

A LEXIKAI PRAGMATIKA MAGYARORSZÁGON (2000–2020)

Pragmatika kerekasztal

Szeged, 2020. október 2.

A Magyar Nyelvtudományi Társaság Pragmatika Tagozata
Pragmatika Centrum Országos Kutatóközpont
SZTE BTK Általános Nyelvészeti Tanszék

Furkó Péter (Károli Gáspár Református Egyetem):

A diskurzusjelölők automatizált annotálása szemantikai mezők mentén

A diskurzusjelölők morfoszintaktikai szempontból heterogén, funkcionális csoportot alkotó elemek, melyek forrásszófajai között a kötőszóktól (*viszont, mégis, mindazonáltal, valamint*) kezdve a határozó(szó)kon keresztül (*ily módon, így*) a mondatszókig (*persze, dehogy, óh*) a nyelvi elemek széles skáláját megtaláljuk. Közös jellemzőjük, hogy szemantikai üresedésen mentek keresztül, tehát jellemzően minimális szójelentésben kódolt információval rendelkeznek, emellett szintaktikailag általában nem integráltak, melynek következtében elhagyásuk esetén sem sérül a mondat jólformáltsága. Kontextuális jelentésüket és funkcionális spektrumukat a kurrens szakirodalom leginkább pragmatikailag kezeli: a koherencia- és személyközi viszonyok széles körét fejezhetik ki, fontos szerepük van a megnyilatkozások közti kapcsolatok (például megerősítés, ellentét, ok-okozat) jelzésében, a kapcsolatteremtésben és a hallgató figyelmének irányításában, emellett udvariassági stratégiaként vagy a beszélő attitűdjének, bizonyosságának, illetve bizonytalanságának kifejezésére is szolgálhatnak, valamint kontextustükröző vagy -teremtő szerepet látnak el (például informális, kollokvialis hatás, a beszélő szociokulturális háttérét tükröző szerep).

Az előadásban az USAS (UCREL Semantic Analysis System) szemantikai címkézése alapján végzett kutatások eredményeit mutatom be. Az USAS olyan egyértelműsítési módszerek (*disambiguation methods*) összesített eredményei alapján történik, mint például a szófaji címkézés (*POS tagging*), az általános valószínűség-rangsorolás, az összetett kifejezések kibontása, a diskurzusdomén azonosítása és a kontextuális szabályok figyelembevétele. Az USAS a lexikai elemek szemantikai címkézését a korábbi tesztelések szerint 91%-os pontossággal végezte el, vagyis a szemantikailag gazdag és szemantikailag kiüresedett (pl. indexikális) elemekre egyaránt 9%-os hibaarányt jelentettek informális beszélt nyelvi korpuszok alapján. A kutatás során az USAS címkézését manuális annotáció révén teszteltük szintén beszélt nyelvi korpuszok alapján és annotátorközi egyezés számítása révén. Az eredmények azt mutatják, hogy a diskurzusjelölők célzott annotációja során is hasonló hibaarány figyelhető meg, mint az egyéb, kontextusfüggetlen(ebb) szemantikai tartalommal rendelkező lexikális elemek esetében. Tehát várakozásainkkal ellentétben a diskurzusjelölők, szemantikai üresedésük ellenére, az automatizált szemantikai annotáció számára nem jelentenek komoly kihívást.