Vonzatkeretek vizsgálata orvostudományi tárgyú, angol nyelvű szabadalmi szövegeken

Klausz Ágnes, Vincze Veronika, Nagy Ágoston, Almási Attila

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

Kivonat: Orvostudományi tárgyú, angol nyelvű szabadalmi szövegekben előforduló igék s főnevek vonzatkereteit vizsgáltuk. Az előfordulási gyakoriságuk alapján összeállítottunk egy kifejezetten az orvostudományi tárgyú szabadalmi szövegekre jellemző vonzatkerettárat, amely hasznosítható a hasonló tárgyú szövegekre alkalmazandó szintaktikai és szemantikai elemzők építésében.

1 Bevezetés

Az ALL és a Szegedi Tudományegyetem egy közös projekt keretében vállalta egy szemantikus keresőrendszer kifejlesztését, amely elsődlegesen az angol és magyar nyelvű szabadalmakban való keresést célozza meg. A rendszer kialakításához a szabadalmi szövegek sajátságai miatt a meglevő nyelvi elemzők testre szabása szükséges, ezért célunk volt egy olyan igei és főnévi vonzatkerettár kialakítása, melyet a későbbiek során egyéb, orvostudományi tárgyú (szabadalmi) szövegek elemzéséhez is fel tudunk használni mind szintaktikai, mind szemantikai szinten.

Megvizsgáltuk, hogy a különféle igék a különféle vonzatkereteikkel milyen gyakran fordulnak elő ezen orvostudományi szakszövegekben. Az eredményt egy általános célú szótárban (az online Google Dictionaryben [2]) található igékkel és vonzatkereteikkel hasonlítottuk össze. Arra voltunk kíváncsiak, hogy a szótárban található igék és vonzatkereteik mennyire fedik le a 60 szabadalomból álló mintakorpuszunkban szereplőket, azaz egy általános célú szótár vonzatkeretei mennyire alkalmazhatók egy speciális tematikájú szövegre.

2 Igék (és vonzatkereteik) kigyűjtése a szabadalmakból, illetve egy általános célú szótárból

Ebben a részben az igék és a vonzatkeretek kigyűjtésének lépéseit ismertetjük a szabadalmi szövegekből és a rendelkezésre álló szótárállományból.

2.1 A gépileg beazonosított igék kézi ellenőrzése

Első lépésként – a Stanford elemzőt [6] használva – gépileg beazonosítottuk a szabadalmi szövegekben az igéket, majd az igének minősített elemeket kézzel is ellenőriztük, amire több okból is szükség volt. Egyrészt a POS-tagger időnként olyan szóalakokat is igének jelölt, amelyeknek egyik lehetséges szófaji kódja valóban ige, azonban az adott szövegkörnyezetben más szófajú szóként fordultak elő. Másrészt pedig arra is volt példa, hogy a gépileg megtalált szavak ugyan igei alakban fordultak elő, azonban főnévi vagy melléknévi szerepük volt. Főnévi szerepben az ún. *gerund*ként (magyarra *-ás*, *-és* végű főnévként fordítandó, angolban *-ing* végződésű igealakként) fordultak elő, s állhattak alanyként, tárgyként (pl. *a method comprising administering a pharmaceutical composition*), és esetenként határozóként is (pl. *a method for inhibiting thrombosis*, *capable of reducing lung volume*). Melléknévként szerepelhettek a főnévi szerkezet előmódosítójaként – egyrészt *-ing*-es alakban (folyamatos melléknévi igenévként, pl. *a protecting group*), másrészt az ige 3. alakjának formájában (past participle, pl. *protected amino group*, *alkylene-substituted amino*).

A fentebb említett különböző elemek esetében el kellett döntenünk, hogy igeként kezeljük-e őket. Figyelembe véve a szemantikai és szintaktikai sajátosságaikat, különböző módokon jártunk el. Mivel a gerundnak például egyaránt van főnévi és igei jellege is, szóba jöhetett az igeként történő kezelése. És emellett is döntöttünk, hiszen a gerund alakok automatikusan öröklik annak az igének a vonzatkereteit, amelyekből képezve lettek, tehát egy igei vonzatkerettár építése szempontjából releváns információkat hordoznak.

Azonban azokat a particípium alakokat, amelyek előmódosító funkciójú melléknév szerepét töltötték be (*a protecting group, protected amino group*), nem vettük fel az igei vonzatkerettárunkba. Ugyanis – bár ezek is öröklik az ige eredeti vonzatait – ezen szószerkezetek esetében a szintaktikai viszonyt kifejező prepozíció a felszínen nem jelenik meg (pl. a *treat with heat* szerkezet *heat treated*-ként jelenik meg), és az elmaradó prepozíció kezelése problémákat vethet fel az elemző számára. Másrészt pedig a melléknév és az azt megelőző tárgy gyakran kötőjellel van egymáshoz kapcsolva (*electron-withdrawing groups*), vagyis ezekben az esetekben már összetett szónak, vagyis egyetlen lexikai elemnek is lehet tekinteni őket.

A kézi ellenőrzés során egyéb esetek is voltak, melyekben nem volt evidens, hogy egy adott szóalakot igeként célszerű-e kezelni vagy sem. Ilyenek voltak bizonyos utómódosítók igéből képzett elemei (pl. *a method comprising administering a pharmaceutical composition*), többszavas kifejezések igei elemei (pl. *as follows*), az alany és állítmány nélküli mellékmondatok, azaz melléknévi szószerkezetek *-ing*-es alakja (pl. *when treating…*), a szenvedő szerkezet maradványaként álló, s alanykomplementumként funkcionáló igei 3. alakok (past participle) (pl. *when administered to*). Ezekben az esetekben egyedi elbírálást alkalmaztunk. Vagyis ha úgy ítéltük meg, hogy ezen kifejezések szignifikánsan magas számban fordulnak elő a szabadalmi szövegekben, akkor felvettük őket a vonzatkerettárunkba. Így jártunk el például a külön szótári tételt is alkotó, lexikalizálódott elemekkel kapcsolatban, (pl. *as follows*, *provided that*, *according to*), amelyek leggyakrabban kötőszóként vagy elöljárószóként funkcionálnak.

2.2 Vonzatkeretek kigyűjtése

Az igék kézi ellenőrzése után a vonzatkeretek kigyűjtése következett – szintén kézileg, (ezt gépileg – a szabadalmi szövegekre testreszabott nagy pontosságú szintaktikai elemző híján – nem lehetett megoldani). A vonzatkeret fogalmát – praktikussági okokból – tágan értelmeztük: az ige *kötelező* vonzatainak összességén kívül az egyéb, szorosan összetartozó elemekből álló kifejezéseket is idevettünk (amelyeket alább részletesebben tárgyalunk), és felvettük a kerettárunkba, hiszen célunk volt egy, a szintaktikai és szemantikai elemzéshez gyakorlatban jól használható eszköz kialakítása.

A vonzatkerettárunk összeállításakor elsősorban természetesen az ige kötelező bővítményeire fókuszáltunk. Az igéknek a tranzitív és nem tranzitív alakjait egyetlen igének és egy elemnek tekintettük a vonzatkerettárunkban, annak ellenére, hogy különböző a vonzatkeretük, pl. a *substitute* ige lehet tárgyas és tárgyatlan is. Tárgyas formájában a vonzatkerete: *V N,* vagy *V N for N;* tárgyatlan formájában: *V*, vagy *V for N.* Ezeket a vonzatkereteket tehát mind felvettük a *substitute* igéhez.

Mivel minden angol igének, így a szabadalmakban szereplő összes igének is van (nyelvtani) alanya, ezt a vonzatot default elemnek tekintettük, s nem vettük fel egyetlen ige vonzatkeretéhez sem.

Kérdést vetett fel, hogy a (közel) azonos jelentéssel bíró és formailag is csak minimálisan eltérő alakú prepozíciókat (pl. *combine together/together with, depend on/upon*) különálló vonzatkeretként célszerű-e kezelni. Mivel az automatikus szintaktikai elemzés nem szemantikai jellemzőkből indul ki, úgy döntöttünk, hogy különálló vonzatkeretekként kezeljük őket.

Hasonló kérdéskörbe tartozó problémát vetett fel a *from* prepozíció esetenkénti megjelenési formája: a *remove*, ill. a *vaporize* vonzataként néhány esetben *therefrom*-ként jelent meg (*drying said plasticized granules to remove substantially all the solvent* ***therefrom***), ami a *from there* szerkezet módosult formája. A *therefrom* megjelenési alakot nem vettük fel külön vonzatkeretként, mivel a *from that* szinonimájaként kezelendő. Az elemző algoritmus implementálásakor emiatt a *therefrom* és *thereof* szóalakokra fokozott figyelmet kell fordítani, mert a tapasztalatok alapján a Stanford parser tévesen főnévnek tekinti e szóalakokat, valójában pedig határozószavak, és nyelvtani szerepüket tekintve PP-k, a *there* alkotóelem pedig anaforikusan utal vissza egy korábbi összetevőre.

A kötelező bővítményeken kívül olyan szókapcsolatokat is felvettünk a vonzatkerettárunkba, amelyek ugyan nem kötelező vonzatai az igének, azonban megítélésünk szerint kiemelkedően jellemzőek a szabadalmakra. Ilyenek voltak bizonyos szabad határozók (pl. célhatározói *to infinitivus* alakok).

A vonzatkeretek kigyűjtése után megszámoltuk, hogy az adott ige az egyes vonzatkereteivel hányszor fordul elő és meghatároztuk, hogy ez az előfordulási szám gyakorinak számít-e a többi vonzatkeret előfordulásához képest viszonyítva (pl. megnéztük, hogy a *consist* ige hányszor fordul elő összesen, és ebből hányszor fordul elő *in* + főnév vonzattal). Erre azért volt szükség, mert a különböző igék összességében nem azonos gyakorisággal (és nem azonos számú vonzatkerettel) fordultak elő a korpuszban, így nem tudtunk meghatározni egy általános érvényű küszöbértéket, amely felett gyakorinak minősítünk egy adott vonzatkeretet.

2.3 Igék és vonzatkereteik kigyűjtése a Google Dictionaryből

Következő lépésként a szabadalmakból kigyűjtött igéket a Google Dictionaryből is kigyűjtöttük, vonzatkereteikkel együtt. (A vonzatkerettárunkban is a szótár jelöléseit követtük, mely szerint – az általános használattól eltérően – a V-ed az ige 3. alakját (past participle) jelöli). Az internetes szótár nem volt egészen következetes a vonzatkereteket illetően, hiszen olyan szerkezeteket is különálló vonzatnak vett, amelyek valójában ugyanannak a vonzatnak különböző (szabályszerűen képezhető) alakjai. Pl. az ige + főnév (V N) vonzatkeret és az '*-ing*'-es alak + főnév (V-ing N) különböző vonzatkeretként fordul elő, holott ez a kettő valójában ugyanaz a vonzat (hiszen az *-ing*-es alak automatikusan képezhető az elsőből). Így a második képletet (V-ing N) redundánsnak tekintettük, s ezért nem vettük fel külön vonzatkeretként a kerettárba.

A szenvedő szerkezetet jelölő vonzatkeret ('be' V-ed) szintén redundáns elemként jelent meg a Google Dictionary vonzatkerettárában (hiszen ez is automatikusan előállítható az alapértelmezettnek tekintett aktív igei szerkezetekből), azonban – az *-ing*-es alakokkal ellentétben – ezeket különálló vonzatkeretnek tekintettük, mert a passzív igei alak eléggé szabadalomspecifikus; ezenkívül bizonyos esetekben a 'be' V-ed 'by' alakot is felvettük jelentéstani okokból, pl. *characterized by, substituted by.*

3 Az igei vonzatok két halmazának összevetése, orvostudományi szakszövegekre alkalmazható vonzatkerettár összeállítása

A következő fázisban összevetettük a szabadalmak igei vonzatkereteit a Google Dictionaryből nyert vonzatkeretekkel, és megvizsgáltuk, hogy mennyire feleltethetők meg egymásnak. Mint az várható volt, a kettő nem volt tökéletes fedésben. Háromféle eset fordult elő: a szabadalmakban szereplő igék

a) a korpuszban szereplő vonzatukkal együtt megtalálhatók voltak a Google Dictionaryben is (pl. *adhere*, *impregnate*, *regard*). (Némely esetben ugyan a szabadalmakban szereplő amerikai angol helyesírású szó helyett a brit angol helyesírású verziót találtuk meg (pl. *analyse* vs. *analyze*), de ezeket természetesen találatnak tekintettük.)

b) szerepeltek ugyan a Google Dictionaryben, azonban a korpuszban előforduló vonzatkeretük(kel) nem. Ezekben az esetekben be kellett illesztenünk a kerettárba egy-egy új vonzatkeretet (pl. a *bind* ige 'to' + főnév vonzatkeretét); s voltak esetek, amikor több új vonzatkeretet is fel kellett vennünk a listára (pl. a *combine* ige esetében ötöt).

c) egyáltalán nem szerepeltek a Google Dictionaryben. Ezek túlnyomó többsége orvosi/kémiai terminus technicus volt, pl. *acidify*, *benzofuse*, *coprecipitate*. (Azonban olyan általánosabb jelentésű igékkel is találkoztunk a korpuszban, melyeknek a szótárból (igeként) történő hiányzása némileg meglepő volt: pl. a *potentiate* ige hiánya, ill. a *passage* szóalak kizárólag főnévként történő szereplé-

se). Ezeket az igéket természetesen egy az egyben felvettük az igei listára a korpuszban szereplő vonzatukkal.

Mivel a b) és c) pontban leírt esetekre számos példa előfordult, evidenssé vált az – amit sejteni lehetett előre is –, hogy az orvostudományi szabadalmi szövegeknek megvan a saját szakszókincsük, és bizonyos nyelvtani fordulatok is elsődlegesen rájuk jellemzők és nem a köznyelvre, vagyis általános célú szótárt nem lehet megfelelően alkalmazni orvostudományi szabadalmi szövegekre. (Ez nyilván jelentős információ a szintaktikai (és szemantikai) elemző kialakításához).

Ennek fényében tehát a Google Dictionaryből nyert vonzatkerettárat jelentős mértékben ki kellett egészítenünk a szabadalmi szövegekből kigyűjtött vonzatokkal, s ezáltal kialakítottunk egy, specifikusan az orvostudományi szabadalmakra alkalmazható vonzatkerettárat.

4 Eredmények

Az elkészült vonzatkerettár 220 igét tartalmaz, melyeknek összesen 1498 vonzatkeretük lett felvéve (ebből 93 nem szerepelt a Google Dictionaryben, ezeket a szabadalmak szövege alapján illesztettük be).

A köznyelvi szóhasználathoz hasonlóan a szabadalmi szövegekben is a legtöbb ige egy vonzatkereten belül egy vagy két vonzattal rendelkezik, s a háromvonzatos ige (pl. *inject N through N to N*) ritka.

Ha viszont a vonzat*keretek* számát vizsgáljuk, a következőket találjuk. A szabadalmakban előforduló igék Google Dictionaryben szereplő megfelelőit tekintve a legtöbb vonzatkerettel rendelkezők a következők: *come* (24), *make* (24), *take* (22), *stand* (21), *leave* (20). Viszont a nagyszámú vonzatkeretek ellenére újabbakkal kellett kiegészítenünk ezen igéknek a szabadalmakra testre szabott vonzatkeretlistáját, hiszen a referenciaszótárunk vonzatkeretei csak elenyésző mértékben fedték le a szabadalmakban szereplő vonzatkereteket: a legtöbbjük esetében csak egy vagy két olyan vonzatkeretet találtunk a Google Dictionaryben, amely a szabadalmakban találhatóval megegyezett. Azonban arra is volt példa, hogy a referenciaszótárunkban található nagyszámú vonzatkeretből egyik sem egyezett meg a szabadalmakban találhatókkal. Például a *take* a Google Dictionaryben huszonkétféle vonzatkerettel szerepel, de a szabadalmakban csak egy 23. vonzatkerettel fordul elő: (*be taken together (with) N*). Tehát – többek között – a fentebbi igék esetében egy vagy két vonzatkerettel ki kellett egészítenünk a Google Dictionaryből nyert – és egyébként gazdag – vonzatkeretlistát.

Ezzel szemben a szabadalmi szövegekben az igék jóval kevesebb ténylegesen előforduló vonzatkeretét figyelhettük meg. Hat vonzatkerettel két ige rendelkezik: az *add* és a *combine*. Öt vonzatkerettel a *comprise* és a *form* igék, néggyel a *define*, *select* és *determine*, három vonzatkerettel 11 ige, két vonzatkerettel 46 ige, 1 vonzatkerettel (amely általában egyetlen tárgyi vonzatot tartalmaz és így az ige tranzitív voltára utal) 172 ige rendelkezik.

1. táblázat: A legtöbb vonzatkerettel rendelkező igék vonzatai.

add:	*be V-ed to N*	**combine**:	*V N with N*
	V N		*V with N*
	V to N		*V together*
	V N to N		*be V-ed together with N*
	be V-ed in N		*be V-ed with N*
	V to N N		*be V-ed*

A (szabadalmi szövegekben) a legtöbb vonzatkerettel rendelkező, fentebb említett *add* és *combine* ige a Google Dictionaryben is viszonylag nagy számú vonzatkerettel rendelkezett (9, illetve 6), azonban mivel ezek nem vágtak egybe a szabadalmakban előforduló vonzatkeretekkel, a vonzatkeretlistánkat ki kellett egészíteni (az *add* ige vonzatkereteit kettővel, a *combine* igéét pedig öttel).

A szótárba összesen 16 darab új, vonzatkeretes igét kellett felvenni: ezek olyan szavak voltak, amelyek – többségükben kémiai, illetve orvosi szakszavak lévén – nem voltak megtalálhatók a Google Dictionaryben. Ilyen volt például az *admix* (*admix N with N*), *solubilize* (*solubilize N*) vagy *anellate* (*be anellated with N*).

39 ige esetében fordult elő, hogy a szabadalmakban a Google által hozzájuk rendelt vonzatkereteik nem szerepeltek, de valamilyen más, azaz új vonzatkerettel viszont igen. Ilyen igékre példa a *prescribe*, amely a szabadalmakban *prescribe to N N*, vagy az *engineer*, amely *be engineered to N* alakban fordult csak elő. A köznyelvben leggyakoribbnak tekinthető igék, például a *take* esetében is ez volt a helyzet, amint már fentebb utaltunk erre.

A legtöbb új vonzatkeretet a *combine* kapta, egészen pontosan ötöt, pl. a *be combined together with N* alakot. Ezen kívül három új vonzatkeretet kellett felvenni a *define* (pl. *be defined as*), *determine*, *rack* és *select* igékhez. A többi igét legfeljebb kettő új vonzatkerettel kellett kibővíteni.

5 Megfigyelések a vonzatkerettáron

5.1 Kompozicionalitás

A korpuszban előforduló igéket és vonzatkereteiket érdemes például a kompozicionalitás szempontjából megvizsgálni. (Minden igei vonzatkeretet kigyűjtöttünk függetlenül attól, hogy azok az igével kompozicionális szerkezetet alkotnak-e.) Az itt előforduló vonzatkeretek legtöbbször kompozicionális szerkezetet alkotnak az igével (vagyis az összetétel jelentését egyértelműen meghatározza az összetevőinek (az igének és vonzatának) jelentése és az összetétel módja), pl. *dilute with N, be added to N, impart from N to N.*

Azonban előfordultak nem kompozicionális szerkezetek is: például a *stand for* elöljárós ige ‘jelent’, ‘helyettesít’ értelemben nem kompozicionális: *R2 and R3 independently* ***stand for*** *H, C1-6 alkyl, C2-6 alkenyl.*. Ezt az elöljárós igei alakot a *stand* ige vonzatkerettárába vettük fel (*V for N*). A kevés ilyen jellegű példa arra utal, hogy az (orvostudományi) szabadalmi szövegekre valószínűleg nem jellemzőek a nem kompozicionális igei szerkezetek (melyek – az angol nyelvben – lehetnek idiómák, illetve a elöljárós igék (‘phrasal verbs’)).

5.2 Módbeli segédigék

Módbeli segédigékkel kapcsolatosan azt figyeltük meg, hogy pl. a segédigeként és főigeként egyaránt funkcionálni képes *do* és *have* igéket tekintve eltérőek a tapasztalatok: a *do* kizárólag segédigeként szerepelt, míg a *have* kizárólag főigeként fordult elő a szabadalmi szövegekben. A *do* mint főige előfordulási hiánya – legalábbis részben – szintén a kompozicionalitás kérdésével lehet összefüggésben. Ugyanis főigeként általános szövegekben igen gyakran nem kompozicionális szerkezetekben (pl. *do away with*), vagy félig kompozicionális szerkezetekben fordul elő (pl. *do a favour*), amely szerkezetek viszont – mint fentebb említettük – határozottan nem jellemzőek e szabadalmi szövegekre. A *have* segédigeként történő előfordulásának hiányát pedig az magyarázhatja, hogy ilyen funkciójában olyan igeidőket, illetve -módokat (pl. a különféle befejezett igeidők, műveltető) fejez ki, melyek szintén nem jellemzik a szabadalmi szövegeket.

5.3 Egyéb jellemzők

Az általános nyelvvel szemben a tudományos szövegekre erőteljesebben jellemző további jelenség a vonzatok sorrendiségével kapcsolatos. Például a *prescribe* ige két vonzata általában a következő sorrendben szokott az ige után állni: *V N to N* (felír vmit vkinek), vagy a *V N N* (felír vkinek vmit). Azonban a szabadalmi szövegekben megfigyelhető, hogy a hosszabb és komplikáltabb tárgy a könnyebb érthetőség kedvéért a (*to* prepozícióval kifejezett) részeshatározó mögé kerül: *V to N N* (pl. *prescribing to the patient a therapeutically effective amount of quazepam*). (Az angol nyelvészeti terminológiában *heavy NP shift*-nek nevezik ezt a jelenséget.)

A fentebbieken kívül a jövőben még érdemes lenne megvizsgálni például azt, hogy a vonzatkerettárba jelenleg fel nem vett, előmódosítói szerepű, particípiumos szerkezetek és vonzataik hogyan építhetők be a vonzatkerettárba – a kötőjelezeléssel összefüggésben (pl. *diabetes-associated disorders*); illetve a többszavas igei kifejezések (vagy félig kompozicionális szerkezetek, l. [10], pl. *come into contact with N*) kezelési módját is érdemes tovább fejleszteni.

6 Összevetés más igei vonzatkerettárakkal

Az angol nyelvre már készültek korábban is igei vonzatkerettárak, illetve olyan korpuszok, amelyek tartalmaznak a vonzatkeretre vonatkozó információt. Ilyen például a VerbNet [3, 4, 5], a Proposition Bank [8] és a FrameNet [1]. A Proposition Bank a Penn Treebank szintaktikai szerkezeteihez rendel szemantikai szerepeket, a VerbNet a kibővített Levin-féle [7] igeosztályok szintaktikai kereteit, az argumentumok szemantikai szerepeit és a rájuk vonatkozó szelekciós megkötéseket tartalmazza, a FrameNet pedig a szemantikai keretek felől közelítve adja meg az adott keretbe illeszkedő igéket és azok argumentumainak szintaktikai és szemantikai tulajdonságait.

Noha a fenti adatbázisok is részletes információkat tartalmaznak az igei vonzatkeretekre nézve, mégsem ezeket választottuk vizsgálatunk alapjául, mivel ezek elsődle-

gesen a szemantikai szerepekre koncentrálnak, minket pedig elsősorban a szintaxis érdekelt. Azonban az egyes igékhez tartozó bejegyzések összevetése mindenképpen hasznos tanulságokkal szolgálhat. Példaként tekintsük a *substitute* igét!

Az általunk kialakított vonzatkerettárban a *substitute (helyettesít)* igének két vonzata szerepel: a) *valamit*: a régi entitás, melyet lecserélünk, és b) *valamivel*: az új entitás, amellyel helyettesítjük a régit (*V N for N*).

Nézzük meg, hogy a tematikus szerepekre koncentráló adatbázisok milyen kategóriákkal dolgoznak, és ott milyen jellemzőkkel jelenik meg a *substitute* ige.

A FrameNet a tematikus szerepeket alapvető és opcionális alcsoportokra osztja. A *substitute* igével kifejezett esemény jellemzésére a következő alapvető szerepeket határozza meg: ágens (aki a cselekvést végrehajtja), új entitás (amellyel az ágens betöltet egy szerepet), régi entitás (amely korábban betöltötte az adott szerepet). Az esemény opcionális szereplőiként pedig olyan szereplőket, illetve szerepeket nevez meg, amelyek szabad határozóként funkcionálnak (vagyis nem kötelező vonzatai az igének), pl: eszköz, mód, szerep, hely, cél, ok, idő stb. A Proposition Bank négy szerepet (argumentumot) határoz meg: ágens, egyes számú téma (Theme1), kettes számú téma (Theme2), és kedvezményezett / beneficiens; s nem jelöli meg ezek közül a kötelezőeket. A VerbNet a *substitute* igének szintén két kötelező vonzatát jelöli meg: téma 1 (THEME 1) és téma 2 (THEME 2).

A fentebbi vonzatkerettárak két fontosabb szempontból térnek el a szabadalmakra készített vonzatkerettárunktól. Egyrészt tárgykörükben térnek el egymástól: a fentebbi adatbázisok általános doménben alkalmazhatók, míg az általunk készített kerettár specifikus doménre készült. Másrészt pedig míg ez utóbbi a szintaxisra helyezi a hangsúlyt, a fentebbi vonzatkerettárak a szemantikai információkra fókuszálnak. Ez utóbbiakat a későbbiekben érdemes lehet beépíteni a szabadalmakra készített vonzatkerettárban szereplő argumentumok reprezentációjába. Mivel a kötelező bővítmények és a tematikus szerepek között egy az egyhez megfeleltetés figyelhető meg, vagyis minden kötelező vonzatnak egy és csakis egy tematikus szerepe lehet, viszonylag gyors és egyszerű a kettő közötti megfeleltetés.

Amennyiben csak a főbb szemantikai szerepekre szeretnénk koncentrálni, célszerű a Proposition Banket, illetve a VerbNetet használni, amelyeknek az az előnye is megvan, hogy – mivel kevesebb adattal operálnak e rendszerek – gyorsabb megoldásokat kaphatunk. Amennyiben azonban részletesebb szemantikai reprezentációra törekszünk, az összetettebb rendszerű FrameNetet érdemes használnunk. Ez azért is lenne előnyösebb a számunkra, mert olyan elemekhez is szeretnénk tematikus szerepet hozzárendelni, amelyeket a fentebb említett két másik rendszer nem tartalmaz. Ezek az elemek a szabad határozók, melyek tematikus szerepének leelemzése hosszadalmasabb folyamat, hiszen – a kötelező bővítményekkel ellentétben – egy-egy szabad határozónak többféle tematikus szerepe is lehet.

7 Főnévi vonzatkerettár

Az igei vonzatkerettáron kívül a főnévi vonzatkerettár is elkészült a szóban forgó szabadalmi korpusz alapján, a fenti elveket alkalmazva. A könnyebb kezelhetőség végett a főneveken belül elkülönítettük a perdurantokat (időbeli történést, esemény-

szerűséget jelölő főnevek, l. [9], amelyek több szempontból hasonlítanak az igékhez. Egyrészt hasonló a jelentésük, mivel eseményt fejeznek ki. Másrészt fontos azon jellemzőjük is, hogy szinte bármennyi és bármilyen szabad határozóval rendelkezhetnek. A perdurant jelentésű főneveket szemantikailag is egy kategóriába soroltuk az igékkel a reprezentáció során, hiszen a *method for treating Alzheimer's disease* és a *method for the treatment of Alzheimer's disease* jelentésében nincs különbség.

A vonzatkerettár szempontjából azért volt fontos megkülönböztetni a perdurant főneveket a nem perdurant főnevektől, mert az utóbbiaknál csak a Google Dictionaryben szereplő vonzatokat illesztettük, míg az előbbieknél szabad prepozíciós szerkezeteket is megengedtünk. Ez sokat javított a program hatékonyságán, mert volt olyan főnév is, amelynek 4 bővítménye is volt, ez pedig a *storage*:

storage (1) of the composition (2) for ten days (3) in an open Petri dish (4) at 40°C.±2°C.

Ezen esetekben, ha csak a vonzatkerettárat vennék alapul, akkor a (2-4) bővítményeket az előtte álló igéhez tettük volna. Általában véve igaz, hogy bármilyen főnévnek lehet *of* prepozícióval kezdődő vonzata, ezért azokat alapértelmezés szerint kivettük a vonzatkerettárból. Kevés olyan nem perdurant jelentésű főnévvel találkoztunk, ami szabadalomspecifikus lett volna. Ezek egyike volt a nagyon gyakran előforduló *means*, amelynek a Google Dictionary szerint csak *to*+inf vonzata lehet, de a szabadalmakban gyakran előfordult a *for* is.

A főnévi vonzatkerettárban 117 db főnév található összesen 162 vonzatkerettel.

8 Összegzés

Ebben a munkában beszámoltunk egy orvostudományi szabadalmak szövegein alapuló igei és főnévi vonzatkerettár létrehozásáról. Kiindulási alapnak egy általános célú szótárt, a Google Dictionary vonzatkereteit tekintettük. A vonzatkerettár létrehozása során kiderült, hogy léteznek szabadalomspecifikus igék, illetve szabadalomspecifikus vonzatkeretek, melyeket az általános célú szótár nem tartalmazott, így ezeket külön fel kellett vennünk, azaz az általános célú szótár csak korlátozottan használható a szabadalmak elemzésére. A vonzatkerettárakat a későbbiekben szeretnénk szemantikai jellegű információval is bővíteni, és ezáltal a vonzatokhoz tematikus szerepeket társítani. Az elkészült adatbázis eredményesen hasznosítható a szabadalmi szövegekre fejlesztett szintaktikai és szemantikai elemző fejlesztésében.

Köszönetnyilvánítás

A kutatást – részben – a MASZEKER kódnevű projekt keretében az NKTH támogatta.

Bibliográfia

1. Baker, C. F., Fillmore, C. J., Lowe, J. B.: The Berkeley FrameNet project. In: Proceedings of the COLING-ACL. Montreal, Canada (1998)
2. http://www.google.com/dictionary
3. Kipper, K., Dang, H.T., Palmer, M.: Class-Based Construction of a Verb Lexicon. In: AAAI-2000 Seventeenth National Conference on Artificial Intelligence (2000)
4. Kipper, K., Korhonen, A., Ryant, N., Palmer, M.: Extending VerbNet with Novel Verb Classes. In: Fifth International Conference on Language Resources and Evaluation (LREC 2006). Genoa, Italy (2006)
5. Kipper, K., Palmer, M., Rambow, O.: Extending PropBank with VerbNet Semantic Predicates. In: Workshop on Applied Interlinguas, held in conjunction with AMTA-2002 (2002)
6. Klein, D., Manning, C. D.: Accurate Unlexicalized Parsing. In: Proceedings of the 41st Meeting of the Association for Computational Linguistics (2003) 423–430
7. Levin, B.: English Verb Classes and Alternations: A Preliminary Investigation. University of Chicago Press, Chicago, IL (1993)
8. Palmer Martha, M., Gildea, D., Daniel, Kingsbury Paul, P.: The Proposition Bank: an annotated corpus of semantic roles. Computational Linguistics Vol. (2005) 31 No. 1(1): (2005) 71–105
9. Ungváry R.: Az ontológiák legfelső generikus szintje, a csúcsfogalmak természetes rendszere és a DOLCE kritikája. In: Alexin Z., Csendes D. (szerk.): MSzNy 2006 – IV. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2006) 85–96
10. Vincze, V., Csirik, J.: Hungarian Corpus of Light Verb Constructions. In: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010). Coling 2010 Organizing Committee, Beijing, China (2010) 1110–1118