TECH_08_A2/2-2008-0092

MODELLALAPÚ SZEMANTIKUS KERESŐ RENDSZER KIDOLGOZÁSA

IDŐKÖZI SZAKMAI BESZÁMOLÓ

3. SZAKASZ

A SZEMANTIKUS NYELVÉSZETI ERŐFORRÁSOK ÁTTEKINTÉSE

4.7.1 melléklet

Alkalmazott Logikai Laboratórium
Szegedi Tudományegyetem

Verziókövetés

dátum	változtatás	szerző
2011-12-13	Első verzió	Gyarmathy Zsófia (ALL)
2011-12-16	Kiegészítés az 5. fejezettel és apróbb változtatások	Gyarmathy Zsófia (ALL)

Tartalomjegyzék

1. Kívánalmak a lexikális erőforrásokkal szemben

A szemantikus lexikonban a szabadalmi szövegek feldolgozásához három különböző típusú kifejezésről kell információt tárolnunk:

- **predikátumok**: ezek olyan kifejezések, melyek argumentumokat vesznek fel (pl. a „treat" ige); tipikusan igék, de főnevek és melléknevek is lehetnek.
- **szakkifejezések**: a jelen esetben elsősorban az orvosbiológiai és kémiai szakterminusokat értjük ezek alatt (pl. „citric acid"); ezeket jórészt névelemként felismeri az elemző.
- **nem predikatív kifejezések**: tipikusan főnevek, melléknevek és határozószók; csak úgynevezett tartalmi szavak, tehát a zárt osztályt alkotó, úgynevezett funkciószavakat (pl. névelők, prepozíciók) nem a lexikonon belül tároljuk.

Mindegyik típusú kifejezéshez más erőforrást érdemes felhasználnunk, mivel a lexikális erőforrások eltérnek hangsúlyukban, mind a lefedett kifejezések körét, mind az eltárolt információ típusát tekintve. Az egyes kifejezéstípusok szerint az alapvető szinoníma és hiper-, illetve hiponíma viszonyok kódolása mellett a következő kívánalmakat támasztjuk a lexikonnal szemben:

- **predikátumok**: A (tágan értett)[1] vonzatkeret kódolása, és az egyes (tágan értett) vonzatok szemantikai argumentumoknak való megfeleltetése.
- **szakkifejezések**: kiterjedt orvosbiológiai és kémiai szakterminusok, alternatív írásmódokkal.
- **nem predikatív kifejezések**: minél nagyobb fedést kell elérni (vagyis minél több kifejezés legyen a lexikonban).

Az alábbiakban a potenciális erőforrások tulajdonságait és hangsúlyait vizsgáljuk, elsősorban a fenti szempontok tükrében.

[1] A szakirodalomban gyakran csupán a kötelező bővítményeket tekintik vonzatnak; itt a legtöbb szabad határozót is „vonzatnak" tekintjük, hasonlóan a FrameNet „Valence Unit" fogalmához.

2. Lexikális erőforrások nem-predikatív kifejezésekhez

2.1. Medical Subject Headings (MeSH)

A National Library of Medicine által kifejlesztett MeSH (Medical Subject Headings) tezauruszt **orvosbiológiai szaklexikonként** érdemes felhasználni.

Előnyei:

- az online használat mellett le is tölthető és szabadon felhasználható üzleti célú alkalmazásra is
- évente frissítik, így szakmailag naprakész tartalom
- relatíve kiterjedt, 177 000 terminust tartalmazó szótár az orvosbiológiai szövegekhez
- fogalmi hierarchia, szinonímák és alternatív írásmódok, továbbá rövid magyarázó definíciók, amit ha szükséges, szintén fel lehet a későbbiekben használni (pl. szótári definíció alapú közelségszámításra)
- már sikeresen használták orvosbiológiai szövegek feldolgozásában, például Rosario and Hearst (2001) összetett főnevek tagjai közti relációk automatikus kikövetkeztetésében.

A MeSH fahierarchiája alkalmas lehet többek között a következők megsegítésére:

- a keresés során a szakterminusok esetén szinonímák, hipernímák és hiponímák megtalálására
- a predikátumok szelekciós restrikcióinak meghatározására és így az egyes argumentumok megkülönböztetésére (pl. a *treat* 'kezel' with-es vonzata egyaránt lehet betegség vagy gyógymód is különböző vonzatkeretekben, amely kettőt így el lehet különíteni)
- az összetett főnevek közötti relációk automatikus meghatározására (pl. az *aspirin treatment*, illetve a *cancer treatment* esetében más a tagok közti szemantikus kapcsolat)
- beágyazott koordinációk esetén a felsorolási szintek elkülönítésére (a hierarchiában közelebb eső kifejezések nagyobb valószínűséggel kapcsolódnak össze felsorolásban, mint a távolabb esők)

A MeSH-en kívül vizsgált többi jelölt a szaklexikon szerepének betöltésére kevésbé alkalmas, mivel vagy *a)* korlátozottabb a kezelt terminustípusok köre (például kizárólag a kémiai vegyületeket kezeli), vagy *b)* kizárólag online használható, és nem tölthető le, vagy *c)* felhasználása korlátozott.

2.2. WordNet

A Princeton University által fejlesztett WordNet talán a legismertebb általános lexikális erőforrás, melynek fejlesztése már több évtizede folyik.

Előnyei:

- Nagy kiterjedtségű (több, mint 200 000 szó-jelentés párt tartalmaz).
- Finom megkülönböztetésű szinonímahalmazok, a szinonímahalmazok között hiper- és hiponímaviszonyok.
- Melléknevek közt egy fontos szintaktikai különbségtétel kódolva van: a csak predikatív, csak posztnominális és csak prenominális használat jelzése.
- A nyelvtechnológiai kutatásokban és alkalmazásokban egyaránt kiterjedt a használata, és már számos felhasználható eszköz és algoritmus létezik, amely a WordNet-ben található információt különböző célokra képes kinyerni (például jelentésegyértelműsítés, szóhasonlóság-számítás).

Hátrányai:

- Mivel általános tezaurusz, ezért a szakkifejezések kezelésére nem alkalmas.
- A predikátumok vonzatkeretéről, valamint a szemantikai argumentumokról nem tartalmaz információt, így a predikátumok kezelésére nem alkalmas.
- Túl finom megkülönböztetéseket tesz, főként igék között, így különböző szinonímahalmazokba kerülnek szinonim jelentésű szavak. Ezen segíthet a WordNetben kézzel feltöltött „wngroups”, amely a hasonló jelentésű igéket vonja össze, azonban ennek kiterjedtsége kicsi.
- A különböző POS-tagű szavak, így például a főnevek és az igék külön vannak tárolva, így a hasonló (különösen az eseményszerűség-) jelentésű szavak messze kerülhetnek egymástól. Segítséget nyújthatnak ugyan a

különböző POS-tagú szavak közötti WordNet-relációk (pl. „derivationally related"), valamint egy nominalizációs lexikon (pl. Nomlex, ld. Szpektor and Dagan 2009), ez azonban még korlátozott fedést biztosít csak, és kidolgozása munkaráfordítást igényel.

A WordNet-et jellemzői alapján tehát a nem szaknyelvi, nem predikatív kifejezések szótáraként érdemes felhasználni.

3. Lexikális erőforrások predikátumok kezeléséhez

3.1. PropBank

`http://verbs.colorado.edu/propbank/framesets-english/`

Jellemzők

- Az argumentum-jelölés igénként specifikus, de ugyanazon címkék használatosak; az első argumentumcímkék nagyjából konstans szemantikai szerepet kódolnak minden esetben:

  Arg0 ≈ (Dowty-féle) prototipikus ágens
  Arg1 ≈ prototipikus páciens

  A VerbNet-beli thematikus szerepek alapján a következőképp oszlik meg a címkék megfeleltetése:

  - Arg0 = Agent (85%), Experiencer (7%), Theme (2%), ...
  - Arg1 = Theme (47%), Topic (23%), Patient (11%), ...
  - Arg2 = Recipient (22%), Extent (15%), Predicate (14%), ...
  - Arg3 = Asset (33%), Theme2 (14%), Recipient (13%),...
  - Arg4 = Location (89%), Beneficiary (5%), ...
  - Arg5 = Location (94%), Destination (6%)

- A kiindulási pont egy tényleges korpusz annotálása, nem pedig egy előre kialakított nyelvészeti koncepció (szemben pl. a VerbNettel). Meglepő jelentések is megjelenhetnek így.

- Egy példa: *coat*

  Roleset id: coat.01 , *cover, apply something to a surface*, vncls: 9.8, framenet: .

Roles:
Arg0: entity causing covering, agent (vnrole: 9.8-Agent)
Arg1: covered (vnrole: 9.8-Destination)
Arg2: covering, coat (vnrole: 9.8-Theme)

Example: passive with covering
person: ns, tense: future, aspect: ns, voice: passive, form: participle

The carpets won't be glued down, and walls-1 will be coated [-1] with nontoxic finishes.*

ArgM-MOD: will
Rel: coated
Arg1: [*-1]
Arg2: with nontoxic finishes

Pozitívum:

- **Különleges argumentumok** is jelölve vannak:

  ArgM-ek (Arg#>5) típusai:

  - TMP: mikor?
  - LOC: hol?
  - DIR: hova?
  - MNR: hogyan?
  - PRP: miért?
  - REC: magát, magukat, egymást...
  - PRD: egy másik argumentumra hivatkozó, vagy azt módosító
  - ADV: egyéb

  A számozott, ige-specifikus argumentumok speciális igéket (pl. *elítél*) is enged reprezentálni, és/mivel:

- **Nincs önkényes döntés-kényszer** pl. Patient és Theme között. (Másrészről: keresés esetén egy explicit thematikus-szerepet jelölő lexikonban Patient esetén Theme-re is kereshetünk automatikusan, ez nem megoldhatatlan.)

- Külön jelölték a **light verb-öket és a szimmetrikus igéket**, ami nekünk szerencsés lehet annak kiszűréséhez, hogy melyik ige nem predikátumot jelöl, hanem például egy főnévi predikátumot „támogat" (pl. *make a promise*).

- A Penn Treebank-ből vett példák annotálva, ami tanulhatóság, tesztelés szempontjából jó lehet, vagyis van **annotált korpusz** hozzá.

Negatívumok:

- A VerbNet-hez hasonlóan **nem túl nagy méretű** (kb. 5 ezer ige).
- **Csak igéket** tartalmaz.
- **A számozott, ige-specifikus argumentumok** (főleg a magasabb argumentumszámoknál) **nem összehasonlíthatóak**. Pl. Arg2 már nagyon eltérhet igénként – ez keresés és következtetés szempontjából nem jó.

  Ehhez kapcsolódóan pedig a **hasonló típusú argumentumok** (pl. LOCATION) **különböző argumentumhelyre** (pl. Arg5, Arg6) kerülhetnek. Ez is nehezíti a következtetést és keresés-kiszélesítést.

Ez utóbbi hátrányok a MaSzeKer szempontjából döntőek, így a PropBank kiesik, mint a predikatív kifejezések szótára.

3.2. VerbNet

`http://verbs.colorado.edu/~mpalmer/projects/verbnet.html`
`http://verbs.colorado.edu/verb-index/vn/reference.php`

- Statisztika: 272 alapszintű igeosztály, 3769 lemma, 5257 igei jelentés. 55 szintaktikai restrikció mondatkomplementumokra, 23 thematikus szerep.

  A MeSH-en kívül a többi, jelen tanulmányban felsorolt lexikonhoz hasonlóan **általános** lexikon (nem domain-specifikus).

- Igeosztályok felállítása, melyekben az igék hasonló szintaktikai / szemantikai kapcsolatokat mutatnak.

  Egy osztályban a következők vannak specifikálva:

  - Az igeosztályhoz tartozó igék (**WordNet-jelentésükkel**).
  - Argumentumok **thematikus szerepekkel** (**szelekciós megszorításokkal** együtt, pl. +élő, +hely, +absztrakt)
  - Az osztály által használt **keretek**, amik meghatározzák a **szintaktikai** megjelenését az argumentumoknak.

  Az egy VerbNet-osztályba való tartozás feltételei nyelvészeti háttérrel megalapozottak, Levin (1993) munkájára építve:

1. **egységes szintaktikai viselkedés**: azonos szintaktikai keretek, vonzatkeret-alternációk. Pl. fill-9.8:

   Anna filled the box with books.
   Books filled the box.

2. **hasonló jelentés**: Levin szerint szemantikai predikátumokkal (pl. mozgás, állapotváltozás) jellemezhetők.

Példák az egyes osztályok által mutatott vonzatkeretalternációkra:

a. Sharon sprayed water on the plants.
b. Sharon sprayed the plants with water.

a. * Gina filled lemonade into the pitcher.
b. Gina filled the pitcher with lemonade.

a. Carla poured lemonade into the pitcher.
b. * Carla poured the pitcher with lemonade.

Számtalan ilyen alternáció engedélyezése/tiltása alapján alakulnak ki az igeosztályok.

- A nyelvészetben hagyományosan használt thematikus szerepeket alkalmazza (a **számozott** verziók a **szimmetrikus** igéknél használatosak, pl. a beszélget két argumentuma Actor1 és Actor2 lenne):

  – Actor – Actor1 – Actor2 – Agent – Asset – Attribute – Beneficiary – Cause – Destination – Experiencer – Extent – Instrument – Location – Material – Patient – Patient1 – Patient2 – Predicate – Product – Proposition – Recipient – Source – Stimulus – Theme – Theme1 – Theme2 – Time – Topic – Value

---

- Egy példa: a *coat* ige a fill-9.8 alá tartozik, melynek a következők a jellemzői:

  – Linkek más erőforrásokhoz a *coat*-ra: FrameNet-be 2; WordNet-be 3; Grouping-ba 1.

  – **Argumentumok**:
  Agent [+animate]
  Theme [+concrete]
  Destination [+location & -region]

  – **Szintaktikai keretek**:

NP V NP PP.theme
example: "Leslie staffed the store with employees."

syntax: Agent V Destination with Theme
semantics: motion(during(E), Theme) not(location(start(E), Theme, Destination)) location(end(E), Theme, Destination) cause(Agent, E)

NP V NP PP.theme
example: "Leigh swaddled the baby in blankets."
syntax: Agent V Destination in Theme
semantics: motion(during(E), Theme) not(location(start(E), Theme, Destination)) location(end(E), Theme, Destination) cause(Agent, E)

NP-LOCATUM V NP
example: "The employees staffed the store."
syntax: Theme V Destination
semantics: location(E, Theme, Destination)

NP V NP.destination
example: "Leslie staffed the store."
syntax: Agent V Destination
semantics: motion(during(E), ?Theme) not(location(start(E), ?Theme, Destination)) location(end(E), ?Theme, Destination) cause(Agent, E)

Pozitívumok:

- **Szelekciós restrikciók** is vannak benne, ami nem annyira gyakori a lexikális adatbázisoknál.
- **Összekapcsolás WordNet-tel** (megfeleltetés WordNet-beli szó-jelentés pároknak). Ez például jelentésegyértelműsítés szempontjából igen hasznos: a WordNet-re már bevált egyértelműsítő algoritmusok vannak, így azokat lehet akár a VerbNet-beli jelentésegyértelműsítésre is használni.
- Különböző típusú vonzatok, így **mondatbővítmények típusai** reprezentálva vannak a legújabb VerbNet-ben, sőt, egyes konstrukciókban lexikális elemek is reprezentálva vannak a szintaktikai keretben, pl:

  NP V S_ING „Success requires working long hours."
  NP V S_INF „I needed to come."
  IT V THAT S „It matters that they left."

- Az úgynevezett **kontroll** is reprezentálva van, pl. a subject control (amely esetben a főmondat alanya egyben a beágyazott mondat alanya is) „sc"-ként így:

  NP V S_ING
  Example: „They confessed stealing the money." Syntax: Agent V Topic <+be_sc_ing>

- Már kifejlesztett eszköz arra, hogy egyszerűen integrálható legyen bármely természetes nyelvfeldolgozó applikációba: `http://verbs.colorado.edu/verb-index/inspector/`
- A későbbiekben esetleges időreprezentálás esetén hasznos lehet, hogy a Moens and Steedman (1988)-hoz hasonló **időszerkezeti információt** is meghatároz az egyes igékre: a leírásában szereplő alap szemantikai predikátumok (pl. motion, contact, transfer-into) a felkészítő (during(E)), végső (end(E)), vagy utószakaszban (result(E)) igazak-e (vagy hamisak).

Negatívumok:

- **Csak igéket** tartalmaz.
- **Nem akkora a mérete**, mint pl. WordNetnek, így főleg a speciálisabb (A61K-hoz erősen kötődő) lexémák/literálok hiányozhatnak, illetve csak más jelentésben lehetnek reprezentálva.

Mivel a MaSzeKer-ben egyelőre nem lényeges sem az időreprezentáció, sem az alappredikátumokkal leírt jelentésreprezentáció, ezek az előnyök kisebb súllyal esnek latba. Ezzel szemben fontos lenne az igéken kívül más kategóriájú predikátumokat is kezelni (főként főneveket), illetve a keresés finomítása szempontjából az általánosabb thematikus szerepek helyett ideálisabb lenne a szituációkhoz kapcsolódó szereprelációk használata.

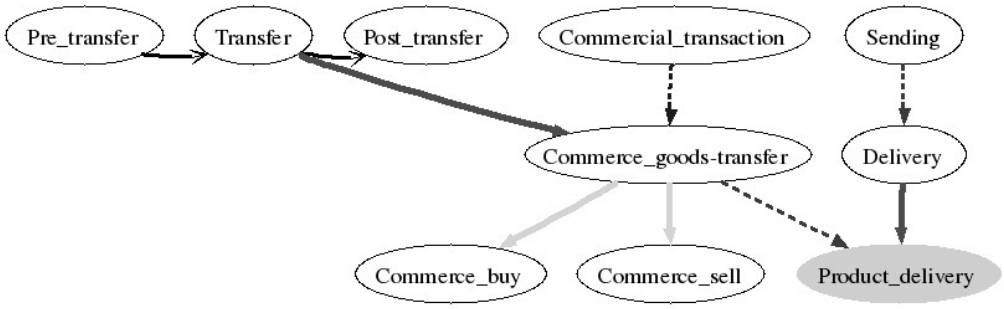
3.3. FrameNet

`http://framenet.icsi.berkeley.edu/`

3.3.1. Jellemzők

- A lexikonhoz annotált korpusz is tartozik.
- **Statisztika**:
  - 11 600 „lexical unit” (vagyis szó+jelentés pár), melyből 6,800 teljesen annotált;
  - 960 szemantikus „frame”, azaz egyfajta szcenárió, amelyhez a lexical unit-ok (LU-k) hozzá vannak rendelve;
  - 150,000 annotált példamondat.
- XML formátumban letölthető. Szabadon felhasználható Creative Commons Attribution-Only licensz alatt.

- „**Szemantikai keretek**”, vagyis egyfajta szcenáriók, „scriptek” képezik az alapját, pl. a vásárlás szcenáriója, fogalmi struktúrája:



  - Az argumentumokat (itt: **frame element**-eket, azaz FE-ket) ezen frame-eken belül definiálják.
  - A hiperníma- és szinonimareláció is a szemantikai kereteken át jelenik meg: azok a szavak „szinonímák” (tág értelemben), amelyek ugyanazt a frame-et „hívják elő”. A frame-ek között pedig több reláció is van, például:
    * **Öröklődés**: Az alárendelt a felérendelt **altípusa**, és minden felérendelt-beli frame element hozzákapcsolható egy alárendeltbeli frame elementhez.
    * **Alkeret**: Az alárendelt keret egy a felérendelt által leírt **összetett esemény részeseménye**, például a letartóztatás a büntetőfolyamatnak.

- **Predikátumok = igék, főnevek, melléknevek is:**
  - *[Item Colgate's stock]* ***rose*** *[Difference $3.64] [Final-value to $49.94]*
  - *...the* ***reduction*** *[Item of debt levels] [Value-2 to $665 million] [Value-1 from $2.6 billion]*
  - *[Sleeper They] [Copula were]* ***asleep*** *[Duration for hours]*

- **Frame elementek esetében** kódolja a szemantikai funkciót (pl. Cél), a grammatikai funkciót (pl. tárgy), valamint a nyelvtani kategóriát (pl. NP, vagyis főnévi csoport).

  Minden predikátumhoz megadja a frame elementjei mellett a lehetséges nyelvtani konstrukciókat, amikben megjelenhet, pl. *coat* esetében az egyik ilyen lehetőség:

Szemantikai fun.:	Agent	Goal	Purpose	Theme
Nyelvtani kat.:	CNI	NP	VPto	PP[with]
Nyelvtani fun.:	–	Ext	Dep	Dep

Példa erre a konstruckióra: *[$_{\text{Goal}}$ Each pellet of ammonium nitrate] was **COATED** [$_{\text{Theme}}$ with a hydrocarbon wax] [$_{\text{Purpose}}$ to reduce the effects of hygroscopicity].* (Az ágens nem jelenik meg.)

Előnyök:

- Nem csak igéket tartalmaz, hanem **főnévi és melléknévi predikátumokat is**.
- A frame-szemantikai megközelítés révén jól illeszkedik a MaSzeKer-beli koncepciónak megfelelő **„kontextus"-alapú szemantikus reprezentációba**, különösen ami az argumentumokat illeti.
- Ugyan kissé limitáltabban, mint a VerbNet, de tartalmaz **szelekciós restrikciókat**, vagyis megszorításokat az argumentumok (FE-k) szemantikai típusára.
- A „szinoníma"-fogalom itt kiterjedtebb, szcenário-alapú, nem a hagyományosan értett szinoníma-fogalom, ami a keresés szempontjából jelen esetben hasznos.

Hátrányok:

- Probléma VerbNethez és PropBankhez hasonlóan: **limitált számú, és nem a szabadalmi szövegekre szabott**, így a lexikon feltöltése / kiegészítése mindenképpen szükséges (amely lényegében csak humán erőforrással lehetséges).

  Érdemes kiterjeszteni például VerbNettel vagy WordNettel való összekapcsolással, ehhez ld. pl. Shi and Mihalcea (2005), illetve ld. pl. Shen and Lapata (2007) a szinonimákon keresztüli kiterjesztésre.
- A VerbNethez és PropBankhez hasonlóan a predikátumokat ragadja meg, így általános, **főnévi lexikonnak nem használható**.
- A frame-et azonosító információ **(jelentésegyértelműsítés) nélkül nehéz feladat az argumentumok címkézése** – ld. Giuglea and Moschitti (2006). Azonban például ők is javasolnak rá módszert, miként segíthető ez a feladat, méghozzá VerbNet-osztályokkal való összekapcsolással.

- A jelentésegyértelműsítő, azaz a **WSD hibázása nagyobb kárt okozhat** – a frame-specifikus szereprelációk miatt nagyobbat tévedhet a parser, ha rossz frame-hez köti az adott kifejezést, mint ha általánosabb szereprelációkat használnánk. Megoldás lehet például több frame együttes megengedése, majd a legvalószínűbb kiválasztása, ld. pl. Shen and Lapata (2007), ahol ilyen módszert követnek.

A MaSzeKer-ben az előnyöket és hátrányokat figyelembe véve a FrameNet-et alkalmazzuk a predikátumok lexikonjaként. Ehhez tehát a számunkra megfelelő formára kellett hozni a FrameNet lexikon adatait, és a hátrányok kiküszöbölésére kell törekednünk.

3.3.2. Jelentésegyértelműsítés (WSD)

Mint a hátrányok kapcsán láttuk, ez kulcsfontosságú a FrameNet jó használhatóságához: jó WSD modul esetén valószínűleg kiválóan használható a FrameNet a MaSzeKer-ben, míg egy nem elégséges WSD modul esetében rossz eredményekre vezethet.

A Burchardt et al. (2005) által kifejlesztett „WordNet Detour to FrameNet”[2] hasznos eszköz lehet a jelentésegyértelműsítés megsegítéséhez. Ez az algoritmus egy WordNet szó-jelentés párhoz megadja a hozzá tartozó legvalószínűbb frame-eket, így akár elegendő lehet egy WordNet-re már kidolgozott WSD-algoritmus is a megfelelő frame azonosítására.

3.3.3. A FrameNet és a MaSzeKer parser összehangolása

A FrameNet két kategóriában kódolja a szintaktikai információkat, ezeknek a megfeleltetését a MaSzeKer parsere által használt jelölésekhez már kidolgoztuk:

1. **GF – Grammatical Function:** A komplemens (beleértve egyes szabad határozókat!) szintaktikai viszonya a fejhez (avagy az ő kifejezésükkel: targethez). Ezeket targettípusonként (ige, főnév, melléknév, adverbium, prepozíció) elkülönítve adják meg, minden target esetében más meghatározást kap. A legalapvetőbb típusok az Ext (External), Obj (tárgy), Dep (Dependent, minden nemtárgyi komplemens). Azt is jelölik, ha különböző okokból nem jelenik meg egy konstituens a felszínen, ezek _NI típusú GF-ek, eltekinthetünk tőlük.

   Néhány példa a FrameNet – parser megfeleltetésre (igei predikátum esetén):

[2] `http://www.coli.uni-saarland.de/~albu/cgi-bin/FN-Detour-short.html`

FrameNet GF	**parser vonzat**
Obj	obj-complement
Dep	obj2-complement VAGY [P]-complement

2. **PT – Phrase type:** A konstituens (fő összetevőjének) szófaja. A különböző prepozíció-típusokat, ugyanúgy, ahogy a MaSzeKer parsernél, az adott prepozícióval jelölik (pl. PP[with]).

   Néhány példa a FrameNet – parser megfeleltetésre (igei predikátum esetén):

FrameNet PT	**parser POS-tag**
Poss	POS vagy PRP$
N	NN
PP	IN

Az egyes „lexical unit”-ok (a predikátumként funkcionáló kifejezések) lehetséges vonzatkereteit úgynevezett Valence Pattern-ek kódolják a GF-ek és PT-k segítségével. (Mivel korpuszalapú gyűjtésről van szó, ezek leginkább a fő argumentumok beazonosítására jók, a szabad határozókhoz azonban érdemes lehet az egyes valence patternekben megjelenő realizációikat bármely olyan másik VP-ben is elfogadni, ahol az adott pozíciót nem foglalja el másik FE.) Ezeket össze lehet vetni a MaSzeKer által használt vonzatkerettárral a GF- és PT-megfeleltetések segítségével, így a két vonzatkerettár kiegészítheti egymást.

3.4. Erőforrások főnévi predikátumok kezelésére

A későbbiekben ezen erőforrások lehetőséget biztosíthatnak a FrameNet fedésének bővítésére, azonban a VerbNet-hez és PropBank-hez hasonlóan nem triviális a FrameNetnek, illetve az alábbi erőforrásoknak sem a szó-jelentés párjainak, sem azok argumentumainak megfeleltetése, amely alaposan kidolgozott algoritmus és kiterjedt kézi ellenőrzés nélkül számos hibát is behozhat a lexikonba, így ezt csupán azesetben érdemes megtenni, ha a potenciális nyereség meghaladja a potenciális hibák és befektetett munka árát.

3.4.1. NomLex

`http://nlp.cs.nyu.edu/nomlex/index.html`

- Nominalizációk kezelésére: argumentumot felvevő főnevek (és vonzatainak) összekapcsolása az azonos jelentésű igével és annak vonzataival.
- Kb. 1000 főnév, némelyik több igei kapcsolattal (pl. deduction: ← deduct / deduce)
- Van elméleti megoldás a light verbök (itt és a FrameNet-ben: support verb-ök) kezelésére, ld. Macleod (2002), ez azonban még nincs beépítve a jelenlegi verzióba.
- Letölthető és szabadon felhasználható.

3.4.2. NomBank

`http://nlp.cs.nyu.edu/meyers/NomBank.html`

PropBank célját és módszerét követi, de az igék helyett a főnevekre koncentrál.

Előnyei

- PropBank-hez hasonlóan ahol lehetséges, **linkek vannak a VerbNet-beli jelentésekhez**.
- **Nem csak nominalizációkat kezelnek** (állításuk szerint kb. az argumentumot felvevő főnevek fele része ilyen), mint Nomlex, hanem pl. relációs főneveket (pl. apa), illetve melléknevek nominalizációit (pl. „incompetence”, „ability”, „wisdom”), és partitív (X of X) konstrukciókban szereplő főneveket (pl. „barrage”, „clump”, „variety”).
- A **light (=support) verb-öket** az annotációban kezelik.

Hátrányai

- PropBank-hez hasonlóan nem tartalmaz explicit vonzatkeret-információt, csak a példamondatok alapján megállapítható az egyes argumentumok nyelvtani funkciója.
- PropBank-hez hasonló szerepreláció-reprezentáció (számozott, predikátum-specifikus argumentumok).

4. A nem szaknyelvi lexikonok összehasonlítása

A **Verbnet – WordNet – FrameNet** lexikonok több szempontból különböznek egymástól, többek közt abban, hogy milyen szavakat (lemmákat) kezelnek és tárolnak, milyen elven csoportosítják a szavakat, milyen relációkat kódolnak és mik között állnak fenn az adott relációk, valamint hogy milyen típusú információt tárolnak az egyes szavakról és csoportokról.

Lemmák:

1. **VerbNet osztályokban**: igei lemmák
2. **WordNet synsetekben**: igei, főnévi, melléknévi, határozói lemmák, de elkülönült hierarchiában
3. **FrameNet frame-ekben**: igei, főnévi, melléknévi, határozói lemmák, akár egy csoporton belül is

Csoportok:

1. Egy **VerbNet osztály**: azonos **szintaktikai** viselkedés ÉS hasonló **jelentés**
2. Egy **WordNet synset**: közel azonos **jelentés**, NEM vizsgálja a szintaktikai kereteket
3. Egy **FrameNet frame**: azonos **szituáció**, lemmánként (LU) a szintaxis ELTÉRHET, teljesen más jelentésük (pl. ellentétek), nyelvtani kategóriájuk (pl. ige, főnév) is lehet

Relációk:

1. **VerbNet osztályok**: öröklődés
2. **WordNet synsetek**: öröklődés, antoníma, mereológia
3. **FrameNet frame-ek**: öröklődés, perspektíva, „alframe”, inchoatív és kauzatív, megelőzés

Fókusz:

1. **VerbNet**: kifinomult szintaxis, körülbelüli szemantika, de nem szinonimitás
2. **WordNet**: kifinomult szinonimitási viszonyok, figyelmen kívül hagyott szintaxis és mélyebb szemantikai összefüggések
3. **FrameNet**: kifinomult szituációelemzés = szemantika, relatíve pontos szintaxis, lényegében figyelmen kívül hagyott szinonimitás

5. A nem szaknyelvi lexikonok összehangolása

A fent felsorolt nem domain-specifikus erőforrások által tárolt információk egyesítésére és így a különböző lexikonok együttes kihasználására is volna lehetőség, amelyet megkönnyítene, hogy egy félig automatikus megfeleltetés már létezik közöttük (beleértve az argumentumstruktúra összehangolását): ez a University of Colorado által fejlesztett **Unified Verb Index**, ld. alább.

Azonban egyelőre nem éri meg a lexikonok összehangolását megcélozni, mivel, ahogy láttuk, a FrameNet a jelen céloknak az elérhető erőforrások közül a legjobb mértékben megfelel, míg a többi lexikonból nem nyernénk annyi plusz hasznot, amely jelen esetben indokolná a sok munkát igénylő és potenciálisan új hibákat behozó erőforrás-egységesítést. Mindazonáltal bizonyos feladatokhoz érdemes lehet mégis bizonyos információkat a FrameNet-en kívül egyéb erőforrásokból bevonni: fentebb a jelentésegyértelműsítés kapcsán már beszéltünk a WordNet – FrameNet kapcsolatról, illetve lentebb az ú.n. szelekciós restrikciók kapcsán merül majd fel a FrameNet-en kívül más erőforrások használata.

5.1. Unified Verb Index

`http://verbs.colorado.edu/verb-index/`

A University of Colorado által fejlesztett félautomatikus egységesítő erőforrás, amey a WordNet, a VerbNet, PropBank és a FrameNet szó-jelentés párjai, valamint emelett az utóbbi három esetében a szemantikai argumentumok közötti megfeleltetéseket tárolja.

Statisztika: 5726 VerbNet link, 4592 PropBank link, 4186 FrameNet link.

Példa a VerbNet– FrameNet kétfázisú megfeleltetésére:

1. fázis: VN-class (osztály) és FN-frame (keret) sok-a-sokhoz típusú megfeleltetése:

```
<vncls class='9.1-2' vnmember='put' fnframe='Placing'
 fnlexent='5355' versionID='vn2.0' />
```

2. fázis: VerbNet thematikus szerepek és FrameNet Frame Element-ek megfeleltetése:

```
<vncls class='9.1' fnframe='Placing'>
    <roles>
```


```
      <role fnrole='Agent' vnrole='Agent'/>
      <role fnrole='Cause' vnrole='Agent'/>
      <role fnrole='Goal' vnrole='Destination'/>
      <role fnrole='Theme' vnrole='Theme'/>
    </roles>
  </vncls>
```

Természetesen **különböző problémák elkerülhetetlenek** a megfeleltetéssel: például előfordul, hogy egy adott szó-jelentés pár hiányzik valamelyik lexikonból (pl. a VerbNet *put-9.1*-es osztályából a *sling* nem szerepel a FrameNet-ben), vagy nem hiányzik ugyan, de részben más argumentumokkal szerepel (pl. a VerbNet-beli *put-9.1*-es osztály Agent argumentuma a FrameNet „place" frame-jének Agent és Cause argumentumának is megfelel). Ezen problémák gyakran még kézi megfeleltetés esetén sem megkerülhetőek. Emellett az is előfordul, hogy a Unified Verb Index-hez használt algoritmus rosszul végezte el a megfeleltetést, és még nem javították a hibát kézzel.

5.2. Szelekciós restrikciók

A szelekciós restrikciók egy predikátum argumentumainak szemantikai típusát szorítják meg. Segíthetnek beazonosítani, hogy egy adott NP vagy PP melyik prediktáum melyik argumentuma lehet, azonban érdemes óvatosan bánni velük, mivel gyakran megsérthetőek, így inkább heurisztikának jó.

Mivel a **VerbNet-ben sokkal több** szelekciós restrikció van felsorolva, mint FrameNet-ben, mindenképpen érdemes előbbi erőforrást (is) kihasználni, különösen, mivel igen nagy számú ige esetében a Unified Verb Index-en keresztül megvan a megfeleltetés a VerbNet osztályok és argumentumok, valamint a FrameNet frame-ek és frame elementek között.

Ahhoz, hogy a szelekciós restrikciókat fel tudjuk használni, először is a főnevek kezelésére kijelölt erőforrás, a **WordNet megfelelő csomópontjainak** kell megfeleltetni azokat. Ezek a megfeleltetések sokszor egy-a-sokhoz típusúak (ehhez ld. még Shi and Mihalcea 2005), vagyis egy szelekciós restrikciót több WordNet csomópont is lefed. Emellett a fedés fontosságát szem előtt tartva gyakran a szelekciós restrikció által megköveteltnél (pl. *elongated*) sokkal általánosabb WordNet csomópontot kell felvenni (pl. object#1).

A VerbNet-beli szelekciós restrikciók WordNet-csomópontokhoz rendelését kézileg elvégeztük, néhány példa ebből (ahol egy adott WordNet-synsetnek megfelelést +-szal, az azt kizárást −-szal jelöltük, valamint megengedtünk AND és OR típusú koordinációt):

VerbNet szel.r.	**WordNet**
abstract	−physical_entity#1
int_control	+animate_thing#1 OR +machine#1
organization	+group

A VerbNet mellett érdemes magának a FrameNet-nek a restrikcióit is kihasználni ugyanilyen módon. Ezek némelyikét ugyan csak Lexical Unit-oknál, és nem Frame Element-eknél (avagy predikátumoknál, nem argumentumoknál) használják a FrameNet-ben (pl. a body_of_water-t), de a legtöbb hasznos lehet az argumentumok helyes beazonosításánál. A FrameNet-beli szelekciós restrikciók WordNet-csomópontokhoz rendelését is elvégeztük kézzel, íme néhány példa:

FrameNet szel.r.	**WordNet**
Physical_entity	+entity#1
Region	+geological_formation OR +body_of_water
Duration	+time_unit#1 OR +time_period

Összefoglalva, **a szelekciós restrikciók kihasználása érdekében a következő kapcsolatokra van szükségünk** az egyes erőforrások között:

i) VerbNet szelekciós restrikciók – WordNet synsetek,
ii) FrameNet szelekciós restrikciók – WordNet synsetek,
iii) VerbNet – FrameNet szó-jelentés párok.

Az első kettőt az szükségelteti, hogy a főneveket és mellékneveket (amelyek a tipikus argumentumok) a WordNet-ben kezeljük, így az argumentumok típusának beazonosítására a WordNet tud szolgálni. A harmadikra azért van szükség, hogy a VerbNet-beli predikátumok szelekciós restrikcióit a FrameNet-beli predikátumoknak tudjuk megfeleltetni a megfelelő argumentumokra.

Emellett természetesen még szükség lesz **saját restrikciók** kidolgozására, elsősorban a betegségek és a kémiai vegyületek, anyagok területén, amelyeket egy NER-felismerő lát el címkékkel. Így elkülöníthető a „treat a patient with [an illness]” és a „treat a patient with [a kind of medicine]”. A NER-tagek mellett még például a következő WordNet-csomópontokhoz (és a megfelelő szaklexikonbeli csomópontokhoz) tartozó restrikciókra érdemes rátanulni:

- pathological state
- drug OR medical care OR care#1
- chemical substance

Hivatkozások

Burchardt, A., K. Erk, and A. Frank (2005). A WordNet detour to FrameNet. *Sprachtechnologie, mobile Kommunikation und linguistische Resourcen 8*, 408–421.

Giuglea, A.-M. and A. Moschitti (2006). Semantic role labeling via FrameNet, VerbNet and PropBank. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pp. 929–936. Association for Computational Linguistics.

Levin, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago, IL: University of Chicago Press.

Macleod, C. (2002). Lexical annotation for multi-word entries containing nominalizations. In *Proceedings of Third International Conference on Language Resources and Evaluation (LREC 2002)*, Spain.

Moens, M. and M. Steedman (1988). Temporal ontology and temporal reference. *Computational Linguistics 14*(2), 15–28.

Rosario, B. and M. Hearst (2001). Classifying the semantic relations in noun compounds via a domain-specific lexical hierarchy. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing (EMNLP-01)*, pp. 82–90.

Shen, D. and M. Lapata (2007). Using semantic roles to improve question answering. In *Proceedings of EMNLP-CoNLL*, pp. 12–21.

Shi, L. and R. Mihalcea (2005). Putting pieces together: Combining FrameNet, VerbNet and WordNet for robust semantic parsing. *Computational Linguistics and Intelligent Text Processing*, 100–111.

Szpektor, I. and I. Dagan (2009). Augmenting WordNet-based inference with argument mapping. In *Proceedings of the 2009 Workshop on Applied Textual Inference*, pp. 27–35.