alapján lett felépítve: a PWN csomópontjait automatikusan magyar synsetjelöltekhez kapcsolták és a relációkat átvették. Az alapstratégia az volt, hogy egy kétnyelvű angol-magyar szótár magyar szócikkeit hozzákapcsolták a PWN 1.6 főnévi/melléknévi synsetjeihez.

A HuWN létrehozása gyakorlatilag azt jelentette, hogy a PWN synseteket magyarra fordították. Azonban, mivel nincs teljes átfedés a nyelvek fogalmai között, kulturális, életkörülmények és egyéb tényezők eltéréséből adódóan a nyelvek gyakran csak rájuk jellemző fogalmakkal rendelkeznek, s ezeknek más nyelvekben csak hozzávetőleges megfelelőik vannak, és nem fordíthatók, fejezhetők ki egyetlen szóval [4].

Így a PWN építési elvek teljes átvételének és alkalmazásának negatív következményei lettek volna a HuWN-re; egyrészt kevésbé tükröződött volna a magyar lexikalizáció, másrészt a PWN konceptuális szerkezetének egy az egyben magyarra

[1] Kiterjesztéses modell

történő átültetése további nehézségeket okozott volna, különösen a többnyelvű alkalmazásokra tekintettel [7].

Azért, hogy ne legyenek „lyukak" a fában, azaz a magyar és angol wordnet a lehető legnagyobb mértékben átfedjen, meg kellett találni az ilyen synsetek megfelelő kezelésének módját. Bevezettük a *non-lex* címkét olyan synsetek jelölésére, melyek (szó szintjén) nem léteznek az adott nyelv lexikonjában. Ezek a synsetek körülírás formájában tartalmazzák az angol synsetnek megfelelő fogalmat, de definíciót és példát nem.

POS: n NL: yes
ID: ENG20-04138222-n BCS: 3
Synonyms: (hajó jobb oldala):0
Domain: aeronautic

NL jelöli a *non-lex*-t; a synsetnek nincs definíciója, példája, értelmező szótárbeli linkje és literálja.

Alább statisztikákat közlünk a HuWN nem lexikalizált synsetjeit illetően. Látható, hogy a HuWN egészét tekintve minden huszadik, a BCS részt tekintve pedig minden tizenkettedik synset nem lexikalizált.

1.táblázat: (Technikai) nem lexikalizált synsetek a HuWN-ben

	HuWN	BCSHu
Synsetek	42 292	8 446
Nem lexikalizált	1 999	463
Technikai nem lexikalizált	454	271
Nem lexikalizált synsetek % -a	5,799	8,69

Most pedig megadjuk azokat a kritériumokat, amelyek alapján egy synset a non-lex synset kategóriába sorolható. Először, lehetséges, hogy a fogalom az adott nyelvben nem fordul elő (különösen kulturális különbségeknek köszönhetően). Másodszor, a fogalom kifejezhető produktív vagy kompozicionális szerkezetekkel (pl. melléknév + főnév szerkezetekkel), azaz nincs mód arra, hogy egyetlen szóval fejezzük ki őket. Harmadszor, a fogalom több más, egyetlen szóval kifejezhető fogalmat foglal magában, így a másik nyelvben csupán egy listával fejezhető ki. Negyedszer, úgy tűnik, hogy a PWN több következetlenséget vagy hibás definíciót, hipernima relációt tartalmaz, melyeket a HuWN építői nem kívántak követni és ehelyett a problémás synseteket *non-lex* címkével látták el.

3.1 A nem lexikalizált synsetek típusai

A nem lexikalizált synsetek hat fő osztályba sorolhatók, melyekre példákat alább láthatunk.

3.1.1 Kulturálisan meghatározott fogalmak

Ezek a fogalmak a kultúrák, életstílus, földrajzi elhelyezkedés stb. különbségeiből fakadnak. Mivel a magyar és amerikai kultúra, (népi) hagyományok és társadalmi háttér igen eltérő, vannak olyan fogalmak, melyeknek vannak ugyan szó szerinti megfelelőik a másik nyelvben, ahogy az alábbi példákból is látszik, azonban nem tükrözik az eredeti szavak által előhívott érzéseket, hangulatokat, azaz, azt, ami az anyanyelvi beszélő eszébe jut, amikor hallja őket [11].

Példák a magyar nyelvből:

- **Luca széke** – *Luca's chair* (az angol fordítás semmit sem árul el a kapcsolódó népi hiedelemről);
- **Máglyarakás** – *stake* (a magyarban ez egy sütemény, melynek jelentése nem adható vissza az angol szóval).

Példák az angol nyelvből:

- ***Anglia*** – Anglia latinul (a magyarban nincs megkülönböztetés, mivel a magyarban az England megfelelője Anglia);
- ***Sassenach*** – angol személyt jelölő skót terminus; nincs lexikalizált magyar megfelelője.

3.1.2 Gyűjtőfogalmak

A nem lexikalizált synsetek egy másik csoportja olyan elemeket tartalmaz, amelyeknek nincs megfelelőjük az adott nyelvben. Igen gyakran bizonyos, ebbe az osztályba tartozó gyűjtőfogalmakat csak körülírással vagy lista megadásával lehet kifejezni a másik nyelvben. Például:

Learned profession:1, a jog- , orvos- és teológia tudományának gyűjtőneve, melyet a magyar nem tud kifejezni egyetlen szóval, csak a három területet tudjuk felsorolni.

Ami a **drug:1**-et illeti, a HuWN-ben nincs egyszavas megfelelője, mivel a magyarban jól elkülönül a gyógyszer a kábítószertől, bár az utóbbit használják orvosi értelemben olyan anyagok jelölésére, melyeknek nagyon erős és tartós fájdalomcsillapító hatásuk van.

3.1.3 Fosztóképzővel ellátott synsetek

A nem lexikalizált synsetek egy másik, alappéldája a fosztóképzővel képzett melléknevek/főnevek olyan prefixumokkal, mint a *non-*, *in-*, *un-* stb. Néhány esettől eltekintve, az ilyen fosztóképzővel képzett lexikai egységek magyar megfelelőit negatív határozókkal képezzük, és ezek együtt nem alkotnak lexikalizált synseteket; például: *unattractive* – nem vonzó; *ill-timed* – rosszul időzített; *incongruity* – meg nem egyezés stb.

3.1.4 Melléknév + főnév szerkezetek

A magyarban bizonyos PWN-ben található fogalmakat melléknév + főnév szerkezetekkel fejezünk ki és ezeket nem tekintjük lexikai egységeknek, mert vagy produktívak, vagy pedig jelentésük teljesen kompozicionális.

Például az **Englishman:1**/**Englishwoman:1** (*English male* 'angol férfi' *English woman* 'angol nő') nem lexikalizált egységek a HuWN-ben, mert a magyarban nincs nyelvtani nem. Másrészt az *Englishman* magyar megfelelője, az 'angol' bekerülhetett volna a HuWN-be. Ugyanakkor az **Englishwoman:1** magyar megfelelője, az 'angol nő' nem vehető fel a HuWN-be.

A HuWN sajnos nem túl következetes e tekintetben. Lásd pl. **Scotsman:1**-t, melyet megfelelően 'skót'-nak vettek fel. A magyarban a 'skót', 'angol', 'magyar' szavaknak nincs neme, e szavak mégis elsősorban az adott nemzet hímnemű tagjára utalnak és nőnemű párjukat a 'nő' hozzáadásával kapjuk meg. A 'skót nő' összetételt azonban már produktív szerkezetnek (melléknév + főnév) és nem többszavas kifejezésnek tekintjük (, mely a magyarban a fenti szerkezetek feltétele a HuWN-be való bekerülésre), ezért nem vettük fel a magyar wordnetbe.

3.1.5 Nyelvtani különbségek

Némely esetben a nem lexikalizált synset nyelvtani különbségekből adódik. A **people:1**-nek (embercsoport) konceptuális szinten van, de lexikai szinten nincs megfelelője a magyarban: például a *200 people* magyarra a 'kétszázan' szóval adható vissza, ahol az esetrag az angol főnévnek felel meg.

Példa a nem lexikalizált melléknevekre a HuWN-ben a **comfortable:1**, **uncomfortable:2** synsetek. A HuWN-be nem lehetséges felvenni a cselekvés ágensét és experiensét egy synsetbe, ami viszont a PWN-ben gyakran előfordul.

3.1.6 Átvételek

Idővel bizonyos nem lexikalizált fogalmak lexikalizálódnak. E folyamat egyik tipikus területe a technológia, melynek fogalmai egyre gyorsuló ütemben terjednek világszerte. Néhány évvel ezelőtt, amikor a HuWN épült, pl. az *RV* (*recreational vehicle*) *non-lex* címkét kapott, ám most már teljes jogú lexikalizált synsetként felvehető lenne a HuWN-be.

3.2 Technikai nem lexikalizált synsetek

A wordnetépítés során gyakran előfordult, hogy két hipernima relációban lévő angol synsetnek egy magyar megfelelője volt; a két fogalom csak a konceptuális szinten különül el, lexikai szinten azonban nem találunk két külön szót. Ez azzal a következménnyel járna a HuWN-re, hogy a magyar szó önmaga hipernimája lenne. Ez volt a fő oka annak, hogy bevezettük a technikai nem lexikalizált (*t non-lex*) címkét.

A *t non-lex* címkét a következő esetekben használjuk: szófaji eltérés, azonos literálok hipernima relációban, azonos literálok *similar_to* relációban.

3.2.1 Eltérő szófaj

Különbségeket a két nyelv lexikonjában is találunk. Némely esetben a synset megfelelője a célnyelvben más szófajú, de a wordnetekben megengedett négy szófaj egyike. Például az *afraid* szó az angolban melléknév, viszont a magyarban a 'fél' igével adható vissza. Ezekben az esetekben vettük hasznát az ún. *eq_xpos_synonym* relációnak, mely eltérő szófajok közt jelöl szinonímiát és a magyar synset pedig *t non-lex* címkét kapott.

3.2.2 Azonos literálok hipernima relációban

A *t non-lex* címkézés második esete két azonos literál hipernima relációban lévő synsetekben. A címkézést azzal indokolható, hogy automatikusan könnyebb lehetséges hibákat azonosítani. Ha ugyanaz a literál *x* és *y* synsetben is megjelenik és azok hipernima relációban vannak, akkor valószínű, hogy az annotátor hibázott.

Az is a wordnetépítés egyik alapelve, hogy a fogalmat helyettesíteni lehet a hipernimájával, ezért ésszerűnek tűnt, hogy a hiponimát nem vettük fel a HuWN-be.

Lásd a következő példát:

1 **curtain:1**	**függöny:2**
2 **drop curtain:1**	(függöny) *t non-lex*

Ebben az esetben a HuWN *t non-lex* synsetjének van egy szinonimája a 'színházi függöny', mely egy kollokáció és teljes joggal felvehető lett volna a wordnetbe. A hiponima helyzetben lévő azonos literál törlésének szabályának felfüggesztésével egy kéttagú synsetet kapunk ('függöny', 'színházi függöny'). Az a különös ebben a synsetben, hogy a két tag nem valódi szinonima, mivel nem minden esetben felcserélhetők:

*Előadás után a **függöny** leereszkedett.*
*Az egész várost felkutattam megfelelő anyagért **színházi függöny** készítéséhez.*

Az első mondatba csak a 'függöny' illeszkedik megfelelően, a 'színházi függöny' furcsán hangzik; a melléknév ('színházi') felesleges. A második esetben azonban ez annyiban módosul, hogy a melléknévi rész használata nélkül a 'függöny' (**curtain:1** a PWN-ben) általánosabb jelentése is előfordulhat.

3.2.3 Azonos literálok központi és szatellit synsetekben

Az ontológia melléknévi részében is alkalmaztuk a *t non-lex* címkét. Mivel építése az antonim párokon és a hozzájuk asszociáció révén kapcsolható, szinonim szatellit synseteken alapul, lehetséges, hogy amíg angolban eltérő szó szerepel a központi és szatellit synsetben, addig a magyarban mindkét helyen ugyanaz a synset jelenik meg. A wordnetépítés szabályai nem engedik meg, hogy azonos literálok szerepeljenek a központi és szatellit synsetben (vö. a hiper- és hiponima azonossága). Ebből következően ismét azt az eljárást követtük, hogy a központi synset lexikalizált marad és a specifikusabb szatellit synset kapja a *t non-lex* címkét.

Például a {**wide:1**; **broad:1**}'s szatellit synsetje a {**heavy:5**; **thick:5**}, de a magyarban a 'széles' mindkettőt lefedi, ezért a központi synset a {**széles:2**}, a szatellit synset pedig a {**széles:0**}.

A *t non-lex* címkével ellátott synseteknek – szemben a *non-lex* synsetekkel – van definíciója, példája és, a legtöbb esetben, ÉKSz-linkje is. Azért választottuk ezt a megoldást, mert ezek a synsetek létező fogalmak a magyarban, szavakkal kifejezhetők, és csak a wordnet szerkezetének köszönhető, hogy a *t non-lex* címkét kell alkalmaznunk.

4 Nem lexikalizált synsetekhez kapcsolódó wordnet hibák

Itt a PWN és HuWN néhány problémás synsetjét mutatjuk be megoldásaikkal együtt.

4.1 Problémák a fában

Bizonyos esetekben a synset és hipernimája nincs összhangban. Például a **location:1** PWN synset definíciója a következő: *a point or extent in space* ('térbeli pont vagy kiterjedés'); egyik hiponimája a **bilocation:1,** melynek definíciója**:** *the ability (said of certain Roman Catholic saints) to exist simultaneously in two locations* ('az a képesség (, melyet bizonyos római katolikus szentekről állítanak), hogy valaki egy időben, két helyen van jelen' (unique beginner synset: **entity:1**). Szerintünk a reláció nem megfelelő, mert a definíciók nem összeegyeztethetők és csak úgy tűnik, hogy szabályszerű hiper-hiponima párt alkotnak. Ehelyett a *bilocation*t az **ability:2**, **power:3**/**képesség:2**-höz kellene kapcsolni éppen PWN-ben szereplő definíció alapján vagy pedig a **phenomenon:1**/**jelenség:1**-hez. Ha a PWN szerkezetét meg akarjuk őrizni a HuWN-ben, a synsetet *non-lex*-nek kellene címkézni és egy új synsetet kellene létrehozni a megfelelő hipernima alatt (**képesség:2** vagy **jelenség:1**).

A PWN kritikátlan másolásának következményei helytelen synset relációk is lettek: pl. **alsó állkapocs:1/lower jaw:1 → állkapocs:2/jaw:1** hipernima relációban vannak, noha a megfelelő a *holo_part* ('része') reláció lenne.

4.2 Lexikalizált synsetek *non-lex* címkével

Bizonyos esetekben – meglátásunk szerint – a HuWN annotátorai vétettek hibát. Például a **labor:1** jelenleg egy *non-lex* synset, miközben teljes joggal lehetne lexikalizált a 'fizikai munka' kollokációval fordítva. Hasonlóképpen a **seating:1, area:1**-t is fel lehetett volna venni mint 'ülőhely'.

A synsetek egy másik csoportja a HuWN-ben – melyet helytelenül *non-lex* címkével láttak el – az, melyben a literálok birtokos esetben vannak (**rear:2**/'hátulja'; **front:2**/'eleje').

4.3 Lexikalizáltként felvett non-lex synsetek

A non-lex synsetek egy érdekes példája a **bow and arrow:1**/**íj és nyílvessző:1**. Meglátásunk szerint a synsetet helytelenül jelölték lexikalizáltnak, mivel – bár két része egy egységet alkot – a kilövőszerkezet és a lövedék nem alkotnak egy fogalmat a magyarban.

A PWN kritikátlan másolásának másik példája egy teljességgel nem létező (bár lehetséges) synsethez, a **fúvóeszköz:1**/**blower:1**-hez vezet a magyarban.

A PWN-ben, úgy tűnik, vannak olyan synsetek, melyek nyilvánvalóan nem alkotnak egységes fogalmat. A **small/large definite/indefinite quantity**, **creating from raw materials**, **sound property**, **change of integrity**, **creating by removal** stb. synseteket *non-lex*-nek tekintjük.

4.4 Öröklési problémák

Bizonyos synseteknek két vagy több hipernimája van a fában. Arra kívánunk rámutatni, hogy csak abban az esetben szabad megengedni a több hipernimát, ha a hiponim synsetek a hipernima összes jellemzőjét örökölhetik. Példa lehet erre a **relaxant:1**, melynek két hipernimája van (*drug* vagy *treatment*). A fában a synset a **treatment:1**-től terjed egészen az **act:2** legfelső szintű fogalomig. A fenti esetben a synset nemcsak a *drug*, hanem a *treatment* tulajdonságait is örökli, ami ahhoz az ellentmondáshoz vezet, hogy (hiponimája,) a *Valium* egyszerre entitás és emberi tevékenység.

5 A non-lex problémák lehetséges megoldásai

A magyar wordnetben található non-lex synsetek nagy száma felveti a wordnetépítési elvek felülvizsgálatának kérdését. A non-lex synsetek tulajdonképpen nem képezik részét az adott nyelvnek, és a nagyszámú non-lex elemet tartalmazó wordnetek aligha tükrözik megfelelően az adott nyelv fogalmi hierarchiáját. Azért, hogy megoldjuk ezeket a problémákat, azt javasoljuk, hogy csökkentsük a non-lex synsetek számát a következőkben ismertetendő módszerekkel.

5.1 Hiponima nélküli non-lex synsetek

Azt javasoljuk, hogy a hiponima nélküli non-lex synseteket töröljük a fából. Mivel a hipernimák minden kontextusban helyettesíthetik hiponimáikat, ez az eljárás nem ássa alá bizonyos fogalmak kifejezhetőségét. Ez a következő példák esetében lehet hasznos:

1 **freedom:1**	**szabadság:1**
2 **liberty:1**	(szabadság)

Magyarban nincs jelentéskülönbség a két PWN-fogalom közt, így a fában lejjebb elhelyezkedő non-lex synsetet törölni kell. Ez a megoldás egyéb kultúra- és földrajz-specifikus synsetek esetében is alkalmazható.

5.2 Gyűjtőfogalmak

Azokat az gyűjtőfogalmakat, amelyeket vissza lehet adni egy lista megadásával, egyszerűen törölni kell a fából és összes hiponimáit a hipernimájához kell csatolni. Például:

cycling:1 (kerékpározás, motorozás)

Ebben az esetben a 'kerékpározás' és 'motorozás' fogalmakat két külön synsetbe kell felvenni és a **sport:1** alá kell bekötni.

5.3 A fa újraépítése

Bizonyos esetekben a fa újraépítése tűnik a legmegfelelőbb megoldásnak. Legelőször is, hadd mutassuk be a problémát az alábbi PWN-ből és HuWN-ből vett farészlettel (a magyar átírások megfelelnek a PWN definícióinak):

1 **building:1**	**épület:1**
2 **place of worship:1**	(istentisztelet helye) *non-lex*
3 **church**	(keresztény templom) *non-lex*
temple:1	(nem keresztény templom) *non-lex*

A PWN-ben a **church:2** és a **temple:1** azonos szintű hiponim synsetjei **a place of worship:1**-nek, és jelenleg nincs lexikalizált megfelelőjük a magyar wordnetben. Azért, hogy „megszabaduljunk" három non-lex synsettől, azt javasoljuk, hogy a 'templom' synsetet (, mely magyarban valamely vallás istentiszteleti helyének, épületének felel meg), hipernima pozícióba kell helyezni párhuzamosan a **place of worship:1**-gyel. A másik két PWN synsetnek a magyarban nincs megfelelője, így helyük üresen marad.

1 **place of worship:1**	1 **templom:1**
2 **church:2**	(-)
temple:1	(-)

5.4 Többszavas kifejezések integrálása

A következő példa elgondolkodtatott az alapvető wordnetépítési elvekről:

1 **gutter:2**, **sewer:3**, **toilet:3** ('WC, ablak, csatorna; kidobhatod az ablakon')

A *misfortune resulting in lost effort or money* (’kárba veszett erőfeszítés vagy pénz’) jelentésű synsetet az annotátorok nem találták lexikalizálható elemnek. Ez arra a tényre vet fényt, hogy a HuWN sokkal inkább lexikai wordnet, mintsem konceptuális. Gyakran a magyar wordnet építői inkább a szóalakra figyeltek, mint a fogalomra, ezért nincs a PWN synsetnek lexikalizált megfelelője a magyarban. Azonban a fő gond az, hogy az angol literálok egy többszavas kifejezés részei (ebben az esetben egy idiómáé), melyeket mint (konceptuális) egységet (, azaz synsetet) lehetett volna felvenni. Mivel a legtöbb többszavas kifejezésnek megvan a megfelelője a másik nyelvben, a megfelelő synsetet könnyebben meg lehet találni.

A probléma megoldására azt javasoljuk, hogy a teljes idiómát vegyük fel egy lexikai egységként a wordnetek igei részében (az idiómák jellemzően komplex predikátumok), melyeket aztán könnyen lehet párosítani anélkül, hogy a névszói összetevők megfelelőit kellene keresnünk a másik nyelvben. Ezek alapján a következő synsetek állnak elő:

be in the gutter, go down the sewer, be in the toilet ’lehúzhatja a WC-n’, ’kidobhatja az ablakon’

Az idióma felvétele mint nyelvi egység sokkal hasznosabb a többnyelvűség szempontjából, mert így könnyebb azok megfelelőit megtalálni a másik nyelvben mint egyes részeit, másrészt pedig az egész idióma felvételre kerül, s nemcsak főnévi, igei vagy melléknévi részei[2]. Egyúttal az idiómák részeihez kapcsolódó non-lex synseteket is fel lehet számolni.

7 Az eredmények értékelése

A non-lex elemek kulturális vagy konceptuális különbségeket tükröznek és így nyelvek közti hasonlóság megállapítására szolgálhatnak. A magyar wordnet jelen formájában tartalmaz non-lex elemeket, de amennyiben valamikor sor kerül a felülvizsgálatára, érdemes lenne bizonyos elemeket törölni vagy lexikalizált elemként felvenni (ha hibásan non-lex synsetként lettek jelölve), így a HuWN igazán tükrözni tudná a magyar nyelv konceptuális hierarchiáját.

Azonban a *non-lex* jelölés több szakterületen is hasznos lehet, pl. a pszicholingvisztikában, ahol különböző nyelvek beszélői mentális fogalmainak hierarchiáját vetik össze – a non-lex synsetek expliciten jelzik ezeket a különbségeket. A kultúraspecifikus synseteknek a néprajz vehetné hasznát. A nyelvi különbségekből adódó non-lex synsetek (pl. fosztóképzős melléknevek) pedig hozzájárulhatnának az elméleti és kontrasztív nyelvészet kutatásaihoz.

A fentiekre alapozva tehát azt javasoljuk, hogy a magyar wordnetet két változatban kellene létrehozni: az egyiket, amennyire csak lehetséges, a PWN-hez kellene kötni, így megőrizve annak hierarchiáját (non-lex synsetekkel); a másiknak nem kellene non-lex elemeket tartalmaznia, hogy a magyar nyelv hierarchiáját tükrözze. A két verziót így a kutatási céloknak megfelelően lehetne felhasználni.

[2] E szófajok és a határozószavak alkotják a wordneteket.

8 Összegzés

Ebben a dolgozatban bemutattuk a két, HuWN-be bevezetett *non-lex* címkét (*non-lex* és *t non-lex*) és megvizsgáltuk, hogy mi áll a non-lex jelenség mögött: elsősorban kulturális és/vagy nyelvi különbségekre vezethetők vissza. Megpróbáltunk megoldásokkal is szolgálni a szükségtelen synsetek törlésével vagy a fa újrarendezésével.

Bár az adott nyelv hierarchiáját ábrázoló wordnetnek nem volna szabad non-lex elemeket tartalmaznia, mégis hasznosnak bizonyulhatnak különféle kutatási területek (pszicholingvisztika, néprajz stb.) szempontjából. Így azt javasoljuk, hogy amennyiben sor kerül a magyar wordnet revíziójára, a non-lex elemeket törölni kellene és így a magyar konceptuális hierarchiát tükröző wordnetet kapnánk, melyet elsősorban magyar nyelvű kutatásokra lehetne felhasználni, az eredetileg kiadott verzió pedig többnyelvű kutatások referencia-adatbázisaként szolgálhatna.

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.

Bibliográfia

1. Alexin, Z., Csirik, J., Kocsor, A., Miháltz, M., Szarvas, Gy.: Construction of the Hungarian EuroWordNet Ontology and its Application to Information Extraction. In: Proceedings of the Third International WordNet Conference. South Jeju Island, Korea (2006) 291–292
2. Alonge, A., Bloksma, L., Calzolari, N., Castellon, I., Marti, T., Peters, W., Vossen P.: The Linguistic Design of the EuroWordNet Database. Computers and the Humanities. Special Issue on EuroWordNet Vol.32, No. 2–3 (1998) 91–115
3. Bhattacharyya, P., Fellbaum, C., Vossen, P. (eds.): Principles, Construction and Application of Multilingual Wordnets. Proceedings of the Fourth Global WordNet Conference. Narosa Publishing House, Mumbai, India (2010)
4. Derwojedowa, M., Piasecki, M., Szpakowicz, S., Zawisławska, M., Broda, B.: Words, Concepts and Relations in the Construction of Polish WordNet. In: Proceedings of the Fourth Global WordNet Conference (2008) 167–68
5. Miháltz, M., Hatvani, Cs., Kuti, J., Szarvas, Gy., Csirik, J., Prószéky, G., Váradi, T.: Methods and Results of the Hungarian WordNet Project. In: Tanács, A., Csendes, D., Vincze, V., Fellbaum, C., Vossen, P. (eds.): Proceedings of the Fourth Global WordNet Conference. University of Szeged, Szeged (2008) 311–320
6. Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.: Introduction to WordNet: an On-line Lexical Database. International Journal of Lexicography Vol.3, No.4 (1990) 235–244

7. Raffaelli, I., Tadić, M., Bekavac, B., Agić, Ž.: Building Croatian WordNet. In: Tanács, A., Csendes, D., Vincze, V., Fellbaum, C., Vossen, P. (eds.): Proceedings of the Fourth Global WordNet Conference. University of Szeged, Szeged (2008) 349–359
8. Tanács, A., Csendes, D., Vincze, V., Fellbaum, C., Vossen, P. (eds.): Proceedings of the Fourth Global WordNet Conference. University of Szeged, Szeged (2008)
9. Tufiş, D. (ed.): Romanian Journal of Information Science and Technology. Special Issue on BalkaNet Vol.7, No.1–2 (2004)
10. Tufiş, D., Cristea, D., Stamou, S.: BalkaNet: Aims, Methods, Results and Perspectives. A General Overview. Romanian Journal of Information Science and Technology. Special Issue on BalkaNet Vol.7, No.1–2 (2004) 9–43
11. Zidoum, H.: Towards the Construction of a Comprehensive Arabic WordNet. In: Tanács, A., Csendes, D., Vincze, V., Fellbaum, C., Vossen, P. (eds.): Proceedings of the Fourth Global WordNet Conference. University of Szeged, Szeged (2008) 531–544