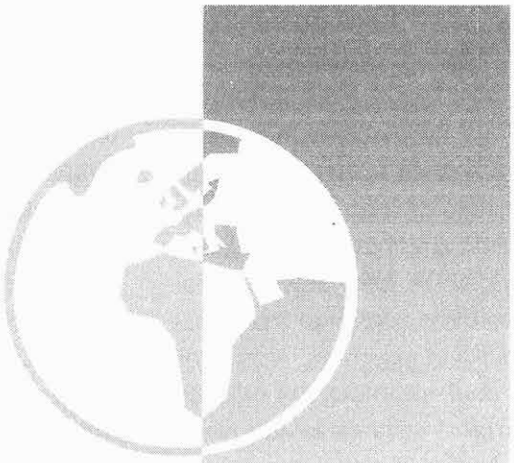

KITEKINTÉS

Kísérlet az internetes keresők nyelvi lehetőségeinek felmérésére

TÓTH Erzsébet

1. A tesztelés célja és motivációja

Írásomban egy olyan tesztelés eredményeiről szeretnék beszámolni, amely *Judit Bar Ilan* és *Tatyana Gutman* 2002 novemberében végzett vizsgálatához kapcsolódik. Ők elsődlegesen morfológiai szempontból elemezték az orosz, francia, magyar és a héber nyelvű lekérdezések értelmezését az angol nyelvű keresőknél és a helyi fejlesztésű keresőknél. Kutatásuk során mindketten arra koncentráltak, hogy a keresők mennyire képesek figyelembe venni a kiválasztott nyelvek egyedi sajátosságait és milyen hatékonyan válaszolnak a nem angol nyelvű lekérdezésekre. Eredményeikben a keresők találatainak a számát közölték a lefuttatott

lekérdezésekkel együtt.[1] Ezen szakirodalmi előzmény alapján végeztem egy tesztelést 2005 júliusa és szeptembere között. Vizsgálatommal az volt a célom, hogy feltérképezzem az angol és a magyar nyelvű keresők nyelvi lehetőségeit a különböző megadott lekérdezésekre. Az angol nyelvű keresőkön angol és magyar nyelvű lekérdezeseket egyaránt lefuttattam, a magyar nyelvű keresőkön azonban csak magyar nyelvű kereséseket végeztem, mert azok főként magyar nyelvű weboldalakat indexelnek, kivéve az Origo-Vizsla és az Altavizsla szolgáltatásokat. Arra a kérdésre kerestem a választ, hogy ezek a keresőszolgáltatások tettek-e bármilyen erőfeszítést a lekérdezések pontosabb nyelvi értelmezése érdekében, vagy továbbra is megoldásra váró feladatok maradnak a számukra. A tesztelés eredményei teljes mértékben alátámasztották az említett szerzők korábbi megállapítását, mely szerint: a magyar nyelvű keresők csak bizonyos mértékig veszik figyelembe a magyar nyelv egyedi sajátosságait. Ezzel szemben az angol nyelvű keresők komoly hiányosságokat mutatnak ezen a téren, hiszen főként pontos szóalakra és kifejezésre keresnek és egyszerű minta-megfeleltetést végeznek.

2. A tesztelés módszere

A tesztelés mindkét nyelv esetében a megfelelő szavak, kifejezések felkutatásával kezdődött. Körültekintően választottam ki a tesztelésre szánt keresőkifejezések halmazát. Ezután próba lekérdezéseket végeztem mindegyik keresőn, annak ellenőrzésére, hogy a kiválasztott keresőszavak egyértelműen kifejezik-e a vizsgált nyelvi problémát és, hogy megfelelnek-e a tesztelés célkitűzéseinek. A keresőszavak kiválasztásánál alapvetően arra törekedtem, hogy olyan kifejezéseket válasszak, amelyek mindkét nyelv speciális nehézségeit tükrözik. Döntésemben az előzetes megfigyeléseimre támaszkodtam. Természetesen a tulajdonneveket figyelmen kívül hagytam, hiszen azok nem fordulnak elő más morfológiai alakban a szabadszövegben. A vizsgálat megkezdése előtt áttekintettem mindegyik keresőszolgáltatás tájékoztató anyagát, hogy pontosan tájékozódjam mindazokról a fontos jellemzőkről, sajátosságokról, amelyek relevánsak lehetnek a tesztelés szempontjából az adott keresőre nézve. A lefuttatott lekérdezésekre az első 100 találatot vizsgáltam meg. Egyedüli kivételt képezett a csonkolás vizsgálata a magyar nyelvben, ahol lekérdezésként egy szűkebb találathalmazt eredményező kifejezést adtam meg, a *májmételyt* (=parazita, féreg). Csonkolásnál rögzítettem az egyes keresők által lekérdezett találatok számát. Elsődlegesen a találatok leírását vettem alapul, de ahol indokolt volt azok tartalmába is belenéztem. A kereséseket 2005 júliusa és szeptembere között hajtottam végre. Nem értékeltem a találatok relevanciáját, mert a keresők nyelvi képességeinek a felmérése állt a tesztelés középpontjában.

A tesztelésbe összesen három angol nyelvű keresőt vontam be: a Google-t, az AltaVista-t és az AlltheWeb-et [2,3,4]. A választásom azért esett erre a három angol nyelvű keresőre, mert ezek lehetővé teszik a magyar nyelvű weboldalakra történő keresést. Ezenkívül a Google kereső magyar nyelvű lekérdezési felületet is biztosít a felhasználók számára. Összesen öt jelentősebb magyar nyelvű keresőt teszteltem, amelyek a következők voltak: Heuréka, Origo-Vizsla, Kurzor, Góliát és az Altavizsla. Ezek közül egyedül az Altavizsla kereső esetében nem találtam semmiféle tájékoztató segédletet. A keresők kiválasztásánál fontos kritérium volt számomra, hogy azok megbízhatóan működjenek, vagyis rövid válaszidő után szolgáltassanak megfelelő mennyiségű találatot a lekérdezésekre.

3. A kiválasztott vizsgálati szempontok ismertetése

A teszteléshez Bar Ilan és Gutman közös tanulmányában található szempontokat vettem alapul, amelyek az alábbiak voltak: morfológiai elemzés

(stemming), stopszavak és ékezetek kezelése, csonkolás, szinonimákra történő keresés. Morfológiai elemzés (stemming) alatt azt értem, amikor a kereső megtalálja egy lekérdezés toldalékolt alakjait.

A morfológiai elemzésnél vizsgáltam tehát, hogy az adott kereső megtalálja-e egy keresőkifejezés többes számú alakját vagy sem (vagyis keres-e a lekérdezett kifejezés többes számú alakjára). Lekérdezéseim a következők voltak: *dog-dogs, ház-házak, kocsi-kocsik, kutya-kutyák.*

Az első két magyar példánál megfigyelhető, hogy a tővégi magánhangzó nem változik a többes számú alakban az egyes számúhoz képest, míg a kutya-kutyák esetében igen. Választottam egy fosztóképzővel ellátott főnevet is – *tisztességtelen* – a morfológiai elemzés elvégzésére. Tettem ezt azzal a céllal, hogy megfigyeljem a keresőszolgáltatás ennek az összetett kifejezésnek milyen más egyéb szóalakjait találja meg, azaz végez-e valamilyen morfológiai vizsgálatot erre a kifejezésre.

A stopszavak kezelésénél azt ellenőríztem, hogy a keresőkifejezés megjelenik-e a határozott és határozatlan névelőkkel együtt a találatokban. A lekérdezésekre kapott találatokban azt tanulmányoztam, hogy a keresők külön keresnek-e a megadott névelőkre vagy sem. Ennek a kérdésnek az eldöntése egyértelmű volt, mivel a megadott keresőkifejezés a névelővel együtt többnyire ki volt emelve a találatleírásokból, amikor az adott kereső keresett a névelőkre. Azonban a másik esetben, amikor a névelők ki voltak zártak a keresésből, csupán a keresőkifejezés szerepelt kiemelve a találatleírásokban. A következő lekérdezéseket vizsgáltam meg az angol nyelvben: *a dog* (=egy kutya), *an aunt* (=egy nagynéni), *the car* (=az autó). A magyar nyelv vonatkozásában pedig *a ház, az ember, egy kocsi* lekérdezéseket használtam. A stopszavak esetében gyakori kérdőszavakra, illetve számjegyekre nem kerestem.

Az ékezetek kezelését kizárólag a magyar nyelvben vizsgáltam meg, hiszen az nem releváns vizsgálati szempont az angol nyelv esetében. Itt nevezetesen két keresőkifejezést használtam a *kertem* és az *alma* lekérdezéseket. Arra a kérdésre kerestem a választ, hogy ezeknek a keresőkifejezéseknek az ékezetes változatait is lekérdezi-e a kereső a találatai között vagy sem.

Csonkolásnál gondot jelentett számomra, hogy a magyar nyelvű keresők közül egyedül a Heuréka kereső állította magáról, hogy képes csonkolni egy keresőkifejezésre [6]. Az összes többi magyar nyelvű kereső tájékoztató segédletében nem találtam erre vonatkozó információt. Minden esetben a * karaktert használtam csonkolásra a keresőkifejezés után. Az angol nyelv esetében az *Olympi** keresőkifejezést adtam meg azzal a céllal, hogy megtaláljam az összes olimpiai játékokról szóló oldalt az alábbi kifejezésekkel, pl.: *Olympic* (=olimpiai), *Olympics* (=olimpia, olimpiai játékok), *Olympia* (=földrajzi név), *Olympian* (=olimposzi), stb. A magyar nyelv esetében szándékosan törekedtem egy szűkebb, behatároltabb találathalmaz vizsgálatára annak érdekében, hogy a lekérdezés ragozott szóalakjait könnyebben találjam meg és le tudjam ellenőrízni azok tényleges előfordulását a találathalmazon belül. Ezért a májmétely nevű parazita ragozott szóalakjaira kerestem a *májmétely** lekérdezéssel. Továbbá megvizsgáltam a különböző keresőknél, hogy a *májmétely** lekérdezésre kapott találathalmazok ténylegesen tartalmazzák-e a keresőkifejezés ragozott szóalakjait, vagy sem. Végeredményül azt kaptam, hogy ezek a találathalmazok nem tartalmazták a lekérdezés ragozott szóalakjait egyik kereső esetében sem. Ekkor szükségesnek láttam külön egyenként, manuálisan is keresni a *májmétely* lekérdezés ragozott szóalakjaira, hogy ellenőrízhessem léteznek-e olyan ragozott szóalakok, amelyeket a keresőknek vissza kellett volna keresniük, de azt mégsem tették meg.

Végül pedig azt elemeztem, hogy a keresők visszakeresik-e egy lekérdezés szinonimáit, vagy sem. A szinonimák hogyan jelennek meg a ka-

pott találatokban, például ki vannak-e emelve a találatleírásokból vagy sem, ill. a megadott keresőkifejezéssel együtt fordulnak-e elő a találatleírásokban, vagy attól függetlenül is szerepelnek. Az angol nyelv esetében az alábbi lekérdezéseket vizsgáltam meg ilyen szempontból: *car* (=autó), *glasses* (=szemüveg). A magyar nyelvben pedig a *kutya* és a *vetélkedő* keresőkifejezések szinonimáira kerestem.
Tesztelés közben felfigyeltem egy hibára az AltaVista és az AlltheWeb esetében. Amikor a keresést leszűkítettem kizárólag a magyar nyelvű találatokra, ekkor angol, spanyol és francia nyelvű oldalak is megjelentek az első 100 találat között az *alma* lekérdezésre. Tehát a találatok nyelvi szűkítése nem működött kielégítően ennél a két keresőnél. Ugyanez a hiba nem fordult elő a Google kereső és a többi magyar nyelvű kereső esetében.
A továbbiakban az angol és a magyar nyelvű keresők nyelvi lehetőségeit tekintem át a korábban ismertetett vizsgálati szempontok alapján.

3.1. Az angol nyelvű keresőknél talált nyelvi megoldások

A morfológiai elemzés nem működött az angol és a magyar nyelvben a Google és az AlltheWeb keresők esetében. A vizsgált angol nyelvű keresők közül egyik sem végzett morfológiai elemzést a *tisztességtelen* lekérdezésre, hanem csak a megadott keresőkifejezés pontos szóalakjára kerestek. Egyedül az Altavista kereső esetében működött a morfológiai elemzés megfelelően az angol nyelvben. Tehát az automatikusan megtalálta a *dog* kifejezésnek a többes számú alakját is. Azonban a morfológiai vizsgálat a magyar nyelvben ugyanúgy nem működött az Altavista esetében, mint a másik két vizsgált keresőnél sem.
A stopszavak kezelésében a Google kereső jeleskedik a többi keresőhöz képest, mivel az teljes mértékben kizárja a keresésből az *a, an, the* névelőket az angol nyelvű lekérdezésekre. Ezenkívül kizárja az *a* határozott névelőt a keresésből, de nem zárja ki az *az, egy* névelőket a keresésből a magyar nyelvű lekérdezésekre. Az Altavista és az AlltheWeb keresők nem zárják ki a határozott és a határozatlan névelőket a keresésből az angol és a magyar nyelv esetében. Tehát a hiányosságuk egyértelmű ezen a területen.

Az ékezetek kezelése problémás mind a három keresőnél, mert azok megtalálták a *kértem* kifejezést tartalmazó weboldalakat a *kertem* lekérdezésre, valamint az *álma* kifejezést tartalmazó oldalakat az *alma* lekérdezésre.
Mind a három kereső hatékonyan valósította meg a csonkolást az angol nyelvben. Egyrészt megtalálták azokat a hosszabb kifejezéseket, amelyekben előfordult a csonkolt *Olympi* kifejezés. Másrészt lekérdezték azokat a weboldalakat, ahol a csonkolt kifejezés együtt jelent meg egy másik, hosszabb változatával. A Google kereső a csonkolt *Olympi* kifejezést az alábbi hosszabb szavakban fedezte fel:

Olympic (=olimpiai), *Olympia* (=földrajzi név), *Olympiad* (=olimpiász). Találatai között a csonkolt kifejezés az alábbi hosszabb változatokkal együtt fordult elő: *Olympics* (=olimpia, olimpiai játékok), *Olympian* (=olimposzi).
Az Altavista szolgáltatás a csonkolt *Olympi* kifejezést az alábbi hosszabb szavakban találta meg: *Olympic* (=olimpiai), *Olympics* (=olimpia, olimpiai játékok), *Olympia* (=földrajzi név), *Olympiad* (=olimpiász), *Olympian* (= olimposzi). Találataiban a csonkolt kifejezés az alábbi hosszabb változatokkal együtt jelent meg: *Olympia* (=földrajzi név), *Olympic* (=olimpiai), *Olympics* (=olimpia, olimpiai játékok). Az AlltheWeb kereső a csonkolt *Olympi* kifejezést az alábbi hosszabb szavakban fedezte fel: *Olympic* (=olimpiai), *Olympics* (=olimpia, olimpiai játékok), *Olympiad* (=olimpiász), *Olympia* (=földrajzi név), *Olympian* (=olimposzi).

Keresési eredményeiben a csonkolt kifejezés az alábbi hosszabb változatokkal együtt jelent meg: *Olympic* (=olimpiai), *Olympics* (=olimpia, olimpiai játékok), *Olympiad* (=olimpiász), *Olympia* (=földrajzi név), *Olympian* (=olimposzi).
A csonkolás viszont nem működik ezeknél a keresőknél a magyar nyelvben, hiszen megfigyelhető volt mindegyik esetben, hogy a *májmétely** lekérdezésre kapott találathalmazok nem tartalmazták a lekérdezés ragozott szóalakjait, ezért külön egyenként, manuálisan kellett azokra keresnem. Ezekben a keresőeszközökben a következő ragozott szóalakokat találtam meg: *májmételynek* (ahol a „-nek" a részeshatározós eset és a birtokos eset ragja), *májmételyt* (ahol a „-t" a tárgyeset ragja), *májmételyek* (ahol a „-k" a többes szám jele), májmételyről (ahol „-ről" az előljárós eset ragja). A *májmétely* lekérdezés csonkolt és csonkolatlan változataira kapott találatszámokat, valamint a ragozott szóalakokra nyert találatszámokat az alábbi táblázatban ismertetem.

	Google	**Altavista**	**Alltheweb**
Csonkolás	*májmétely** (106 találat), *májmétely* (140 találat), *májmételynek* (1 találat), *májmételyt* (6 találat), *májmételyek* (8 találat), *májmételyről* (3 találat).	*májmétely** (100 találat), *májmétely* (94 találat), *májmételynek* (1 találat), *májmételyt* (5 találat), *májmételyek* (6 találat)	*májmétely** (62 találat), *májmétely* (62 találat), *májmételynek* (1 találat), *májmételyt* (4 találat), *májmételyek* (1 találat).

Ebben a táblázatban az Altavista keresőnél minimális eltérés figyelhető meg a lekérdezés csonkolt és csonkolatlan változataira lekérdezett találatszámokban. A Google szolgáltatásnál a lekérdezés csonkolt változatára kapott találatszám alacsonyabb a csonkolatlan változat találatszámához képest. Az AlltheWeb keresőnél ugyanaz a találatszám fedezhető fel a lekérdezés csonkolt és csonkolatlan változataira. Ez utóbbi két eset azt tükrözi számunkra, hogy a csonkolás egyáltalán nem működik ezeknél a keresőknél a magyar nyelvben, mert alapértelmezés szerint a lekérdezés csonkolt változatának nagyobb találathalmazt kell eredményeznie a csonkolatlan változat találathalmazához képest. Továbbá az is észrevehető a táblázatban, hogy mindegyik szolgáltatás csak kevés találatot nem kérdezett le a témával kapcsolatban. Ezek a figyelmen kívül hagyott találatok a lekérdezés ragozott szóalakjainak a segítségével voltak megtalálhatóak.

Az angol nyelvű keresők mindegyike keresett a megadott lekérdezés szinonimáira az angol nyelvben, azonban ennek a feladatnak a megvalósításában találtam hibás nyelvi értelmezésből eredő hibákat. A *car* (=autó) és a *glasses* (=szemüveg) lekérdezésekre kapott találatokat mindegyik kereső vonatkozásában az alábbi táblázatban foglalom össze:

	Google	Altavista	Alltheweb
Lekérdezett találatok	pl.: *~car* – automobile (USA), automotive (USA), auto (USA), motor (nem szinonima!), vehicle (nem szinonima!), racing (nem szinonima!); pl.: *~ glasses* – eyeglasses, glassware, sun-glasses (nem szinonima!), goggles (nem szinonima!).	pl.: *car* – automobile (USA), automotive (USA), auto (USA), vehicle (nem szinonima!); pl.: *glasses* – spectacles, eyeglasses, glassware, sun-glasses (nem szinonima!), reading-glasses (nem szinonima!), goggles (nem szinonima!).	pl.: *car* – automobile (USA), automotive (USA), auto (USA), vehicle (nem szinonima!); pl.: *glasses* – eyeglasses, glassware, sun-glasses (nem szinonima!), reading-glasses (nem szinonima!), goggles (nem szinonima!).

A *car* keresőkifejezésnek az alábbi szinonimái fordultak elő a találatokban: *automobile (USA)* (=autó, automobil, gépkocsi, gépjármű), *automotive (USA)* (=autó, gépkocsi, automobil, gépjármű), *auto (USA)* (=autó, kocsi).

Továbbá a *motor* (=motor, jelzői értelemben is használatos kifejezés, aminek a jelentése autó(s)-, motoros), *racing* (=verseny-, versenyzés) és a *vehicle* (=jármű, közlekedési eszköz) kifejezések is szinonimaként jelentek meg, azok azonban jelentésükből adódóan nem tekinthetők a *car* kifejezés tényleges szinonimáinak. A *glasses* keresőkifejezésnek a következő szinonimái jelentek meg a találatokban: *eyeglasses* (=szemüveg), *spectacles* (=szemüveg, pápaszem). Ezenkívül a *sun-glasses* (=napszemüveg), *reading-glasses* (=olvasószemüveg) és a *goggles* (=motor-szemüveg, védőszemüveg) kifejezésekre is kerestek az angol nyelvű keresők, ezek inkább szemüvegfajtáknak tekinthetők és nem pedig a *glasses* keresőkifejezés szinonimáinak. Mind a három kereső megtalálta a *glasses* lekérdezés egyes számú alakjának a szinonimáját is a *glass-ware*-t, aminek a jelentése „üvegáru". Itt fontos megjegyeznem, hogy az egyes számú alak, azaz a *glass* kifejezés jelentései a következők: „üveg, pohár, üvegedény, üvegáru".

A táblázatban látható, hogy mindegyik kereső megtalálta a *car* keresőkifejezés amerikai angolban használatos szinonimáit. Azonban a Google kereső három nem szinonimának minősülő kifejezést is lekérdezett. Az Altavista és az AlltheWeb külön-külön egy nem szinonimának tekinthető kifejezést fedezett fel. A másik esetben az Altavista kereső két szinonimát talált meg a *glasses* lekérdezésre és három nem szinonimát az egyes számú alak szinonimáján kívül. A Google és az AlltheWeb egyetlenegy szinonimát kérdezett le az egyes számú alak szinonimáján kívül. A Google két nem szinonimát talált meg, míg az Alltheweb három nem szinonimát fedezett fel.

A Google keresőnél külön a ~ (tilde karakter) segítségével kerestem egy adott keresőkifejezés szinonimáira. Itt a megtalált szinonimák kiemelve szerepeltek a találatok

leírásában és azok többnyire együtt fordultak elő a lekérdezéssel. Ezzel szemben a másik két keresőnél a szinonimák nem voltak kiemelve a találatok leírásából és azok mindig együtt fordultak elő a keresőkifejezéssel. Mind a három keresőnél kerestem a *kutya* és a *vetélkedő* keresőkifejezések magyar nyelvű szinonimáira, ezekre a lekérdezésekre azonban nem kaptam értékelhető szinonimákat. A vizsgált keresők csupán a lekérdezés pontos szóalakjára kerestek és nem vették figyelembe annak a szinonimáit a magyar nyelvben.

3.2. A magyar nyelvű keresőknél talált nyelvi megoldások

A magyar nyelvű keresők közül kizárólag a Heuréka keresőről valószínűsíthetjük azt, hogy végez morfológiai elemzést a lefuttatott lekérdezésekre, mert a *tisztességtelen* keresőkifejezésre az alábbi szóalakokat fedeztem fel a találatok leírásában, pl.: *tisztesség, tisztességes, tisztességért, tisztességgel, tisztességtelen, tisztességtelenül*. A *ház, kocsi, kutya* lekérdezések többes számú alakja mindig együtt fordult elő azok egyes számú alakjával a találatok leírásában. A megadott keresőkifejezések nem voltak kiemelve a találatleírásokból, ami megnehezítette annak az eldöntését, hogy a Heuréka kereső keres-e külön a lekérdezés többes számú alakjára vagy sem. Az összes többi kereső esetében nem működött a morfológiai elemzés, vagyis azok nem kerestek automatikusan egy adott lekérdezés többes számú alakjára. Ezenkívül a *tisztességtelen* kifejezésre egyikük sem végzett morfológiai elemzést, mivel azok csak a megadott keresőkifejezés pontos szóalakjára kerestek.

A stopszavak kezelése eléggé megosztott képet mutat a vizsgált keresőknél, mert az Origo-Vizsla és a Kurzor keresők nem zárják ki a határozott és a határozatlan névelőket a keresésből [7, 8]. A Heuréka kereső esetében nem lehetett egyértelműen megállapítani, hogy a határozott és a határozatlan névelők ki voltak-e zárva a keresésből, mert a keresőkifejezések a névelőkkel együtt nem voltak kiemelve a találatok leírásából. A kapott találatoknál megfigyelhető volt az is, hogy azokban a keresőkifejezések sokszor a megadott névelők nélkül fordultak elő. Ebből arra lehetett következtetni, hogy a Heuréka kereső igyekezett kizárni a névelők használatát a keresésből. A Góliát és az Altavizsla keresők kizárták az *a, az* határozott névelőket a keresésből, de nem zárták ki az *egy* határozatlan névelőt. A keresőkifejezések kiemelésének a problémája egyedül a Heuréka keresőnél jelentkezett, a többinél ez nem okozott gondot.

Az ékezetek kezeléséről elmondható, hogy azt a keresők nagy része sikeresen oldotta meg. Az egyedüli kivételt képezte az Origo-Vizsla kereső, ahol nem működött az ékezetek kezelése. Tehát megtalálta a *kértem* kifejezést tartalmazó weboldalakat is a *kertem* lekérdezésre és az *álma* kifejezést tartalmazó oldalakat is az *alma* lekérdezésre. A Heuréka keresőnél az ékezetek kezelése csak az ékezethelyes keresési lehetőség kiválasztásánál működött jól. Azonban az automatikus ékezetkezelési lehetőség kiválasztásánál az már nem működött megfelelően. A Kurzor, Góliát és az Altavizsla keresők esetében hatékonyan működött az ékezetek kezelése.

A csonkolás azonban nem működött egyik kereső esetében sem. Sokszor a *májmétely* lekérdezés csonkolt változatára nem kaptam egyetlenegy találatot sem. A Heuréka, Góliát és az Altavizsla keresők esetében ugyanazt a negatív eredményt kaptam. Emögött az elfogadható magyarázat az lehet, hogy a kereső nem képes értelmezni a csonkolást mint lekérdezési műveletet. Csonkolásnál fennállt egy másik eset is, amikor ugyanannyi találatot kaptam a lekérdezés csonkolt és csonkolatlan változataira egyaránt. Ez megint csak azt tükrözi, hogy a csonkolás nem működik egyáltalán, hiszen alapértelmezés szerint a lekérdezés csonkolt változatának bővebb találathalmazt kell eredményeznie a csonkolatlan változat találathal-

mazához képest. Ez utóbbi eset állt fenn az Origo-Vizsla és a Kurzor keresőknél. A *májmétely* lekérdezés csonkolt és csonkolatlan változataira kapott találatszámok, valamint a ragozott szóalakok a rájuk kapott találatszámokkal együtt megtekinthetők az alábbi táblázatban.

	Heuréka	Origo-Vizsla	Kurzor	Góliát	Altavizsla
Csonkolás	*májmétely** (0 találat), *májmétely* (36 találat) *májmételyek* (1 találat)	*májmétely** (35 találat), *májmétely* (35 találat) *májmételynek* (2 találat), *májmételyt* (1 találat), *májmételyek* (3 találat)	*májmétely** (30 találat), *májmétely* (30 találat) *májmételyt* (2 találat), *májmételyek* (1 találat)	*májmétely** (0 találat), *májmétely* (30 találat)	*májmétely** (0 találat), *májmétely* (30 találat)

A Góliát és az Altavizsla keresőknél nem kaptam egyetlenegy találatot sem a lekérdezés ragozott szóalakjaira[5]. Ezek a szolgáltatások valószínűleg nem indexelik azokat a weboldalakat, amelyek relevánsak lehetnek egy ilyen típusú lekérdezés számára.

A keresőkifejezések szinonimáira történő keresések csupán a Heuréka és az Origo-Vizsla keresők esetében voltak eredményesek. Az összes többi kereső nem keresett egy lekérdezés szinonimáira a magyar nyelvben. Ezzel a két keresővel az alábbi szinonimákat kaptam a *kutya* lekérdezésre: *eb, öleb*. A *vetélkedő* keresőkifejezésre a következő szinonimákat találtam meg: *kvíz, kvízjáték, verseny, agytorna*. Mindkét keresőnél a szinonimák nem voltak kiemelve a találatok leírásából, ami kevésbé egyértelművé teszi az ilyen típusú lekérdezéseket. A szinonimák többnyire együtt fordultak elő a lekérdezéssel a találatok leírásában, és azok csak néhol szerepeltek önállóan. A magyar nyelvű keresők közül kizárólag a Heuréka keresőnél találkozunk olyan beépített tezaurusszal, amely fogalmi relációival támogatja a keresési témával kapcsolatos további lekérdezési lehetőségeket.

4. Összegzés

A vizsgálat eredményei rámutatnak arra, hogy az angol nyelvű keresők sokkal jobban kezelik és értelmezik az angol nyelvű lekérdezéseket, mint a magyar nyelvűeket. Ezekben a keresőeszközökben a csonkolás és a szinonimákra történő keresés eredményesen működik az angolban, de meglehetősen problémás a magyar nyelvben. Ezenkívül megfigyelhető az is, hogy azok hiányosságokat mutatnak hasonló területeken, például nem kezelik helyesen az ékezeteket, ami fontos felhasználói elvárás a magyar nyelvű lekérdezések vonatkozásában. Ez a hiányosság egyben tükrözi a hátrányukat abban, hogy nem képesek releváns információkat szolgáltatni a magyar felhasználók számára. A Google kereső nem veszi figyelembe a határozott és a határozatlan névelőket az an-

gol nyelvű lekérdezéseknél, azonban még nem oldotta meg ezt a kérdést a magyar nyelvben. Az Altavista kereső megtalálja az angol nyelvű lekérdezések többes számú alakját, ezért ez az egyedüli olyan szolgáltatás, ahol a morfológiai elemzés hatékonyan működik az angolban. Az angol nyelvű keresők közül egyik sem küzdött meg ezzel a problémával a magyar nyelvben. Elmondható a Google és az Altavista szolgáltatásokról, hogy azok azonos teljesítményt nyújtottak a lekérdezések értelmezésében, utánuk pedig az AlltheWeb kereső következik a rangsorban. A kapott eredmények alapján levonhatjuk azt a következtetést, hogy ezeknek a keresőknek még nagyobb hangsúlyt kellene fektetniük a morfológiai elemzésre, a stopszavak és az ékezetek kezelésére a jövőbeli fejlesztésükben.

A magyar nyelvű keresők többsége helyesen kezeli az ékezeteket, ezáltal azok egy lényeges felhasználói követelményt biztosítanak. A magyar nyelvű keresők nyelvi megoldásait értékelve megállapíthatjuk, hogy a Heuréka kereső nyújtotta a legjobb teljesítményt a morfológiai elemzés és a szinonimákra történő keresés területén. Teljesítménye elfogadható volt két másik területen, a stopszavak és az ékezetek kezelésében. Ez után következnek a Góliát és az Altavizsla keresők, amelyek nagyon hasonló teljesítményt nyújtottak a tesztelés során. Mindketten pontosan kezelték az ékezeteket. Azonban a stopszavakat illetően fejleszteniük kell a jelenlegi működésüket, mert az még nem teljesen tökéletes. A rangsorban az Origo-Vizsla és a Kurzor keresők az utolsó helyre kerültek ugyanazzal a teljesítménnyel. Az Origo-Vizsla szolgáltatás számos területen mutat hiányosságokat, mint például morfológiai elemzés, csonkolás, ékezetek és stopszavak kezelése. A Kurzor kereső hiányosságokkal rendelkezik az alábbi területeken: morfológiai elemzés, stopszavak kezelése, csonkolás, szinonimákra történő keresés. A vizsgálat eredményei alapján kijelenthetjük, hogy majdnem mindegyik magyar keresőnek fejlesztenie kell a saját teljesítményét a csonkolás és a morfológiai elemzés területén.

Felhasznált irodalom

1. Bar-Ilan, J. – Gutman, T.: How do search engines respond to some non-English queries? In: Journal of Information Science vol. 31. no. 1. 2005. pp.13-28.
2. AlltheWeb tájékoztató segédlete URL: http://www.alltheweb.com
3. Altavista tájékoztató segédlete URL: http://www.altavista.com
4. Google tájékoztató segédlete URL: http://www.google.com
5. Góliát tájékoztató segédlete URL: http://www.goliat.hu
6. Heuréka tájékoztató segédlete URL: http://www.heureka.hu
7. Kurzor tájékoztató segédlete URL: http://www.kurzor.hu
8. Origo-Vizsla tájékoztató segédlete URL: http://www.origo.hu

Új információk s Könyvtári Intézet honlapján

Elkészült a (könyvtár)építésre vonatkozó szabványok jegyzéke, megtekinthető: (http://www.ki.oszk.hu/konyvtarepiteszet/index.html) címen.

A Technikai eszközök menüpontban a Biztonságtechnikai eszközök kínálatat az árnyékolástechnikai eszközökkel, illetve az azokat forgalmazó cégek adataival egészítették ki. (http://www.ki.oszk.hu/konyvtarepiteszet/index.html)

A Hungarológiai kézikönyvek jegyzékét felfrissítették a 2005-ben megjelent kötetek adataival (http:// www.ki.oszk.hu/hungarologia/index.html).

(Hölgyesi Györgyi (Könyvtári Intézet) tájékoztatójából, 2006. máj. 3.)