Gradus Vol 7, No 1 (2020) 87-97
ISSN 2064-8014

# SZEMANTIKUS SZEREPEK AUTOMATIKUS CÍMKÉZÉSE TERMÉSZETES SZÖVEGEKBEN

# SEMANTIC ROLE LABELING ON NATURAL TEXTS

*Subecz Zoltán[1], Nagyné Csák Éva [2]*

[1] Informatika Tanszék, GAMF Műszaki és Informatikai Kar, Neumann János Egyetem, Magyarország
[2] Gazdasági szaknyelvek Tanszék, Kereskedelmi, Vendéglátóipari és Idegenforgalmi Kar, Budapesti Gazdasági Egyetem, Magyarország
https://doi.org/10.47833/2020.1.CSC.001

***Kulcsszavak:***
információkinyerés
adatbányászat
szövegbányászat
gépi tanulás
eseménydetektálás

***Keywords:***
information extraction
data mining
text mining
machine learning
event detection

***Cikktörténet:***
Beérkezett 2019. december 10.
Átdolgozva 2020. január 7.
Elfogadva 2020. január 20.

***Összefoglalás***

*Jelen tanulmányunkban bemutatjuk gazdag jellemzőtéren alapuló gépi tanuló megközelítésünket, amely automatikusan képes magyar nyelvű szövegekben szemantikus szerepek címkézésére függőségi elemző alkalmazásával. Munkánkban a vállalati vásárlások, tulajdonváltozások keretével foglalkoztunk. Jellemzőkészletünkben felszíni, morfológiai és a függőségi elemzés alapján kinyert jellemzőket használtunk fel. Ezen alapjellemzőket kiegészítettük a jellemzőkből számolt statisztikai arányokkal is. Megvizsgáltuk, hogy a modell hogyan teljesít egy gyakori célszóra önállóan, és a célszavak keretekbe összefoglalt csoportjára is.*

***Abstract***

*In this study we introduce a machine learning-based approach that can automatically label semantic roles in Hungarian texts by applying a dependency parser. In our study we dealt with the areas of purchases of companies and news from stock markets. For the tasks we applied binary classifiers based on rich feature sets. In this study we introduce new methods for this application area. Having evaluated them on test databases, our algorithms achieve competitive results as compared to the current English results.*

## 1. Bevezetés

Az Információkinyerés egyik fontos feladata a névelemek felismerése mellett az *események* detektálása [15,16]. A szövegekben lévő események felismerése, analizálása, és hogy hogyan viszonyulnak egymáshoz időben, fontos a szöveg tartalmának megismerésében. Az események detektálása mellett fontos azok szemantikus kapcsolatainak, vagy *szemantikus szerepeinek* megtalálása is (*szemantikus szerepek címkézése*, Semantic Role Labelling, SRL). Az események és azok szemantikus szerepeinek detektálását a természetes nyelvfeldolgozás sok területén lehet hasznosítani. Például az összegzéskészítés, gépi fordítás és a válaszkeresés területein.

Munkánkban a *szemantikus szerepek címkézésével* foglalkoztunk. Ez a szemantikus kapcsolatok azonosítását jelenti egy *szemantikus kereten* belül (semantic frame). A keretek eseményeket írnak le azok szereplőinek szintaktikai és szemantikai megkötésein keresztül. Munkánkban a *vállalati vásárlások, tulajdonváltozások* keretével foglalkoztunk.

[1] Kapcsolattartó szerző: subecz.zoltan@gamf.uni-neumann.hu

A *szemantikus szerepek címkézése* napjainkban a természetes nyelvfeldolgozás (NLP) egyik legdinamikusabban fejlődő területe. Angol nyelvű szövegekre általában *konstituensfa alapú* szintaktikai elemzőt használnak az elő-feldolgozásnál az angol nyelv erősen konfiguratív tulajdonsága miatt, ahol is a legtöbb mondatszintű szintaktikai információt a szórenddel fejeznek ki. Ezzel ellentétben a magyar nyelv gazdag morfológiával és szabad szórenddel rendelkezik. A *függőségi fákkal* dolgozó elemzők különösen jól használhatóak szabad szórendű nyelvek elemzésére, így a magyarra is, ezek ugyanis könnyebben teszik lehetővé az egymással nem szomszédos, de összetartozó szavak összekapcsolását is. Ezért mi a magyar nyelvű szövegeinkre *függőségi fákkal dolgozó elemzőt* használtunk a *Magyarlanc* programcsomag segítségével [20]. A szövegek szavakra bontására, a szavak morfológiai elemzésére, szófaji egyértelműsítésére, és mondatok függőségi nyelvtan szerinti szintaktikai elemzésére is ezt alkalmaztuk.

A *függőségi nyelvtan* (Dependency grammar) a modern szintaktikai elméletek egy olyan csoportja, amelyik a függőségi kapcsolatokon alapul (ellentétben a konstituensfa-alapú kapcsolatokkal). A függőségi reprezentáció a szavakat a közöttük lévő kapcsolatok alapján kapcsolja össze és egy fastruktúrában ábrázolja. A fa minden csomópontja egy szót reprezentál, a gyerekcsomópontok azon szavak, amelyek függnek a szülőcsomóponttól és az ágat a kapcsolattal címkézzük. A fő ige a gyökér elem. Ha egy kapcsolati-egység több szót tartalmaz, akkor ezek a szavak egy részfát alkotnak a fő fán belül.

A mondat szavai közt fennálló kapcsolatok reprezentálásának másik elterjedt módja a *konstituens (constituent) fák* alkalmazása. Ebben egy konstituens egy szó, vagy szavak csoportja, ami egy egységként funkcionál egy hierarchikus struktúrán belül. Szavak csoportjai (eredeti sorrendben) egységeket alkotnak. A frázisok (phrase) több mint csak szavak csoportja. Olyan szavak csoportja, amelyek együtt töltenek be egy speciális szerepet a mondaton belül. Ezen szócsoportok együtt mozgathatóak, vagy helyettesíthetőek, miközben a mondat jól olvasható és nyelvtanilag helyes marad. A konstituens reprezentáció célja a mondatok ilyen rész-frázisokra való bontása. A természetes nyelvekben a frázisok egymásba ágyazva helyezkednek el, így a mondatokat fastruktúrában tudjuk ábrázolni.

A *szerepek* a legegyszerűbb esetekben a célszó *szintaktikai kapcsolatai* voltak, de nem mindig. Sokszor a keresett szerep távol helyezkedett el a függőségi fában a célszótól, gyakran a mondat másik felében. És olyan is volt, hogy a szintaktikai kapcsolat alapján várt helyen nem a keresett szerep volt. Ez utóbbi gyakran a szintaktikai elemző hibájából adódott. Így a feladat a függőségi fában a célszótól távolabbi szerepek megkeresése és a közelebbi hamis pozitív jelöltek kiszűrése volt.

## 2. Kapcsolódó munkák

A *szemantikus szerepek címkézése* napjainkban a természetes nyelvfeldolgozás (NLP) egyik legdinamikusabban fejlődő területe. Angol nyelvű szövegekre sok módszer született már, ezek általában konstituensfa alapú szintaktikailag elemzett mondatokat használnak, és mondat szinten vizsgálják az eseményeket.

Kezdetben az SRL munkákban csak igékkel foglalkoztak, az igéket önállóan vizsgálták és általános szerepeket kerestek (például Agent, Patient, Instrument). A PropBank korpusz [13] szövegeit használták fel, amiben angol nyelvű szövegekhez vannak kiemelt igék annotálva a hozzájuk tartozó szemantikus szerepekkel. A 2004-es és 2005-ös CoNNL feladatokban foglalkoztak ezzel a témával [2,4].

Később az igéket már nem önállóan vizsgálták, hanem tématerületenként csoportosították azokat (keretek). Az általános szerepek mellett már vizsgáltak domén-specifikus szerepeket is. Ehhez a FrameNet korpusz [7] szövegeit használták fel, amiben angol nyelvű szövegek vannak szemantikus szerepek szerint annotálva. Ezek is elsősorban igékkel foglalkoznak, de keresnek nem igei cél-szavakra is. Egy fontos alaptanulmányt készített D. Gildea és D. Jurafsky [8] az SRL témában. A Senseval-3 task [11] egyik része is a FrameNet-re alapozott SRL feladat volt. Az ACE program is más NLP feladatok mellett SRL témával is foglalkozik [1].

Xue és társa [19] a jelöltek számának csökkentésére mutatott be egy módszert. A jelöltek számát jelentősen csökkentették, miközben a fedést magasan tartották.

Koomen és társai [10] és Toutanova és társai [18] a szerepek azonosítása után a szerepek közötti kapcsolatokkal, függőségekkel foglakoztak. Azt vizsgálták, hogy a megtalált kifejezések hogyan lehetnek együtt a célszónak a szerepei.

Surdeanu és társai [17] és Pradhan és társai [14] számos SRL alapú rendszer kimenetét kombinálták egy rendszerbe.

Carreras és társai [2,3] és Surdeanu és társai [17] munkáiban, ha egy mondaton belül több célszó található, akkor ezeket nem csak egymástól függetlenül kezelték, hanem közös szerepeket is kerestek hozzájuk.

Johansson és társa [9] angol nyelvű szövegekre a konstituens alapú elemzés helyett függőségi elemzést használt.

Szemantikus szerepek címkézésére magyar nyelvű szövegekre is készültek már munkák. Farkas és társai [3] a szemantikuskeret-illesztésre *szabály alapú* módszert használtak. Mi gépi tanulásos módszert alkalmaztunk ugyanerre. A szabályalapú módszerrel ellentétben a gépi tanulásos módszer nem igényel annyi erőforrást és elő-feldolgozást, és automatikusan alkalmazható más doménekre is. Ehmann és társai [4] pszichológiai témájú szövegeken szemantikus szerepek címkézésénél csak két általános szerepet keresnek: az ágens és a recipiens szerepeket (cselekvő, elszenvedő). Mi a vállalatfelvásárlások kereten belül nem csak a két általános szerepet, hanem több domén-specifikus szerepet is címkéztünk. A következő szerepeket vizsgáltuk: Vevő, Eladó, Áru, Ár, Idő. Csak az igei és főnévi igenévi célszavak szerepeit kerestük.

Az angol nyelvű szövegekre általában *konstituensfa alapú* szintaktikailag elemzett mondatokat használnak. Az előző pontban ismertetett okok miatt mi a magyar nyelvű szövegeinkre *függőségi fákkal dolgozó elemzőt* használtunk a *Magyarlanc* programcsomag segítségével [20].

## 3. Szemantikus keretek és a szemantikus szerepek

Sok információkinyerő rendszer manapság *tárgykör (domén)* specifikus *keretekkel* dolgozik. Egy-egy tárgykör eseményeit célszerű egy kereten belül vizsgálni, hiszen ugyanazok a *szerepek* tartoznak minden eseményhez, ami egy adott csoporthoz tartozik. Például egy repülőjegy foglalásokat feldolgozó rendszer a következő *szerepeket* használhatja: indulási időpont, érkezési időpont, célállomás, indulási állomás, távolság, ár. Az előző rendszer *célszavai* lehetnek például: foglal, lefoglal, előjegyez, vált. Ha a célszavakat önállóan dolgozzuk fel, akkor csak kevés tanító adattal tudunk dolgozni. A célszavak *keretekben történő csoportosítása* jelentősen csökkenti ezt a problémát, hiszen a több célszó tanító adatai összeadódnak.

Munkánkban a *vállalati vásárlások, tulajdonváltozások keretével* foglalkoztunk, *igei és főnévi igenévi célszavakhoz* kerestük ki a szereplőket. A következő igei célszavakat vizsgáltuk meg az adott kereten belül: *vesz, vásárol, szerez, bekebelez, gyarapít, ad, átruház, értékesít, forgalmaz*. Valamint e célszavak minden igekötős, módbeli és időbeli változatát is. A célszavakhoz a mondatokon belül a következő szerepeket kerestük meg: *vevő, eladó, áru, ár, idő*.

**Példák** a szerepekre a vállalati vásárlások tárgykörben. A példákban vastag betűvel vannak kiemelve a ***célszavak*** és szögletes zárójelben a *szerepek* találhatóak. Alsóindexben szerepel az adott szerep típusa.

*[A svéd Electrolux]$_{Eladó}$ **eladja** [motorgyártó részlegét]$_{Áru}$ [az olasz Appliance Components Companies részvénytársaságnak]$_{Vevő}$ - tájékoztatott az Electrolux.*

*[A Deutsche Börse AG]$_{Vevő}$ pénteken bejelentette, hogy teljesen **megveszi** [a luxemburgi Clearstream elszámolóházat]$_{Áru}$.*

A példákban látszik, hogy egy szerep általában *több szóból* áll és a mondatok általában nem tartalmazzák mind az öt szerepet.

## 4. Felhasznált korpusz és programcsomagok

### 4.1. Felhasznált Korpusz

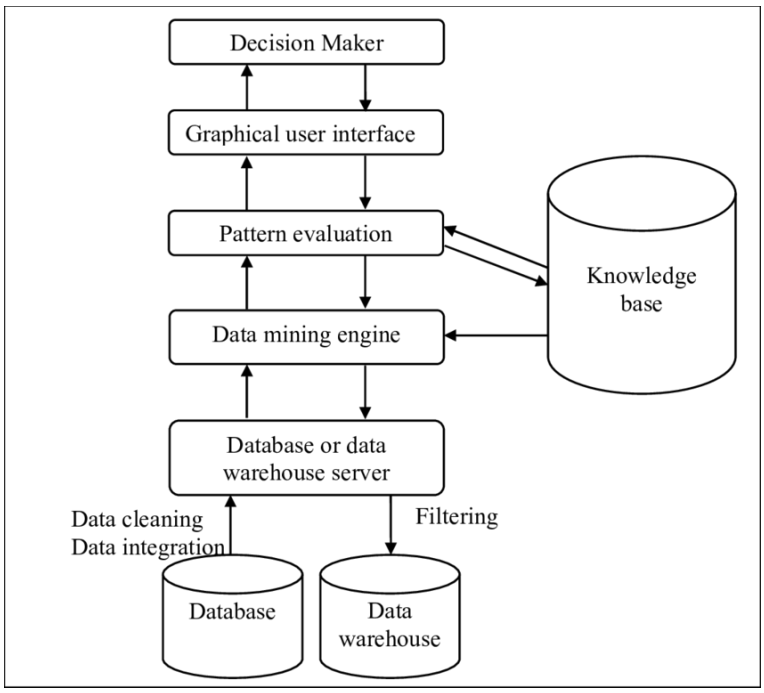
Az alkalmazásunk teszteléséhez a *Szeged Korpusznak* a rövidhírek csoportjának egy olyan változatát használtuk fel, amelyikben annotálva vannak a vállalati vásárlásokra a szemantikus

szerepek. Ezek közül 1000 mondatot használtunk fel. A tanításhoz és kiértékeléshez 10-szeres keresztvalidációt alkalmaztunk.

### 4.2. Felhasznált programcsomagok

A feladatokat *bináris osztályozásra* vezettük vissza. Az osztályozáshoz a *Weka* programcsomagnak[1] a J48-as döntési fa elemzőjét használtuk fel. A Weka adatbányászati feladatokhoz készített gépi tanuló algoritmusok gyűjteménye. A feladathoz felhasználtuk még a Magyarlanc 2.0 programcsomagot is. [20] A csomag magyar szövegek mondatra és szavakra bontására, a szavak morfológiai elemzésére, majd szófaji egyértelműsítésére, és mondatok függőségi nyelvtan szerinti szintaktikai elemzésére alkalmazható.

Egy tipikus adatbányászati szoftver architektúráját láthatjuk az 1. ábrán:




*1. ábra. Egy tipikus adatbányászati szoftver architektúrája*
*forrás: https://www.researchgate.net*

## 5. Magyarlanc programcsomag elemzésének bemutatása

A Magyarlanc a bemenetére érkező mondatoknak elkészíti az előző pontban leírt elemzését. A mondat minden szavához külön sorba elkészíti az elemzést (2. ábra). Minden szóról megadja a következő információkat: *sorszám, szó, lemma, szófaj, morfológiai kódok*. A sor végén megadja, hogy az adott szó melyik szóval van *szintaktikai kapcsolatban*, és hogy milyen a *kapcsolat típusa*. A szintaktikai kapcsolatok alapján a mondatok egy *függőségi fát* alkotnak.

Az elemzés után megjelenítettük vizuális elemzővel a szintaktikai kapcsolatok alapján a mondat függőségi fáját a program online elemzőjével[2] (3. ábra). Az elemzés és a vizuális ábrázolás egymásnak megfelelően megadja a szavak közötti *szintaktikai kapcsolatokat*.

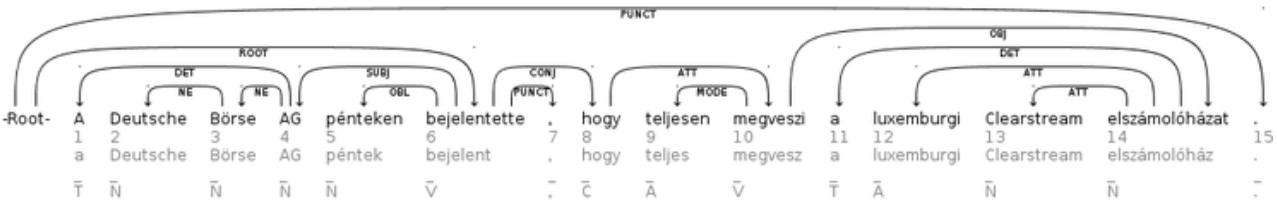
**Példa:**

*A Deutsche Börse AG pénteken bejelentette, hogy teljesen megveszi a luxemburgi Clearstream elszámolóházat.*

[1] http://www.cs.waikato.ac.nz/ml/weka/
[2] http://www.inf.u-szeged.hu/rgai/magyarlanc-service/

```
1     A       a       T       SubPOS=f                4       DET
2     Deutsche        Deutsche        N       SubPOS=p|Num=s|Cas=n|NumP=none|PerP=none|NumPd=none     3       NE
3     Börse   Börse   N       SubPOS=p|Num=s|Cas=n|NumP=none|PerP=none|NumPd=none     4       NE
4     AG      AG      N       SubPOS=p|Num=s|Cas=n|NumP=none|PerP=none|NumPd=none     6       SUBJ
5     pénteken        péntek  N       SubPOS=c|Num=s|Cas=p|NumP=none|PerP=none|NumPd=none     6       OBL
6     bejelentette    bejelent        V       SubPOS=m|Mood=i|Tense=s|Per=3|Num=s|Def=y       0       ROOT
7     ,       ,       ,       _       6       PUNCT
8     hogy    hogy    C       SubPOS=s|Form=s|Coord=p 6       CONJ
9     teljesen        teljes  A       SubPOS=f|Deg=p|Num=s|Cas=w|NumP=none|PerP=none|NumPd=none       10      MODE
10    megveszi        megvesz V       SubPOS=m|Mood=i|Tense=p|Per=3|Num=s|Def=y       8       ATT
11    a       a       T       SubPOS=f                14      DET
12    luxemburgi      luxemburgi      A       SubPOS=f|Deg=p|Num=s|Cas=n|NumP=none|PerP=none|NumPd=none       14      ATT
13    Clearstream     Clearstream     N       SubPOS=p|Num=s|Cas=n|NumP=none|PerP=none|NumPd=none     14      ATT
14    elszámolóházat  elszámolóház    N       SubPOS=c|Num=s|Cas=a|NumP=none|PerP=none|NumPd=none     10      OBJ
15    .       .       .       _       0       PUNCT
```

*2. ábra. A Magyarlanc programcsomag elemzése*




*3. ábra. A mondat függőségi fája*

Az elemzésekből látszik, hogy a függőségi elemző egy *szabályos elemző fát* készít. A fa legfelső eleme a *Root*. A fa *csomópontjaiban* vannak a mondat szavai, az *ágak* a szavak közötti *szintaktikai kapcsolatokat* reprezentálják. A fában kiemelt szerepe van az *igék*nek. A *főige* (a példákban a *bejelentette*) általában a Root alatt helyezkedik el, a szintaktikai kapcsolatokon keresztül ehhez kapcsolódnak a többi elemek.

Ha a szerep *több szóból* áll, akkor ezek a szavak egy *részfát* alkotnak a mondat fáján belül. A részfa a *kiemelt szaván* (fejszó, headword) keresztül kapcsolódik a fa többi részéhez.

Van, amikor a szerep kiemelt szava (headword) a célszóhoz kapcsolódik közvetlenül. Ilyen esetben könnyebb megtalálni a szerepet. Van, amikor a szerep kiemelt szava nem a célszóhoz kapcsolódik közvetlenül. A példamondatnál a vevő kiemelt szava (AG) nem kapcsolódik szintaktikailag a *megveszi* célszóhoz az elemzőfában, hanem a *bejelentette* igén keresztül. Ilyenkor nehezebb megtalálni a szerepet. Minél közelebb van a szerep a célszótól a mondaton vagy az elemzőfán belül, annál nagyobb a valószínűsége a szerep azonosításának.

Bár a Magyarlanc program elkészíti a mondatoknak a szintaktikai elemzését, de a példákon is láttuk, hogy a szintaktikai kapcsolat típusából nem következik a szemantikai szerep. Például a *vesz* célszónak az alanya a vevő, az *elad* célszónak az alanya az eladó. Így a szintaktikai kapcsolatok mellett több más tulajdonságot is meg kell figyelni a mondatban. A feladatot megnehezíti, hogy a Magyarlanc elemző is hibával dolgozik, így ez a hiba és a hibákból eredő hamis döntések megjelennek a mi eredményeinkben is. Jobb eredményeket kaptunk volna, ha szövegeink kézzel lettek volna annotálva ezen szempontok szerint.

## 6. Az osztályozás bemutatása

A célszavakhoz a következő szerepeket vizsgáltuk: *vevő, eladó, áru, ár, idő*. Minden bemeneti mondatnál adott volt a célszó. A feladat az adott szerep megkeresése volt.

Az osztályozóknál a *jelöltek* a függőségi elemzőfa csomópontjai voltak. Egy mondaton belül általában egy csomópont a keresett szerep kiemelt szava (head word). Az osztályozásnál ezek a *true* esetek, a többi csomópont pedig a *false* eset.

Az osztályozáshoz *bináris osztályozót* használtunk. Az osztályozó az adott mondatnál bejelöli a keresett szerepet. Az osztályozónak *nem adtuk meg*, hogy az adott mondat tartalmazza-e az adott szerepet, vagy sem. Voltak olyan mondatok is, amelyek nem tartalmazták a keresett szerepet. (1. táblázat)

A kiértékelésnél *szigorú szabályt* alkalmaztunk: csak azt a döntést fogadtuk el, amelyik pontosan az annotált szerepet jelöli meg. Sem az ezt tartalmazó fákat, sem ennek a részfáit nem fogadtuk el pozitív döntésnek. Ha ennél enyhébb szabályt alkalmaznánk, akkor magasabb eredményeket kapnánk.

### 6.1. Jellemzőkészlet

A tanító és a kiértékelő halmazon a *jelöltekhez jellemzőket* vettünk fel. Az SRL feladatokban használt általános jellemzőket [8] mi is alkalmaztuk. Ezeken kívül újakkal is kibővítettük a jellemzőkészletünket. Ehhez felhasználtuk a *függőségi elemzőfát* is, a jelölt és a célszó viszonyát a függőségi fában, mert ez gyakran egy fontos tulajdonsága az adott szerepnek.
A jelöltekhez a *következő jellemzőket* választottuk ki:

**Felszíni jellemzők**: *Bigramok, trigramok*: A vizsgált szavak végén lévő 2-es, 3-as betűcsoportok. *Pozíció*: a jelölt a célszó előtt vagy után áll a mondatban. *Távolság-mondatban*: a jelölt és a célszó szótávolsága a mondaton belül.

**Morfológiai jellemzők**: Mivel a magyar nyelv igen gazdag morfológiával rendelkezik, ezért számos morfológiaalapú jellemzőt definiáltunk. Jellemzőként definiáltuk az eseményjelöltek MSD-kódját felhasználva a következő morfológiai jegyeket: *típus(SubPos), mód(Mood), eset(Cas), idő(Tense), személy(PerP), szám(Num), határozottság(Def). Szófaj, Lemma*: a jelölt és a célszó szófaja és lemmája.

**Jellemzők az elemzőfa alapján-1**: Ide azokat a jellemzőket soroltuk, amelyeket az SRL feladatokhoz általában felhasználnak [8]. A jelölt és a célszó viszonyát vizsgáltuk a függőségi elemzőfában. Mindkettő egy-egy csomópont az elemzőfában. *Szófaj-útvonal*: Egymás után írtuk a jelölt és a célszó közötti csomópontok szófaját, feljegyezve azt is, hogy az elemzőfában felfelé, vagy lefelé haladtunk az adott kapcsolatnál. Például: C↑S↑V↑C↑V↑V↓V↓N↓N↓A. *Uralkodó-kategória-szófaja*: A jelölt és a célszó közötti útvonalon megkerestük a legmagasabban fekvő csomópontot, és feljegyeztük a hozzá tartozó szó szófaját.

**Jellemzők az elemzőfa alapján-2**: Itt az egyéni, új jellemzőket soroltuk fel. *Jelölt-célszó-távolság-elemzőfában*: A jelölt és a célszó csomópontjai közötti csomópontok száma az elemzőfában. *Lemma-útvonal*: Mint a Szófaj-útvonal, de itt a jelölt és a célszó között végigmenve a csomóponti szavak lemmáját jegyeztük fel. Például: Budapesti↑Értéktőzsde↑honlap↑közöl↓megvásárol. *Szintaktikai-kapcsolat-útvonal*: Az előzőhöz hasonlóan itt azt vettük fel, hogy a jelölt és a célszó között az elemzőfában milyen szintaktikai kapcsolatokon keresztül tudunk eljutni. Például: ↑COORD*SUBJ↓ATT↓INF↓OBJ↓ATT. *Jelölt-alatti-részfában-van-e-névelem*: A Magyarlanc program az elemzésében jelöli, ha talált névelemeket a mondatban. Mivel a vállalati tulajdonváltozások témakörében gyakran találkozunk vállalati névelemekkel, ezért felvettük, hogy a jelölt, vagy az alatta levő részfa tartalmaz-e névelemet? *Jelölt-alatti-részfában-névelem-távolság*: az előzőhöz hasonlóan megadtuk a részfában azt a mélységet, ahol először találtunk névelemet.

**Az osztályozás leírása**: Minden szerephez külön osztályozót készítettünk. Minden osztályozó esetén a tanító és a kiértékelő halmazon minden *jelölthez* felvettük a bemutatott *jellemzőket*. Ezen jellemzők kialakításához felhasználtuk a korpusz mondatainak kézi annotációját és az elemzőfák által adott elemzéseket. A tanító halmaz jelöltjeit és a hozzájuk kigyűjtött jellemzőket beolvastuk az osztályozóba a gépi tanításhoz. Az osztályozó a jellemzők alapján megismerte és megtanulta a „rejtett összefüggéseket” és szabályokat alakított ki. A kiértékelésnél a tanított osztályozó elemezte a kiértékelő halmaz mondatait, amelyekkel eddig nem találkozott. Az elemzés során minden jelölthöz bejelölte, hogy az adott szerephez tartozónak tartja vagy nem.

### 6.2. Statisztikai arány felhasználása az osztályozásnál

A jelöltekhez a jellemzőket *két módszer* alapján választottuk ki. *Első módszernél* az előző részben bemutatott alapjellemzőket használtuk fel. *Második módszernél* az alapjellemzők helyett a tanító adatokon a jellemzőkészletből számított statisztikai arányokat használtuk fel: a tanító halmaz alapján megszámoltuk minden jellemző esethez, hogy hány alkalommal fordult elő és ebből hányszor volt a jelölt *pozitív*. Ezek alapján kiszámítottuk a hozzá tartozó pozitív-arányt. Például ha a *Jelölt-lemma* jellemzőnél a *jelölt-lemma = Corp.* eset 11-szer fordult elő és ebből 7-szer volt *pozitív* eset (4-szer pedig negatív), akkor hozzá a 0,64-es pozitívarány tartozott. Ebben az esetben az osztályozónak a jelöltekhez nem az alap-jellemzőt, hanem a hozzá tartozó arányt adtuk meg. Az előző példánál *Jelölt-lemma-arány* = 0,64. Ezzel *jelentősen csökkentettük az osztályozó vektorterének méretét* az első módszerhez képest és így a futási időt is. Ez a kidolgozási idő-szakban hasznos volt. *Harmadik esetben* az előző két módszer jellemzőit együtt használtuk fel.

**A statisztikai-arány jellemzők hatása az osztályozás eredményére**. Megvizsgáltuk, hogy az előzőleg bemutatott *statisztika-arány jellemzők* hogyan befolyásolják az osztályozási eredményeinket. Először az osztályozást lefuttattuk csak a statisztikai-arány jellemzőkkel, majd csak az alapjellemzőkkel és végül a két jellemzőcsoporttal együtt. Volt, amikor a statisztikai-arány eset önállóan jobban teljesített, mint az alapjellemzőkkel eset önállóan. Volt, amikor fordítva. De a *legjobb eredményt* akkor kaptuk, amikor az alapjellemzőket és a statisztikai arány jellemzőket együtt használtuk.

### 6.3. Vektortér méretének csökkentése

A *vektortér méretét csökkentettük* a következő módszerrel: csak azokat a *jellemző-előfordulásokat* vettük fel az osztályozáshoz, amelyek a tanító halmazon *legalább háromszor* szerepeltek. Ezzel *jelentősen csökkentettük a futási időt* és csak az osztályozás szempontjából jelentéktelen jellemző-előfordulásokat hagytuk ki.

### 6.4. Célszavak csoportosítása a kereten belül

Először a modell viselkedését egy gyakori célszóra önállóan néztük meg. Ehhez a *vásárol* célszót választottuk ki.

Majd a célszavakat csoportosítottuk. A vásárlásokkal kapcsolatos mondatoknál a *vevő* és az *eladó* szerepek viselkedését meghatározza, hogy az adott célszónál az alany általában vevő vagy eladó. Ezért a célszavakat két csoportra bontottuk a következő egyszerű módszerrel. A *vevő-centrikus* csoportba azok a szavak kerültek, amelyeknél az alany általában a *vevő*: vesz, vásárol, szerez, bekebelez, gyarapít. Az *eladó-centrikus* csoportba pedig azok, amelyiknél az alany általában az *eladó*: ad, átruház, értékesít, forgalmaz. Ez a felosztás segítette a *vevő* és az *eladó* szerepek megtalálását. Egy harmadik esetben pedig nem végeztünk csoportosítást.

### 6.5. Baseline mérések

A Baseline módszereket a *döntési fa legfontosabb feltételei alapján* állítottuk össze. Azokat a jelölteket vettük pozitívnak, amelyekre teljesül: Az *Áru szerepnél* azokat, amelyek tárgy (OBJ) szintaktikai kapcsolatban vannak a célszóval. Az *Ár szerepnél* azokat, amelyeket egy előre elkészített pénznemek lista tartalmazott. Az *Idő szerepnél* azokat, amelyeket a következő lista tartalmazott: évszámok 1990-2014-ig, hónapnevek, napnevek, sorszámok 1-31-ig.

A *vevő-centrikus célszavaknál* a *Vevő szerepnél* és az *eladó-centrikus célszavaknál* az *Eladó szerepnél* azokat, amelyek alany (SUBJ) kapcsolatban vannak a célszóval. A *vevő-centrikus célszavaknál az Eladó szerepnél* azokat, amelyek végén a következő trigramok állnak: tól, től, ból, ből. Az *eladó-centrikus célszavaknál a Vevő szerepnél* azokat, amelyek részes eset (DAT) kapcsolatban vannak a célszóval. A következő eredményeken látni fogjuk, hogy gépi tanulási modell *jóval felülteljesítette a Baseline modellünket*.

### 6.6. Statisztikai adatok

Mondatok száma összesen: 1000 db

Azon mondatok száma, amelyek tartalmazzák az adott szerepet:

A fontosabb statisztikai adatokat a következő táblázat tartalmazza (1. táblázat):

*1. Táblázat. Statisztikai adatok (db)*

| *Célszavak* | *Mondatok száma* | *Vevő* | *Eladó* | *Áru* | *Ár* | *Idő* |
|---|---|---|---|---|---|---|
| kiemelt: vásárol | 265 | 263 | 107 | 276 | 104 | 99 |
| Vevő-centrikus | 548 | 531 | 222 | 573 | 214 | 208 |
| Eladó-centrikus | 452 | 261 | 374 | 459 | 82 | 115 |
| csoportosítás nélkül | 1000 | 783 | 579 | 1025 | 299 | 312 |

Az osztályozónak *nem adtuk meg*, hogy az adott mondat tartalmazza-e az adott szerepet, vagy sem. (Az Áru szerep azért nagyobb, mint a mondatok száma, mert volt olyan mondat, ahol több áru szerepelt.)

# 7. Eredmények

## 7.1. Baseline mérések eredményei

A Baseline mérések eredményeit a következő táblázatban láthatjuk (2. táblázat):

*2. Táblázat: Baseline mérések eredményei*

| *Szerep* | *Pontosság* | *Fedés* | ***F-mérték*** |
|---|---|---|---|
| Vevő-centrikus célszavak | | | |
| Vevő | 48,24 | 59,73 | **53,37** |
| Eladó | 54,77 | 72,13 | **62,26** |
| Áru | 73,25 | 73,25 | **73,25** |
| Ár | 67,33 | 96,02 | **79,16** |
| Idő | 34,74 | 57,89 | **43,42** |
| Eladó-centrikus célszavak | | | |
| Vevő | 78,18 | 44,10 | **56,39** |
| Eladó | 42,63 | 47,50 | **44,93** |
| Áru | 77,47 | 72,97 | **75,15** |
| Ár | 62,64 | 93,44 | **75,00** |
| Idő | 23,95 | 46,51 | **31,62** |

## 7.2. Eredmények a *vásáro*l kiemelt célszóra

A *vásárol* célszóra az eredményeket a következő táblázatban láthatjuk (3. táblázat):

*3. Táblázat: Eredmények a vásárol kiemelt célszóra (%)*

| *Szerep* | *Pontosság* | *Fedés* | ***F-mérték*** |
|---|---|---|---|
| Vevő | 69,88 | 49,77 | **57,63** |
| Eladó | 82,10 | 60,30 | **68,70** |
| Áru | 80,72 | 77,11 | **78,70** |
| Ár | 90,38 | 83,02 | **85,78** |
| Idő | 78,75 | 52,82 | **61,27** |
| Átlag: | 80,37 | 64,60 | **70,71** |

## 7.3. Eredmények a vevő-centrikus célszavakra

A vevő-centrikus célszavakra az eredményeket a következő táblázatban láthatjuk (4. táblázat):

*4. Táblázat: Eredmények a vevő-centrikus célszavakra (%)*

| *Szerep* | *Pontosság* | *Fedés* | ***F-mérték*** |
|---|---|---|---|
| Vevő | 76,01 | 57,33 | **65,09** |
| Eladó | 79,57 | 66,15 | **71,58** |
| Áru | 79,18 | 80,93 | **79,94** |
| Ár | 87,78 | 80,07 | **82,47** |
| Idő | 83,13 | 63,89 | **71,26** |
| Átlag | 81,13 | 69,67 | **74,07** |

A 3. és a 4. táblázat eredményeit összehasonlítva látható, hogy ha a hasonló viselkedésű célszavakat *egy csoportban kezeltük*, akkor majdnem minden esetben jobb eredményeket értünk el, mint ha a célszavakat önállóan vizsgálnánk. Ennek oka, hogy a több célszó több mondatot és jelöltet ad meg és a több jelölt jellemzőiből általánosabb szabályokat tudott készíteni az osztályozó. A modell legjobban az *Ár* és az *Áru* szerepekre, leggyengébben pedig a *Vevő* szerepre teljesített.

### 7.4. Eredmények az eladó-centrikus célszavakra

Az eladó-centrikus célszavakra az eredményeket a következő táblázatban láthatjuk (5. táblázat):

*5. Táblázat: Eredmények az eladó-centrikus célszavakra (%)*

| *Szerep* | *Pontosság* | *Fedés* | ***F-mérték*** |
|---|---|---|---|
| Vevő | 74,59 | 66,82 | **70,13** |
| Eladó | 68,97 | 48,51 | **56,35** |
| Áru | 85,92 | 82,16 | **83,64** |
| Ár | 83,64 | 63,87 | **71,58** |
| Idő | 76,38 | 51,78 | **59,86** |
| Átlag | 77,90 | 62,63 | **68,31** |

Az eladó-centrikus esetben a modell legjobban az *Ár* és az *Áru* szerepekre, leggyengébben pedig az Eladó szerepre teljesített.

### 7.5. Eredmények a célszavak csoportosítása nélkül

A célszavak csoportosítása nélküli esetre az eredményeket a következő táblázatban láthatjuk (6. táblázat):

*6. Táblázat: Eredmények a célszavak csoportosítása nélkül (%)*

| *Szerep* | *Pontosság* | *Fedés* | ***F-mérték*** |
|---|---|---|---|
| Vevő | 76,93 | 60,11 | **67,05** |
| Eladó | 72,04 | 50,39 | **59,13** |
| Áru | 83,62 | 80,24 | **82,01** |
| Ár | 88,47 | 76,26 | **81,77** |
| Idő | 85,44 | 64,53 | **73,14** |
| **Átlag** | **81,30** | **66,31** | **72,62** |

A *célszavak csoportosításától* azt vártuk volna, hogy a Vevő-centrikus cél-szavaknál a Vevő szerepre, az Eladó-centrikus célszavaknál pedig az Eladó szerepre jobb eredményt kapunk, mint a csoportosítás nélküli esetben. Ez az Eladó szerepre nem teljesült. Ennek egyik oka, hogy az Eladó-centrikus mondatokban az Eladó szerep sokszor távolabb volt a célszótól az elemzőfában. Másik oka, hogy a Vevő-centrikus célszavaknál az Eladó szerepre jó eredményt kaptunk (71,58-es F-

mérték) a *JeloltVegenBigram* és a *JeloltVegenTrigram* jellemzők hatására. Ez a jó eredmény javította erősen a csoportosítás nélküli esetben is az Eladó szerep eredményét. Így a jobb eredményt a csoportosítás nélküli esetre kaptunk (*72,62-es F-mérték*).

### 7.6. Az eredmények összehasonlítás a kapcsolódó munkákkal.

Angol nyelvű szövegekre Gildea és társa [8] sok keretre és azokon belül sok szerepre végezték el a feladatot. Elsősorban igékkel foglalkoznak, de keresnek nem igei célszavakra is. Ezek átlagolt eredményére 63%-os F mértéket kaptak. Eredményeink (*72,62% F-mérték átlag*) jónak számítanak annak ellenére, hogy mi csak egy keretet és ahhoz csak öt fő-szerepet vizsgáltunk, és csak igei és főnévi igenevekhez kerestünk szerepeket.

## Összegzés

Munkánkban bemutattunk gazdag jellemzőtéren alapuló gépi tanuló megközelítésünket, amely automatikusan képes magyar nyelvű szövegekben szemantikus szerepek címkézésére függőségi elemző alkalmazásával. A *vállalati vásárlások, tulajdonváltozások keretével* foglalkoztunk. Ezen a kereten belül 1000 annotált mondatot dolgoztunk fel és a következő szerepeket kerestük: *Vevő, Eladó, Áru, Ár, Idő*. *Jellemzőkészletünkben* felszíni, morfológiai és a függőségi elemzés alapján kinyert jellemzőket használtunk fel. Ezen alapjellemzőket kiegészítettük a jellemzőkből számolt *statisztikai arányokkal* is. Megvizsgáltuk, hogy a statisztikai jellemzők hogyan befolyásolják a modell hatékonyságát. Megvizsgáltuk, hogy a modell hogyan teljesít *egy gyakori célszóra önállóan*, és a *célszavak keretekbe összefoglalt csoportjára* is. A mérésekhez célszavainkat csoportosítottuk több szempont szerint. Bár munkánkban a vizsgált szövegek kevesebb témát fedtek le, mint az angol nyelvű szövegekre bemutatott munkák, de eredményeink jónak számítanak a bemutatott angol munkák eredményeivel összehasonlítva.

## Köszönetnyilvánítás

*Köszönettel tartozunk a kutatás támogatásáért, amely az EFOP-3.6.1-16-2016-00006 „A kutatási potenciál fejlesztése és bővítése a Neumann János Egyetemen” pályázat keretében valósult meg. A projekt a Magyar Állam és az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával, a Széchenyi 2020 program keretében valósul meg.*

## Irodalomjegyzék

[1] D. Ahn: The stages of event extraction. Proceedings of the ARTE '06 Proceedings of the Workshop on Annotating and Reasoning about Time and Events, (2006) 1-8

[2] X. Carreras, L. Màrquez: Introduction to the CoNLL-2004 shared task: semantic role labelling. CONLL '04 Proceedings of the Ninth Conference on Computational Natural Language Learning, (2004) 152-164

[3] X. Carreras, L. Marquez, G.Chrupała: Hierarchical recognition of propositional arguments with perceptrons. In Proceedings of the Eighth Conference on Computational Nat-ural Language Learning (CoNLL-2004), Boston, MA., 106–109

[4] X. Carreras, L. Màrquez: Introduction to the CoNLL-2005 shared task: semantic role labelling. CONLL '05 Proceedings of the Ninth Conference on Computational Natural Language Learning, (2005) 89–97

[5] Ehmann B., Lendvai P., Mihăltz M., Vincze O., László J.: Szemantikus szerepek a narratív kategoriális elemzés (NARRCAT) rendszerében. IX. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, 2013, p. 121-123

[6] Farkas R:, Konczer K., Szarvas Gy.: Szemantikus keret illesztés és az IE-rendszer automatikus kiértékelése. II. Magyar Számítógépes Nyelvészeti Konferencia (2004). Szeged, Szegedi Tudományegyetem, 49-53.

[7] C.L. Fillmore, J.Ruppenhofer, C.F. Baker: Framenet and representing the link between semantic and syntactic relations. In Churen Huang and Winfried Lenders, editors, Frontiers in Linguistics, volume I of Language and Linguistics Monograph Series B.Institute of Linguistics, Academia Sinica, Taipei (2004) 19–59.

[8] D. Gildea, D. Jurafsky: Automatic labelling of semantic roles. Computational Linguistics Journal, Volume 28 Issue 3, September 2002, (2002) 245-288

[9] R. Johansson, P.Nugues: Semantic structure extraction using nonprojective dependency trees. In Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), Prague, Czech Republic, 227–230

[10] P. Koomen, V. Punyakanok, D. Roth, W.Yih: Generalized inference with multiple semantic role labelling systems. In Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005), Ann Arbor, MI., 181–184

[11] K. C. Litkowski: SENSEVAL-3 TASK: Automatic Labelling of Semantic Roles. Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (2004)
[12] L. Márquez, X. Carreras, K.C. Litkowski and S. Stevenson: Semantic Role Labelling: An Introduction to the Special Issue. Computational Linguistics 34(2), (2009) 145-159.
[13] M. Palmer, D. Gildea, P. Kingsbury: The Proposition Bank: An annotated corpus of semantic roles. Computational Linguistics, 31(1):71–105. (2005)
[14] S. Pradhan, K.Hacioglu, W.Ward, J.H. Martin, D.Jurafsky: Semantic role chunking combining complementary syntactic views. In Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005), Ann Arbor, MI., 217–220
[15] Z. Subecz: Detection and Classification of Events in Hungarian Natural Language Texts. Proceedings of the 17th International Conference, TSD 2014, Brno, Czech Republic (2014) Springer Lecture Notes in Computer Science Volume 8655, 2014, pp 68-75
[16] Subecz Z, Nagyné Cs.É..: Igei események detektálása és osztályozása magyar nyelvű szövegekben. X. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, (2014) 237-247
[17] M. Surdeanu, L. Marquez, X. Carreras, P.R.Comas: Combination strategies for semantic role labelling. Journal of Artificial Intelligence Research (JAIR), 29:105–151. (2007)
[18] K. Toutanova, A.Haghighi, C.Manning: Joint learning improves semantic role labelling. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, Ann Arbor, MI. (2005) 589–596
[19] N. Xue, M.Palmer: Calibrating features for semantic role labelling. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Barcelona, Spain. 88-94
[20] Zsibrita J., Vincze V., Farkas R.: magyarlanc 2.0: szintaktikai elemzés és felgyorsított szófaji egyértelműsítés. IX. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, (2013) 368-374