

Magyar hiedelmek hierarchikus struktúrája

Dominich Sándor¹, Darányi Sándor², Szlávik Zoltán³

¹Veszprémi Egyetem, Magyarország, email: dominich@dcs.vein.hu

²Högskolan i Borås, Sweden, email: daranyi@hb.se

³Queen Mary, University of London, United Kingdom, email: zolley@dcs.qmul.ac.uk

A számítógépes korpusznyelvészet egyik fontos területe a klaszterezés. Jelen dolgozatban elvégezzük a Magyar Hiedelemszövegek AdatBázis (MHAB) osztályozását az interakciós eljárás felhasználásával. Eredményeink alapján megállapíthatjuk, hogy a magyar hiedelmek párokba szerveződő hierarchikus szerkezetet alkotnak.

Bevezetés

A nyelvészet és a számítógép-tudomány kapcsolata mintegy öt évtizedes múltra tekint vissza, amikor amerikai kutatók felvetették a gépi fordítás (Yngve, 1957) és információ-visszakeresés (Salton, 1965) lehetőségét. Az alapötlet mindkét esetben azonos volt: mivel a számítógép tetszőleges jelrendszer elemzésére programozható, a természetes nyelvi jelekből álló rendszerek vizsgálata is megvalósítható számítógép segítségével. Nem kell tehát mást tenni, mint a szóban forgó nyelv nyelvtanát és szókészletét 'betáplálni' a számítógépbe. Ahhoz, hogy ezt megtehessek, a nyelvtan szabályait formalizálva, a matematika szabályaihoz hasonlóan kell megadnunk (Chomsky, 1956, 1959; Kay, 1965; Hays, 1966). Jóllehet a grammatikának ilyen formalizálása és alkalmazása nem váltotta be a hozzá fűzött reményeket, egyrészt elindítója volt annak a kutatási irányzatnak, amely a természetes nyelvek grammatikájának formalizálási lehetőségeit vizsgálja, másrészt elvezetett a hatvanas évek elejétől „számítógépes nyelvészet” (computational linguistics) terminussal illetett diszciplína létrejöttéhez.

A számítógépes nyelvészet elméleti és alkalmazott interdiszciplináris tudományág, amely a természetes nyelvek számítógépes feldolgozásával (natural language processing) foglalkozik. A megszokott klasszikus nyelvészeti területeken (hangtan, alaktan, mondatn, jelentésn) kívül felöleli a fordítás, az automatikus kivonatolás, az automatikus indexelés, az információ-visszakeresés, az automatikus osztályozás (klaszterezés), az automatikus kivonatolás, a mesterséges intelligencia-kutatás, a párbeszédes rendszerek vizsgálatának egyes kérdéseit is. A szóalakok belső szerkezetének megállapítása mind elméleti, mind pedig

gyakorlati szempontból alapvető feladat, ugyanakkor az alaktani elemzés a természetes nyelvi szövegeknek számítógépes feldolgozásra alkalmas reprezentációi (adatstruktúrák) létrehozásának előfeltétele. Ez a reprezentáció utóbbi teszi lehetővé az automatikus indexelés, a számítógépes szótárkészítés, a számítógépes mondattani elemzés, a számítógépes információ-visszakeresés automatizált megvalósítását biztosító algoritmusok implementálását (Prószéky, 1999).

A természetes nyelvi szövegekből álló adatbázisok a szövegkoherencia és a jelentés vizsgálatában új perspektívát jelentenek, továbbá a — fizikából ismert tudományos igényű és jellegű — kísérletezés lehetőségét teremtették meg. A szövegek különböző célú számítógépes vizsgálata külön diszciplína, a korpusznyelvészet kialakulásához vezetett. A korpusznyelvészet elsősorban lexikográfiai jellegű kérdéseket vizsgál, a hagyományos szótárkészítővel szemben azonban a korpuszokat nemcsak arra használja, hogy belőlük példákat merítsen, hanem rendszeres vizsgálatnak veti alá őket, vagyis gondosan szemügyre veszi a szavak statisztikai jellemzőit (előfordulásának számát, eloszlását, törvényszerűségeit, együtt-előfordulást). A nyelvi adatbázis szavainak alaktani elemzése nyomán index-kifejezéseket állapítunk meg, ezeknek segítségével a célnak (pl. keresés, osztályozás) megfelelő számítógépes korpusz-reprezentációt (pl. mátrix, vektor, gráf, neuron) állítunk elő, ez vezet el a használt algoritmus, számítógép és szoftver által megkövetelt adatszerkezethez, amellyel a kívánt vizsgálatokat el lehet végezni (van Rijsbergen, 1979; Salton and McGill, 1983; Dominich, 2001).

A természetes nyelvi szövegekből álló adatbázisokhoz kapcsolódó egyik fontos gyakorlati alkalmazás az automatikus osztályozás (klaszterezés). Ennek kétféle jelentése van: (i) a szövegeknek előre megadott kategóriák segítségével megfelelő osztályba történő besorolása), és (ii) a szövegeknek besorolása automatikusan felismert (tehát előre meg nem adott) kategóriákba. Jelen dolgozatnak az a célja, hogy ez utóbbi értelemben elvégezze a magyar hiedelemszövegek osztályozását az interakciós módszer felhasználásával, és megállapítsa azok szerkezetét.

Interakciós eljárás

Az interakciós) információ-visszakeresési módszerben (angol rövidítése: I^2R = Interaction Information Retrieval; Dominich, 1994) a felhasználó által feltett kérdés „hatást” gyakorol a keresendő dokumentumhalmazra: részben megváltoztatja a dokumentumok közötti kapcsolatokat, azaz kölcsönhatásba lép velük. A dokumentumok (a későbbiekben: objektumok) nem egymástól elszigetelt egységeket képeznek, hanem egy súlyozott, összekapcsolt hálózatot,

amit a felhasználó kérdése a válasz megadása előtt részlegesen átalakít. Az I^2R matematikai modellje a mesterséges neurális hálózatok elméletének alapvető állapotegyenletén nyugszik. Az objektumok mesterséges neurális hálózatnak felelnek meg, amelyben az egyes objektumokat aktivitási szinttel rendelkező mesterséges neuronok modelleznek. A kérdés szintén egy neuronnak felel meg, ez beépül a hálózatba – mint egy új objektum –, és így a hálózat részlegesen átszerveződik: új kapcsolatok alakulnak ki a kérdés-neuron és az eredeti objektum-neuronok között, továbbá az eredeti hálózatban kialakult kapcsolatok egy része módosulhat. Minden objektumot egy-egy 'súlyvektor' (számtömb) reprezentál. Az o_i objektum vektorát jelölje $t_i = (t_{ik})$, $k = 1, \dots, n_i$. Két objektum közötti kapcsolatot két típusú súllyal jellemezhetünk (1. ábra). Az első egy index-kifejezés *gyakorisága* egy adott objektumban (w_{ijp}). Ezt a gyakoriságot a következőképpen definiáljuk:

$$w_{ijp} = \frac{f_{ijp}}{n_i}, \quad p=1, \dots, n_j \quad (1)$$

ahol f_{ijp} a t_{jp} index-kifejezés előfordulásainak száma az o_i objektumban, n_i pedig az o_i objektum összes index-kifejezéseinek száma. A másik súlysúly az *inverz objektum gyakoriság* (w_{ikj}), ami egy index-kifejezés 'visszaverődése' az objektum tartalmáról. Ezt a következő módon határozzuk meg:

$$w_{ikj} = f_{ikj} \log_2 \frac{2n}{df_{ik}} \quad (2)$$

ahol f_{ikj} a t_{ik} index-kifejezés előfordulásának száma az o_j objektumban, df_{ik} azon objektumok száma, amelyekben megtalálható a t_{ik} , n pedig az objektumok száma. Két objektum közötti kapcsolat erősségét ezekből a gyakoriságokból a következőképpen számítjuk ki:

$$\sum_{p=1}^{n_j} w_{jpi} + \sum_{k=1}^{n_i} w_{jik} \quad (3)$$

1. ábra. Két objektum közötti kapcsolatok az interakciós eljárásban.

A visszakeresési eljárás aktiváció indítását és terjesztését jelenti a hálózatban. Induláskor a kérdés kapcsolódását vizsgáljuk a többi objektummal. Az aktivitás terjedése a WTA (Winner-Takes-All: a nyertes visz mindent) stratégia alapján történik: mindig a legnagyobb súllyal

kapcsolódó objektum kerül kiválasztásra. Az aktivitás addig terjed, amíg egy, előzőleg már nyertes objektumhoz nem jutunk. Ekkor találtunk egy *öngerjesztő kört*, ennek elemei válaszok a kérdésre.

Példa. Adott az alábbi három dokumentum (2., 3. és 4. ábrák):

D_1 = Az információ-visszakeresés egy nagyon gyorsan és dinamikus fejlődő tudományág.

D_2 = Napjainkban a kereskedelmi keresők a Boole-féle információ-visszakeresés modelljét használják a leggyakrabban. Ez a megoldás implementálható a legkönnyebben.

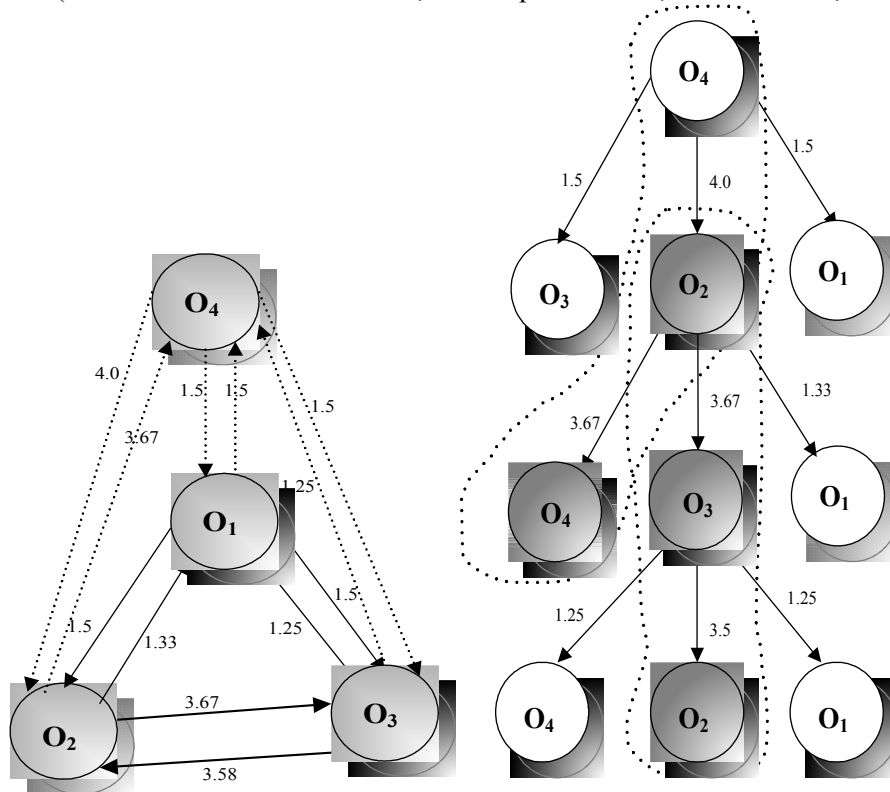
D_3 = Az implementálás során fontos a memória és a lemez megfelelő használata. Az információ-visszakeresés implementálása nem egyszerű feladat.

Indexeljük a dokumentumokat a következőképp:

$o_1 = (t_{11} = \text{információ-visszakeresés}, t_{12} = \text{tudományág})$

$o_2 = (t_{21} = \text{információ-visszakeresés}, t_{22} = \text{Boole-féle}, t_{23} = \text{implementálás})$

$o_3 = (t_{31} = \text{információ-visszakeresés}, t_{32} = \text{implementálás}, t_{33} = \text{memória}, t_{34} = \text{lemez})$



2. ábra.

Eredeti objektum-hálózat.

3. ábra

A kérdés beépülése utáni hálózat.

4. ábra.

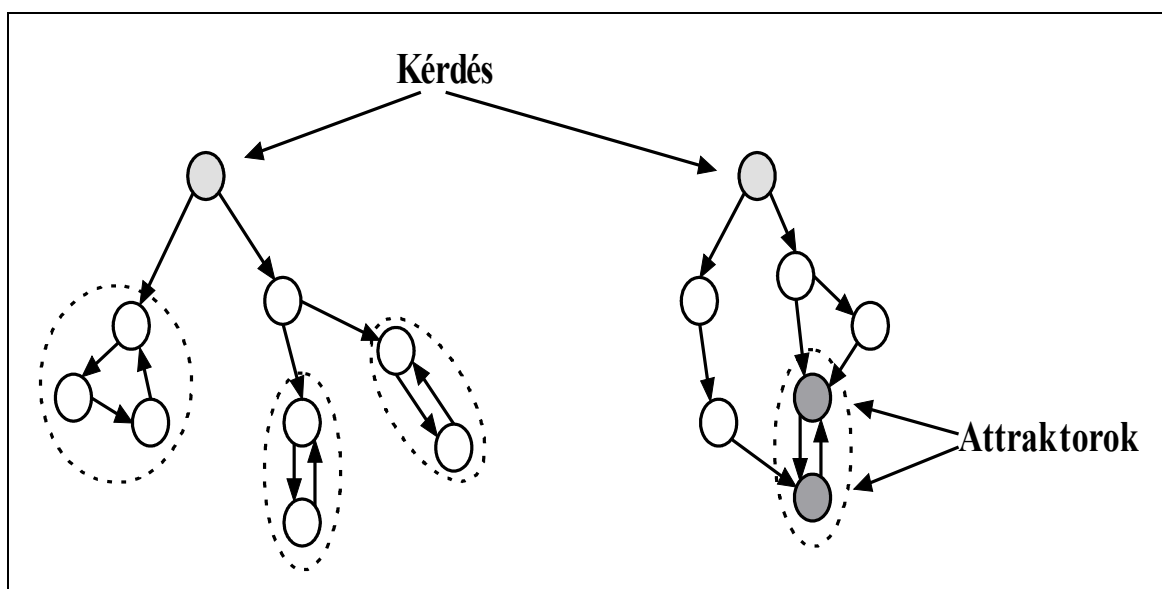
A keresés folyamata.

A kérdés legyen a következő: $Q = \text{Használják a Boole-féle információ-visszakeresés modelljét?}$ A kérdésből indexeléssel a következő objektumot kapjuk, ami beépül az objektumhálózatba: $O_4 = (t_{41} = \text{információ-visszakeresés}, t_{42} = \text{Boole-féle})$. A 3. ábra a kérdés beépülése utáni hálózatot szemlélteti. Az új súlyokat szaggatott, a régi, de megváltozott súlyokat vastag vonallal jelöltük. Az aktivitás a kérdés-objektumtól, o_4 , indul, és halad objektumról objektumra (4. ábra). A legnagyobb súllyal az o_2 objektum kapcsolódik, így a WTA stratégia elvén ez lesz a következő objektum, amelyre az aktivitás áterjed. Ezután vizsgáljuk az o_2 kapcsolatait a többi objektummal. Itt két objektum is azonos súllyal szerepel (o_3 és o_4): elágazás történik. Mivel az o_4 objektum már ki volt nyertes, megvan az első öngerjesztő kör. A másik irányba terjesztve az aktivitást: az o_3 objektum vizsgálatakor az o_2 objektum lesz a nyertes, ő visz mindent. Ezzel bezárul egy újabb kör, és elhal az aktivitás. A kölcsönhatás eredménye két öngerjesztő kör, melynek tagjai adják a kérdésre a választ. Így a válasz-objektumok a következők: o_2 és o_3 . (o_4 nem

szerepel a válaszok közt, hisz az maga a kérdés.) o_3 nem tartalmaz Q-beli szavakat, mégis válasz, mivel más kulcsok révén erősen kapcsolódik az o_2 objektumhoz. Ez a tulajdonság az FR modell sajátossága.

Attraktorok

Az FR eljárásban a keresés azokat az objektumokat adja vissza válaszul, amelyek egy öngerjesztő körben vannak. Különböző kérdések esetén különböző válaszokat kapunk, tehát a válasz-objektumokból álló körök is kérdésenként változhatnak. Az *attraktor*-objektumok olyan objektumok, amelyek meglehetősen sok – szélsőséges esetben minden – kérdés esetében elvonzzák az aktivitást. Ennek következtében a válasz-körök különböző kérdések esetében is azonosak lesznek, ugyanazt a (gyakran nem releváns) választ kapjuk vissza. Az attraktor-objektum mindig szerepel a válaszok között (5. ábra).



5. ábra. Öngerjesztő körök attraktor nélkül (balra) és attraktorokkal (jobbra). Az attraktorok magukhoz vonzzák az aktivitást.

Az interakciós eljárást alkalmazni lehet attraktorok felismerésére a következő lépésekben:

- elég nagy számú, index-kifejezéseket tartalmazó kérdést generálunk véletlenszerűen,
- terjesztjük az aktivitást minden egyes kérdés esetében,
- figyeljük azonos körök(attraktorok) esetleges kialakulását,
- okokat és hatásokat vizsgálunk, tárgyalunk.

Magyar Hiedelemszövegek Adatbázis (MHAB) informatikai előkészítése

Az MHAB 2704 magyar hiedelem-szöveget tartalmaz (Darányi, 2001). Vannak köztük rövid szövegek, amelyek tömören egy hiedelmet tartalmaznak, és vannak olyanok is, amelyek az adott hiedelmet kis történeten keresztül mutatják be. Néhány példa a hiedelmekből:

- *Aki mákor vet, az ne szóljon senkinek.*

- *Aki aratáskor a tarlót vízzel leönti s a vizes helyre lép, akkor seb lesz a lábán.*

- *Alig lehettem 10 éves de úgy emlékszem mintha most történt volna. A tehenünk véres tejet adott azt monták, hogy a boszorkányok megfejik. Hát jóvan gondotam magamban majd én meglessem, hogy ki feji még a tehenünket. Szótam a szogánok a bátyámnok és így hárman kifeküdtünk az istállóba. Egyszer aztán megcsapott engem egy hideg levegő és abban a pillanatban erős nyomást éreztem a mellemén és jól hallottam hogy a tehen zúg vagyis a teje. Fölakartam kenit de nem tudtam még megmozduni sē, szóni akartam de azt se tudtam hát vártam. Egyszercsak hallom, hogy fölugrik a bátyám és kapja a vasvillát és belevágja az istálló ajtóba és ekezd kiabányi, hogy megvan a boszorkány. Kigyűttek a szüieim a lármárá kerestük, de sēhol sēm találtunk sēnkit az egész istállóban, és a tehen tōgye üres vót és nedves. Másnap aztán együtt hozzánk az öreg Tolam néni egy kis sóér még hagymáér, mer a boszorkánynak ha valahol megsúrják el köll mēnni oda másnap sóért és foghagymáér mer csak akkor gyógyul még a sebe. De bizony az öreg Tolam néni megjártó mert amint e panaszotta hogy a hatábó esett a vasvilla mingyá tudtuk hogy ő fejte még a tehenünket, és a só még a foghagyma mellé bátyám jó everte, de többet nem is gyűtt még felénk se.*

- *Ha a kész ruha leesik a földre, akkor az tetszetős lesz, ha a tű eltörik benne, akkor mēnyasszony lesz bent valaki, ha megsúrja tűvel a kezét a ruha varrásakor az, aki varrja, akkor a ruha jól fog állni (sikerni fog), ha elszakad a varrás alatt, akkor megbetegszik benne valaki, ha ollóval bevágja valaki, akkor pör keletkezik belőle, ha még nincs a ruha kiprobávo és úgy kerül az ágyra, akkor szintén bajosan sikerül, és végül ha pénteken lett szabva, akkor se lesz kifogástalan.*

- *Ha azt akarod, hogy a bajszod hama nyőjön, akkor régge amind kinyitod a szēmed, minnyá a két újjadda éjnyalló bekenyéd a bajszod helit.*

Számítógépes nyelvtechnológiai szempontból az MHAB sajátosságai közül ki kell emelni a következőket. A szövegek ASCII, azaz szöveg formátumban szerepelnek, ezért az MHAB viszonylag könnyen adaptálható, azaz lehetővé teszi a kívánt számítógépes adatszerkezetek kialakítását és szükséges algoritmusok alkalmazását. Mind mai írásmód, pl.

Ha kis gyermeknek komoly baja van, akkor szenes vízzel mossák meg. A meleg vízbe 9 db. szenet tesznek, megkenik a vízzel a gyermek homlokát, és ezt mondják: Ha férfi, kalap alá; ha leány, párta alá; ha asszony, fejkötő alá, az atya, fiú, szentlélek nevében. Ámen.

mind pedig régebbi vagy tájszólás jellegű írásmód és szóhasználat, pl.

Ha a tehenet merrontya a boszorkány, vésznek egy új fēlliteres cserepbēgrét; abba belētésznek ecs csomaócskát a tehen gannajjábó. Azután szöget vernek a kény belsejébe s erre felakasztyák a bēgrét. Etteó aszt meggyön a tehen haszna.”

jellemző az MHAB állományára. Viszonylag sok a szóalak (pl. a számítógépes nyelvtechnológiában elterjedten vizsgált angol nyelvvel összehasonlítva):

asszony, asszonnak, asszony, zasszony, háziasszony, asszonyról, fehérnép, asszonyhoz, gazdasszony, kisasszony, gazdasszonyok, fehérnép, asszonyt, asszonnyal, asszonyok, fehérnépek, asszon, háziasszonyok, vászoncselédnek, asszonyokhoz, gazdaasszony, asszonyokról, gazdasszonynak, gazdasszonya, ételvivőasszony, asszonya, asszonynak, asszonyoknak, háziasszonynak, asszonyai, asszonyra

Az automatikus szövegfeldolgozás első lépéseként a stop-lista meghatározására került sor. 1,551 stop-szó azonosítása történt meg, manuálisan, azaz olyan szóé (névmás, határozó, jelző, ige, múlt idejű alak, ritkán használt szó, ragozott alak), amely nem vagy alig hordoz jelentést a hiedelemre nézve. A stop-lista néhány részlete a következő:

abba, abban, abbó, abból, abbú, abbüő, addig, ahány, ahanyadik, ahányadik, ahányan, ahányat, ahányszor, ahányszori, ahhoz, ahogy, ahol, ahon, ahonneét, ahonnét, ahova, ahová, ahun, ahuon, ajánlatos, ajánlják, akár, akárhogy, akármelyik, akármilyen, akármit, aképen, ..., aki, akié, akiébe, akiért, akihez, akijé, akik, akiknek, akin, akinek, akinél, akire, akiről, akit, akitől, akivel, akki, akkinek, akkire, ..., zén, zett, zije, zik, zis, zisnagyon, zni, zsémb, ztem, zzen, zönt, örvend, örvendetes, örvendetesebbet, összefügg, összefüggő, övéket

A stop-lista szavait automatikusan, C++ programozási nyelven megírt számítógépes programok segítségével eltávolítottuk az MHAB-ból. Az eltávolítás után 14.421 szóalak maradt az eredetileg szereplő összes 15.972 szóalakból. Az automatikus nyelvtechnológia következő lépéseket az azonos jelentéssel felruházható, de különböző alakú szavaknak azonos töre való redukálása (lemmatizálás, stemming) képezi. Jóllehet léteznek a magyar nyelvre kifejlesztett stemmerek (Morphologic, Szószablya), a hiedelemszövegek változatos, különleges (fentebb érzékeltetett) szóhasználata, régi homonimák miatt azokat az MHAB-ra nem vagy csupán igen alacsony hatékonysággal lehet alkalmazni. Ezért a szótőre való visszavezetés manuálisan valósult meg. A szótő-lista néhány részlete a következő:

<i>#agyon</i>	<i>#úrvacsora</i>
<i>#ajak</i>	<i>#üsző</i>
<i>#ajtó</i>	<i>#üt</i>
<i>#ajándék</i>	<i>#ütőér</i>
<i>#akadály</i>	<i>#üveg</i>
<i>#akar</i>	
<i>#alsónemű</i>	
<i>...</i>	
<i>#csal</i>	
<i>#család</i>	
<i>#csütörtök</i>	
<i>...</i>	
<i>#úrfelmutatás</i>	
<i>#úrnapja</i>	

A stop-listán szereplő szavaknak automatikus, C++ programozási nyelven írt számítógépes programok segítségével való törlése után maradt szavak tövesítése 2.602 szótót eredményezett. Ezek a szótövek képezik az *index-kifejezéseket*, amelyek segítségével minden hiedelemszöveget a benne előforduló index-kifejezések előfordulási számainak számtömbjeként, 'vektoraként' ábrázoltunk. A valamennyi hiedelem-vektort oszlopokba és egymás mellé rendezve kapjuk a *kifejezés-dokumentum* (term-by-document) *matrixot*, TD-t. A TD mátrixnak az *i*-ik sorában és *j*-ik oszlopában szereplő eleme az *i*-ik index-kifejezésnek a *j*-ik hiedelemszövegben való előfordulási száma. A TD mátrixot automatikusan, C++ programozási nyelven írt számítógépes program segítségével állítottuk elő, 2.602 sora és 2.704 oszlopa van (1. táblázat).

1. táblázat. MHAB (Magyar Hiedelemszövegek Adatbázis) TD mátrixának részlete. Sorszám=index-kifejezés sorszáma, oszlopszám=hiedelemszöveg sorszáma. Pl. az első index-kifejezés a huszonkettedik számú hiedelemszövegben egyszer fordul elő.

	21	22	23	24	25	26	27	28	29	30
1	0	1	0	0	0	0	0	0	0	0
2	0	1	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0
9	1	0	0	0	0	0	0	0	0	0

A 2. táblázat az MHAB-nak számítógépes nyelvtechnológiai szempontból fontos statisztikai jellemzőit mutatja.

2. táblázat. MHAB (Magyar Hiedelemszövegek Adatbázis) nyelvtechnológiai statisztikája.

Hiedelemszövegek száma	2.704
Index-kifejezések száma	2.602
Index-kifejezések maximális száma/szöveg	263
Index-kifejezések minimális száma/szöveg	1
Index-kifejezések átlagos száma/szöveg	12
Index-kifejezések átlagos számának szórása	11
Index-kifejezés maximális előfordulási száma/szöveg	16

Szövegek maximális száma ugyanazon index-kifejezéssel	386
Szövegek minimális száma ugyanazon index-kifejezéssel	0

A hiedelemszövegekhez mint objektumokhoz a következő fájlok felelnek meg:

- *.i2k kiterjesztésű fájl, amely tartalmazza a szöveg kulcsszavainak számát, valamint magukat a kulcsszavakat enterrel elválasztva.
- *.htm fájl, az interakciós kereső az ezekre a HTML oldalakra mutató linkeket ad vissza válaszul; az állományok tartalmazzák a hiedelmek sorszámaikat és magukat a hiedelemszövegeket.
- a kisebb fájlokat egy objbase.i2d állomány fogja össze, amely az objektumok számát, illetve a hozzájuk tartozó fájlok listáját tartalmazza kiterjesztés nélkül.

A szövegeknek keresésre alkalmas átalakítását, az indexelést több kis segédprogram végezte, amelyek C++ nyelven íródtak. Az átalakítás kis szakaszokra való bontását az egyes lépések könnyebb ellenőrizhetősége, illetve szükség könnyebb történő módosítása indokolta, de a program futási ideje sem volt elhanyagolható szempont. Ezen műveletek eredménye egy olyan fájl lett, amelyben a hiedelmek sorszámaikat követően a kulcsszavaik találhatóak. A *.htm HTML oldalakat egy harmadik segédprogram hozta létre az eredeti fájlból. A szövegeket tartalmazó htm fájlok neveinek meg kellett egyezniük a hozzájuk tartozó i2k kulcsfájlok neveivel, ami a keresőprogram által megszabott követelmény volt.

Vizsgálatok, eredmények

Tesztkérdések

A tesztkérdések kiválasztása két lépésből áll. Elsőként ki kell jelölni a témát, amelyet keresni kívánunk, majd pontosan meg kell határozni a keresőkérdéseket. Mivel a kereső a hiedelmek között keres, ezért a kérdéseket is ebből a témakörből kellett kiválasztani. Általában a felhasználók strukturálatlan kérdéseket alkalmaznak, ezért részesültek előnyben az egyszerűbb keresőkérdések: egy, két illetve három kifejezésből álló kérdésekkel végeztük a vizsgálatokat. Ahhoz, hogy választ is adjon a kereső, olyan keresőkérdéseket kellett feltenni, amelyek legalább egyik kifejezése előfordul legalább egy objektum kulcsszavai között. A dokumentumok indexkifejezéseiből ezért egy lista készült, amelyben minden indexkifejezés

csak egyszer szerepelhetett (kulcsszavak.txt). A fájl létrehozása az egyszerűség és gyorsaság kedvéért egy PHP script segítségével történt.

A kísérletek menete

Első körben egyszavas keresőkérdéseket adtunk a keresőnek: a kulcsszólistát tartalmazó szövegfájl (kulcsszavak.txt) minden egyes kulcsszava keresőkérdés lett. Mivel előzetesen 2602 db kulcsszót állapítottunk meg, ennek megfelelően 2602 keresést futtattunk le. A kulcsszavak szövegfájlból való kiválasztását, a kereső megfelelő paraméterekkel történő meghívását, valamint az eredmények eltárolását egy PHP script végezte.

A második lépésben két-két kifejezésből álltak a keresőkérdések. A kérdéseket szintén egy PHP script generálta. Ebben az esetben is 2602 keresést végeztünk el. Az első kulcsszó mindig a szövegfájlból vett kulcsszó volt (minden keresésnél a következőt választottam ki, így alakult ki a 2602 keresőkérdés), a második pedig egy véletlenül kiválasztott kulcsszó volt ugyanebből a listából.

A harmadik kör kérdéseit az előzőhöz hasonló módon generáltuk: az első szót a listából választottuk (minden lépésben a következőt), a másik kettő ezek közül véletlen generálással jött létre.

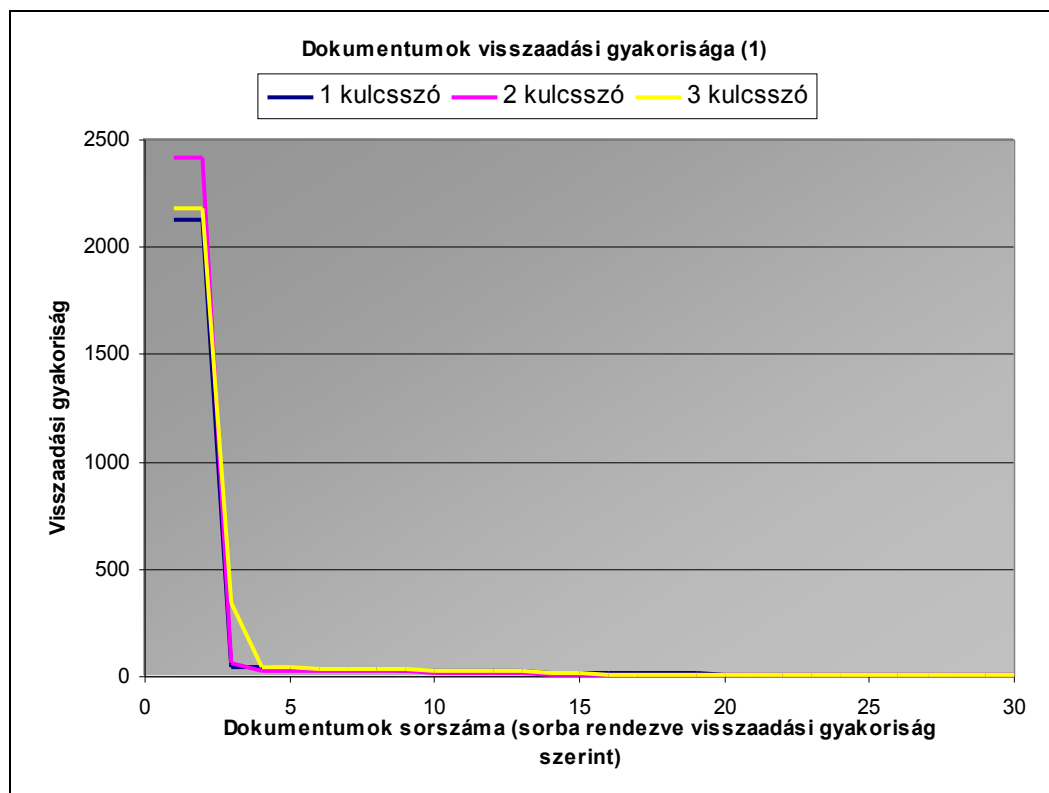
A három lépés azonos keresésszáma azért volt indokolt, mert így az eredményeket könnyebben össze lehet hasonlítani.

A kísérletek menetét a következő pszeudo-kód segítségével foglalhatjuk össze:

```
FOR i=0 TO 2
  DO
    FOR j=1 TO kulcsszavak_szama
      Adjuk be a j. kulcsszót a keresőnek + i db véletlenszerűen
      választott kulcsszót
    ENDFOR
    IF van nagyon gyakori válaszobjektum THEN Töröljük azt az
    objektumbázisból
    WHILE van kiugróan gyakori válasz-objektum
  ENDFOR
```

A kísérletek első szakasza

A 2602 kérdésből rendkívül sok esetben (1 kérdésnél 2130-szor, kettőnél 2422-szer, háromnál 2184-szer) kaptuk vissza ugyanazt a két dokumentumot, amelyek sorszáma 771 és 2691 (ezek alkotnak egy öngerjesztő kört, azaz klasztert). Átlagban néhány száz eset kivételével pedig csak ezt a két dokumentumot kaptuk válaszul. Az ezektől különböző, visszakapott dokumentumok pedig csak nagyon ritkán (mind 50 alatt) fordultak elő a kérdések válaszlistáiban. A három kérdéskör meglehetősen hasonló eredményt adott, ezeket a 6. ábra mutatja.



6. ábra. Az egyes dokumentumok visszaadási gyakorisága a létrehozott objektumbázisban. Látható, hogy egy, két és három kulcsszavas keresőkérdés esetén is két objektum visszaadási gyakorisága feltűnően nagy volt.

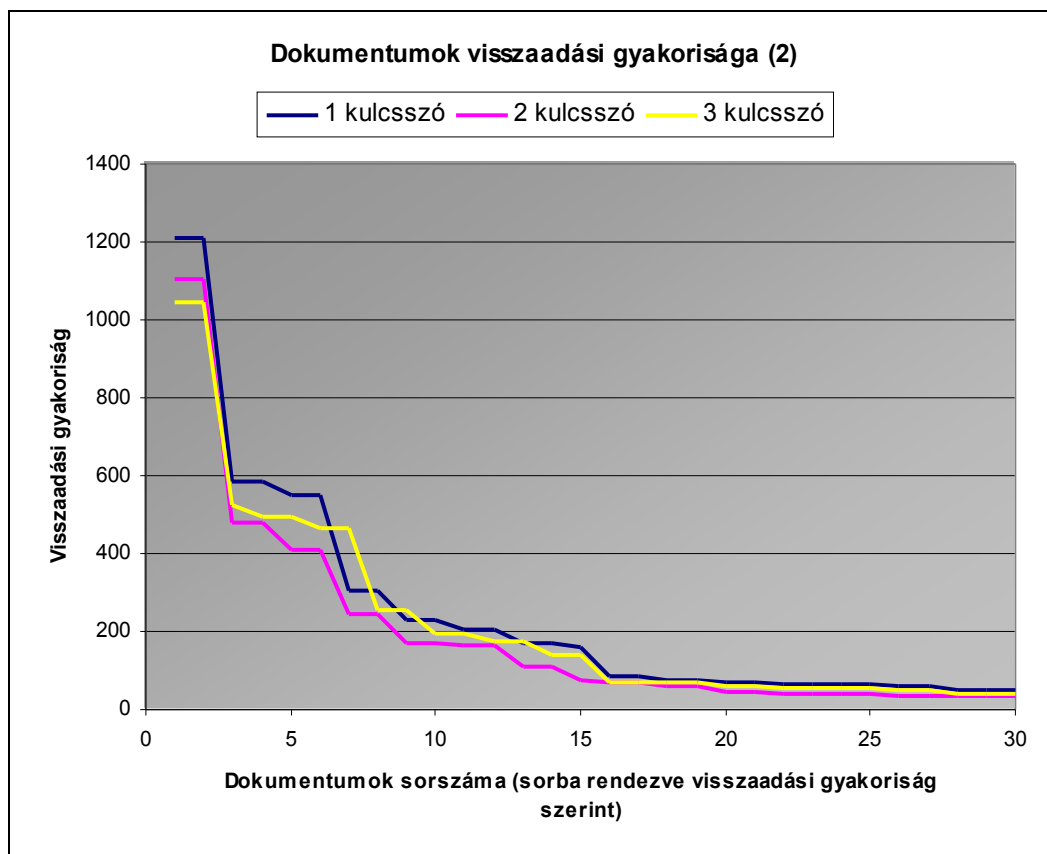
Ezen kísérleti eredmények alapján arra a következtetésre jutottunk, hogy ez a két dokumentum *attraktorként* viselkedik.

A kísérletek további szakaszai

A következő lépésben az MHAB-ból eltávolítottuk a két attraktor-objektumot, majd a kereső újra megkapta az előző sorozat 2602 kérdését.

Az így kapott eredmények is meglehetősen hasonlítottak egy, kettő illetve három kulcsszavas kérdések esetében: ismét volt két objektum (2541 és 2542 sorszámmal), amelyeket

kiemelkedően többször kaptunk vissza válaszul, mint a többi dokumentumot. Ebben az esetben a különbség már nem volt az előzőhöz hasonlóan nagyon látványos, de e két objektumot még mindig több, mint kétszer annyiszor kaptuk vissza mindhárom esetben, mint a következő leggyakoribbat (7. ábra).



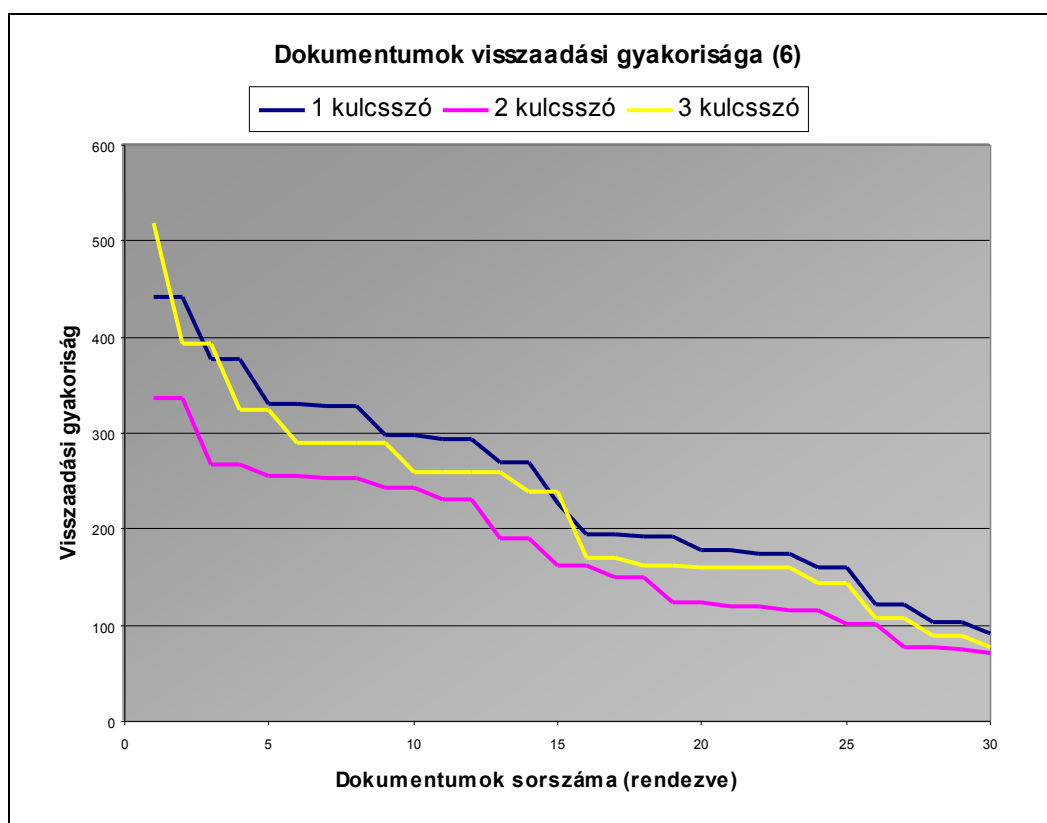
7. ábra. Dokumentumok visszaadási gyakorisága az első attraktorpár objektumbázisból történt kivétele után, egy, két és három kulcsszavas kérdéseket alkalmazva.

Mivel ez a kettő nemcsak a többi dokumentumhoz képest volt gyakori válasz, hanem mindhárom esetben minimum 1000-szer visszakaptuk, arra következtettünk, hogy ezek is elvonzzák az aktivitást, attraktorszerűen viselkednek. Ennek megfelelően eltávolítottuk a 2541 és 2542-es számú objektumokat a hiedelembázisból.

Az újra megismételt kísérletsorozat eredményeképpen szintén kettő darab hiedelem-objektum válaszgyakorisága volt kiemelkedően nagy, ezek az 1846 és 2540-es sorszámot viselik. Ebben a sorozatban is megfigyelhető, valamint a korábbiak során is az volt tapasztalható, hogy az attraktor-objektumok nemcsak gyakran fordulnak elő a válaszokban, hanem pontosan ugyanannyiszor és mindig párban.

A kísérletsorozatok negyedik körében három hiedelem-pár válaszgyakorisága volt meglehetősen magas: 2543-2544, 354-2423, 2538-2556. Ezek visszaadási gyakorisága már nem volt 1000 felett, viszont a többi objektum válaszgyakoriságától jelentősen eltért, magasabb volt (itt is körülbelül kétszerese: ~700, a következők pedig 300).

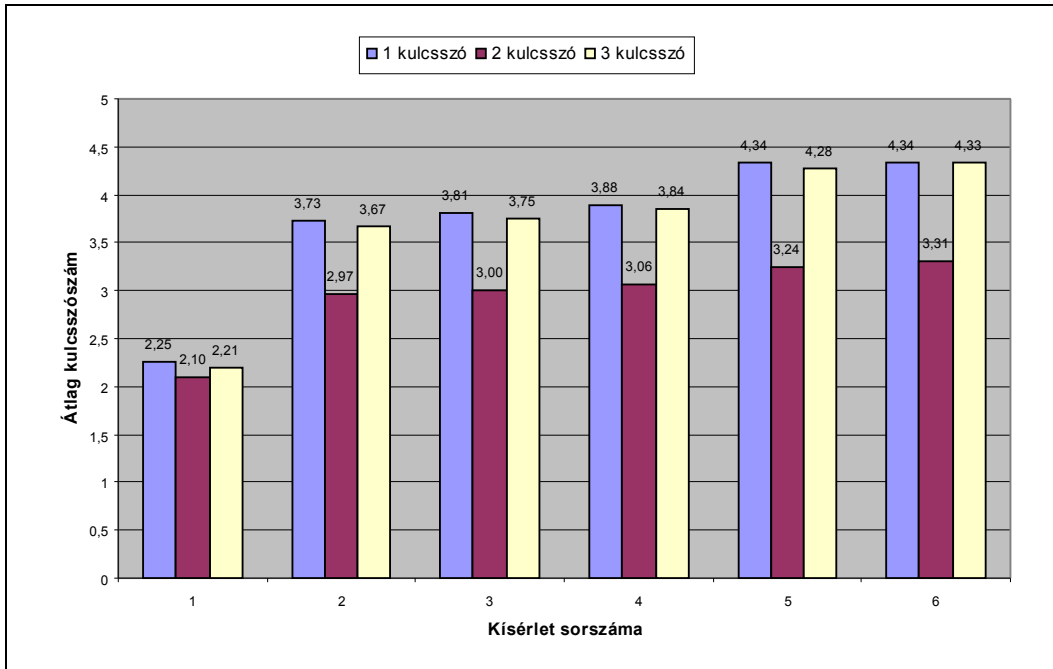
Az ötödik körben is volt még egy kiemelkedő objektum-pár: 140 és 1895 sorszámmal, ezt követően azonban már nem volt olyan objektum(pár), amelynek válaszgyakorisága a többi objektumétól erősen eltért volna (8. ábra), így a kísérletsorozatot ezen a ponton befejeztük.



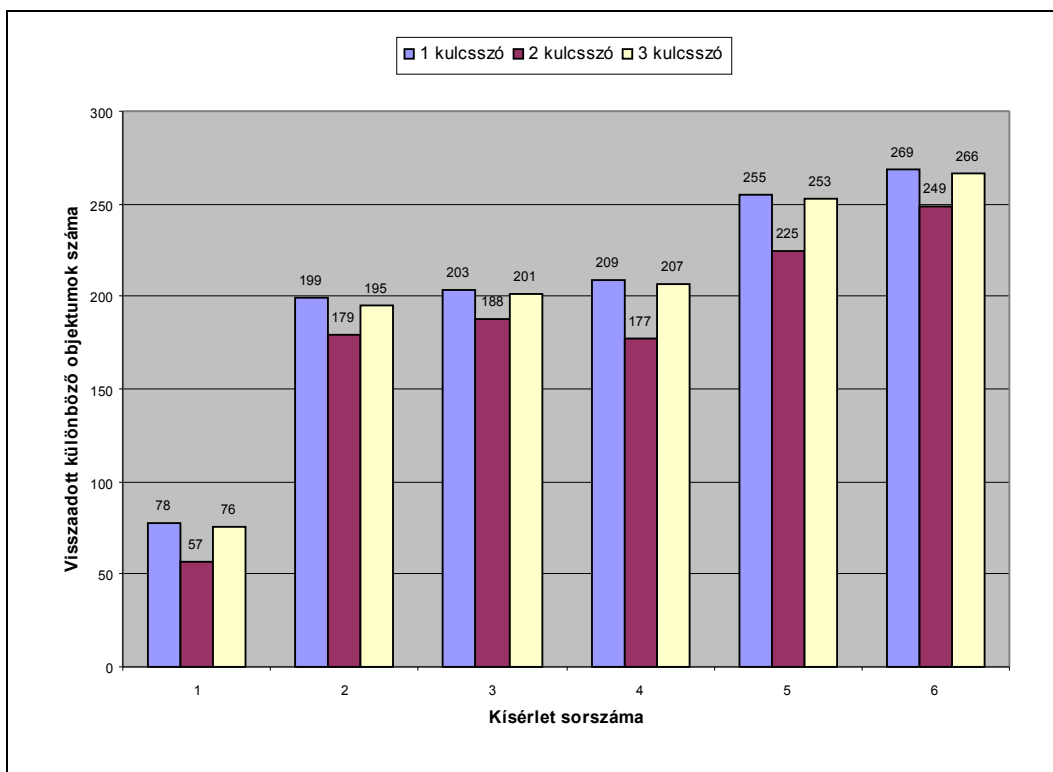
8. ábra. A kísérletsorozat utolsó lépésének eredménye: már nincs kiugróan nagy gyakoriságú dokumentum a három eset egyikében sem.

A válaszok számának vizsgálata

Az 9. ábra szemlélteti, hogy a kísérletek előrehaladtával mindig nő az egy kérdésre adott átlagos válaszszám, az attraktor-dokumentumok egyre kevésbé vonzzák el az aktivitást, egyre jobban „érvényesülnek” más hiedelemszövegek is. Ezt a megállapítást támasztja alá az is, hogy minél több attraktort távolítunk el az objektumbázisból, annál változatosabbak lesznek a válaszok.



9. ábra. Átlagos válaszsámok. Látható, hogy a 6 kísérletsorozat során egyre nő az átlagos válaszsám.



10. ábra. Válaszobjektumok gyakorisága, változatossága. Láthatjuk, hogy minél kevesebb attraktor marad az objektumbázisban, annál több más objektum jut szerephez.

Ilyen módon az attraktorokat úgy képzelhetjük el, mint amik a többi objektum „felett állnak”. Ezt igyekszik látványosabban bemutatni a következő alfejezet.

Az attraktor-körökhöz vezető utak vizsgálata

Láttuk, hogy a kérdésből kiindulva, mint egy idegi hálózat esetén, az aktivitás objektumról objektumra (idegsejtről idegsejtre) terjed, és mindig a legerősebb kapcsolatban lévő másik objektumra terjed át. A folyamat addig folytatódik, amíg egy öngerjesztő körhöz nem érünk, és ekkor a körben lévő objektumokat adjuk vissza válaszul.

Minden kérdés esetén tehát megfigyelhető egy út, amelyen haladva előbb-utóbb elérünk egy olyan objektumhoz, amelyről az aktivitás egy, már az útban szereplő objektumhoz ér vissza (kör).

A keresőprogram forrásának kis átalakítása után a program nemcsak a keresés eredményeit, hanem az azokhoz vezető utakat is visszaadta, így lehetővé vált az attraktorokhoz vezető utak vizsgálata is.

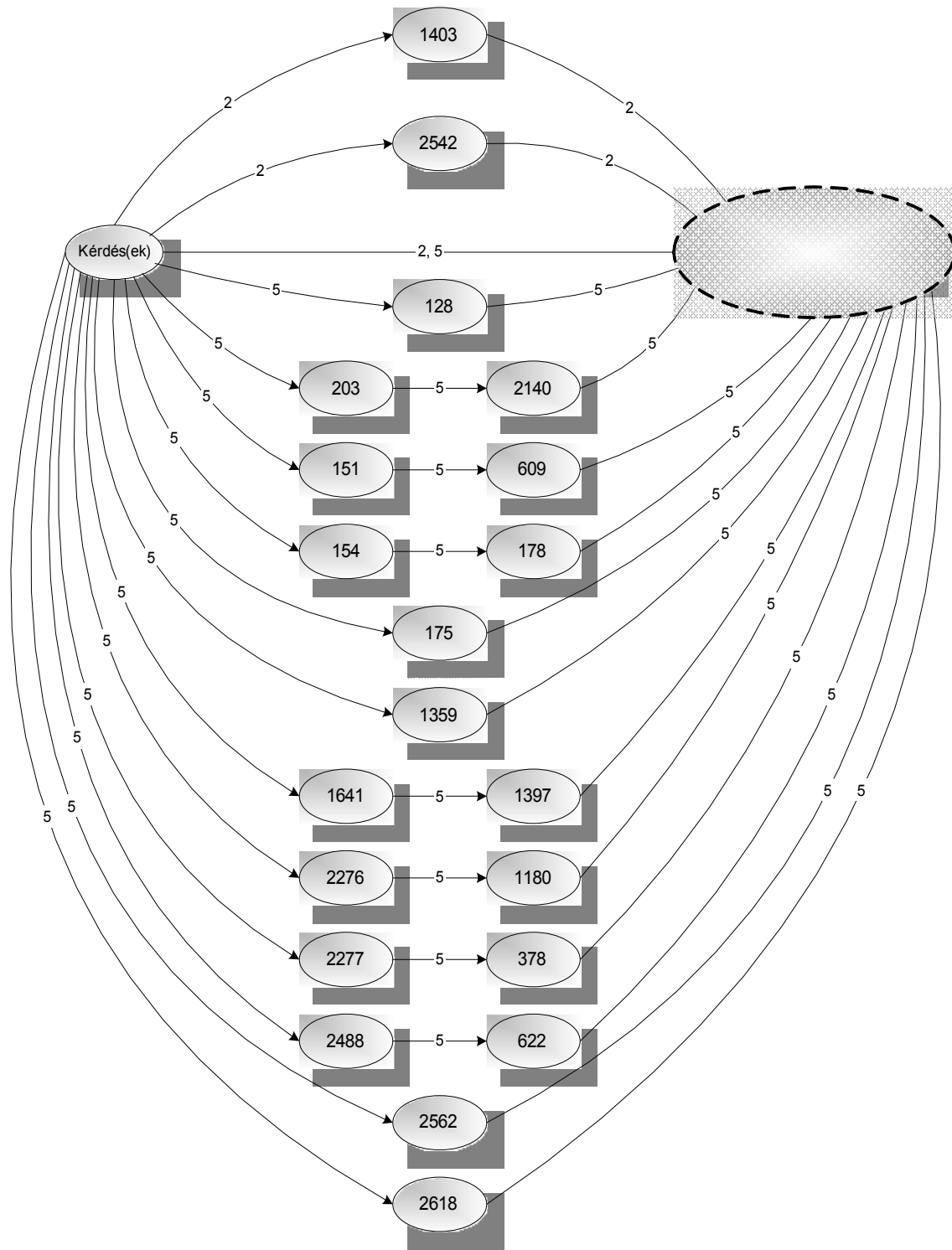
Néhány út bemutatását mutatja a 11. ábra és a . Az első ábrán az eredeti, módosítatlan adatbázison történt keresést követhetjük végig néhány példán keresztül, a második ábrán pedig az első attraktor-pár kivétele utáni adatbázis vizsgálata látható. Az egy- két illetve három kulcsszavak keresőkérdés esetén tapasztalt eredmény-hasonlóság miatt elegendő az egy kulcsszavas kérdésekkel végzett kísérletek bemutatása.

A második ábrán öt kérdés útját követhetjük végig. Ezek a kérdések a következők voltak: hisz, beszél, család, termés, kedd.

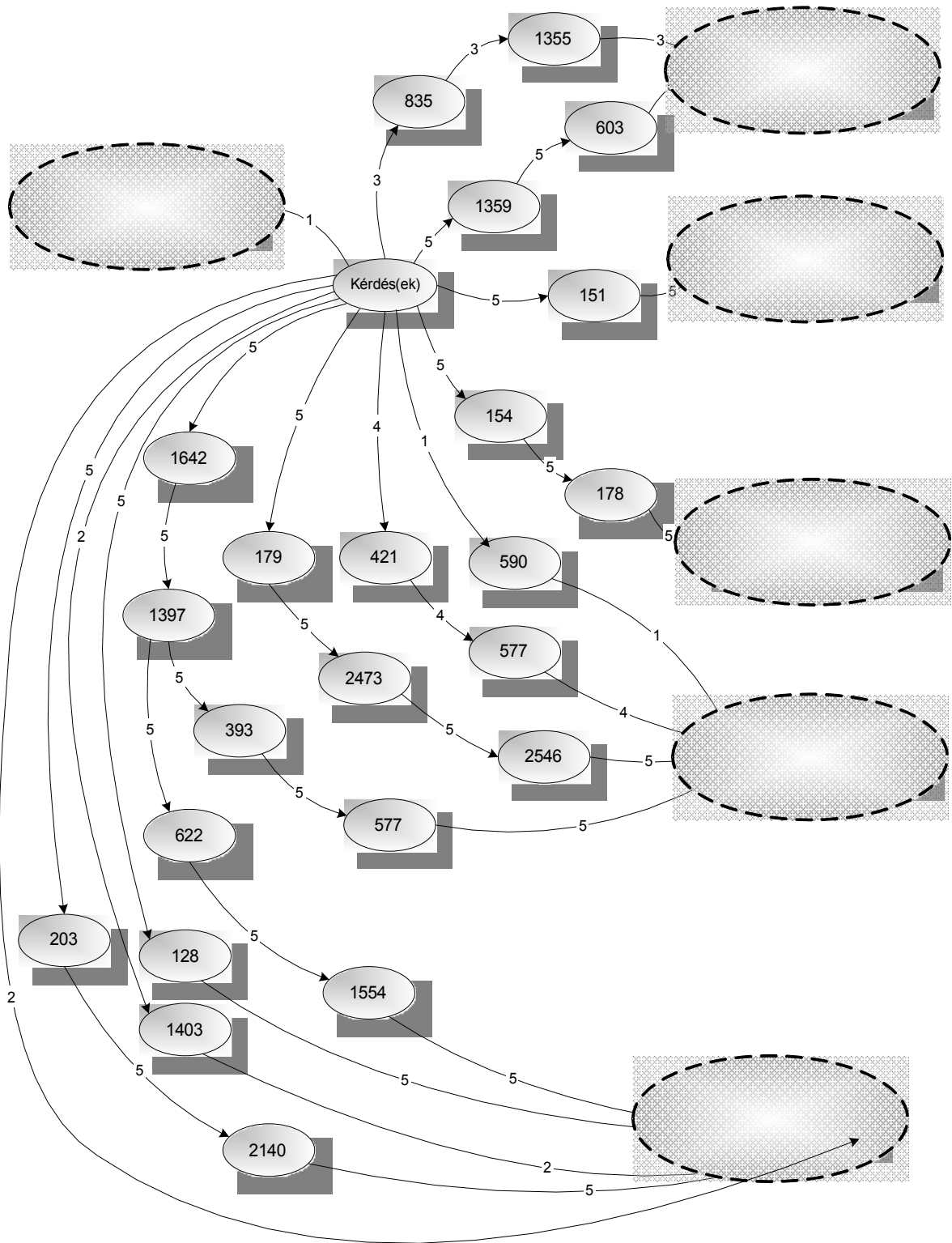
Az első ábrán az átláthatóság kedvéért csak a 2. és 5. kérdésből kiinduló utak láthatók, az 5. kérdésnek sem minden útját láthatjuk ugyanezen okból.

A diagram csomópontjai az „útba eső” dokumentumok számait tartalmazzák. Minden út értelemszerűen a kérdésből indul ki. Az éleken lévő nyilak az aktivitás terjedésének irányát jelölik, a mellettük lévő szám pedig annak a kérdésnek a száma, amelyből kiindulva ezen az úton haladunk (ha több kérdés esetén is fennáll a megfelelő irányú kapcsolat két dokumentum között, akkor az átláthatóság miatt nem az élek többszörözöttek, hanem az élhez tartozó számok vannak felsorolva, vesszővel elválasztva). Az öngerjesztő körök szaggatott vonallal vannak körülhatárolva, ezek objektumai lesznek a válasz-objektumok.

Ha egy csomópontból több él is kiindul, amelyek mellett ugyanaz a szám szerepel, akkor ez azt jelenti, hogy az adott dokumentum több másikkal is „legszorosabb” kapcsolatban van, azaz a dokumentumból kiindulva több maximális súly is van.



11. ábra. Két kérdés(ek)ből kiinduló utak az attraktoros dokumentumbázisban. Az ellipszisekben lévő szám az „útba eső” objektum sorszama, a nyilak a legnagyobb súlyú objektum irányába mutatnak, a nyilakon található számok a keresőkérdés sorszámát jelzik, a szaggatott vonallal körülhatárolt objektumokat kapjuk vissza válaszul.



12. ábra. 5 kérdés útja az eredeti attraktorok nélküli objektumbázisban. A jelölések a 11. ábra jelöléseivel egyeznek meg.

Mindkét ábrán megfigyelhető, hogy a válasz-körök mindig párosak, két objektum alkotja őket (két elemű osztályok). Ebből következik, hogy mindig legalább két választ kapunk

eredményül. Csak akkor kapnánk egy választ, ha egyetlen egy dokumentum lenne nagyon szoros kapcsolatban a kérdéssel, ilyet azonban nem tapasztaltuk a hiedelem adatbázis esetében egyetlen kulcsszavas kérdésnél, de két és három kulcsszó esetén sem.

A 771. és a 2691. objektumok elvonnák az aktivitást, és mivel egymással kölcsönösen a legszorosabb kapcsolatban vannak, az öngerjesztő kör előáll, ezeket kapjuk vissza válaszul. Ha ezt a két objektumot eltávolítjuk, akkor már az ábrázolt öt kérdés esetén is látható, hogy megjelent az aktuális attraktorpár (2541-2542), amelyekhez a legtöbb út vezet (az ábrán 5). Az is megfigyelhető ugyanakkor, hogy a válaszok között olyan dokumentumpárok (1846-2540, 354-2423, 140-1895) is megjelentek, amelyek a későbbi kísérletek során feltűnően gyakori válaszok lesznek, attraktorként fognak az aktuális objektumbázis esetében viselkedni.

Szintén megfigyelhető, hogy a gráfban szinte csak a kiindulópontnál, a kérdés-objektumból van elágazás egy konkrét kérdés esetén. Az, hogy az „úton” később nincs elágazás, annak köszönhető, hogy a maximális súlyok egyszeresek, tehát egy objektummal nincs több, ugyanolyan szoros kapcsolatban lévő másik két (vagy több) objektum (legalábbis a kiválasztandó maximális súlyoknál).

A kulcsszavak és válasz-hiedelmek száma közötti kapcsolat

A visszaadott dokumentumokat vizsgálva kiderült, hogy az első két attraktor-objektumnak van magasan a legtöbb kulcsszava (506 és 402 db). Az őket követő legtöbb kulcsszót tartalmazó objektumban már csak 171 indexkifejezés van, de a száz kulcsszónál többet tartalmazó dokumentumok a tesztek során kivétel nélkül attraktorként viselkedtek (7 ilyen objektum van a hiedelem bázisban). Ez azt mutatja, hogy nemcsak a vektortér modellben, hanem az P^2R modellben is van összefüggés a kulcsszavak száma és az objektum visszaadásának gyakorisága között. A két fő attraktor-hiedelem szövege alább olvasható.

771.: „A gyűjteményben két ízben is megkísérleltem bemutatni azt a társalgási modort, folyamatot, amely a falu egyszerű gyermekei között van, persze a beszéd tárgyát a gyűjtés szempontjából választva meg. A szavak lejegyzésénél lehetőleg kiejtés- és hangzási hűségre törekedtem. Hely: Rátközberencsen a Jóni András háza. Augusztus van. A vājogból rakott spór vígon dúrozsol a pítvarban. Jóni Andrásné a vacsora készítés körül forog, közben-közben élénken felel komja-asszonya szavaira, ki a küszöböt nyergelve tartja őt szóval. Vendég-asszony: Hogy pattog a za tűz kifelé, Komámasszony! Még valami haragos vendégfog jönni. Jóniné: Tán igaz a! Én sose hittem a zijet. Aszt mongyák, ha a zember szemódöke viszket, új embert lát; ha a szeme viszket, sírni fog. Aszt elhiszem, hogy ha szarka csörög a ház tetején, hogy akkor vendég jön. V[endégasszony]: Hát mit mongyék a zember, komámasszony!?! Nekem a multkor a zállam viszketett, asztán csakugyan szakállas vendégem jött, a Ferenc bácsi. J[óniné]: Nem tán!? V.: Úgy-a!... Hát még a zén anynyokkom mijen babonás! Asz monga, hogy nem szabad kimondani, ha a zember szeme ugrál, mert leesik eg csillag. Meg oszt asz se szabad kimondani, ha a zember csillagot lát leesni, mer Isten se mond ki ránk semmit, pedig mindent tud. J.: Igen azt, én is hallottam pujakoromba, hogy ha a csillag esik le a zégrül, meghal valaki, s ha valaki látytya a leeső csillagot, ne mongya senkinek, hanem harapja meg a mutató uját. Ha pedig rámutatott a csillagra, akkor egött harapja ketté. V.: Hátosz mijjén? J.: Hát így tartik! V.: Na de a mán szent igaz, hajja! hogy ahol valamék butor recseg, ott meghal valaki. J.: Aszt is mongyák, hogy ha este a szemetet kiviszik, kiviszik vele a gazdát is. V.: Errül a sepregétésrül jut eszembe, hogy ha akkor sepregetnek, mikor vedég van, akkor aszt a vendeget akarják kiseperni. J.: E mán lehecséges, mikor hogy? V.: Osztán meg, Komámasszony! A minap egy este a szemetet végig sepertem a házfögggyén snem tuttam aludni egész écaka. Azóta úgy vigyázok az esteli sepregetéssel, csap a pór [spór] alá seprem. J.: De a nagyon jól van, Kománé! hogy a zember nem tud aludni akkor se, ha a kanál a fázékba marad mosatlan. Legalább igyekszik a zember felmosogatni. V.: Hát asz hallotta már, hogy ha tanyér leesik a faarül, meghal a házbúl valaki. J.: Hallottam. V.: Hát aszt, hogy akkor is meghal valaki, ha a kuvikk (bagoly) madár a háztetején kuvikkol. J.: A vitte el a drága Jani fijacskám! V.: Mos veszem észre, hogy komámasszonynak két tyukja is ül. J.: Ül ül! De nem tudom mi lesz mán ebbül is. Tugygya a mulkor gabona virágra köt (akkor, mikor a g[abona] virít) hát alig maratt belüle, ami van, azok is nagyon sipógós. Mos meg ezek a gonosz köjkök valami fityülöt hoztak, hat nem tudom

a nagy sipolásra nem felé bele? V.: Bizon pedig még csak fődete szabad ijenkor hozni a házba, mer bele ful. Tugygya, megjártam; a tejet méretlen attam másnak, aszt nem elvitte a tehen hasznát. J.: Én is ugy attam eccer tejet napszálat után, én is úgy jártam. De mos mán mindig teszek bele egy kis sót. V.: Napszálat után nem is jó semmit se kiadni a házból. Én tojást se adok addig másnak ültetni, míg magam nem ültetek. Van is hozzá szerencsém! Hál Istennek! J.: Szabó kománé panaszkodott it a mulkór, hogy nekünk atta a dinynyeje hasznát, mert akkor adott, mikor ű még nem vetett. V.: Hallott ijen csudát, valaki megverte a kis fíjamat, nagyon beteg. Ugy tartják, hogy szemmel. szóval, vagy bamulással az kicsit fiókákat meg lehet verni. (betegek leszek) J.: Hát mér nem monta neki hogy seggibe a szemed, a két első fogad törjön ki benne! V.: Iszen csak láttam vón! J.: Azojan embernek, vagy asszonynak, akinek a szemödöke összeër nem jó a kis puját mutatni, de semmit se, míg kicsi. J.: Szenesvizejje meg, komámasszony! Guggojjon le a pór élébe, oszt ne szójék egy szót se sénkibe. Eggy csupor vízbe olvassék 9 parást ijen módon se nem 9, se nem 8, se nem 7,se nem 1. Ha ezek leülnek, akkor megvan verve a gyermek. Ijen esetbe mossa meg, komám asszony a szenesvízzel a 2 szemét, a két tenyerét, a két talpát; a kis ujjával tegyék a viszbül a gyerek szájába. Ezek után türújje meg a zinge fonákjával, és a zenes vizet oncsa a zajtó sarkára. E biztosan segít. De mondom ne szójék senkihe, míg ezt csinájja. V.: Asz mongyák a zis használ, ha a zember megköpködi a megnézett, vagy megbámutt puját, kis libát, vagy csirkét. J.: Használ, használ Komám-asszony. A zisnagyon jó, ha annak a hajával füstolik meg, aki megverte. J.: Hát csak szenes vizejje meg. Vagy ha megijesztette vón valaki, akkor meg öncsek ölmet. A zén emberem is ugy ijgett meg eccer, vagy 10 esztendős lehetett. Mindig beszéli, hogy anynyokom egy sejem kendőt terített rá, persze ütöt lefektette. Átán szivére a sejem kendő fölé egy tál vizet tett. A tálát leburította egy rostával s a rostán keresztül megolvastott ölmet öntött a vízbe. Osztán hajjék csudát e za vèn Szabóné ömlött ki, ojan vót, mint egy boszorkán. V.: Oszt csak ugyan attal ijgett meg. J.: Attul bizon, komám asszony! Mikor oszt ez kijött a zöntéssel, a zén gazdámnak minden baja elmúlt. Itt a vendégasszony elsiet, hogy gyermekét kikuruzsolja. Én Jóninét megkértem, hogy mongyék nekem egy pár babonát, mert senki sem tugygya úgy mint ű. Ekkor jegyeztem el a következőket:”

2691: „Hát vannak bízsbányosok. Ha nem velem esett vaóna még, magam sè hinném, – kezdte a szót Lörincz János. Hát úgy vaót, hogy árva gyerek vaotam. 10-11 esztendeős koromba Csèrmelybe szakattam hidaónak. Otteé szagátam hét esztendejig. Mikor haza kerütem, asszámára (szinte, azt lehet mondani, már-már) legénsorba vaotam. A biraóho szegeöttem kocsisnak. Szombat estve azt a többi legén, hogy gyereónk a lányasházba (fonóház). Én szabadkozottam, hon nem ösmerék senkit, – nem ménék. De oszt annyit beszeétek, hogy meégis csak elméntem. Ahol lételepszéonk, a lányosház gazdasszony behoz él litèrès-furma üveg pályinkát még ek kis porciaós üvegét, oszt minden legént sorra kénál a pályinkabaó, csak engém nem. Ahogy rámkerü a sor aszongya hogy, hohó fíjam, tè új embèr vagy, nekèd meézeset hozok. Hogy oszt mindel legény itt, az asszony kifordol a házbaó s nemsokára telyi hozza a porciaóst. De bennè a pályinka, mereó èegy hab. Kéná vélè, de én sèhossè akartam meffognyi. De aggyit ereósködött rajtam, hogy meégis elvèttem. Ráköszöntöm, aszt s megiszom. Vaót annak mindèn èzi, csak pályinka nem. Im hagy ki nem gyött beleölem. Vaót annak az asszonnak ès semmi hitvány lánya, olyan akar èt tolyúsèpreó. Rá sè neézètt el legèn sè. Én még attaó a pereteó kezdve ögy belèbolondoltam, hogy meé akkor is elöttem vaót, ha behútam a szèmèm, még ha száz méröfeódre vaotam teólè is. Nem vaót nekem többet sè èvelèm sè nappalom. Hijjába mèntem estve haza a letereósebb dologbaó. Bekaptam az ételèm, vettem a jaószág eleè s mèntem a lányho. Mindeég èfè vetètt haza teólè. A sok èccakázás úgy mészszürt, hogy allyig vaót jártányi ereóm. Pirongatott a gazdám is, hogy aszongya, meé tészèm sèmmire magam? Èleég èccèr egy hètèn szombatestve elmènyyi a szereteómhó! De nem ért as semmit. Pegy magam is belàttam, hogy nem üt jaóra ki as soram. Még is fogattam hit alatt, úgy magamba, mindèn régvèl: hon no nem ménék hozzá letalabb ègy hétyig. De mikor elgyött az estve, meégènt csak szállottam vaóna hozzá, ha szárnyom lett vaóna. – Nappal mindeè eszèmèn vaotam. El-el teépelöttem, hogy mitèveó légyek? Èccèr oszt arra gyöttem, hogy bemènek Egerbe, a cselèdszèrzeóbe e elszegeódk valahova messzire, ahonneèt, ha akarok sè tudok mindèn estve hazabolondolnyi. – Úgy is lett. Bemèntem Egerbe pèntèkèn régvèl s meé an nap deébe kiszzegeóttettek Kerecsèndre a jágèerho. Kocsis léttem ottè is. A gazdám jaó embèrnek mutatkozott, kijárt ételèm italom böcseóletèsen. De hijjába vaót mindèn: nem tuttam èn sè ènnyi, sè innya, sè alunnyi. Csak abba vaotam, hogy még kell bolondulnyi. Betegnek tèttem magam. No igaz, hogy ev vaót a letkönynebb, mert nem vaót egyebem: csak a csontom, még a beóróm. Oszt èsz szaó anyni, mint száz: haomannap estve maá itthon vaotam Maónosbèbe. Itthon bijony a Krisztussa nè légyeék annak a huncut asszonnak! Mèr azután is ott kèllètt nekèm lényyi mindèn estve a lányná ná, meèt ha vasvilla esett vaóna is. Oszt hogy napró-napra lèjjebb estem, panasoltam a sorom fínek-fának. Eccer aszt ègygyik cimborám, hogy Bèra Lajos tud forditanyi a bajomon, – neki szaóllyak. Úgy is cselekèttem. Elmondtam neki az egeész èstaóriát. Eò sozt mingya félígèrkezett, hogy ha jaórtartom borral, segít rajtam. Én még aszontam neki, hogy fizeték, amennyit megbír innya. Nem bánom, ha jaz esztendei berèm utánna megyèn is. Hogy aszt így meéègyèselteünk aszongya a Bèra Lajos – No, szombateste ereggy el a lyánho. Beszeéess vélè, ahogy másszor szoktal. De nè járj nagyon soká, mer èn az úccába vállak. Mikor kikiseér a lyány, ahogy ki leépsz a pitarajtaón, nyú fél az ereszbe s húzz ki beleólè kész szá zsófut, úgy, hogy aly lyány eszre nè végèyè. Az ègygyik szálat bocsádd lè magad mellett mingyá, a másikat még, mikor az udvar közepire eèrsz, visszakèzbeó hagyítod a kert felè, oszt gyere ègyènést hozzám. Én üt tèttem mindènt, ahogy parancsóta. Azután mènteónk nyútvá asz zsidáóho s úgy belaktaónk borral, hogy èn világom sè tuttam. Másnap régvèl hogy félèrzèk, eszèmbè jut az ab bizonyos lyány. Hát olyan úttallat fogott el, minthacsak pondreés kutyát láttam vaóna. S azaóta sè tudok ránèznyyi sèmmifèle csomotájára.”

Több kulcsszava a 771. dokumentumnak van, és az attraktor-körhöz ezen az objektumon keresztül jutunk el többször. A 771. objektum eredményként való visszaadása nem teljesen hibás eredmény. A dokumentumot elolvasva ugyanis nem egy, hanem több hiedelemmel találkozunk, amelyek egy konkrét történeten keresztül kerülnek megemlítésre. Így több hiedelem között nagyon szoros kapcsolat van, olyannyira, hogy ha a keresőkérdésünk egyikkel kapcsolatos, akkor a többit is visszakapjuk ezen a nagy dokumentumon keresztül. Lehet, hogy ezen hiedelmek között akkor is lenne kapcsolat, ha külön objektumok lennének. E nagy dokumentummal lényegében „egymáshoz láncoltunk” néhány hiedelmet, a hiedelem-

objektumok közötti kapcsolatok vizsgálata, a visszaadandó hiedelmek kiválasztása viszont a keresőprogram feladata kell, hogy legyen.

A hiedelemszövegeket tovább vizsgálva megállapítható, hogy a következő néhány hiedelem (772, 773, stb.) kapcsolatban áll ezzel a nagy dokumentummal, de csak olyan értelemben, hogy ugyanott gyűjtötték be a szövegeket.

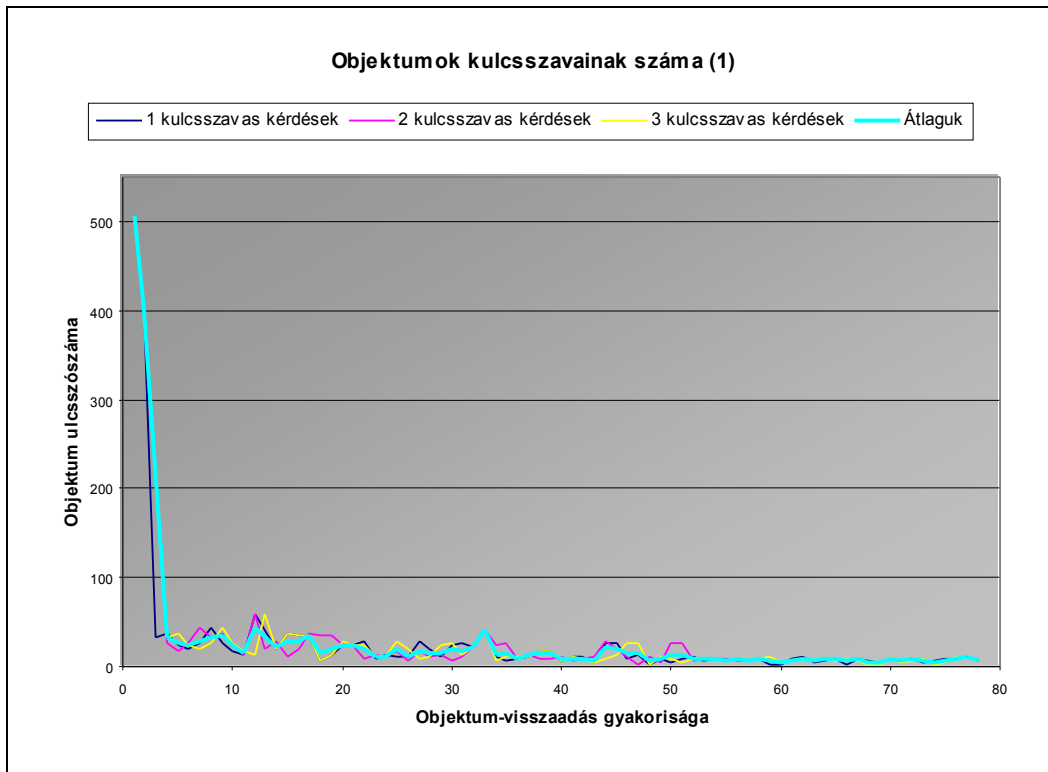
Valószínű, hogy eme terjedelmes objektum létrejötte tehát a dokumentumok kialakításakor elkövetett „hiba” eredménye, mivel egy dokumentumnak egy hiedelmet kellene tartalmaznia, és itt ez nem valósult meg. Az, hogy a dokumentumot a történet alapján választották ki, hibás lépésnek bizonyult, ami a keresés során feltűnő módon jelentkezett: a 2691-es számú hiedellemmel együtt „domináns” hiedellemmé vált.

Az említett objektummal párban előforduló 2691. számú objektum esetén is hasonló a helyzet. Ebben az esetben is egy történettel találkozunk, ugyanakkor csak egyetlen hiedelemről van szó benne, ellentétben az előzővel.

A probléma megoldása itt is, akárcsak az előző esetben, manuálisan oldható csak meg. Ki kell szűrni azokat a mondatokat, szövegrészeket, amelyek csak „körítésként” szolgálnak, nem tartoznak szorosan egy konkrét hiedelemhez, valamint szét kell választani adott esetben a több hiedelmet tartalmazó dokumentumokat. Azonban ehhez át kell nézni az összes szöveget, mert nem csak ebben a kettőben fordulhatnak elő ilyen hibák, ami rendkívül sok időt venne igénybe.

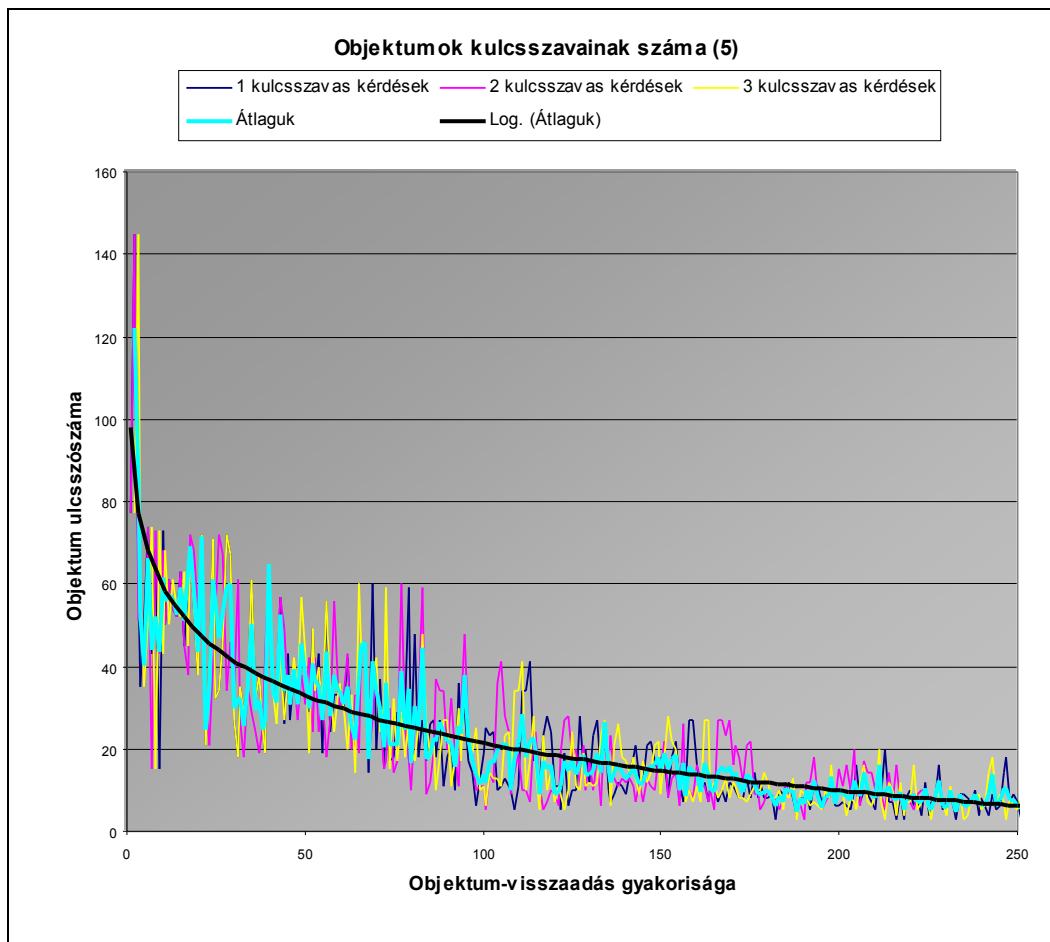
Ha a két attraktor közül az egyiket eltávolítjuk, nagy változás nem következik be a keresés során, tapasztalatom szerint a megmaradt objektum továbbra is attraktor marad, csak egy másik objektummal párban alkotnak mindig egy öngerjesztő kört. A kísérletek későbbi szakaszában előforduló dokumentumpárok esetén hasonló a helyzet, ezért is távolítottuk el őket mindig párban.

A kapott eredményekből megállapítható, hogy nagyobb valószínűséggel kapjuk vissza azokat a hiedelem-szövegeket, amelyek több kulcsszó tartalmaznak. Azt is láthatjuk azonban, hogy a visszaadás valószínűsége nem csak az indexkifejezések számától függ, ekkor a függvény monoton lenne.



13. ábra. A visszaadott dokumentumok kulcsszavainak száma. Kisebb sorszámmal a gyakrabban visszakapott dokumentumok szerepelnek. Jól látható az eredeti adatbázis két attraktorának kiemelkedő kulcsszómennyisége.

A kiemelkedő dokumentumpár eltávolítását követően minden kísérletsorozat esetén ugyanazt tapasztaltuk: Több kulcsszó esetén nagyobb az esély, hogy egy dokumentumot válaszként visszakapjunk. Erre mutat példát a 14. ábra, amelyben a 250 leggyakoribb válaszobjektum kulcsszavainak száma látható egy, két és három kulcsszavas keresőkérdés esetén. Az ábrán feltüntettem a három kísérletsorozatból származó kulcsszóátlagokat, és egy logaritmikus trendvonalat, amellyel a függvény viszonylag jól közelíthető.



14. ábra. A visszaadott dokumentumok kulcsszavainak száma a kísérletsorozatok 5. körében. Itt is látható, hogy a több kulcsszót tartalmazó dokumentumok gyakrabban lesznek a válasz-objektumok között.

Az I^2R modellel végzett osztályozás eredménye

A hiedelemszövegeken végzett vizsgálataink eredményeiből megfigyelhető, hogy az öngerjesztő körök párokból állnak, mi több, szinte kivétel nélkül állandóak ezek a párok, nem csak az egyes lépésekben kiemelt attraktor-párok esetében. Ez a hiedelem az MHAB egyik jellegzetes tulajdonsága.

Ezek az objektumok szoros kapcsolatban állnak egymással, sok közös kulcsszavuk van. Az objektumokhoz tartozó hiedelmek tehát valószínűleg hasonlóak, vagy akár ugyanazt a hiedelmet tartalmazzák, ezért sorolja az I^2R modell őket egy osztályba. Ezt a néhány, közelebről is megvizsgált objektum-pár is alátámasztja. Az 1982. és 2149. szövegek például a szemmel veréssel kapcsolatos nagyon hasonló hiedelmeket tartalmaznak, vagy a 322. és 1371. dokumentumok is rokonságban állnak, mindegyik a kihullott foggal kapcsolatos.

322: „Ha a kitört vagy a kiesett fogunk helyett új erős fogat akarunk magunknak a következő szavak kíséretébe kell azt a hátunk mögé dobni:

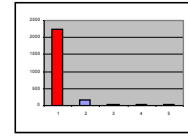
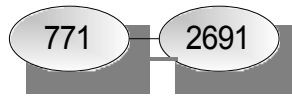
*Cine cine eger adj fogat
Adjá nékem vas fogat
Én is adok csont fogat.
Aki ezt megteszi, annak a kitört foga helyébe jó erős fog fog majd nôni.*

1371: „Ha gyermeknek kihúzzák a fogát, akkor ezen igék mondása közben: <Egér, eger adj vas fogat, én is adok csont fogat>, dobja el a háta mögé, de utána ne nézzen, ne keresse, mert akkor rossz foga nô.”

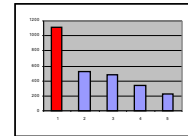
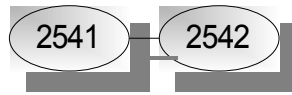
A megtalált motívum, a versecske, szinte szó szerint ugyanaz, de a hiedelmek egy kicsit különbözőek. Ez utalhat arra, hogy a két hiedelem alapja egy és ugyanaz, csak az idők során tájegységtől függően átalakult. Láthatjuk tehát, hogy a hiedelemszövegek páronként összekapcsolhatók, két dokumentumból álló osztályokat eredményezett a vizsgálatsorozat. A kísérletek során az is bebizonyosodott, hogy egyes párok (az attraktor-párok) bizonyos értelemben felette állnak a többieknek, amit úgy is értelmezhetünk, hogy egy hierarchia magasabb szintjén helyezkednek el. Ily módon ezen magyar hiedelemszövegek *hiedelem-dokumentumok hierarchikus struktúrájaként* modellezhetők, amit a mutat. Vizsgálataink eredményeképpen hat ilyen szintet tudunk egyértelműen megkülönböztetni. A struktúra legfelső szintjén a legelső körben attraktorként viselkedő hiedelempár áll, a második szinten a második lépésben feltűnően nagy válaszgyakoriságú pár van, stb. Az egyes hierarchiaszinteken belül is definiálhatunk bizonyos sorrendet a párok között (kivéve értelemszerűen az egy párt tartalmazó szinteket). A szinten előrébb (az ábrán balrább) lehet helyezni azokat a párokat, amelyek visszaadási gyakorisága nagyobb, így alakult ki az ábrán látható 4. szint attraktor-párjainak sorrendje. Az legalsó (6.) szinten ilyen sorrendezés csak megközelítőleg végezhető el, mivel az első néhány párt kivéve a válaszgyakoriság sorrendje nem egyezik meg mindhárom (egy, két és három kulcsszavas) kísérletsorozat esetében.

Hierarchiaszintek

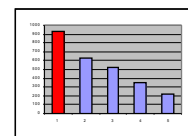
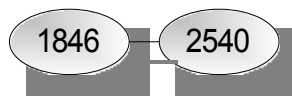
1. szint



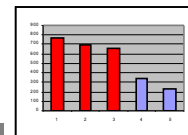
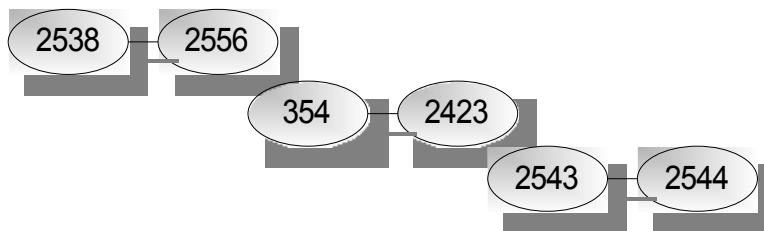
2. szint



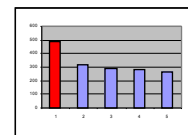
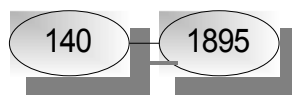
3. szint



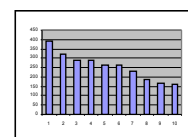
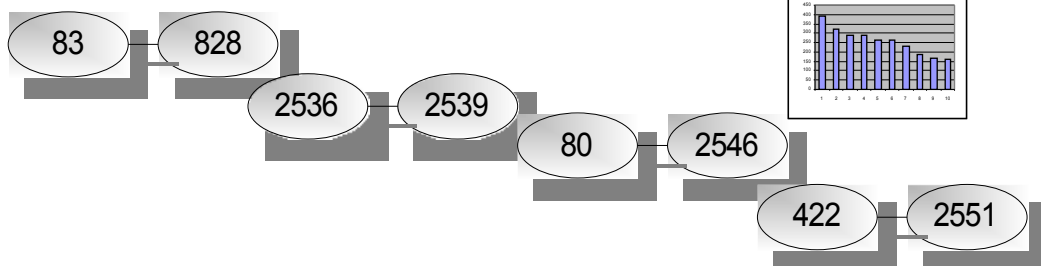
4. szint



5. szint



6. szint



...

15. ábra. Magyar hiedelemszövegek hierarchikus szerkezete. Az egyes szinteken az aktuális kísérletsorozat attraktorai láthatók (értelemszerűen kivéve az utolsó szintet), vonallal összekötve a párokat figyelhetjük meg, a kis diagramok pedig a visszaadási gyakoriság függvény kis részei, kiemelve rajtuk az adott szinten attraktorként viselkedő dokumentumok válaszgyakoriságát.

Összefoglalás

Elvégeztük az MHAB Magyar Hiedelemszövegek Adatbázisnak egy számítógépes nyelvtechnológiai vizsgálatát, nevezetesen a hiedelemszövegek osztályozását az I^2R (Interaction Information Retrieval) eljárás segítségével. Az I^2R módszer bemutatását követően leírtuk azokat a számítógépes nyelvtechnológiai műveleteket, amelyeknek az MHAB adatbázist alávetettük, és amelyeknek során megkaptuk a vizsgálataink elvégzéséhez szükséges kifejezés-dokumentum mátrixot. Az attraktor-hiedelmek általában sok index-kifejezést tartalmaznak, ez megkönnyítheti felismerésüket, azonban az attraktor-jelenség kialakulásában szerepet játszik az index-kifejezéseknek az adatbázisban való gyakorisága is. Amennyiben nincs attraktor-hiedelem az adatbázisban, tehát olyan, amelyik feltűnően sokat szerepel a válaszok között, akkor érvényesül az I^2R modell osztályozó képessége, aminek segítségével akár motívumokat vagy nagyon hasonló hiedelmeket azonosíthatunk. Legfontosabb megállapításunk az, hogy **a magyar hiedelemszövegek hierarchikus struktúrába szerveződnek**. Az azonos hierarchia-szinten található hiedelmek nagyon hasonlóak, ez akár egyfajta, szemantikai redundanciának is tekinthető, és további vizsgálat tárgyát képezheti.

Felhasznált szakirodalom

- Chomsky, N. (1956). Three Models for the Description of Language. *IRE Transactions on Information Theory*, vol. 2-3, September, pp: 113-124.
- Chomsky, N. (1959). On Certain Formal Properties of Grammars. *Information Control*, vol. 2, June, pp: 137-167.
- Csajághy, György (2000). A magyar népzene ősrétegeről és néhány ősi hangszeréről. In: Kőrösi Csoma Sándor és a magyarság keleti eredete. Sepsiszentgyörgy: Trisedes Press Rt. (pp. 338-386)
- Darányi, S., Faculty of Journalism, Library and Information Science (2001), Magyar hiedelemszövegek tartalmi térképezése, Proceedings of Conference on Cultural innovation and information retrieval: The digitalization and automated content exploration of Finno-Ugric cultural heritage (pp. 68-85.)
- Darányi, S., Faculty of Journalism, Library and Information Science (2001), Magyar hiedelemszövegek tartalmi térképezése, Proceedings of Conference on Cultural innovation and information retrieval: The digitalization and automated content exploration of Finno-Ugric cultural heritage (pp. 68-85.)
- Dominich, S. (1994). Interaction Information Retrieval. *Journal of Documentation*. 50(3); (pp. 197-212)
- Dominich, S. (2002). Connectionist interaction information retrieval. *Information Processing and Management*. Elsevier Science Ltd.
- Ferenczi, István (2000). Délibábos ábránd készítette-e Kőrösi Csoma Sándort Belső-Ázsiába? In: Kőrösi Csoma Sándor és a magyarság keleti eredete. Sepsiszentgyörgy: Trisedes Press Rt. (pp. 115-162)
- Hays, D. G. (1966, ed.) *Readings in Automatic Language Processing*, American Elsevier, New York.

-
- Kay, M., Ziehe, T. (1965). Natural Language in Computer Form. In: Hays. D. G. (ed.) *Readings in Automatic Language Processing*, American Elsevier, New York, 1966. pp: 33-50.
- László, J. (2002). Dinamikus weboldalak, CGI programozás Windows és Linux rendszereken. Budapest: ComputerBooks
- László, J. (2002). Dinamikus weboldalak, CGI programozás Windows és Linux rendszereken. Budapest: ComputerBooks
- Prószéky-Kiss G. (1999). Számítógéppel emberi nyelven.
- Salton, G. (1965). Automatic Phrase Matching. In: Hays. D. G. (ed.) *Readings in Automatic Language Processing*, American Elsevier, New York, 1966. pp:169-188.
- Salton, G., and McGill, (1983). *Introduction to modern information retrieval*. McGraw Hill, New York.
- Van Rijsbergen, C.J. (1979). *Information Retrieval*. Butterworths, London.
- Van Rijsbergen, C.J. (1987). Információ visszakeresés. Budapest: Múzsák Közművelődési Kiadó.
- Yngve, V.H. (1957). A Framework for Syntactic Translation. *Mechanical Translation*, vol. 4, no. 3, pp: 59-65.