Igék szemantikai szelekciójának és tematikusszerep-kiosztásának korpuszalapú vizsgálata

Tartalom

1. Bevezetés ... 1
2. Elméleti háttér ... 2
3. Problémafelvetés, hipotézisek ... 3
4. Módszertan ... 4
5. Adatok, adatfeldolgozás ... 5
5.1. Vizsgált igék és főnevek kiválasztása ... 5
5.2. Korpusz ... 7
5.3. Számítógépes adatfeldolgozás ... 8
6. Eredmények ... 11
6.1. Szemantikai szelekció ... 11
6.1.1. Két klaszterre osztás ... 12
6.1.2. Három klaszterre osztás ... 15
6.1.3. Négy klaszterre osztás ... 17
6.1.4. Öt klaszterre osztás ... 19
6.2. Tematikusszerep-kiosztás ... 20
6.3. Összevetés ... 22
6.4. Értékelés ... 24
7. Összefoglalás, perspektívák ... 25
Irodalomjegyzék ... 25

1. Bevezetés

A kompozicionalitás elve szerint ahhoz, hogy egy összetett kifejezés jelentését megtudjuk, nincs másra szükségünk, mint az összetevőinek jelentésére és ezen összetevők kapcsolódási módjára (l. például Jackendoff 1997: 48–50). Ahhoz tehát, hogy egy mondat jelentését megismerjük, csak az egyes mondatrészek jelentéseit és összekapcsolódásuk módját kell tudnunk. Honnan tudjuk viszont, hogy milyen szavak azok, amelyek egyáltalán megjelenhetnek mondatrészként? Mi határozza meg az egyes elemek megjelenésének lehetőségét?

Dolgozatomban ezzel a témával foglalkozok, szűkebben az igék mellett alanyi pozícióban előforduló főnevek megjelenését meghatározó tulajdonságokkal, amelyek megszabják, hogy mi lehet és mi nem lehet az egyes igék mellett alany. A kutatás arra irányul, hogy ezen tulajdonságok közül kettőt, a szemantikai szelekciót és a tematikusszerep-kiosztást lehet-e vizsgálni számítógépes nyelvészeti eszközökkel, korpuszalapú módszerekkel.

Ezen felül a korpuszomban lévő adatok alapján nem felügyelt gépi tanulás, k-means klaszterezés segítségével kísérletet teszek arra, hogy a vizsgált elemeket automatikus módon olyan csoportokra bontsam, amelyek értelmezhetők és magyarázó erejűek lehetnek a vizsgált jelenségekkel kapcsolatban. Amennyiben ez lehetségesnek bizonyul, úgy olyan feldolgozási módszereket leszek képes bevetni az igék és főnevek ilyen szempontú jellemzésére, amely gyorsan nagy mennyiségű adatot képes feldolgozni és kevés kutatói beavatkozással is megfelelően működik.

Jelen dolgozat 2. fejezetében bemutatom a téma releváns elméleti hátterét, kitérve a régensek szemantikai szelekciós tulajdonságaira és tematikusszerep-kiosztására. Szintén ebben a fejezetben indoklom a korpuszalapúságot jelen kutatás esetén. A 3. fejezetben felvázolom az alapvető problémafelvetést a kutatás mögött, illetve ismertetem a munka hipotéziseit is. Ezt követően a 4. fejezetben bemutatom a módszereket, amelyeket az elemzés során használtam, amit az 5. fejezetben a konkrét adatgyűjtés és -feldolgozás ismertetése követ. A 6. fejezetben a kutatás eredményeit mutatom be, felbontva a két vizsgált aspektus külön-külön értelmezésére, majd összevetve az eredményeket. Végül a 7. fejezetben összegzem a dolgozatomat és a kutatásom további lehetőségeit vázolom fel.

A kutatás során használt programok, fájlok, ábrák és táblázatok a *https://github.com/otdkanon/otdk.git* linken érhetőek el.

2. Elméleti háttér

Jelen dolgozatban az igék tematikusszerep-kiosztását és a szemantikai szelekcióját vizsgálom, ezért fontos megállapítani, hogy mit értünk ezek alatt a fogalmak alatt, illetve hogy milyen kategóriákat vizsgálok ezeken belül. Először lássuk a tematikus szerepeket. Fillmore (1970) úgy jelenített meg valamiféle, a szintaktikai felszíni szerkezeten túlmutató, a predikátum és argumentumai (jelen esetben az alany) kapcsolatát részletező viszonyt, hogy több mondaton bemutatta, hogy bár az elem alakja (*the boy*) és mondatban betöltött alanyi szerepe egyezik, az általuk betöltött szemantikai szerep nagyban különbözött. Fillmore ezen mondatokra tekintve megállapítja, hogy az „alanyiságnak" (subjectness) nincs olyan közös elméleti tulajdonsága, amely meghatározhatná a mondatban betöltött szerepüket. Komlósy (2015) úgy mutatta be a tematikus szerepeket, mint a vonzatok grammatikai tulajdonságait nagyban befolyásoló, egymáshoz hasonló módon viselkedő fogalmi szerepek. Egy régens minden vonzatának kioszt pontosan egy tematikus szerepet, és minden olyan tematikus szerep, amely az ige mellett megjelenik, kiosztásra kerül pontosan egy argumentumra – ezt nevezzük théta-kritériumnak (θ-criterion, Chomsky 1981: 36). Ebben kifejezésre kerül az is, hogy a tematikus szerepeket csak a vonzatok vehetik fel. Jelen kutatásban a vizsgált régensként álló elemek igék lesznek.

Kenesei segédigékkel foglalkozó tanulmányában (2000) megjegyezte, hogy nincs módunk arra, hogy közvetlenül megtudjuk, hogy egy mondatrész kap-e tematikus szerepet, valamint az összes tematikus szerep elkülönítésére sincs megbízható eljárásunk, de az ágens szerep beazonosítása lehetséges. Ennek megfelelően a dolgozatban is ezzel a tematikus szereppel foglalkozok, így itt érdemes definiálni, hogy mit is értünk alatta. Az ágens szerepet Jackendoff (1990) úgy írta le, mint egy olyan tematikus szerep, amelyet egy főnévi csoport tölt be, amelyet szemantikailag az akarat vagy szándékosság jellemez. Ebből következően azt is állítja, hogy az ágensként álló főnévi csoportnak élő szemantikai jeggyel is rendelkezniük kell. Az ágensként azonosítható elemek mondatrészi szerepéről Jackendoff azt állítja, hogy általában alanyként fordulnak elő, azonban megjegyzi, hogy az alany más tematikus szerepet is felvehet, nem csak ágensit. Ez alapján tehát a dolgozatban is az alany mondatrészeket fogom vizsgálni tematikus szerepük szempontjából. Dowty (1991) egy prototipikus ágens szerepű elem tulajdonságait a következőkben állapította meg: az alany által jelölt entitás akaratlagosan vesz részt az eseményben, az esemény vagy állapotváltozás okozója, képes ítéletet hozni, mozgásra képes, illetve az ige által megnevezett eseménytől függetlenül is létezik.

Az előzőekben azt láthattuk, hogy az ágentivitásnak olyan jegyeket tulajdonítottak a különböző szerzők, amelyek szemantikai jellegűek. Ezzel kapcsolatban tehát felmerül, hogy milyen megszorítást gyakorolnak az egyes igék a mellettük előforduló argumentumokra szemantikai szempontból. Komlósy (2015) a szemantikai szelekciót úgy írta le, hogy a szintaktikai generálást követően történik, amikor „kiválasztásra" kerülnek a jelentésük alapján megfelelő elemekből álló mondatok. Ez a megkötés tehát a valóság tárgyaira vonatkozik, csak azokat lehet jól formáltnak tekinteni, amelyeknél nem áll fenn az, hogy „olyan »abszurd« szituációt fog ábrázolni, amelynek bekövetkezése a valóságban lehetetlen" (Komlósy 2015: 302). Resnik (1996) szelekcióval kapcsolatos tanulmányában a következő módon jellemzi a szelekció által kifejtett hatást: a predikátumokhoz általánosságban mérve az aktuális predikátum különböző mértékben befolyásolhatja a mellette megjelenő elemek gyakoriságát. Ha az adott predikátum szelekciója erős, az átlagos eloszlástól nagyban eltérő lesz a különböző elemek megjelenési gyakorisága, ha azonban szelekció gyenge, az eloszlás az átlaghoz közeli lesz.

Végül pedig arról szólva, hogy miért korpuszalapú a kutatás. Alaphipotézisem az, hogy a fent említett jelenségeket vizsgálhatjuk számítógépes nyelvészeti módszerekkel a korpuszban megtalálható adatok alapján. Ennek segítségével nem kell pusztán az intuíciómra hagyatkoznom, hanem valós, nagy mennyiségű adatra alapozhatok. Emellett az elemzéshez is rendelkezésre állnak változatos eszközök, amelyek segítségével könnyebben és gyorsabban statisztikai adatokhoz juthatunk, amelyeket használhatunk az automatikus osztályozáshoz.

3. Problémafelvetés, hipotézisek

A különböző morfológiai és szintaktikai elemzésekre számos számítógépes nyelvészeti eszköz áll rendelkezésünkre az utóbbi időben, ezek jelentősen megkönnyítik az elemzési feladatokat. Ennél azonban nehezebben megoldható azoknak a tulajdonságoknak a számítógépes, automatikus vizsgálata és elemzése, amelyek valamilyen módon a jelentéssel kapcsolatosak. Jelen dolgozatban ezzel foglalkozok, alapvető problémafelvetésem az, hogy lehet-e valamilyen számítógépes nyelvészeti eszköz, módszer segítségével vizsgálni az igék szemantikai szelekcióját és tematikusszerep-kiosztását.

Jelen kutatásban két hipotézist állítottam fel az igék szemantikai szelekciójának és tematikusszerep-kiosztásának számítógépes nyelvészeti vizsgálatával kapcsolatban

1. Az igék alannyal szemben támasztott szemantikai szelekciója összefüggést mutat a korpuszban lévő mondatok alapján számított kvantitatív adatokkal.

2. A vizsgált igék alanyválasztás szerinti klaszterezése segítségével megismerhetjük
    a. az igék jellemző szemantikai szelekciós tulajdonságait,
    b. az igék jellemző tematikusszerep-kiosztását, illetve
    c. ezen két jellemző között felfedezhetünk valamiféle összefüggést.

4. Módszertan

A kutatást korpuszadatokra alapoztam. A korpusz építéséhez olyan mondatokra volt szükségem, amelyek gyakran előforduló igéket tartalmaznak. Ehhez 100 igét választottam ki. Emellett ahhoz, hogy a vizsgálatot megfelelően behatároljam, az ezek mellett előforduló alanyokra is szűkítettem a kutatást. Ez a szűkítés olyan módon valósult meg, hogy a lehetséges alanyok közül csak főneveket vizsgáltam. A kutatás során 54 gyakori főnevet vizsgáltam. A főnevek kiválasztásában a gyakoriságuk mellett az játszott szerepet, hogy milyen szemantikai kategóriába tartoznak.

A korpusz összeállításában arra kellett figyelmet fordítanom, hogy maga a keresés ne tolja el a vizsgált főnevek gyakorisági arányát, mivel az nem tette volna megfelelővé a korpuszt a szemantikai szelekció vizsgálatára. Ha a lekérés során meghatározott számú mondatot kérnék le minden ige-főnév kombinációból, nem kapnék olyan adatokat, amelyekből kimutatható lenne a különböző alanyok megjelenésének valószínűsége. Ennek a problémának a megoldására több lehetőség is felmerülhet, jelen kutatás esetében amellett döntöttem, hogy minden főnevet egyszerre keresek az igék mellett, így látszódik azoknak az egymáshoz viszonyított megjelenési aránya.

A kutatásban k-means klaszteranalízist végeztem. Ez egy egyszerű nem felügyelt gépi tanulási módszer, mely során a klaszterező automatikusan előre megadott k csoportra osztja az N dimenziós adathalmaz elemeit. Jelen esetben a dimenziók számát a szemantikai szelekció vizsgálatánál a szemantikai jegyek száma (5), a tematikus szerepek vizsgálatánál pedig az adja, hogy az egyes mondatokban az alany ágens-e, vagy nem (2). A klaszterezés során minden pont valamelyik csoportba kerül, nincsenek olyan elemek, amelyek egyik klaszterbe sem tartoznak. A létrehozott klaszterek középpontja mindig a klaszterben lévő pontok koordinátáinak átlagán elhelyezkedő pont, amely dinamikusan változik az optimális állapot eléréséig (MacQueen 1967).

A tematikus szerepekkel kapcsolatos kutatási kérdés megválaszolásához létrehoztam egy részkorpuszt a másik, a szemantikai szelekció meghatározására létrehozott korpuszból. Ebben 10000 mondat szerepelt, minden ige esetében 100-100 mondat. Ezeket kézzel annotáltam

olyan szempontból, hogy az adott ige mellett a kiválasztott főnevek ágensek-e az adott mondatban vagy sem. Ebben az elemzésben a tematikus szerep megállapítását intuíciómra alapoztam, majd a szemantikai szelekcióval kapcsolatos vizsgálathoz hasonlóan itt is klaszterezéssel alakítottam ki csoportokat.

5. Adatok, adatfeldolgozás

5.1. Vizsgált igék és főnevek kiválasztása

Ahhoz, hogy megfelelő mennyiségű, értelmezhető eredményeket szolgáltató adatmennyiséggel rendelkezzem, előzetes listákat készítettem a Szeged Dependency Treebank (Vincze et al. 2010) adatai alapján a vizsgálatra alkalmas igékről és főnevekről. A keresési kritériumot úgy állapítottam meg, hogy a talált szavak legyenek elég gyakoriak, így 30 előfordulás volt a feltétel mind az igék, mind a főnevek esetében. Ezen felül a főnevek előfordulásait alanyesetben kerestem, mivel ágens tematikus szerepet ebben az esetben kaphatnak a régensük mellett.

A szűrés után megkaptam azon igéknek és főneveknek a listáját, amelyek a meghatározott mennyiség felett fordultak elő az Szeged Dependency Treebankben. Ebből a listából választottam ki először a 100 igét, amelyek a következőkben olvashatók:

1. A vizsgált igekötőtlen igék: *ad, akad, alakít, állít, beszél, biztosít, csatlakozik, dolgozik, dönt, ébred, emel, énekel, épít, épül, érdekel, érint, érkezik, erősít, ért, fél, figyelmeztet, fogad, fordul, halad, hall, határoz, hisz, hív, hoz, indít, indokol, irányít, iszik, jelent, jut, kér, kezdeményez, kötelez, lakik, lát, létezik, leül, mutat, nevez, néz, nő, örül, pihen, remél, siet, szán, szerez, szól, találkozik, telik, törődik, utazik, ül, választ, változik, zavar*
2. A vizsgált igekötős igék: *átad, átmegy, bevezet, elalszik, elér, elfárad, elfelejt, elindul, elkészít, elkezdődik, elkövet, elmarad, elmúlik, eltűnik, elvesz, felkel, feltűnik, kiegészít, kikerül, kimond, kinyit, kivált, leszáll, megáll, megfelel, meghal, meghalad, megjegyez, megmarad, megnyílik, megöl, megszólal, megtalál, megtanul, megtud, megváltozik, visszaesik, visszajön, vonatkozik*

Az igekötők képesek megváltoztatni az ige argumentumszerkezetét, bár az alanyra gyakorolt hatásuk nem nagy (Gyulai 2021). Azonban az igék azon jellemzője, hogy igekötővel vagy igekötő nélkül állnak, nem játszik meghatározó szerepet a kutatásban elvégzett kategorizálások szempontjából, mindössze a keresés szempontjából számottevő különbség, melyet később, a korpusz készítéséhez tartozó résznél részletesebben kihangsúlyozok. Az

igék kiválasztásánál törekedtem arra, hogy a vizsgált igék jelentésükben különbözzenek egymástól, illetve előreláthatóan legyenek közöttük különböző tematikus szerepű alannyal állóak. Ebben intuíciómra hagyatkoztam, de ez az előzetes megítélés sem játszik szerepet a későbbi kategóriák kialakításában, részben azért, mert nem törekedtem ezeket kimondott csoportokba osztani, részben pedig azért, mert ugyan az én anyanyelvi megítélésem jelen lehetett a kiválasztásban, később az automatikus osztályozást nem befolyásoltam.

A kutatás előzetes tervezése során felmerült, hogy vizsgálható lenne az alanyok ágenssége tulajdonnevek és névmások esetében is, nem csupán köznévi elemeknél. Azonban annak a megállapítása, hogy mely tulajdonnevek gyakoriak, illetve annak az automatikus azonosítása, hogy ezek milyen szemantikai kategóriába tartoznak, nem tűnt megvalósíthatónak jelen kutatás keretei között. Hasonló módon túlmutatott volna ezen kutatáson a névmások szemantikai szerepének automatikus beazonosítása, így csak a köznevekre esett a választás. A dolgozatban ezért csak a közneveket vizsgáltam.

A főnevek listájából történő kiválasztás során először azt figyeltem meg, hogy a keresés eredményeként kapott nyelvi elemeket 5 szemantikai kategóriába lehet osztani: humán (H), intézmény/csoport (I), nem élő (N), absztrakt (A), illetve élő, nem humán (L). Ezek a kategóriák, bár nem annyira részletes osztályozást jelenítenek meg, nagy hasonlóságot mutatnak a WordNet (2010) kategóriáival. A kiválasztott főneveket a következőkben közlöm:

1. Humán*: anya, apa, barát, elnök, ember, férfi, gyerek, lány, nő, személy, tag, tanár*
2. Intézmény/csoport: *bank, cég, család, csapat, csoport, iskola, kormány, könyvtár, ország, osztály, társaság, város*
3. Nem élő: *ajtó, autó, busz, épület, eszköz, fa, gép, ház, könyv, lemez, pénz, számítógép*
4. Absztrakt: *adat, döntés, élet, eredmény, gond, hír, idő, jog, lehetőség, ok, szó, terv*
5. Élő, nem humán: *állat, hal, kutya, ló, macska, madár*

Az utolsó kategóriába tartozó főnevekből a szűrt listán jelentősen kevesebb volt, mint a többi csoportban lévőkből. Összesen ez a 6 főnév volt a feltételeknek megfelelő, így ezekből kevesebbet tudtam vizsgálni. Éppen ennek a kategóriának elemszáma adta meg a többi kategória elemszámát is: a H, I, N és A csoportokba 12-12 főnév tartozik, annak érdekében, hogy arányos legyen a csoportokba tartozó főnevek eloszlása, de ez később nem vált számottevővé, erről továbbiakban még szót ejtek a korpusz adatainak részletes bemutatásánál.

5.2. Korpusz

A kutatás során használt korpusz mondatait a Magyar Nemzeti Szövegtár (Oravecz–Váradi–Sass 2014) (továbbiakban MNSZ2) adatbázisából nyertem. Az előzetesen kiválasztott igéket tartalmazó mondatokra volt szükségem, illetve ezt tovább szűkítve olyan mondatokra, amelyben a választott főnevek alanyként szerepelnek. Annak érdekében, hogy megfelelő pontossággal kapjam meg a mondatokat, a lekéréseket CQL (Jakubíček et al. 2010; Sass 2017) segítségével végeztem. Mint korábban említettem, itt van jelentősége annak, hogy a vizsgált ige igekötős-e vagy sem, mivel a CQL-kifejezés különbözik a két esetben. Lentebb láthatóak a használt kifejezések:

1. igekötőtlen igékkel:
   (meet [lemma="FŐNEVEK" &
   msd="(FF\.)?FN(\._FOK)?(\.(PL|FAM))?(\.PS)?[et]?[123]?i?(\.POS|\.POSi)?\.NOM"]
   *[lemma="IGE" & msd="(IK\.)*IGE\.(_HAT\.) ?[TI]?[MPF]?[et]?[123]?"] -5 5)*
2. igekötős igékkel:
   (meet [lemma="FŐNEVEK" &
   msd="(FF\.)?FN(\._FOK)?(\.(PL|FAM))?(\.PS)?[et]?[123]?i?(\.POS|\.POSi)?\.NOM"]
   *(union (meet [lemma="IGE"&msd="(IK\.)*IGE\. (_HAT\.)?[TI]?[MPF]?[et]?[123]?"]*
   [lemma="IGEKÖTŐ" & msd="IK(\.IK)"]-3 3)*
   *[lemma="IGEKÖTŐ,IGE"&msd="(IK\.)*IGE\.(_HAT\.)?[TI]?[MPF]?[et]?[123]?"]) -5 5)*

A kifejezések olyan mondatokat kerestek, amelyekben a keresett igétől 5 szónyi távolságon belül szerepeltek a keresett főnevek alanyesetben. A főnevekre vonatkozó feltételek csak arra szűkítettek, hogy a megadott lemmával szereplő elemek alanyesetűek legyenek. Az igéknél a kétfajta szűrés különböző feltételeket tett szükségessé. Mindkettőnél megszabtam, hogy ne legyenek -hat/-het képzősek, valamint az igekötőtlen igéknél azt is, hogy ne legyenek igekötősek. Az igekötős igék keresésénél kétféle feltételt szabtam, egyet az elváló igekötőkre, egyet pedig a tapadókra. Az elválók esetében az ige 3 szavas környezetében kellett megjelennie a keresett igekötőnek, míg a tapadóknál az igekötő és az ige együtt szerepeltek lemmaként. A kifejezésekben a konkrét szavak helyén a FŐNEVEK, IGE, illetve IGEKÖTŐ szavak szerepelnek. A főneveket egyszerre kérdeztem le, hogy a megadott mennyiségű mondatban megmaradjon ezeknek az aránya, és megfelelően megállapítható legyen a szemantikai szelekció.

Az egyes igék esetében véletlenszerű mintát kértem le, minden igéhez 4000 mondatot. Ekkora mennyiségű adat esetében akkor is marad megfelelően sok vizsgálható mondat, ha a

lekérdezett adatok közül a feldolgozási lánc egyes szakaszaiban kiesik valamekkora mennyiség. Az MNSZ2-ből történő lekérést követően az eredményeket program segítségével szűrésnek vetettem alá, amely egyrészt a duplikátumokat szűri ki, másrészt pedig az olyan adatsorokat, amelyek nem mondatok, azaz nem nagy kezdőbetűvel kezdődnek és mondatvégi írásjelre végződnek.

5.3. Számítógépes adatfeldolgozás

Az MNSZ2-ből szerzett, tisztított mondatok elemzését a magyarlanc (Zsibrita–Vincze–Farkas 2013) függőségi elemzőjével elemeztettem. Az elemzést követően az eredményként kapott TSV fájlt program segítségével XML fájllá alakítottam, hogy a következőben használt programok számára értelmezhető legyen az adat.

Ezt követően az esetleges elemzési pontatlanságok kijavítására olyan programokon mennek keresztül az adatok, amelyek beazonosítják, hogy az elemző által igének címkézett elemek szerepelnek-e egy, a Szeged Dependency Treebank adataiból készült szólistán. Az első program kigyűjti az ilyen igéket, és ellenőrzi, hogy az elemek felbonthatók-e a listában szereplő igekötőtlen igére és egy igekötőre (például a *kihasznál* ige *ki+használ* formában szerepel ilyen esetben). Az így felismert igéket hozzáadja az ismert igék listájához, valamint létrehoz egy új XML fájlt, amely csak azokat a mondatokat tartalmazza, amelyekben nincs ismeretlen ige. Ezután az így létrejött XML fájlt egy következő program feldolgozza, ennek során az elváló igekötőket párosítja az igéjükhöz. Ha a magyarlanc által javasolt igekötőhöz tartozó ige egy olyan beférkőző ige, amelynek főnévi igeneves bővítménye van, akkor a főnévi igenévhez köti az igekötőt, ha pedig olyan létige, amelynek határozói igeneves bővítménye van, akkor a határozói igenévhez.

Az ilyen módon tisztított és javított adatokat egy másik programon keresztül futtattam át, amely kigyűjti a tagmondatokat, ezen tagmondatok igéjét és az ige melletti maximális összetevőket. Ennek a kimenete egy olyan XML fájl, amely <cp> tagek között felsorolja az egyes tagmondatok reprezentációját, jelöli ezeknek kezdetét és végét, illetve a tagmondatok fejét. A fejen attribútumként jelöli az igekötőt, az elváló igekötő azonosítóját, az ige képzetlen és igekötőtlen tövét, valamint az ige típusát (kijelentő, felszólító stb.). Ezek után a tagmondatban szereplő maximális összefüggő kifejezéseket sorolja föl <xp> tagek között, jelölve annak elejét, végét, és a fejének a szótövét, legfontosabb jellemzőjét (f), és ha irányhármasságként értelmezhető, akkor azt a jellemzőjét (loc). A feldolgozási láncról lásd még (Szécsényi–Virág 2022). A következő ábrán látható, hogy milyen formában jelenik meg egy mondat a feldolgozást követően.

```
<cp>
    <start>alakít.mnszs9w1</start>
    <head pv="0" pvid="0" lemma="alakít" mood="ind">alakít.mnszs9w14</head>
    <end>alakít.mnszs9w16</end>
    <xp>
        <start>alakít.mnszs9w1</start>
        <head lemma="tag" f="nom">alakít.mnszs9w9</head>
        <end>alakít.mnszs9w9</end>
    </xp>
    <xp>
        <start>alakít.mnszs9w10</start>
        <head lemma="szupersejt" f="acc">alakít.mnszs9w12</head>
        <end>alakít.mnszs9w13</end>
    </xp>
    <xp>
        <start>alakít.mnszs9w15</start>
        <head lemma="Észak-afrika" f="TÓL" loc="FROM">alakít.mnszs9w15</head>
        <end>alakít.mnszs9w15</end>
    </xp>
    <xp>
        <start>alakít.mnszs9w16</start>
        <head lemma="Délkelet-ázsia" f="IG" loc="TO">alakít.mnszs9w16</head>
        <end>alakít.mnszs9w16</end>
    </xp>
</cp>
```

1. ábra: Egy tagmondat feldolgozás után

Eredeti tagmondat: Az Afganisztán elleni amerikai offenzíva nyomán elmenekült al-Kaida tagok új "szupersejteket" alakítanak Észak-Afrikától Délkelet-Ázsiáig

A tagmondatok feldolgozását követően program segítségével szűrést végeztem arra vonatkozóan, hogy csak azok a tagmondatok szerepeljenek a korpuszban, amelyekben a vizsgált igék mellett a vizsgált főnevek állnak alanyként. Ennek során csak azok a tagmondatok kerültek be a végleges fájlba, amelyekben a feldolgozott ige és a listában szereplő vizsgált főnevek együtt fordultak elő.

A feldolgozás előtti és azt követő mondatmennyiséget ábrázolva az adatsor legkisebb, legnagyobb és átlag értékét, valamint az összes mondat számát a következő táblázatban láthatjuk:

	feldolgozás előtt (mondatszám)	feldolgozás után (tagmondatszám)
min	419	141
mean	2384,88	724,14
max	3224	1939
összesen	238488	72414

1. táblázat: Az adatok mennyiségének összehasonlítása a feldolgozási sor előtt és után
min: a legkevesebb mondattal álló ige mondatszáma; mean: az összes ige mondatszámának átlaga;
max: a legtöbb mondattal álló ige mondatszáma

Az utolsó feldolgozó lépésben a feldolgozott mondatokat a vizsgált főnevek szemantikai kategóriája alapján automatikusan one-hot kódolással jelöli, hogy összesíthető legyen az egyes igék melletti alanyok előfordulási aránya. Ennek kimenete egy XLSX fájl, amelyből néhány sor látható a következő ábrán.

id	vlemma	np	H	I	N	A	L
siet.mnszs1601	siet	ember	1	0	0	0	0
siet.mnszs1603	siet	állat	0	0	0	0	1
siet.mnszs1605	siet	ember	1	0	0	0	0
siet.mnszs1607	siet	kormány	0	1	0	0	0
...							

2. ábra: A one-hot kódolás kimenete
Magyarázat: id – a mondat azonosítója; vlemma – a mondatban lévő ige lemmája;
np – a mondatban lévő, alanyesetű főnév; H, I, N, A, L – a lehetséges szemantikai kategóriák

A szemantikai szelekciót, az igei argumentumszerkezethez hasonlóan (Szécsényi 2019), az egyes igéknél a kiszámított előfordulási arányok alapján meghatározott vektorként értelmezhetjük, amely az adott ige helyét mutatja meg abban az 5 dimenziós térben, amelyet a vizsgált szemantikai kategóriák határoznak meg. Néhány vizsgált ige vektorait a következő táblázat szemlélteti:

	H	**I**	**N**	**A**	**L**
siet	0,69882	0,23529	0,01647	0,03294	0,01647
pihen	0,38559	0,28390	0,16949	0,06780	0,09322
mutat	0,08965	0,08965	0,04803	0,76307	0,00961
lát	0,59136	0,17608	0,04319	0,17442	0,01495

2. táblázat: Pár vizsgált ige mellette előforduló szemantikai jegyek szerint meghatározott vektora

Az adatfeldolgozási folyamatokat követően a korpuszban 72414 releváns tagmondat maradt, amelyekben az alanyok szemantikai kategóriájára való tekintettel az alábbi táblázatban látható eloszlásban voltak jelen:

	H	**I**	**N**	**A**	**L**	**összesen**
összeg	29608	19113	6090	16078	1525	72414
arány	40,89%	26,39%	8,41%	22,20%	2,11%	100%

3. táblázat: A korpusz adatfeldolgozást követő összegzése

Ahhoz, hogy az alanyok tematikus szerepét is vizsgálni tudjam, kézi annotálást alkalmaztam. Ennek során minden ige esetében 100 mondatot elemeztem, mindegyik mondatnál jelöltem, hogy saját intuícióm szerint a vizsgált főnév az adott mondatban ágens alany-e vagy sem. Az annotálás során az 1. ábrán látható táblázat kiegészített verzióján dolgoztam, amelyben a fent látható oszlopokon kívül még látható volt az eredeti mondat is, hogy eldönthető legyen az alany tematikus szerepe. A tematikus szerep beazonosítása során azt tartottam szem előtt, hogy az alany cselekszik-e, illetve ez a cselekvése szándékos-e.

A fent bemutatott elemzési sort követte a k-means klaszterezés. Ehhez a Scikit-learn Python modult (Pedregosa et al. 2011) használtam. Ennek a programnak a bemenete egy olyan táblázat, amely tartalmazza az egyes igéket és a különböző szemantikai kategóriájú főnevek előfordulási arányát mellettük. Az adatokat ezt követően normalizáltam a modulba beépített normalizálóval, majd a normalizált adatokon végeztem el a klaszterezést. A szemantikai jegyeket követően hasonló módon klaszterezést végeztem a tematikus szerepeket jellemző adatsoron is, ennek a bemenete egy olyan táblázat volt, amelyben az ágensesség kézi annotálásából származó adatok szerepeltek.

6. Eredmények

Az adatfeldolgozás és a teljes elemzés utáni klaszteranalízis eredményeit a jelen fejezetben ismertetem. Minden klaszterezés esetében bemutatom az adatsort, amelyet a program létrehozott, azon ötdimenziós tér néhány kétdimenziós vetületét, amelyben az egyes igék elhelyezkednek, majd ezeket az adatokat értelmezem.

6.1. Szemantikai szelekció

A szemantikai szelekció elemzése során az adatokat többszöri, különböző számú k-means klaszteranalízisnek vetettem alá. A vizsgálat során kettő, három, négy, illetve öt klaszterre oszttattam az adatokat, minden esetben azt tesztelve, hogy milyen értelmezhető csoportokat hoz létre a számítógép nem felügyelt tanítási folyamat segítségével. A k-means analízis

sajátossága miatt, miszerint minden futtatás esetén különböző eredmények születhetnek, a programot többször lefuttattam, hogy megbizonyosodjak arról, hogy az adatok nemcsak az egy adott lefutásra igazak, hanem ilyen és ehhez hasonló arányokban jelennek meg több próbálkozás esetén is. Miután erről meggyőződtem, az eredményként létrejött adatsorokból diagramokat készítettem a releváns tengelyekre vetítve.

A konkrét klaszterezésről szólva, a 4. pontban bemutattam a módszert elméleti síkon, most ezt pontosítva, jelen kutatásra vonatkoztatva a következőket mondhatjuk el: az *N* dimenziós adathalmaz jelen esetben 5 dimenziót jelent, ezeket a dimenziókat az egyes szemantikai kategóriák (H, I, N, A, L) előfordulási gyakorisága adja. A különböző klaszterezések *k* értékei pedig sorban 2, 3, 4, illetve 5 voltak. A továbbiakban az egyes klaszterezések részletes adatait közlöm.

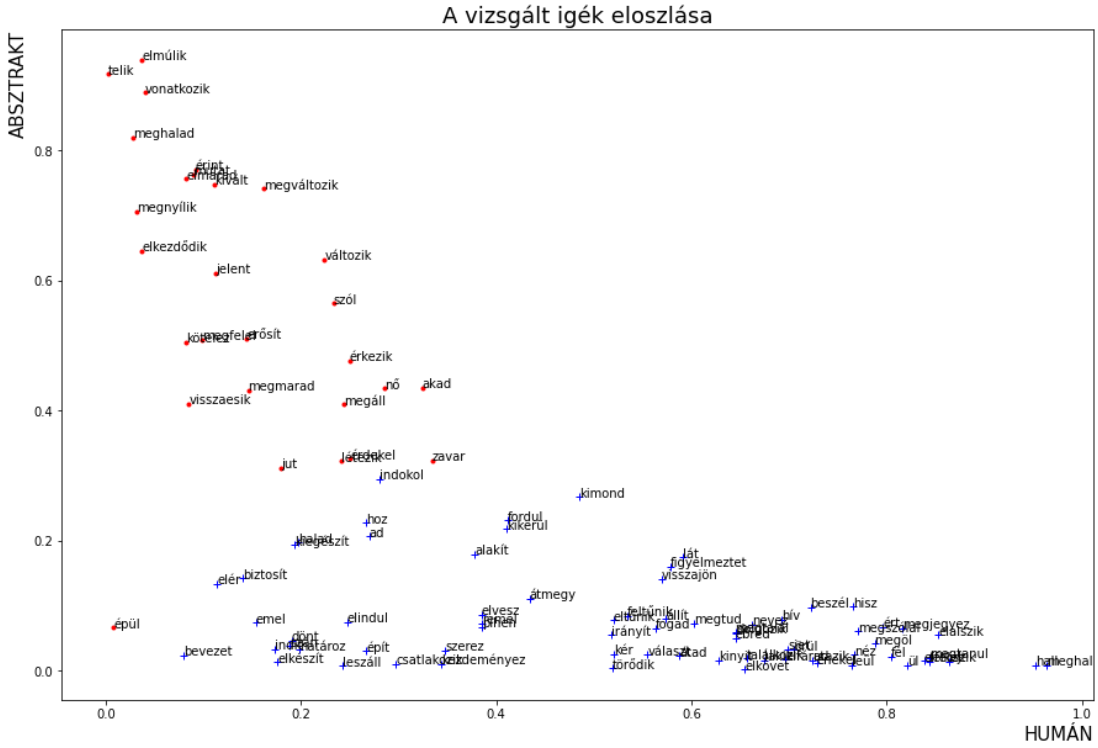
6.1.1. Két klaszterre osztás

Az elemzések során a klasztereknél a középpontjukat adó pont koordinátáit láthatjuk megadva, ezzel egy csoportátlagot megadva a létrejött adathalmazoknak. Minden esetben kiemelve jelenítem meg a táblázat egyes soraiban a sor többi adatához mérten magas számadatokat. A program a következő táblázatban található adatsort adta eredményül:

klaszter	H	I	N	A	L
0	**0,532187**	**0,311938**	0,057474	0,071611	0,026790
1	0,141548	0,142798	0,131952	**0,570873**	0,012829

4. táblázat: A 2 klaszterre osztás eredménye – a klaszterek középpontjai

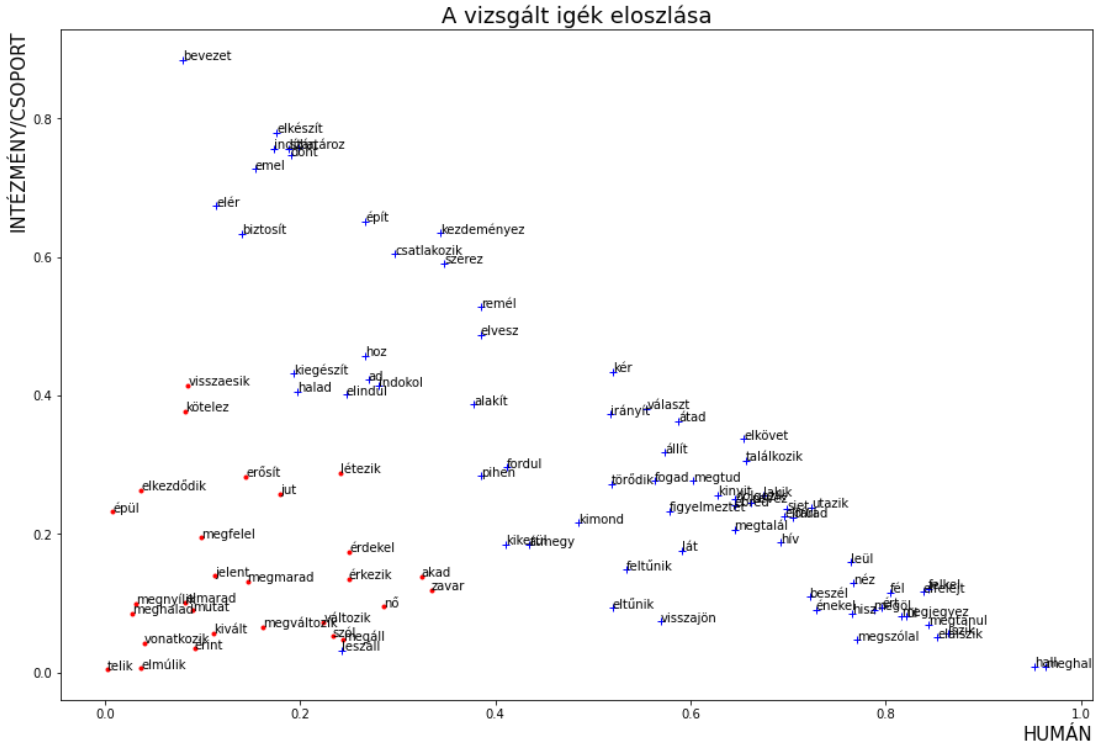
Azt láthatjuk, hogy a 0-val jelölt sorban a többi szemantikai kategóriákhoz mérten a H és az I magas számadatokat mutatnak. Az 1-es sorban pedig az A száma kiemelkedő. Ennek tudatában az ábrázolás során a releváns tengelyeket ezek a kategóriák adják. A következőkben három ábrát mutatok be, minden esetben ezen tengelyek közül valamely kettőre levetítve az adathalmazt.



3. ábra: A 2 klaszterre bontás csoportjainak 2 tengelyre (humán-absztrakt) vetítése

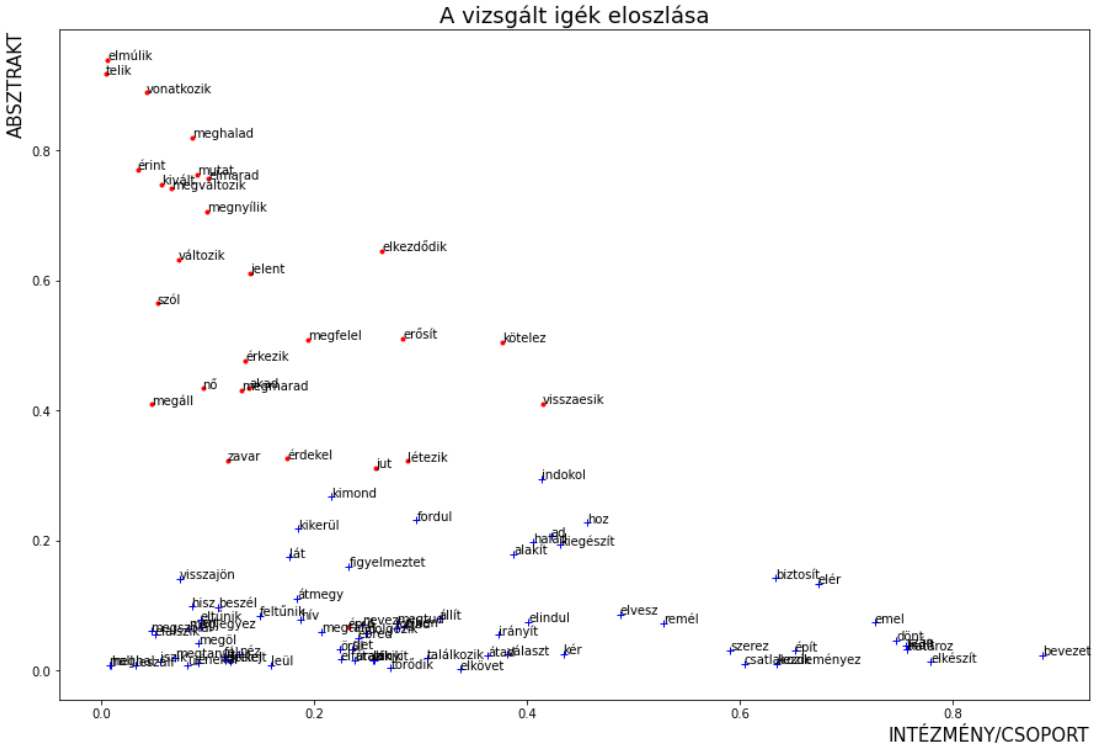
Jelmagyarázat: kék kereszt – 0-val jelölt csoport; piros kör – 1-gyel jelölt csoport

Az ábrát magyarázva: mind az X, mind az Y tengelyen a nagyobb értékek felé haladva olyan igék helyezkednek el, amelyek azzal a szemantikai kategóriájú alannyal állnak gyakrabban, például a 3. ábrán a *meghal* ige a HUMÁN tengely legmagasabb értékéhez nagyon közel helyezkedik el, azaz az alanyai majdnem kizárólag H jegyűek voltak. Az ABSZTRAKT tengelyen ehhez hasonló helyzetben van az *elmúlik* ige. Azok az igék, amelyek mindkét tengelyen alacsony értéket vesznek fel, mindkét ábrázolt jeggyel kevés alkalommal állnak. A 3. ábrán ilyen például a *bevezet* vagy az *épül*.



4. ábra: A 2 klaszterre bontás csoportjainak 2 tengelyre (humán-intézmény/csoport) vetítése

Jelmagyarázat: kék kereszt – 0-val jelölt csoport; piros kör – 1-gyel jelölt csoport



5. ábra: A 2 klaszterre bontás csoportjainak 2 tengelyre (intézmény/csoport-absztrakt) vetítése

Jelmagyarázat: kék kereszt – 0-val jelölt csoport; piros kör – 1-gyel jelölt csoport

A fenti táblázat, illetve az ábrák alapján a két klaszterbe történő osztályozással kapcsolatban a következőket állíthatjuk. A klaszterező képes volt értelmezhető csoportok létrehozására, amelyeket valamely alanytípus(ok) magas előfordulási aránya fémjelez. A létrejött csoportokat az absztrakt kategória sokasága, illetve a humán vagy intézmény/csoport alanyok magas aránya jellemzi. Az így létrejött halmazokban elválik két csoport, az egyikben (0) többségében szándékos cselekvésre nem képes alanyokkal álló igéket találunk, a másikban (1) pedig szándékos cselekvésre képes alanyokkal állóakat. Láthatjuk, hogy ez a különbség hasonlít arra a szembeállításra, amely alapján el tudjuk különíteni az ágensként és nem ágensként álló alanyokat. Azonban nem állíthatjuk, hogy már ez az egy analízis szilárd alapot adhat egy tematikus szerepekre vonatkozó csoportosításnak, hiszen ez az elkülönítés pusztán a cselekvésre való képességet mutatja meg. Nem érdemes azonban ezt a csoportosítást figyelmen kívül hagyni a továbbiakban.

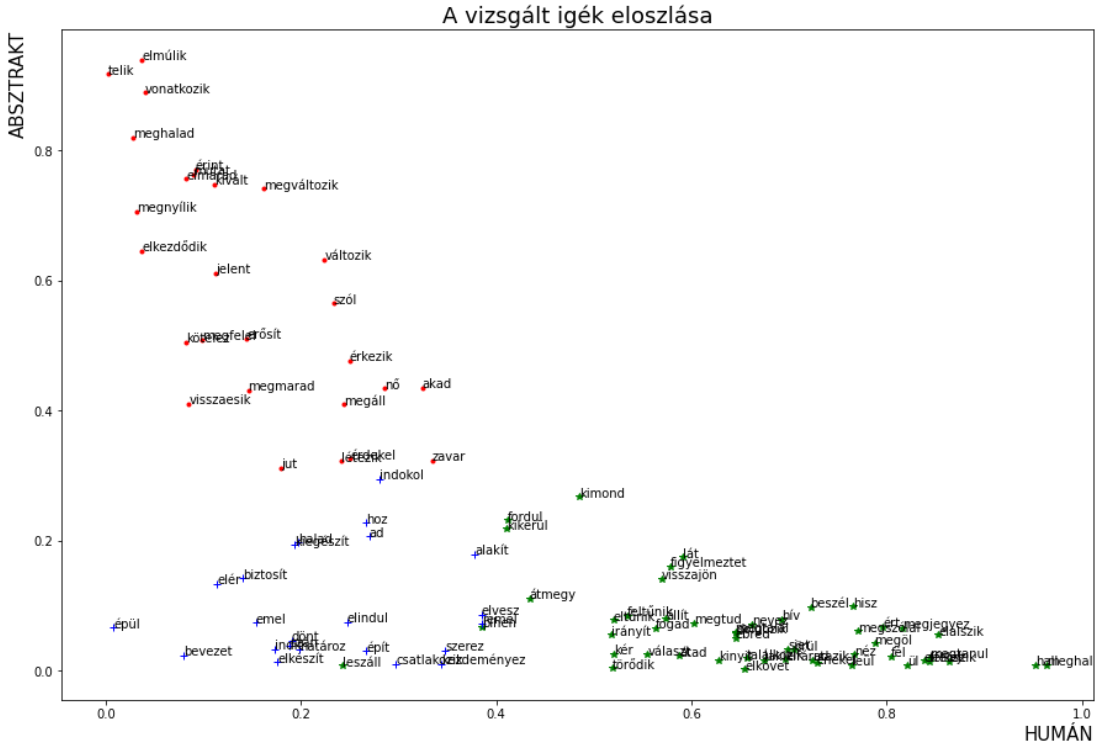
6.1.2. Három klaszterre osztás

A három klaszterre bontás esetén arra számíthatunk, hogy az előzőleg megállapított csoportok finomítása mehet majd végbe, valamilyen kategória dominanciájának kihangsúlyozásával egy harmadik kategória jelenhet meg. A klaszterek súlypontjainak adatai a következő táblázatban láthatóak:

klaszter	H	I	N	A	L
0	**0,229788**	**0,580997**	0,087462	0,096439	0,005313
1	**0,660809**	**0,186573**	0,056342	0,060088	0,036187
2	0,146489	0,139493	0,111260	**0,589554**	0,013204

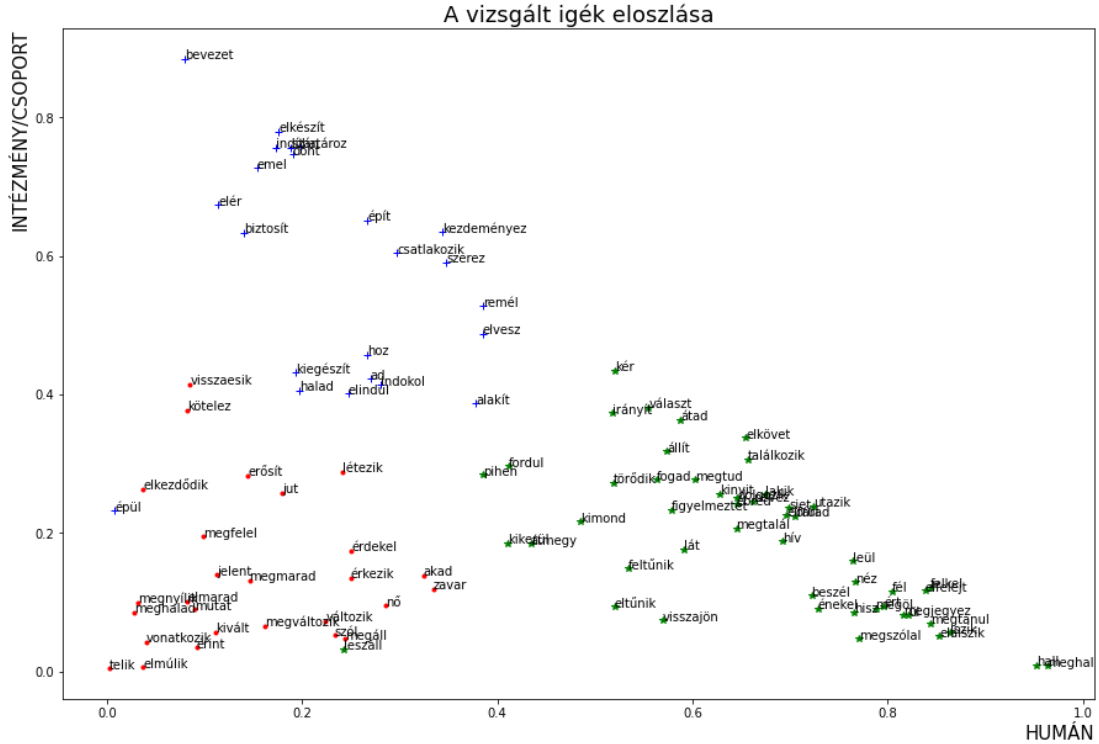
5. táblázat: A 3 klaszterre osztás eredménye – a klaszterek középpontjai

A három klaszter adatsorai közül az egyik már ismertnek tűnhet az előző osztályozásból: a 2-vel jelzett sorban ismét az absztrakt (A) kategóriájú alanyok többsége a meghatározó. A további sorokban a humán (H), illetve az intézmény/csoport (I) többségét láthatjuk, illetve mindkét esetben a másik, többi kategóriához viszonyított hozzávetőleges gyakorisága a jellemző. Az előző osztályozáshoz hasonlóan a következőkben itt is láthatunk három ábrát, amelyek megmutatják a releváns tengelyek mentén a csoportokat.



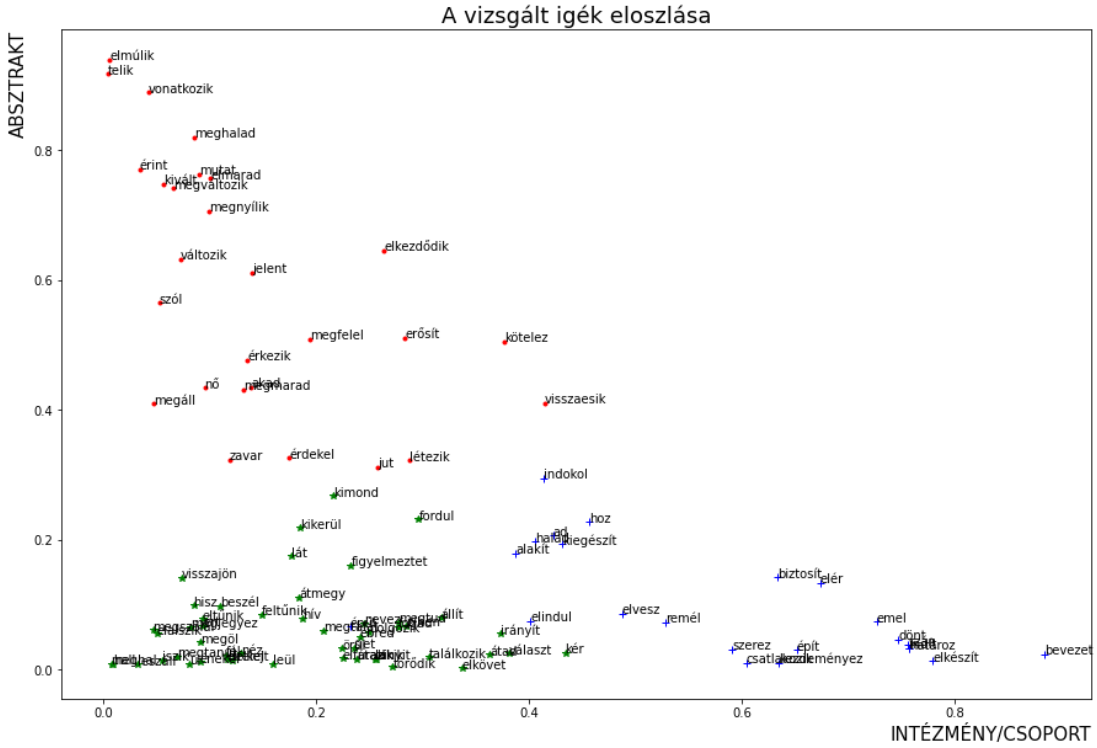
6. ábra: A 3 klaszterre bontás csoportjainak 2 tengelyre (humán-absztrakt) vetítése

Jelmagyarázat: kék kereszt – 0-val jelölt csoport; zöld csillag – 1-gyel jelölt csoport; piros kör – 2-vel jelölt csoport



7.ábra: A 3 klaszterre bontás csoportjainak 2 tengelyre (humán-intézmény/csoport) vetítése

Jelmagyarázat: kék kereszt – 0-val jelölt csoport; zöld csillag – 1-gyel jelölt csoport; piros kör – 2-vel jelölt csoport



8. ábra: A 3 klaszterre bontás csoportjainak 2 tengelyre (intézmény/csoport-absztrakt) vetítése Jelmagyarázat: kék kereszt – 0-val jelölt csoport; zöld csillag – 1-gyel jelölt csoport; piros kör – 2-vel jelölt csoport

A fenti adatokat vizsgálva azt láthatjuk, hogy az absztrakt alanyok által fémjelzett kategória továbbra is megmaradt, és ahogy az előbbiekben már megállapíthattuk, az egyik csoportban továbbra is jellemző a szándékos cselekvésre nem képes alanyok többsége. A továbbiakban létrejött csoportok esetében láthatjuk, hogy a nagy elemszámú alanyfajták alapján jönnek létre csoportok: az 1-es csoportban a humán, a 0-s csoportban pedig az intézmény/csoport alanyok a meghatározóak. Fontos még megjegyezni, hogy bár ezekben a klaszterekben van egy-egy egyértelmű domináns kategória, mindkettőben megfigyelhető, hogy a nem élő (N), absztrakt (A) és élő, nem humán (L) alanyok nagyon alacsony előfordulási arányával szemben a fennmaradó kategória viszonylag magasabb arányban jelenik meg. Ennek tudatában feltételezhetünk kiegészítő valamilyen szintű kiegészítő disztribúciót a két szándékos cselekvésre képes kategória elemei között.

6.1.3. Négy klaszterre osztás

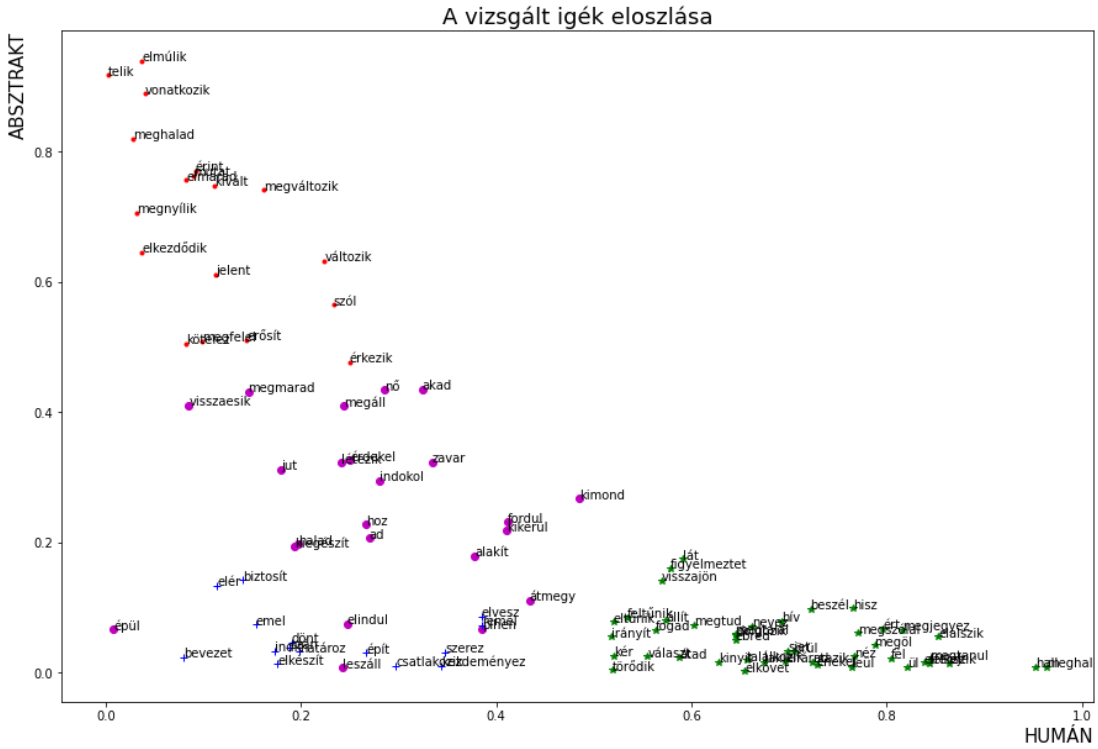
A korpusz felépítéséből viszonylag könnyen értelmezhető a három klaszterre bontás eredménye, tekintve, hogy olyan klaszterek jöttek létre, amelyeket alapvetően a három nagyobb elemszámú szemantikai kategória aránya fémjelzett. Ezután azonban nincs olyan egyértelmű jellemző, amelyre könnyen alapozható lenne egy következő klaszter, így a

következő eredmények további, eddig ismeretlen információkat szolgáltathatnak. A négy klaszterre osztás súlypontjai a következő táblázatban láthatóak:

klaszter	H	I	N	A	L
0	**0.697036**	**0.184827**	0.034281	0.047759	0.036097
1	0.103460	0.116713	0.077483	**0.695120**	0.007224
2	**0.229563**	**0.680872**	0.033994	0.051845	0.003725
3	0.274151	0.261375	0.192164	0.249967	0.022343

6. táblázat: A 4 klaszterre osztás eredménye – a klaszterek középpontjai

Az előzőekben megfigyelt csoportokhoz hasonló adatsorokat látunk a 0, 1 és 2 sorokban, a 3 sorban pedig azt figyelhetjük meg, hogy az összes kategória hasonló arányban jelenik meg (az L-t kivéve). Innentől az ábrák közül már csak egyet-egyet közlök, mivel egy ábrán is látszik már, hogy hogyan különülnek el a klaszterek az előző ábrák ismeretében. Itt tehát az előző két klaszterezéshez hasonlóan az egyik releváns tengelypárra vetített ábrázolás látható:



9. ábra: A 4 klaszterre bontás csoportjainak 2 tengelyre (humán-absztrakt) vetítése

Jelmagyarázat: zöld csillag – 0-val jelölt csoport; piros kör – 1-gyel jelölt csoport; kék kereszt – 2-vel jelölt csoport; lila kör – 3-mal jelölt csoport

Mire következtethetünk a fenti adatok alapján? Fentebb, a 6.1.2. pontban már értelmeztem a három, nagy elemszámú szemantikai kategória által fémjelzett csoportok jelenlétét, ezek ebben a csoportosításban is jelen vannak. Ahogy a 9. ábrán is láthatjuk, a legújabb, az ábrákon lila körrel jelölt 3-as klaszter egy olyan csoport létezését mutatja, amelyben nincs szigorú

szemantikai szelekció, a különböző szemantikai kategóriákba tartozó főnevek hasonló arányban jelenhetnek meg.

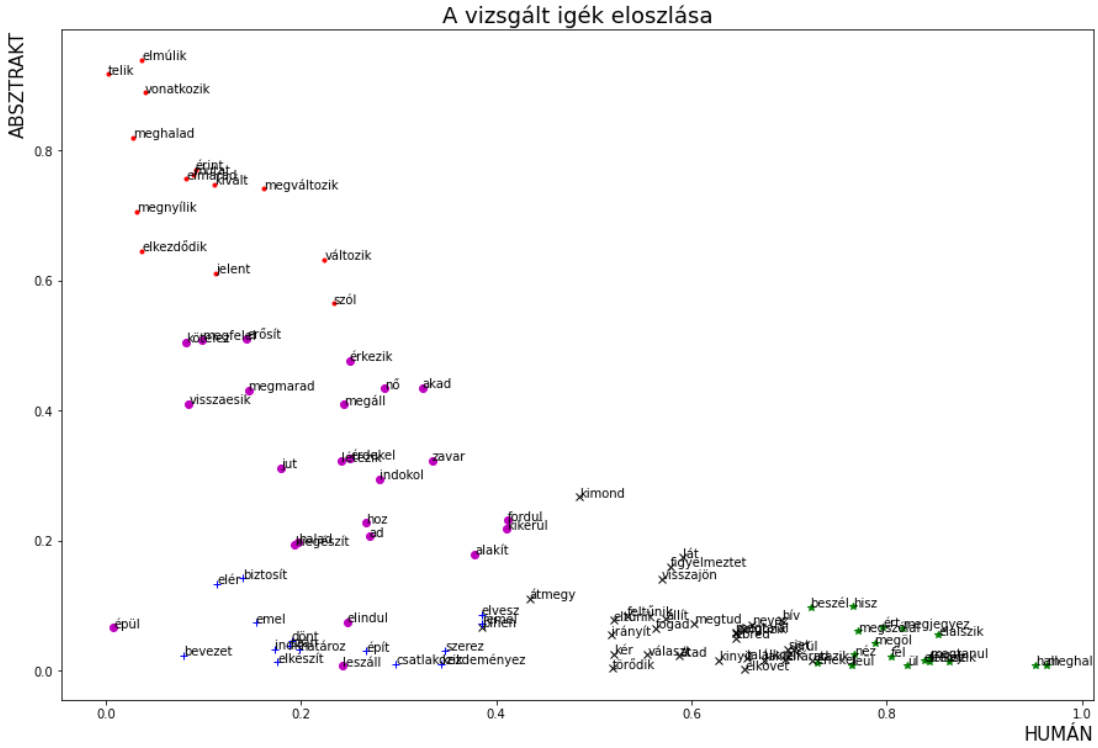
6.1.4. Öt klaszterre osztás

A kutatás tervezési szakaszában az 5 klaszterre bontás jelentősége előreláthatólag az volt, hogy a különböző klaszterek reprezentálják majd az egyes szemantikai kategóriákat. Ez részben a klaszterek elemszámának ismeretében, részben pedig a lenti táblázat adatait figyelembe véve nem tűnik helytálló elméletnek. A következő táblázatban tehát az öt klaszterre bontás során született csoportok középpontjainak koordinátáit láthatjuk:

klaszter	H	I	N	A	L
0	**0.229563**	**0.680872**	0.033994	0.051845	0.003725
1	0.091834	0.079363	0.071767	**0.750693**	0.006344
2	**0.595535**	**0.252029**	0.054544	0.065686	0.032206
3	0.232347	0.263215	0.182930	0.304401	0.017108
4	**0.816935**	0.083764	0.019223	0.035699	0.044379

7. táblázat: Az 5 klaszterre osztás eredménye – a klaszterek középpontjai

Itt is láthatjuk az előzőekhez hasonló adatsorokat a 0, 1, 2 és 3 sorokban, a négy klaszterre bontásban is hasonló eredményekkel szembesülhettünk. A 4-gyel jelölt sorban a korábban említett tendencia, amely szerint a H és az I oszlopok együttesen, magas számban jelennek meg, elmarad, és csak a H oszlopban látunk feltűnően magas számadatot. A következőkben közlöm az utolsó szemantikai szelekcióra vonatkozó klaszteranalízis ábráját:



10. ábra: Az 5 klaszterre bontás csoportjainak 2 tengelyre (humán-absztrakt) vetítése

Jelmagyarázat: kék kereszt – 0-val jelölt csoport; piros kör – 1-gyel jelölt csoport; zöld csillag – 2-vel jelölt csoport; lila kör – 3-mal jelölt csoport; fekete X – 4-gyel jelölt csoport

Milyen újdonságot ad még a csoportosításhoz az öt klaszterre bontás? Van-e valamilyen relevanciája ennek a klaszterezésnek? Láthatjuk, hogy az előző csoportok továbbra is megjelennek, és a már korábban feltételezett kiegészítő disztribúció is látható azokban a csoportokban, amelyekben a H vagy I kategória sokasága a meghatározó. Azonban itt láthatjuk a 4-gyel jelölt csoport megjelenését, amelyben ez a viszony nem áll fenn a két említett kategória között. Így tehát magyarázhatjuk ezt egy olyan csoportként, amelynek a szelekciós szabálya szigorúbb, nem engedi meg az intézmények vagy csoportok megjelenését az alanyi pozícióban.

6.2. Tematikusszerep-kiosztás

Az egyes alanyok ágensségével kapcsolatos kézi elemzés előtt feltételeztem, hogy két csoport lesz majd létrehozható, az egyik csoportban egyértelmű túlsúlyban lesznek az ágens alannyal álló igék, a másikban pedig a nem ágens alannyal állók. Az annotálás során azonban felmerültek olyan igék, amelyek mellett kifejezetten vegyesen álltak ágens és nem ágens alanyok, így a klaszterezés előkészítésénél 3 klaszter létrehozására állítottam be a programot. A szemantikai szelekcióval kapcsolatos klaszterezéshez hasonlóan ennél az elemzésnél is fontosnak tartom közölni az elemzett adathalmaz dimenzióit. Mivel itt két érték volt

lehetséges (igen vagy nem), így a halmaz is 2 dimenziós volt. A létrejött csoportok középpontjai a következő táblázatban láthatóak:

	I	**N**
0	1.372549	98.627451
1	90.820513	9.179487
2	45.80	54.20

8. táblázat: A tematikus szerepek szerinti klaszterezés eredménye – a klaszterek középpontjai

Láthatjuk, hogy valóban létrejöttek az előzetesen feltételezett, többségében ágenses (1) és többségében nem ágenses (0) csoportok, azonban a harmadik csoportban (2) nagyjából fele-fele arányban állnak a különböző tematikus szerepű alanyok. Az egye klaszterek elemszáma a következőképpen alakult: ágenses igékből 39, nem ágenses igékből 51, vegyes tematikus szerepű alanyokkal állókból pedig 10 darab szerepelt a korpuszban.

Ebben az osztályozásban a 0-val és 1-gyel jelölt csoportok nem szorulnak különösebb magyarázatra, a 2-es csoport azonban érdekes lehet további vizsgálatra. A csoportban a következő igék szerepeltek: *ad, átmegy, elindul, erősít, fogad, kiegészít, kötelez, leszáll, megáll, megtanul*. Ezek közül az *ad, átmegy* és *megtanul* igék esetében azt állíthatjuk, hogy különböző jelentések vannak a vegyesség hátterében (például *a zsűri tagjait adja (valamilyen csoport)/ajándékot ad, zebrán megy át/változáson megy át, megtanulják tisztelni egymást/megtanul oroszul*). A másik hét igénél viszont nem ilyen magyarázatot feltételezek. A jelenség illusztrálására bemutatok néhány példát (a következő mondatok ilyen formában nem szerepeltek a korpuszban, több, hasonló mondat egyszerűsített változatai):

1. Az autó *elindul* = Valaki *elindul* az autóval
2. Az adat *erősíti* az állítást = Valaki *erősíti* az állítást az adattal;
3. Ez a hír *fogadta* = Valaki ezzel a hírrel *fogadta*
4. Az eredmény *kiegészíti* a megállapításokat = Valaki *kiegészíti* a megállapításokat az eredménnyel
5. A jog *kötelez* engem = Valaki a joggal *kötelez* engem;
6. A gép *leszáll* = Valaki *leszáll* a géppel;
7. Az autó *megáll* = Valaki *megáll* az autóval

Ez a változás, miszerint az eszközhatározóként szereplő mondatrész átveheti az alanyi pozíciót, ugyanakkor nem veszi át az alany eredeti tematikus szerepét, nem ismeretlen (l. például Bibok 2017), azonban a magyarázata kétséges. Ezt a magyarázatot egy, az ergatív

igékhez hasonló szerkezetben feltételezem. Az ergatív igék esetében az okozó (ágens) megjelenítése nem kötelező, elhagyásakor nem ágens tematikus szerepű tag válik alannyá, az ige pedig formáját tekintve változatlan marad (l. például Radford 2004: 336–343). Az ergatív igékkel álló mondatokra szerepeljen itt néhány példa, hogy meg lehessen vizsgálni a hasonlóságot. A példák (Radford 2004: 338)-tól származnak, alattuk feltüntetve saját fordításom látható:

1. They will *roll* the ball down the hill = The ball will *roll* down the hill
   Ők le fogják gurítani a labdát a dombról = A labda le fog gurulni a dombról
2. He *filled* the bath with water = The bath *filled* with water
   Ő megtöltötte a kádat vízzel = A kád megtelt vízzel
3. He *broke* the vase into pieces = The vase *broke* into pieces
   Ő darabokra törte a vázát = A váza darabokra tört

Az ergatív igék esetében az alanyi pozíciót a korábban tárgy esetű elem foglalja el, itt azonban a parafrazeált mondatokban ezek az elemek eszközhatározóként jelennek meg, ezért nem állíthatjuk, hogy ezek az igék ergatívak lennének, de a hasonlóság feltűnő.

6.3. Összevetés

Ugyan az eddigi eredményekből is megállapíthatók tendenciák, következtetéseket úgy vonhatunk csak le, ha egymásra vetítjük ezeket. Először vizsgáljuk meg, hogy mennyire felelnek meg az egyes klaszterezések csoportjai egymásnak.

Az összehasonlítást a két és öt klaszterre bontással kezdve azt láthatjuk, hogy az összes elem kivétel nélkül, amelyik H vagy I jegyű alanyok többsége által fémjelzett csoportba tartozott, a két klaszteres elosztásban a „cselekvésre képes” csoportban kapott helyet. Hasonlóan egyértelmű a reláció az A jegyű csoportok esetében, az öt klaszteres osztásban ilyen csoportba kerülők mind az ennek megfelelő csoportban voltak a két klaszteres osztásban is. Az egyetlen olyan csoport, amit visszavetítve vegyesnek látunk, sem okoz meglepetést, mivel itt ez a szemantikai szelekció nélküli klaszter. A vegyesség oka az, hogy a két klaszteres osztásban minden elem elhelyezésével olyan klaszterbe kerültek ezek az igék, amelyben nem volt lehetőség annak a megállapítására, hogy nincsen szemantikai megszorítás az esetükben.

A három és öt klaszterre osztásokat vizsgálva hasonló eredményeket kapunk, mint az előző összehasonlításnál. Mivel a három klaszternél a nagy elemszámú csoportok külön csoportokra osztása már megjelent, ezért ennek fényében megvizsgálhatjuk, hogy mennyire voltak állandóak ezek a kategóriák. Az öt klaszterre osztás esetében az I jegyű klaszterbe kerülő

elemek már a három klaszteres osztásnál is ilyen klaszterben voltak. Hasonlóan mindkettő H többségű csoport elemei is ilyen csoportban voltak a hármas osztásnál, valamint az A jegyű csoport is állandó volt. A szelekció nélküli csoport ebben a visszavetítésben is vegyes volt.

Végül a négy és öt klaszteres csoportosítás összehasonlításánál is azt láthatjuk, hogy azok a csoportok, amelyek egyértelmű többséggel rendelkeztek az öt klaszteres osztásnál (0, 1 és 2), a négy klaszteres osztásnál is egyértelmű csoportokba oszthatóak voltak. Kis variációt mutat a másik két klaszter a visszavetítésnél. Az öt klaszteres osztásnál 3-mal jelölt, szemantikai szelekció nélküli csoport elemei javarészt a négy klaszteres osztásnál is ilyen csoportba kerültek, ezeken kívül azonban néhány A többségű ige is ide került. A 4-gyel jelölt, H és I többségű csoport is javarészt egy klaszterbe, a H többségűbe került, illetve egy kevés olyan elem volt még benne, amely a vegyes, szemantikai szelekció nélküli csoportban volt.

A szemantikai szelekció alapján történő klaszterezéseket megvizsgálva azt láthatjuk, hogy kifejezetten állandó csoportok jöttek létre, így azt mondhatjuk, hogy ebből a szempontból a k-means klaszterezés megfelelő eszköz. A továbbiakban megvizsgálom, hogy milyen összefüggéseket mutat a szemantikai szelekció alapján történő klaszterezések és a tematikus szerepek alapján történő klaszterezés eredménye, meghatározhatóak-e a tematikus szerepek a szemantikai eloszlás vizsgálatával.

Elsőként az ágenses klasztert megvizsgálva azt láthatjuk, hogy egyetlen olyan elem se került ezek közé, amelyek szemantikai szelekció alapján olyan csoportban lettek volna, amelynek meghatározó jegye az A szemantikai kategória többsége. Túlnyomó részt olyan klaszterekben voltak, amelyekben a H vagy az I kategória a meghatározó minden alkalommal, a négy klaszterre osztásnál 4, az öt klaszterre osztásnál pedig 3 ige volt a szelekció nélküli klaszterben.

A nem ágenses csoportnál nem ilyen egyértelmű a helyzet, mivel minden szemantikai jegy által fémjelzett klaszterből voltak benne elemek. A relációt megfordítva azonban láthatjuk, hogy az összes olyan elem, amelyik bármelyik szemantikai szelekciós osztályozásnál, az A többségén alapuló klaszterbe került. Ezen felül a legtöbb elem, ami szemantikai szelekció nélküli csoportokba került, szintén ebben a klaszterben található.

A vegyesen ágens és nem ágens alannyal álló igék más klaszterezések alapján való csoportosítását illusztrálja a következő táblázat:

	klaszterek száma				tematikus szerepek
	2	3	4	5	
ad	cselekvésre képes	I	szelekció nélküli	szelekció nélküli	vegyes
átmegy	cselekvésre képes	H	szelekció nélküli	H+I	vegyes
elindul	cselekvésre képes	I	szelekció nélküli	szelekció nélküli	vegyes
erősít	cselekvésre nem képes	A	A	szelekció nélküli	vegyes
fogad	cselekvésre képes	H	H	H+I	vegyes
kiegészít	cselekvésre képes	I	szelekció nélküli	szelekció nélküli	vegyes
kötelez	cselekvésre nem képes	A	A	szelekció nélküli	vegyes
leszáll	cselekvésre képes	H	szelekció nélküli	szelekció nélküli	vegyes
megáll	cselekvésre nem képes	A	szelekció nélküli	szelekció nélküli	vegyes
megtanul	cselekvésre képes	H	H	H	vegyes

9. táblázat: A vegyes klaszterbe kerülő igék egyéb eredményei

Bár nem egyöntetű az eredmény, ha azokat az igéket tekintjük, amelyeket a 6.2. fejezetben az ergatív igékhez hasonlónak neveztem, egy kivételével mind a szelekció nélküli kategóriába kerülnek az ötklaszteres felosztásban.

Azt láthatjuk, hogy bár a megfeleltetés nem egyértelmű az egyes szemantikai szelekciós klaszterezések eredményei és a tematikus szerepek szerinti csoportok között, felfedezhető valamilyen mértékű összefüggés.

6.4. Értékelés

A kutatás hipotéziseit értékelve a következőek. Az 1. hipotézist, miszerint az igék alannyal szemben támasztott szemantikai szelekciója összefüggést mutat a korpuszban lévő mondatok alapján számított kvantitatív adatokkal, igazoltnak tekinthető, mivel a feldolgozást és elemzést követően értelmezhető csoportok jöttek létre a klaszteranalízist követően. A 2a. hipotézisben azzal kapcsolatban kerestem a választ, hogy megismerhetjük-e a klaszteranalízis segítségével az igék szemantikai szelekciós tulajdonságait. Ezt szintén igazoltnak tekinthetjük, hiszen a különböző számú klaszterekre való osztások esetében kifejezetten állandó csoportok jöttek létre, így megismerhettük az egyes igék szemantikai szelekciós tulajdonságait. A 2b. hipotézisben azt vetettem fel, hogy a klaszteranalízis alapján megismerhetjük az igék jellemző tematikusszerep-kiosztását. Az erre vonatkozó klaszterezés eredményei azt mutatják, hogy létrejön három, nyelvre vonatkoztatva is releváns, nem véletlenszerű csoport, így ezt a hipotézist is igazoltnak nevezhetjük. Végül a 2c. hipotézisre vonatkozóan elmondhatjuk, hogy bár az egyes szemantikai szelekció szerinti kategóriák nem feleltethetőek meg egy az egyben az ágencia alapján történő klaszterekkel, felfedezhettünk bizonyos kapcsolatokat az egyes csoportosítások között, így ezt a hipotézist is igazoltnak tekinthetjük.

7. Összefoglalás, perspektívák

Dolgozatomat összefoglalva elmondható, hogy hipotéziseim igazolást nyertek a k-means klaszteranalízis segítségével. Ennek megfelelően kijelenthető, hogy a szemantikai szelekció és a tematikusszerep-kiosztás (legalábbis abban a szeletében, amelyet jelen kutatásban megfigyeltem) vizsgálható számítógépes nyelvészeti eszközök segítségével. A kutatásomat korpuszadatokra alapoztam, különböző számítógépes elemzési módszereket és nem felügyelt gépi tanulást használtam annak érdekében, hogy az igék különböző, alanyukkal szemben támasztott szelekciós tulajdonságait megismerjem.

Jelen kutatás közvetlen folytatása lehet a korpusz anyagának bővítése, kiterjesztése több igére és több gyakori főnévre. Ennek eredménye pontosabb képet adhat a szemantikai szelekció tekintetében, akár a megfigyelendő kategóriák számának növeléséhez is vezethet, valamint a tematikus szerepek tekintetében is, már csak azért is, mert több mondat elemzésével is tisztább képet kaphatnánk az egyes igék melletti alanyok ágensségével kapcsolatban.

Távolabbi folytatási lehetőség a vizsgálat kiterjesztése egyéb mondatrészek és tematikus szerepek kombinációjára, például a tárgy mondatrész és a páciens vagy téma tematikus szerepek kapcsolatának felderítésére. Ezen az elméleti folytatáson felül alkalmazásként is folytatható lehet a kutatás. Ezt a lehetőséget a folyamat automatizálásában látom, mind az igék, mind a főnevek jellemzésével kapcsolatban. Az igék automatikus jellemzését az adhatná, hogy milyen, már címkével ellátott főnevek állnak gyakran a környezetében. A főnevek jellemzésénél pedig felcímkézetlen elemek szemantikai szerepének megállapítására lehetne lehetőség különböző, igelistában szereplő igék melletti előfordulásuk alapján. Ezeken túl konkrét felhasználási lehetőséget látok a módszerek többnyelvű korpuszon történő alkalmazásában, ahol az egymásnak megfelelő igék szemantikai szelekciós és tematikusszerep-kiosztásra vonatkozó tulajdonságainak különbségeit lehetne feltárni, ezzel akár a gépi fordítási folyamatokat is elősegítve.

Irodalomjegyzék

Bibok, Károly 2017. Az eszköz–alany váltakozás lexikai pragmatikai keretben. *Jelentés és Nyelvhasználat* **4**/**1**:35–49. doi:10.14232/jeny.2017.1.3.

Chomsky, Noam 1981. *Lectures on government and binding*. (Studies in generative grammar 9) Dordrecht, Holland ; Cinnaminson, [N.J.]: Foris Publications.

Dowty, David 1991. Thematic Proto-Roles and Argument Selection. *Language* **67**/**3**:547. doi:10.2307/415037.

Fillmore, Charles J. 1970. Subjects, speakers, and roles. *Synthese* **21**/**3–4**:251–274. doi:10.1007/BF00484800.

Gyulai, Lívia 2021. Az igekötők legjellemzőbb argumentumszerkezet-változtató hatásainak korpuszalapú vizsgálata. MTA Nyelvtudományi Intézet. doi:10.18135/ALKNYELVDOK.2021.15.8.

Jackendoff, Ray 1990. *Semantic interpretation in generative grammar*. 6. print. (Studies in Linguistics Series) Cambridge: MIT Pr.

Jackendoff, Ray 1997. *The architecture of the language faculty*. second print. (Linguistic Inquiry Monographs 28) Cambridge, Mass.: MIT Press.

Jakubíček, Miloš – Adam Kilgarriff – Diana McCarthy – Pavel Rychlý 2010. Fast Syntactic Searching in Very Large Corpora for Many Languages. *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*. Tohoku University, Sendai, Japan: Institute of Digital Enhancement of Cognitive Processing, Waseda University. 741–747.

Kenesei, István 2000. Van-e segédige a magyarban? Esettanulmány a grammatikai kategória és a vonzat fogalmáról. In István Kenesei (szerk.) *Igei vonzatszerkezet a magyarban*. Budapest: Osiris.

Komlósy András 2015. Régensek és vonzatok. In Kiefer Ferenc (szerk.) *Strukturális magyar nyelvtan 1. Mondattan*. Budapest: Akadémiai Kiadó. 274–486.

MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics* **5.1** University of California Press:281–298.

Oravecz, Csaba – Tamás Váradi – Bálint Sass 2014. The Hungarian Gigaword Corpus. In Nicoletta Calzolari – Khalid Choukri – Thierry Declerck – Hrafn Loftsson – Bente Maegaard – Joseph Mariani – Asuncion Moreno – Jan Odijk – Stelios Piperidis (szerk.) *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik: European Language Resources Association (ELRA). 1719–1723.

Pedregosa, Fabian – Gaël Varoquaux – Alexandre Gramfort – Vincent Michel – Bertrand Thirion et al. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**/**85**:2825–2830.

Radford, Andrew 2004. *Minimalist syntax: exploring the structure of English*. Cambridge, UK: Cambridge University Press.

Resnik, Philip 1996. Selectional constraints: an information-theoretic model and its computational realization. *Cognition* **61**/**1–2**:127–159. doi:10.1016/S0010-0277(96)00722-6.

Sass, Bálint 2017. Keresés korpuszban: a kibővített Magyar történeti szövegtár új keresőfelülete. In Tamás Forgács – Miklós Németh – Balázs Sinkovics (szerk.) Vol. 9 Szeged: Szegedi Tudományegyetem Magyar Nyelvészeti Tanszék. 267–277.

Szécsényi, Tibor 2019. Argumentumszerkezet-variánsok korpusz alapú meghatározása. Szeged: Szegedi Tudományegyetem, Informatikai Intézet. 315–329.

Szécsényi Tibor – Virág Nándor 2022. Az ige helyhatározói bővítményeinek megkülönböztetése és az argumentumszerkezeti variánsok korpusz alapú szétválasztása. *Magyar Számítógépes Nyelvészeti Konferencia*. Szeged. 535–547.

Vincze, Veronika – Dóra Szauter – Attila Almási – György Móra – Zoltán Alexin – János Csirik 2010. Hungarian Dependency Treebank. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA).

WordNet 2010. Princeton University „About WordNet." Weboldal. *http://wordnet.princeton.edu*.

Zsibrita, János – Veronika Vincze – Richárd Farkas 2013. magyarlanc: A Tool for Morphological and Dependency Parsing of Hungarian. *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*. Hissar, Bulgaria: INCOMA Ltd. Shoumen, BULGARIA. 763–771.