# Felszíni szintaktikai elemzés és a jóindulatú interpretáció elve információ-visszakeresésben

Gyarmathy Zsófia, Simonyi András, Szőts Miklós

Alkalmazott Logikai Laboratórium, Budapest
e-mail:{szots,simonyi}@all.hu, gyzsof@gmail.com

**Kivonat** Tanulmányunkban egy új szintaktikai elemzési megközelítésre teszünk javaslatot, amely egy, a szemantikai predikátum-argumentum viszonyokra építő, orvosi ajánlásokban történő információ-visszakeresést megvalósító rendszerben kerül alkalmazásra. Szakítva a hagyományos, mélyelemzési módszerekkel, egy fedésorientált, „jóindulatú interpretáció-val" kiegészített felszíni szintaktikai elemzést javaslunk. Másszóval nem kívánjuk meg a szintaktikai reprezentáció helyességét, azaz nem törekszünk a pontosságra, hanem sokkal inkább csak a fedésre. A keresés pontosságát ehelyett szemantikai információk és a keresőkifejezés segítségével javítjuk. Bemutatjuk, hogy ezáltal csak kevés esetben lesz rosszabb a pontosság, míg számos jelenség (például kontrolligék, koordinációk, szabad határozók) esetében komoly előnyt jelent a javasolt fedésorientált megközelítés.

**Kulcsszavak:** felszíni szintaktikai elemzés, szemantikus információ-visszakeresés, frame-szemantika, argumentumstruktúra

## 1. Felszíni szemantikai elemzés egy IR-rendszerben

A cukorbetegség hosszútávú kezelését támogató informatikai platform kifejlesztésére létrejött európai REACTION projekt[1] keretén belül egy orvosi ajánlásokban *információ-visszakeresést* (l. [5]) megvalósító rendszert építünk ki, amely számba veszi a predikátumokat és a hozzájuk tartozó argumentumokat. Ezáltal sokkal eredményesebb lehet a keresés (l. pl. [13]), javítva nem csupán a pontosságot (mivel a megfelelő argumentumrelációjú találatok magasabbra lesznek rendezve, azaz megkülönböztethetjük pl. a *Péter kedvenc tanára Mari* és a *Mari kedvenc tanára Péter* mondatokat), hanem a fedést is (mivel megtaláljuk a hasonló eseményeket kifejező mondatokat, valamint a hasonló argumentumstruktúrával rendelkező mondatokat is).

A szemantikai predikátum-argumentum struktúra azonosítása, amit felszíni szemantikai elemzésnek hívnak (l. pl. [9]), legalább az alábbi lépésekből áll (l. pl. [8]):

1. A predikátumok és az argumentumok határának beazonosítása, és az argumentumok predikátumokhoz kapcsolása. Ez alapvetően egy *szintaktikai lépés*

[1] Remote Accessibility to Diabetes Management and Therapy in Operational healthcare Networks, http://www.reaction-project.eu.

a folyamatban, és a korábbi megközelítések a szövegek *teljes* szintaktikai elemzését feltételezték a feladathoz. Még ha természetesen nem is kizárólag mélyelemzés adta szintaktikai jegyeket használtak fel a klasszifikációs algoritmusokban (hanem például olyanokat is, mint a POS-tag, azaz a nyelvtani kategória), minden elemzés hivatkozik az argumentumnak a teljes szintaktikai fában elfoglalt helyére is (erre szolgál például [3] esetében a „Parse Tree Path" jegy, [9] esetében a „Path" jegy stb.).

Ezek az elemzések – az elméleti nyelvészeti hagyományoknak megfelelően – egyértelmű és helytálló predikátum-dependens, illetve fej-argumentum viszonyokat feltételeznek, azaz a folyamat minden szintjén a legpontosabb reprezentációt kívánják meg. Mi ezzel a hagyománnyal kívánunk a REACTION projekt keretén belül szakítani, és a célszövegek esetében csupán *felszíni* szintaktikai elemzést javaslunk, amelynek az esetében *nem* tételezzük fel annak helyességét, azaz a folyamat ezen a pontján pusztán a maximális fedést kívánjuk meg, a *pontosságot nem* (l. további fejezetek).

2. Szükség van még természetesen a predikátum szemantikai típusának beazonosítására. Ez egy jelentés-egyértelműsítési lépés, amely esetünkben – mivel a FrameNet keretet használjuk – *frame-azonosítást* jelent. Erre a fázisra is számos megoldási javaslat van már az irodalomban (például [2]); mi azonban – egyelőre, a maximális fedést biztosítandó[2] – minden olyan frame-et megengedünk egy predikátum esetében, amelynél az fel van sorolva, az ismeretlen predikátumok esetében pedig úgy járunk el, hogy az igékre egy *alapértelmezett*, nagyon általános frame-et veszünk fel, míg főnevek és melléknevek esetén egyelőre nem tekintjük predikátumnak a frame-hez nem rendelt szavakat.
3. Az argumentumok felcímkézése megfelelő szemantikus szerepekkel (*semantic role labelling*), amely lépés természetesen erősen függ a használt szemantikai szerepektől. Általános, a nyelvészetből is jól ismert tematikus szerepekre, amilyeneket például a VerbNet lexikon ([4]) használ, könnyebb statisztikai tanuló algoritmust adni, mivel minden predikátum ugyanazt a szemantikai szerephalmazt használja, s így viszonylag nagy a mintamennyiség. Mi azonban – a rendszer egyéb előnyei miatt, l. [10,12] – a FrameNet [1] frame-szemantikai megközelítését alkalmazzuk, amelyben a szemantikai szerepek frame-specifikusak. Ez a minta szegénysége miatt[3] megnehezíti a statisztikai tanulást.[4]

[2] Jelenleg ugyanis a rendszer többi részének teljesítményét szeretnénk tesztelni, márpedig egy újabb független paraméter nagyon elbonyolítaná a mérést.

[3] Sőt, számolni kell „hibás" mintával is, mivel egyazon szerepnév más-más szerepet takarhat. Noha a legtöbbször az azonos nevű szerepek valójában hasonló általános szereprelációt tükröznek (l. [6]) – például a Goal szerep mint cél szinte minden esetben tekinthető ugyanazon általános szerepnek –, sok esetben mást takar az ugyanazon elnevezésű szerep – például a Patient a tematikus szerepek esetén is megszokott jelentése mellett az orvoslásbeli pácienst is jelölheti egyes egészségügyi frame-ekben.

[4] A probléma természetesen azokkal a predikátum-argumentum párokkal van, amelyek esetén a FrameNet szótár hiányos, és nem specifikálja, hogy az adott szintaktikai vonzat a predikátum milyen szemantikai argumentumának felel meg.

Ennek ellenére már számos FrameNet-alapú statisztikai SLR-algoritmust javasoltak (pl. [14,7]) legtöbbször az alapcikknek számító [3]-ra építve, különböző szintaktikai és szemantikai jegyeket felhasználva különböző klasszifikációs eljárásokkal; illetve vannak olyan javaslatok is, amelyek más erőforrások klasszifikációs eljárásainak kimenetét aknázzák ki a FrameNet típusú szerepek klasszifikációjához (pl. [10]).
Mi azonban egyelőre – egyszerűsítési okokból – a szótárban nem specifikált vonzatok esetében oly módon járunk el, hogy először az adott predikátum más vonzatkereteiben nézzük meg, milyen szemantikai szerepet kap az adott vonzattípus (tipikusan valamilyen prepozíciós bővítmény), majd második lépésben az azonos frame-hez tartozó, hasonló predikátumok vonzatkereteit nézzük át e célból, végül, amennyiben itt sem találtuk meg ezt a vonzattípust, alapértelmezett eseteket alkalmazunk. Természetesen az egyre tágabb körben talált mintához egyre kisebb megbízhatósági valószínűséget allokálunk.

A felszíni szemantikai elemzés a kétezres években lezajlott kiterjedt kutatások (pl. [3]) ellenére továbbra sem kielégítően megoldott, ezért, szemben a teljes szintaktikai elemzést feltételező gyakorlattal, megkíséreljük pusztán felszíni szintaktikai elemzéssel megközelíteni a felszíni szemantikai elemzés feladatát. Mivel ez a fent leírt rendszer első fázisát érinti, ezért alapvetően meghatározza a további lépések sikerességét is.

## 2. Felszíni szintaktikai elemzés

A projekt során a korábbi, MaSzeKer projektben[5] szabadalmi igénypontokra kifejlesztett elemzőrendszert ([12]) alakítjuk át a megváltozott feladatnak megfelelően. A szabadalmi igénypontok szintaxisa *kötöttebb* volt (például nem tartalmaz felszólító módú mondatokat), viszont egy igényponton belül szemantikailag nagyon közel álló entitásokról tartalmazott szemantikailag hasonló állításokat (például egyes kémai anyagok összetevőiről, jellemzőiről). A szabadalmi igénypontokhoz tehát elengedhetetlen a mély szintaktikai elemzés, hogy egészen pontosan beazonosíthassuk az egyes kifejezések közötti kapcsolatot a szemantikai reprezentáció kiépítéséhez.

Ezzel szemben a REACTION projektbeli cukorbetegséggel kapcsolatos ajánlások sokkal közelebb állnak a természetes nyelvhez, mint a kötöttebb szabadalmi szövegek, például vannak bennük „phrasal verb"-ök (pl. *carry out*), birtokos szerkezetek, folyamatos igeidő, névmások és mondatkezdő prepozíciós frázisok, ezen felül pedig messze nagyobb a bennük előforduló nyelvtani szerkezetek és a megfogalmazás változatossága. Emiatt a MaSzeKer-beli szövegekre kifejlesztett dedikált szintaktikai elemző nem tud velük megbírkózni. Mi több, bármilyen mélyelemzés sikertelenségre van ítélve, ha a célszöveg a ***blood ketone monitoring*** *with increased healthcare professional support is preferable to urine ketone monitoring* ***in young adults with type 1 diabetes***, míg a keresőkifejezés a „blood ketone monitoring of adults with type 1 diabetes".

[5] Modell Alapú Szemantikus Kereső Rendszer, TECH_08_A2/2-2008-0092.

Ezért inkább amellett döntöttünk, hogy feladjuk a szövegek teljes szintaktikai elemzését, és ehelyett egyfajta felszíni szintaktikai elemzést végzünk. A felszíni szintaktikai elemzésnek is több lépése van hagyományosan:

1. *POS-tagging*, azaz a szavak nyelvtani kategóriájának megállapítása. Ezen a ponton még nem térünk el a mélyelemzésektől.
2. *Chunking*, azaz az összetevők határainak kijelölése. A mi esetünkben ez alapvetően a MaSzeKer-ben kifejlesztett MagNP-kijelölő modult takarja.
3. *Relációfeltárás*, azaz az összetevők közötti szintaktikai viszonyok megállapítása. A jelen tanulmányban ennek a fázisnak egy újfajta, fedésorientált megközelítését mutatjuk be.

Ezen a területen is a statisztikai tanulóalgoritmusok, azon belül is a kevert tanulóalgoritmusok (ensemble learning) alkalmazása a jellemző [11]. Mi ezzel szemben i) a MaSzeKer elemzőrendszerbe jobban illeszkedő szabályalapú megközelítést alkalmazunk, és ii) ahogy fentebb említettük, az elemzésben nagyobb hangsúlyt fektetünk a fedésre, mint a pontosságra, azaz megengedünk „hibás" predikátum-bővítmény kapcsolatokat is a kialakuló szintaktikai reprezentációban. Így például a *treatment [of a patient] [with diabetes]* esetében a *diabetes* főnévi frázist egyaránt felvesszük a *treatment* és a *patient* bővítményeként, miközben csupán az utóbbi elemzés a helyes. Mivel azonban a keresőkifejezésben minden valószínűség szerint nem fogunk *treatment* és *diabetes* között olyan kapcsolatot találni, amely a *with*-es vonzatnak (eszközhatározó) felel meg, így ezt a hibás elemzést a keresés során nem fogjuk felhasználni.[6]

A mi rendszerünk továbbá hibrid rendszer, amennyiben a *főnévi frázisok szintjééig* – az általunk használt terminológiában a MagNP-k[7] szintjééig – *mélyelemzést* végzünk a szövegeken.[8] Mivel a MaSzeKer során egy jól működő modult fejlesztettünk ki a MagNP-k kijelölésére és szintaktikai elemzésére, ezt egy az egyben át tudjuk venni a REACTION projektbeli ajánlások elemzésére. Ami megoldandó, az a MagNP-k és a predikátumok közti (szintaktikai, szemantikai) viszonyok feltárása. Ez tehát lényegében az egyetlen modul, amit a MaSzeKer projekt során kialakított szintaktikai parszerben meg kell változtatni az ajánlásokbeli keresés céljából.

Ezen a ponton pedig összefonódik és egymást meghatározza a rendszerben a szintaxis, a szemantika és a keresés. Ugyanis a MagNP-k és a predikátumok közötti viszonyok megállapításában sokkal megengedőbbek vagyunk, mint egy mélyelemzés, azaz nagyobb a kötési lehetőség, és megengedjük, hogy egy MagNP több predikátum bővítménye is legyen (akár egyazon nyelvtani funkcióban is),

[6] Természetesen egyes esetekben ez a keresésvezérelte „szűrési" eljárás nem fogja eredményesen elkülöníteni a helyes és a helytelen kapcsolatokat, elfogadván helyteleneket is, azonban ezek aránya a szövegtípustól függ: A REACTION-beli szövegek (cukorbetegségekkel kapcsolatos ajánlások) jellegükből adódóan alkalmasak erre a megközelítésre.

[7] Egy MagNP egy minden utómódosítójától megfosztott főnévi frázis.

[8] Erre a célra újraírószabályokat alkalmazunk, azaz frázisstruktúra-nyelvtant használunk.

valamint hogy egy predikátumhoz több, ugyanolyan nyelvtani funkciójú bővítmény (például tárgy) kapcsolódjon. A több kötési lehetőség közül pedig azokat tartjuk majd meg, amelyek a *keresés*, illetve a szemantika[9] szempontjából a *legideálisabbak*: ezt nevezzük a *jóindulatú interpretáció elvé*nek.

## 3. A jóindulatú interpretáció elve

A jóindulatú interpretáció elvének működését a következő absztrakt példa illusztrálja. Tegyük fel, hogy a keresőkifejezésre felépített szemantikai gráfban megtalálhatók az $A$, $B$ és $C$ csomópontok, és a következő élek:

- $A \xrightarrow{arg1} B$
- $A \xrightarrow{arg2} C$

Továbbá tegyük azt is fel, hogy a (felszíni szintaktikai elemzéssel elemzett) illesztendő szövegre felépített szemantikus gráfban megtalálhatók az $A$, $B$ és $D$ csomópontok, és a következő élek:

- $A \xrightarrow{arg1} B$
- $A \xrightarrow{arg3} B$
- $A \xrightarrow{arg2} D$
- $B \xrightarrow{arg2} D$

Ekkor azt fogjuk „jóindulatúan" feltételezni, hogy az $A \xrightarrow{arg1} B$ él illeszkedik, tehát az illesztendő szöveg részleges találat a keresőkifejezésre. Ez akkor is fenntartható, amennyiben például a $A \xrightarrow{arg3} B$ és a $B \xrightarrow{arg2} D$ élek „hibásan" kerültek be a szemantikus gráfba, a pontosságot figyelmen kívül hagyó felszíni szintaktikai elemzés révén.

Innentől kezdve gyakorlati kérdés, hogy mennyire megszorított, illetve szabad szintaktikai kötési lehetőségek bizonyulnak a keresés pontossága *és* fedése szempontjából legideálisabbnak (megkeresve a legjobb „trade-off"-ot a két mérték között). Elképzelhető – a szövegek jellegétől függően –, hogy egy „anything goes", azaz megkötések nélküli fej-dependens összekapcsolás működik a legjobban, amennyiben megfelelő szemantikai eszközökkel (például szelekciós restrikciókkal) kordában tudjuk tartani az elemzések elburjánzását. Ehhez azonban szemantikai információval gazdagon feltöltött lexikonra van szükség, amely például specifikálja az egyes predikátumok megfelelő argumentumainak a szemantikai típusát (azaz a predikátum szelekciós restrikcióit). Noha a lexikális erőforrások fedése és információgazdagsága terén jelentős előrelépések történtek az elmúlt évtized során is, efféle szemantikai információ meglétére még kevéssé támaszkodhatunk a legtöbb szótári tétel esetén (l. 9. lábjegyzet).

Mi, részben ezért is, első körben egy megszorítottabb megközelítést választottunk, és megfogalmaztunk egy *véges szabályrendszert* arra vonatkozóan, hogy

[9] A lexikonban rendelkezésre álló szemantikai információ (elsősorban az egyes argumentumokra vonatkozó szelekciós megszorítások) jelenleg még elég korlátozott, ezért megszorító hatása egyelőre lényegében elhanyagolható.

az egyes esetekben milyen főnévi frázisokat milyen fejekhez köthetünk, és milyen mondattani szereppel. Ezáltal pusztán a *legreálisabb* elemzési lehetőségeket tartjuk meg (így továbbra is különbséget tudunk tenni a *Péter kedvenc tanára Mari* és a *Mari kedvenc tanára Péter* között szintaktikai szinten is), ám eközben teret hagyunk a jóindulatú interpretáció elvének, ami összességében számos előnnyel járhat a mélyelemzéses megközelítésekhez hasonlítva, ahogy lentebb érvelni fogunk. Ez a szabályrendszer azonban nem a mélyelemzéseknél megszokott formátumú (például újraírószabály) és pontosságú. Olyan típusú szabályok ezek, mint például „egy nem prepozíciós MagNP, ha követi a igét, akkor lehet a direkt és indirekt tárgya annak".[10]

A mély elemzés és az itt alkalmazott felszíni közötti alapvető különség abban áll, hogy az utóbbi „megengedőbb", ennek folytán *több lesz az „igaz pozitív"* találat, mert a kereső megtalál olyan ajánlásokat, amelyeket a mélyelemzés nem, vagy csak nagyon alacsonyra rendelt résztalálatként. Jelentősen javul tehát a rendszer *fedése*. Viszont éppen ezért *több lesz a „téves pozitív"* találat is, mert olyat szövegrészeket is találatnak vesz (egy predikátumhoz kapcsolva nem egybe tartozóakat), amelyek valójában nem azok. Ez csökkenti a rendszer *pontosságát*.

Reményeink szerint azonban az ajánlások esetében ez a pontosságcsökkenés alacsony lesz. Ha például *blood pressure*, *patient* és *high* szerepel egy ajánlásban, igen kicsi (persze nem nulla) a valószínűsége, hogy egy magas páciens vérnyomásáról van szó (*the blood pressure of a patient who is high*), tehát valószínű, hogy a *high* a *blood pressure*-re vonatkozik (*the blood pressure of the patient is high*); ugyanígy feltételezhetően minden egy mondaton belüli információ egyetlen páciensre vonatkozik. A felszíni elemzés tehát azért működhet a REACTION-beli ajánlásokon, mert az ajánlások jellemzően *rövidek*, így emiatt és a szövegtípus sajátossága miatt kicsi az esélye, hogy a keresésben szereplő főnévi és egyéb frázisok „rekombinálása" a szövegen belül sokszor hozna be téves pozitívot.

## 4. A felszíni elemzés előnyei

Az itt felvázolt, jóindulatú interpretáción alapuló, fedésorientált felszíni elemzés számos esetben lehetővé teszi a keresés jobb fedését, illetve kiválthat bonyolultabb dedikált szintaktikai modulokat. Fentebb már a „blood ketone monitoring" példáján bemutattuk, hogy a természetes nyelvben általánosságban is igen sokféle megfogalmazása lehet egyazon gondolatnak, ilyen esetekben pedig bármiféle mélyelemzés kudarcra van ítélve. Egy másik jó példája az itt felvázolt megközelítés előnyének ilyen szempontból a következő célszövegbeli részlet:

(1) *Cataract extraction should not be delayed [in patients with diabetes].*

[10] Mi a MaSzeKer-beli elemzőhöz hasonlóan egy dependencianyelvtant használunk a MagNP-k feletti szinten, ez a választás azonban az itt tárgyaltak szempontjából kevéssé releváns. Egy frázisstruktúra-nyelvtan például azonban alapjaiban összeegyeztethetetlennek tűnik az itteni koncepcióval, már pusztán amiatt, mert egy összetevőnek több szülőnódusa is kellene, hogy lehessen, valamint mert nem folytonos al-fákra is szükség lenne.

Az általunk alkalmazott szabályrendszer jelenleg felveszi az *extraction* fejhez a *patients in*-prepozíciós dependenst, hiszen teljesen reális lehet egy „*cataract extraction in patients with diabetes*" keresőkifejezés, amelyre helyesen, magasra értékelt találatként kapnánk meg a fenti részletet.

Egyes esetekben a mélyelemzésre felépített szemantikai reprezentáció is módosítható, kiegészíthető lehet megfelelő reasoninggel, azonban egy ilyen szintű reasoning modul komoly kihívásokat jelent, és igen kétséges, hogy az ehhez szükséges tudásbázis rendelkezésre áll-e vagy kiépíthető-e reális időkereteken belül.

Azonban ezen általános kérdéskör mellett vannak egyes specifikus jelenségek is, amelyeknek kezelése sokszor külön, dedikált modult igényelne, azonban egy a javasolthoz hasonló megközelítés mellett erre nem lenne szükség. Az alább részletesebben is bemutatott ilyen jelenségek a következők:

1. ECM/raising/controll igék,
2. koordináció,
3. szabad határozók.

A fent bemutatott felszíni elemzési módszerrel az angol *raising, control*, illetve *ECM igék* (azaz lényegében a megosztott argumentumok) esetében nincs szükség külön minimodulra a célból, hogy a főige alanya, illetve tárgya a beágyazott mondat igéjének is alanya legyen, és ezáltal a megfelelő élek megjelenjenek a szemantikus reprezentációban is. Különösen problémásak ezen igetípusok, ha nem is beágyazott mondatbővítményük van, mivel ekkor nem tudnánk általános szabályt alkalmazni. A következő mondat illusztrálja ezt az esetet:

(2) *Intensive management plus pharmacological therapies should be offered [to patients with diabetes].*

Ebben az esetben az „intensive management for patients with diabetes" keresésre az *offer* jellegű igék külön kezelése nélkül csak résztalálatot kapnánk, miközben valójában teljes találat. A fent vázolt jóindulatú megközelítésben azonban a „patients with diabetes" az „intensive management" vonzata is lenne, így magasabb találati értéket kapna a célszöveg erre a keresésre.

Egy másik nyelvi jelenség, amelynek esetében a javasolt elemzési módszer kiválthat egy külön, dedikált modult, a *koordináció*. A természetes nyelvekben igen szerteágazó az ellipszis, és az egyes összetevők koordinálása, ezek azonban – a triviálisabb esetektől eltekintve – komoly kihívást jelentenek a gépi szintaktikai elemzéseknek. Íme egy nem triviális koordinációt tartalmazó példa:

(3) *Sulphonylureas should be considered as first line oral agents in patients who are not overweight, who are intolerant of, or have contraindications to, metformin.*

Ha a keresőkifejezésünk „medications for patients allergic to metformin", a fenti célszöveget mélyelemzés esetén szinte kizárt, hogy megtaláljuk (legfeljebb olyan résztalálatként, ami nagyjából egy kiterjesztett kulcsszavas keresésnek felel

meg). Egy jóindulatú felszíni megközelítéssel kicsivel túlmehetünk ezen, mivel a „metformin" dependense lehet az „intolerant" fejnek (többek között például a „have" és a „contraindications" fejek mellett). Innentől pedig feltételezve, hogy helyes a frame-szemantikai osztályunk, és az „allergic" és az „intolerant" azonos frame-be tartozik, máris sikeresen nagyobb súlyt kap találatként az ajánlás.

Hasonló módon tudunk megküzdeni a *szabad határozók* problémájával. Ezeknek a disztribúciós lehetőségei még a kötöttebb szórendű angol nyelvben is igen szélesek, ami mélyelemzés esetén megnehezíti a megfelelő fejhez kötésüket. Mi több, amint az ismert „*see [a man] [with a telescope]*" példa is mutatja, valódi szerkezeti többértelműség is fennállhat, ami lehetetlenné teszi, hogy az egyetlen pontos reprezentációt megcélzó mélyelemzés *minden* esetben sikeres legyen. Az itt javasolt keretben azonban megengedjük, hogy egyazon prepozíciós bővítmény több fejhez is kapcsolódjon, azaz ilyen esetben az „a telescope" összetevő mind a „see"-nek, mind a „man"-nek dependense lesz, tehát egy esetben sem vesztünk találatot.

## 5. A felszíni elemzés veszélyei és feltételei

A jelen rendszerben a legalapvetőbb problémát természetesen a téves pozitív találatok jelentik. Bár – amint említettük – a cukorbetegséggel kapcsolatos ajánlások rövidek, és emellett sem jellemző rájuk, hogy több, szemantikailag hasonló állítást tartalmaznának, ettől azért egyes esetekben előfordulhat. Ez, sarkítva, azonban valamilyen szinten kikerülhetetlen: ha a keresőkifejezésben az áll, hogy teleszkópos embert nézünk, a célszövegben pedig „*see a man with a telescope*", akkor hiába értelmezendő a célszövegben úgy, hogy teleszkóppal nézzük az illetőt (ez kiderülhet egy hosszas szövegkörnyezetből impliciten), ez a rész óhatatlanul illeszkedni fog a keresőkifejezésre. Azaz mindig lesznek „kezelhetetlen" esetek, a cél csupán ezek számának minimalizálása, aminek eszköze alapvetően egy olyan szabályrendszer megfogalmazása, amely elég restriktív ahhoz, hogy a keresési pontosság elfogadható legyen, míg a fedést lényegében nem rontja.

Van azonban két specifikus nyelvi jelenség, amelynek kezelése elengedhetetlen egy jól működő fedésorientált felszíni elemzéshez. Az egyik a *zárójelben* álló összetevők problémája. Egy példa:

(4) *Obese adults with type 2 diabetes should be offered individualised interventions to encourage weight loss (including lifestyle, pharmacological or surgical interventions) in order to improve metabolic control.*

Ebben a példában problémát okozhat, hogy például a „weight loss" a szabályok alapján (ha a rendszer nem „látja" a zárójelet mint határt) a zárójeles részt kezdő „including"-nak lesz az alanya, hibásan.

A legegyszerűbb megoldás, hogy zárójelen belüli szöveget a szöveg többi részétől *elkülönülten* kell leelemezni szintaktikailag. Az elkülönült szintaktikai elemzés csak ritkább esetekben nem működik, például akkor, ha egy főnévhez tartozó előmódosító kerül zárójelbe, például „*(oral) medications*". A nyereség azonban sokkal nagyobb, mint a veszteség, és később természetesen dedikált modul is kidolgozható a zárójeles kifejezések hatékonyabb kezelésére és kiaknázására.

Problémát okoznak a keresés során a zárójelek mellett még a *többszavas kifejezések* (multi-word expression, MWE) is, mint a *for example*, *in the case of*, *in addition to*. Egyrészt ezeket mint dependenseket és/vagy fejeket hibásan fogja kötni az elemző: például az *in the case of* esetén a *case* valamilyen fej(ek)nek az *in*-es dependense lesz hibásan, míg őhozzá mint fejhez *of*-os dependensként lesz kötve az őt követő főnévi frázis – hibásan. Ezek a szintaktikai reprezentációból azután bekerülnek a szemantikai reprezentációba, így helytelen illesztések történhetnek. Másrészt ezek a kifejezések megakadályozhatják a szintaktikai szabályok helyes alkalmazódását, és így a dependensek helyes kötését: például ha egy szabály a fej és a dependens közötti prepozíciókra hivatkozik, a „*for example*"-beli *for* illeszkedni fog a szabálymintára, pedig a „*for example*" összetett kifejezés egy határozó. Az előfeldolgozás során tehát mindenképpen érdemes a többszavas kifejezéseket kijelölni egy külön modulban.

Végül felmerült olyan probléma, amely kevésbé a szintaktikai, sokkal inkább a *szemantikai* reprezentációt érinti. A fedésorientált elemzés miatt esetünkben a szintaktikai gráfok igen nagyok lehetnek: több élt tartalmaznak, mint egy pontos, „helyes" elemzés, sőt, akár csomópontból is több kerülhet be, mivel argumentummal rendelkező predikátum is több lesz potenciálisan egy ilyen megközelítésben (ez a névszói predikátumokban jelent számszerű növekedést). Azonban elképzelhető, hogy a keresés szempontjából kevéssé releváns csomópontok és élek illeszkedése fog magasra értékelni valójában nem releváns találatokat. Egy példa:

(5) a. Keresőkifejezés: *Elderly patient with diabetes. The patient has mobility problems.*
b. Célszöveg: *All people with diabetes, and people without diabetes with a GFR less than 60 ml/min/ 1.73 m2, should* **have** *their urinary albumin/protein excretion quantified. The first abnormal result should be confirmed on an early morning sample (if not previously obtained).*

Ebben az esetben a „have" mint fej (ráadásul nem is megfelelő értelmű) előfordulása a célszövegben magas relevanciát nyújt az irreleváns célszövegnek. A legmegfelelőbb megoldásnak erre a problémára a kulcsszavas keresés újszerű felhasználása lenne: a keresőkifejezésben a felhasználó által megadott kulcsszavak jelölnék ki a szemantikai gráf lényeges csomópontjait, és az ebből kiinduló élek illeszkedése súlyozottan számítana be a relevanciaszámításba. Márpedig a példabeli keresőkifejezésben a „have" egyértelműen nem lenne kulcsszó, így illeszkedése sem hozna be magas relevanciaszámmal irreleváns találatokat.

Úgy tűnhet, hogy az itt leírt problémák és megoldhatóságuk feltételei súlyos ellenérvet jelentenek a fedésorientált felszíni elemzéssel szemben. Mindezen feltételek fennállása azonban ugyanúgy szükséges egy *mélyelemző* parszert használó keresőrendszerben is, hiszen egy mélyelemző ugyanúgy hibás fej-dependens viszonyt fog feltételezni az „*in the case of*" esetén, ugyanúgy problémái lehetnek a zárójeles kifejezésekkel (ezzel problémával ugyanis a mélyelemzést használó MaSzeKer projekt során is találkoztunk), és ugyanúgy magasra értékelhet egy célszöveget a kevéssé kulcsfontosságú frame-ek és argumentumok illeszkedése. A különbség csupán annyi, hogy a legutolsó probléma a fedésorientált felszíni

elemzés esetén hatványozottan jelentkezik, mivel abban az esetben sokkal több él kerül be a szintaktikai, és ezáltal a szemantikai reprezentációba is.

Az itt felvázolt, jóindulatú interpretációval párosított felszíni szintaktikai elemzés módszere egyértelműen olyan esetekben használható sikerrel, ahol i) a fedés sokkal alapvetőbb fontosságú, mint a pontosság, és ii) a célszövegek megfelelő jellegűek, azaz egységenként relatíve rövidek, és nem tartalmaznak nagyon hasonló jellegű állításokat. Mind a szabadalmak, mind a cukorbetegséggel kapcsolatos ajánlások közötti keresés megfelel az i) pontnak, azonban míg az utóbbi a ii)-at is teljesíti, ennek a feltételnek a szabadalmi szövegek nem tesznek eleget. A szabadalmi szövegekre megfelelő kötöttebb mélyelemző ezzel szemben a cukorbetegséggel kapcsolatos ajánlásokon bukik el azoknak sokkal szabadabb nyelvtani szerkezetei miatt. Fontos tehát a keresési rendszer egyes moduljait minden esetben a tárgynak megfelelően megválasztani.

## Hivatkozások


1. Baker, C.F., Fillmore, C.J., Lowe, J.B.: The Berkeley FrameNet project. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1, ACL'98, Association for Computational Linguistics, Stroudsburg, PA, USA (1998) 86–90
2. Das, D., Schneider, N., Chen, D., Smith, N.A.: Probabilistic frame-semantic parsing. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, HLT '10, Association for Computational Linguistics, Stroudsburg, PA, USA (2010) 948–956
3. Gildea, D., Jurafsky, D.: Automatic labeling of semantic roles. Computational Linguistics, **28**(3) (2002) 245–288
4. Kipper, K., Dang, H.T., Palmer, M.: Class based construction of a verb lexicon. In: AAAI-2000 Seventeenth National Conference on Artificial Intelligence, Austin TX (2000)
5. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, New York, NY, USA (2008)
6. Matsubayashi, Y., Okazaki, N., Tsujii, J.: A comparative study on generalization of semantic roles in FrameNet. In: Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP (2009) 19–27
7. Moldovan, D., Girju, R., Oltenau, M., Fortu, O.: SVM classification of Framenet semantic roles. In: SENSEVAL-3 (2004)
8. Palmer, M., Gildea, D., Xue, N.: Semantic Role Labeling. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers (2010)
9. Pradhan, S., Ward, W., Hacioglu, K., Martin, J., Jurafsky, D.: Shallow semantic parsing using support vector machines. In: Proceedings of HLT/NAACL (2004) 233–240
10. Shi, L., Mihalcea, R.: Putting pieces together: Combining FrameNet, VerbNet and WordNet for robust semantic parsing. In: Computational Linguistics and Intelligent Text Processing (2005) 100–111
11. Stav, A.: Shallow parsing. Seminar in Natural Language Processing and Computational Linguistics (2006)

12. Szőts, M., Gyarmathy, Zs., Simonyi, A.: Frame-szemantikára alapozott információvisszakereső rendszer. In: Tanács, A., Vincze, V., eds.: IX. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2013) 275–288
13. Szpektor, I., Dagan, I.: Augmenting WordNet-based inference with argument mapping. In: Proceedings of the 2009 Workshop on Applied Textual Inference (2009) 27–35
14. Thompson, C.A., Levy, R., Manning, C.D.: A generative model for semantic role labelling. In: Senseval-3 (2003) 397–408