Egy általános modellnek megfelelő szerkezetek kinyerése korpuszból

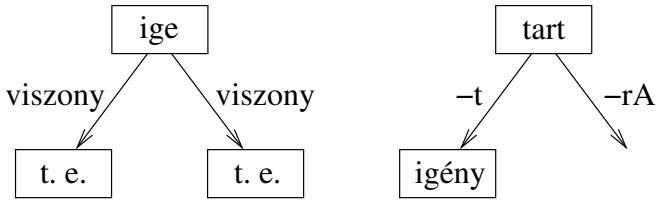
Sass Bálint – MTA Nyelvtudományi Intézet – 2011. június 16. 11h

1. Alapmodell

'*Részt vesz vmiben*', '*beleüti az orrát vmibe*', '*szó van vmiről*', '*hasznot húz vmiből*', '*kétségbe von vmit*', '*kockán forog vmi*', '*górcső alá vesz vmit*'... Bár a számítógépes nyelvészet és a lexikográfia is hajlamos külön, egymástól függetlenül kezelni a kollokációkat és a vonzatkereteket, számos olyan – a fentiekhez hasonló – jellegzetes kifejezés van, ami egyszerre mindkettő. Ezek sokszor idiomatikus jelentést hordoznak, szükség van rájuk egy tanulói szótárban és egy nyelvtechnológiai alkalmazás háttér-adatbázisában egyaránt.

A kollokátumot és a vonzatot ugyanazok a nyelvi eszközök – esetrag, névutó – jelölik. Egy olyan eljárást dolgoztam ki, mely képes ezeket a szerkezeteket korpuszból kinyerni, képes felismerni, elkülöníteni, hogy mikor melyik esetrag melyik szerepnek felel meg. Az eljárás az ábrán látható, erre kihegyezett egyszerű tagmondatmodellre épül:

tagmondat = ige + bővítmények
bővítmény = viszonyjelölő + tartalmi elem



Bal oldal: a modellnek megfelelő általános függőségi fa; jobb oldal: az '*igényt tart vmire*' reprezentációja (a vonzatnak a jobb alsó sarokban hiányzó csomópont felel meg.)

2. Általánosítás

A fenti modell nyelvfüggetlennek tekinthető, a magyaron kívül számos nyelvre előállítható a modell szerinti reprezentáció, és kinyerhetők a fenti típusú igei szerkezetek. Dán és szerb példákat fogunk látni.

A kinyerő algoritmus működésének feltétele lényegében pusztán a fenti gráfstuktúrának megfelelő formájú reprezentáció. Megtehetjük tehát, hogy elvonatkoztatunk az eddigi tartalomtól, és egyéb szerkezeteket próbálunk meg ebben a struktúrában reprezentálni.

Megmutatom, hogy a modell megfelelően módosítva alkalmas főnévi igenévi bővítmények kezelésére, helyhatározó, időhatározó fogalmának egységes kezelésére. Használhatjuk jellegzetes főnévi csoportok (pl.: '*belső fül*', '*szabad szemmel*'), valamint többszintű függőségi struktúrával bíró szerkezetek (pl.: '*gyenge lábakon áll*', '*jelentős szerepet játszik vmiben*') kinyerésére is.

Végül a holland–francia nyelvpáron megmutatom, hogy párhuzamos korpusz és párhuzamos igei szerkezetek (igei szerkezetek és fordításaik) szintén reprezentálhatók a fenti módon. Ebből a reprezentációból a változatlan kinyerő eljárás közvetlenül párhuzamos szerkezeteket gyűjt. Képes arra, hogy aszimmetrikus (azaz a két nyelven eltérő felépítésű) szerkezeteket is párba állítson, mint például a *részt vesz vmiben* holland és francia megfelelőjét. A holland szerkezet felépítése azonos a magyarral: '*nemen deel aan*'; a jellemző francia szerkezet viszont egy egyszerű vonzatos ige: '*participer à*'.