



Hogyan válaszolják meg a keresőmotorok a nem angol nyelvű keresőkérdéseket?

Bevezetés

Izraeli kutatók azt vizsgálták, hogy a leggyakrabban használt keresőmotorok: a *Google*, az *AltaVista* és a *FAST* hogyan boldogulnak a nem angol nyelvű kérdésekkel. Abból indultak ki, hogy bár az Internet közös nyelve az angol, a *Global Reseach* 2004-es eredményei azt mutatták, hogy a felhasználóknak csupán 35,8%-a angol anyanyelvű. 3,8% francia, 2,5% orosz, 14,1% kínai, 9,6% japán, 9% spanyol, 7,3% német, 4,1% koreai, 3,5% portugál, 3,3% olasz, és 1,8% holland anyanyelvű volt. Másrészt, ha a honlapok nyelvét vizsgáljuk, egészen más eredményeket kapunk. A *Cyber Atlas* a *VilaWeb* adatai alapján arról tudósított, hogy 2000-ben a honlapok 65,6%-a volt angol, 2,96%-a francia, 1,88%-a orosz, 0,16% magyar, és 0,16%-a héber nyelvű. Az angol mellett a web legnépszerűbb nyelvei a japán (5,85%), a német (5,77%) és a kínai (3,87%) voltak, ezeket követte a francia (2,96%), a spanyol (2,42%) és az orosz (1,88%). Az *OCLC* adatai szerint 2002-ben a honlapok 72%-a volt angol nyelvű, 3%-a francia és 1%-a orosz. Az angol után a legnépszerűbb volt a német (8%), a japán (6%), a francia és a spanyol (3–3%), az olasz, a holland és a kínai (2–2%).

A felhasználók és a honlapok nyelve közötti eltéréseket különféleképpen lehet magyarázni:

- A nem angol anyanyelvűek azért készítenek olyan nagy számban angol nyelvű honlapokat, hogy az oldal minél szélesebb körben hozzáférhető legyen (hiszen az angol az internet közös nyelve). Ezt igazolják *G. Nunberg* megállapításai.
- Nem minden felhasználó készít honlapot, és lehet, hogy speciális problémák merülnek fel a nem angol nyelvű weboldalak létrehozásánál (ez különösen igaz a héberre, amely nem latin karaktereket használ, és az írás jobbról balra történik).

- Számos többnyelvű honlap létezik, ezért ha az oldalak nyelvek szerinti százalékos megoszlását összeadjuk, meghaladják a 100%-ot.
- Az internetes statisztikákat óvatosan kell kezelni. A különféle statisztikák számára gyűjtött és feldolgozott adatok különböző mintavételezési eljárásokkal és módszerekkel készültek, és a web gyors és dinamikus fejlődésének köszönhetően a 2000-es és 2002-es adatok nehezen mérhetőek össze.

Annak ellenére, hogy az adott számok nem pontosak, egyértelmű, hogy a nem angol nyelvű felhasználókat és oldalakat nem lehet figyelmen kívül hagyni. A felhasználók országos, illetve regionális keresőmotorokkal vagy címjegyzékekkel tudnak nem angol nyelvű információkat keresni, vagy pedig nemzetközi keresőmotorokat használnak. 2002-ben a vezető keresőmotorok közül a Google lehetővé tette, hogy a keresést 35 nyelv egyikére korlátozzuk, a *FAST* (*AlltheWeb*) 49 nyelven, az *AltaVista* 25 nyelven, az *MSNsearch* (*Inktomi*) 15 nyelven és a *WiseNut* 25 nyelven kínálta fel a keresést (a *Teomában* az egy nyelvre történő szűkítésre nem volt lehetőség 2002-ben). Ebből jól látszik, hogy a nagy keresőmotorok igyekeznek a nem angol nyelvű kérdéseket is megválaszolni. Néhány keresőmotornak helyi oldalai is vannak (például a Google-nak 34), emellett a Google, az *AltaVista*, a *FAST* és az *Inktomi* helyi keresőmotorokat is működtetett (néha globális, máskor helyi indexek alapján).

A felhasználók által beírt a keresőkérdésekre időnként csak sematikus egyező válasz érkezik, ha a szimbólumsor valahol megtalálható a szövegben. Máskor sokkal fejlettebb módszereket, morfológiai nyelvi elemzést alkalmaznak (számolnak a többes számmal, az igeidőkkel, elő- és utótagokkal). A főbb keresőmotoroknál az angol az alapértelmezett nyelv (mivel a legtöbb honlap is angolul készül).

A kutatás kiinduló kérdése az volt, hogy a főbb keresőmotorok hogyan támogatják és teszik lehetővé a nem angol nyelvű kereséseket, hogyan válaszolnak ezeken a nyelveken, és milyen mértékben veszik figyelembe a nyelvek speciális jellemzőit (például a ragozást, az ékezeteket, az előtagokat). A kutatók csak a keresőkérdésekre adott válaszokat vizsgálták, a keresési algoritmusokat nem. A válaszok száma általában tájékoztató jellegű, a legkülönlegesebb kérdésekre a válaszok csak az ezres listákban találhatók. Közismert, hogy a felhasználók csak az első tíz–húsz választ veszik figyelembe, ezért a kutatók az eredményeknél csak az első tíz választ vizsgálták.

Összehasonlították a helyi és a nemzetközi keresőmotorok eredményeit. A vizsgálathoz négy nyelvet választottak: az orosz, a franciát, a magyart és a hébert. Mindegyik sok akadályt jelent a keresőmotorok számára; mivel különböző nyelvcsaládba tartoznak (szláv, latin, finnugor, sémi). A francia és a magyar a latin betűkészletet használja hozzáadott ékezetekkel, az orosz cirill betűket, a héber saját betűket használ, az írás pedig jobbról balra történik. Az eredmények bemutatásának célja a felhasználók figyelmeztetése, hogy a jelenlegi keresőmotorok, különösen a nemzetköziek, nem a megfelelő módon válaszolnak a nem angolul feltett keresőkérdésekre.

Módszertan

Minden egyes nyelv számára gondosan kiválasztották a keresőkifejezéseket, majd lefuttatták őket a nemzetközi és a hazai keresőmotorokon. A próbakeresések találatait kifejezetten azért mutatták be, hogy felhívják a felhasználók figyelmét azokra a problémákra, amelyek a vizsgált nyelveken történő keresés során felmerülhetnek. A kereséseket 2002 novemberében végezték. Ahol lehetséges volt, tanulmányozták a keresőmotorok *súgóit*, hogy megismerkedjenek az eszközök lehetőségeivel és tulajdonságaival. Olyan szavakat és kifejezéseket kerestek, amelyek kiemelik az adott nyelvek nehézségeit. Természetesen nem választottak tulajdonneveket, mert ezek szabad szövegekben általában nem szerepelnek különböző morfológiai formákban, valamint ki kellett rekeszteni a legnépszerűbb kérdéseket. Minden keresésnél rögzítették a találatok számát. Minden egyes kérdésnél átvizsgálták az első tíz választ, hogy lássák, a honlapon megfelelő szóalak, vagy csak a keresett kifejezés morfológiai alakjai jelentek-e meg. A válaszok relevanciáját nem vizsgálták, mert a kuta-

tásnak nem ez volt a célja, hanem a keresőmotorok nyelvi lehetőségeinek vizsgálata.

A következő nemzetközi keresőmotorokat vizsgálták: a Google helyi változatait, az AlltheWebet (FAST) és az AltaVistát. Az AltaVistának a korábbi, még a 2002. november előtti verzióját használták, kivéve a héber nyelvű kereséseknél. A Google a kifejezéseket csak pontos formájukban kereste, és a tiltott szavak (pl. névelők) kihagyásán kívül semmilyen morfológiai elemzést nem végzett. Az AlltheWeb a kifejezést csak a megadott formában kereste, az AltaVista pedig érzékeny volt az ékezetekre: ahol a keresett szó ékezetek nélkül szerepelt, ott az összes formát megkereste, ahol pedig volt ékezet, csak a pontos alakra keresett. A Google és az AlltheWeb a leírás szerint csak pontos megfeleléseket keres a keresőkérdésekre, például amikor valaki az „electricité” francia kifejezésre keres, és nem az „életricité”-re vagy az „electricité”-re, csak az első formát kell visszakeresnie, valamint az „électricité”-re történő keresésnél sem szabadna megtalálnia az egyéb előfordulásokat, mint az „electricité” vagy az „electricite”.

Az orosz nyelvhez a három legnépszerűbb helyi keresőt használták: a Yandexet (<http://www.yandex.ru>), a Rambler (<http://www.rambler.ru>), és az Aportot (<http://www.aport.ru>). A Yandex az összes kifejezés minden nyelvtani alakjára keresett, a Rambler szintén minden eltérő formára is keresett a súgójának megfelelően, az Aport pedig csak a szokványos orosz szavaknál használta a morfológiai elemzést, a ritka kifejezéseknél nem.

A francia nyelvhez három nyelvspecifikus eszközt vizsgálták meg: Voliát (<http://www.voila.fr>), az AOL France-t (<http://recherche.aol.fr>) és egy francia–kanadai La Toile de Québec (<http://www.toile.com>) portált, amely viszonylag kevés oldalt tartalmaz. A Voilának meg kellett volna keresnie az adott kifejezés összes alakját (ékezetekkel vagy anélkül), attól függetlenül, hogy a keresőkérdést hogyan fogalmazták meg; a La Toile de Québec számára az ékezetek közömbösek voltak; az AOL France-nak nem volt sem súgója, sem a keresőmotor tulajdonságairól szóló egyéb felvilágosítása.

A magyar nyelvű kereséshez három keresőmotort használtak: az Origo-vizslát (<http://www.origo.hu>), a Startlapot (<http://www.startlap.com>) és a Heurékát (<http://www.heureka.hu>). Az Origo és a Startlap keresőihez találtak megfelelő leírásokat. A Heurékánál választani lehetett a kifejezések éke-

zetek nélküli és ékezetes begépelése, vagy az ékezetek automatikus hozzáadása között – a szóalak leggyakoribb használata szerint. A keresésekhez a pontos alakra történő keresést választották. A Heuréka lehetővé tette a jobb oldali csonkolást a csillag „*” segítségével. Az Origo-vizsla a csonkolást csak négy karakter után engedélyezte, jóllehet úgy tűnt, hogy a szóalakokat automatikusan felismerte, vagyis a „kutya” szónál a „kutyák”-ra is keresett. A sűgőja viszont semmit sem említett az ékezetek értelmezéséről.

A héber nyelvű kereséseket a *Morfix* (<http://www.morfix.co.il>), beépített morfológiai elemzést végző keresőmotor, valamint a legnépszerűbb izraeli portál, a *Walla* (<http://www.walla.co.il>) segítségével végezték. A Morfix négytípusú keresést tett lehetővé: a pontos alakra, minden morfológiai alakra, a kiterjesztett formákra és az azonos töre, illetve a szóra és szinonimáira. A Walla indexelte a héber és az angol oldalakat is. A héber keresőkérdekre nem héber nyelven is keresett, mert minden, nem héber nyelvű indexelt honlapnak volt héber nyelvű összefoglalója. Minden, a Walla által indexelt oldal keresésekor két lehetőség volt: a pontos alakra történő keresésre, illetve a szóelemként való kezelésre. Ez utóbbira semmilyen felvilágosítást nem találtak, de valószínűleg azt jelentette, hogy a keresőmezőbe begépelte betűsor egy szó részeként jelent meg, lehetséges karakterek hozzáadásával jobbról, illetve balról. Ehhez a portálhoz semmilyen leírást vagy sűgőt nem találtak. A Wallában weboldalakra is lehetett keresni (a FAST által működtetett szolgáltatás, amelyet nehéz volt elérni). A keresések – az alapértelmezés szerint – először a tartalomjegyzékben történnek.

Eredmények

Orosz nyelvű keresések

Az első tíz találat elemzése azt mutatta, hogy a vizsgált (helyi) keresőmotorok – dokumentációiknak megfelelően – végeznek morfológiai elemzéseket a keresőkérdeken, míg a nemzetközi keresők nem vették figyelembe a nyelv speciális tulajdonságait, és csak egyszerű sablonos megfeleltetést végeztek a keresőkifejezések és szavak között. Ellentétben a kutatóknak a sűgőkon alapuló

elvárásaival, az AlltheWeb megtalálta az alkalmi végződések is az első tíz válaszbán.

Francia nyelvű keresések

Itt az eredmények még nagyobb csalódást okoztak: a nemzetközi keresőmotorok nemcsak a nyelv sajátosságait hagyták figyelmen kívül (az ékezeteket, az aposztrófós ékezeteket, az egyes és többes számokat), hanem a legtöbb nyelvi eszközt is. A La Toile de Québec egy portál, ennél fogva kicsi volt a lefedettsége. Megmagyarázhatatlan, hogy a Voila több találatot adott a „l'electricité”-re, mint az „électricité”-re, hiszen a „l'electricité” eredményeinek tartalmaznia kellene az „électricité” oldalait is. A nemzetközi keresőmotorok lefedettségében óriási különbség van az előzőekhez képest, ami talán a keresőtér értelmezésének különbségében rejlik: eleinte azt gondolták, hogy a francia oldalak keresése az AlltheWeben, és a francia vonatkozású oldalaké a Google-ban ugyanazt jelenti. Az AltaVista lefedettsége jóval alacsonyabb volt, mint a Google-é és az AlltheWebé, ebből eredően a találatok száma is. Az ékezetek helytelen kezelésének vagy a francia karakterkészlettel történő írás, vagy a speciális HTML karakterek használata (például „´” az „é” helyett) lehet az oka.

Magyar nyelvű keresések

A magyar keresőmotorok figyelembe vették a magyar nyelv sajátos tulajdonságait. Az Origo-vizsla kiterjesztette a keresést a különböző szóalakokra, de ahogy a „kutya – kutyák” példa is sugallta, a kétféle keresés némileg eltérő találati halmazt eredményezett. A Startlap túlzottan kiterjesztettnek tűnt, a nem kapcsolódó szóalakokat is megtalálta, például a „kar” keresése esetén a „zenekar”-t. Könnyebb csupán a felhasználó által begépelte betűláncra keresni, de az olyan morfológiai bonyolult nyelvek esetén, mint a magyar, az egyszerű sablon szerinti megfeleltetés nem elegendő. Meg kellene találni a helyes egyensúlyt a túl sok szóalak megtalálása és a túlszabályozott visszahívás között. A nemzetközi keresőmotorok nem kerestek szóalakokra, ami a magyar nyelv szempontjából komoly probléma. Az AlltheWeb nem is különítette el az ékezetes és ékezetek nélküli magánhangzókat. Az 1. táblázat a magyar nyelvű keresések eredményeit foglalja össze magyar keresőmotorokon, a 2. táblázat ugyanezeket nemzetközi keresőmotorokkal.

1. táblázat

Magyar nyelvű keresések magyar keresőmotorokkal

Keresőkérések	Origo-vizsla	Startlap	Heuréka
kar	705 136 oldal csak önálló szóként az első 100 találat-ban	299 364 oldal különböző alakokban karaoke vagy zenekar	21 782 oldal kar* jobb oldali csonkolással 72 425 oldal
kár	642 999 oldal eltérő az előző eredményektől	216 274 oldal különböző alakok, amelyekben a karak-terlánc megtalálható	14 412 oldal kár* 100 489 oldal
kutya	395 947 oldal kutya*729 184 oldal az első eredmények alacsony relevanciájúak	235 166 oldal	13 681 oldal kutya* 21 526 oldal
kutyák	399 152 oldal nem egyezik az előzővel, de a kutya szó is felbukkan a találatok közt	51 554 oldal	4399 oldal csak az Origo-vizslánál találtak többet a kutyák alakra
falú	410 000 oldal	257 735 oldal	15 767 oldal
falvak	410 620 oldal	37 588 oldal	3983 oldal
javítás	752 185 oldal	136 067 oldal	6956 oldal
kijavítás	18 271 oldal	98 oldal	95 oldal

2. táblázat

Magyar nyelvű keresések nemzetközi keresőmotorokkal

Keresőkérések	Google (a magyar nyelven írt oldalakra keresve)	AlltheWeb (magyarul)	AltaVista (magyarul)
kar	111 000 oldal	138 026 oldal benne a kár is, de nem a vezető helyeken	30 945 oldal
kár	40 700 oldal csak a teljesen azonos formára	142 266 oldal a vezető helyen a kar szerepel a kár helyett, de a találatok száma különbözik az előző keresésétől	10 867 oldal
kutya	41 500 oldal	64 270 oldal	8652 oldal
kutyák	15 600 oldal	20 400 oldal	14 322 oldal több mint az egyes számú alaknál
falú	43 600 oldal	49 548 oldal	12 292 oldal
falvak	12 000 oldal	13 851 oldal	3883 oldal
javítás	18 600 oldal	43 624 oldal	4648 oldal
kijavítás	298 oldal csak a teljesen azonos formára	277 oldal	74 oldal

Héber nyelvű keresések

A héber nyelvű keresőeszközök lefedettsége különböző méretű volt. Úgy tűnt, hogy a Morfix csak néhány honlap keresését végzi el az izraeli webtérben, míg az izraeli oldalak jelentős hányada más doméneken van nyilvántartásban. Érdekes példa az *Izraeli Postahivatal* (<http://www.postil.com>), amely nem szerepel az Izraelen kívül készült honlapok között. A Morfixen

a négy lehetőség közül kettőt használtak ki: a pontos és a morfológiai keresést. A Wallán mind a tartalomjegyzékben, mind a FAST által működtetett keresőmotorral végeztek keresést. Az oldalon teljes és részsavas kereséseket alkalmaztak. A nagyobb keresőmotorok csak a pontos szóalakokra kerestek. A „ház” szó számos előfordulása egyedülálló volt, de még ebben az esetben is, amikor a Morfix a morfológiai alakokat is megtalál-

ta, a találati arány körülbelül 60%-ra emelkedett. Az igény az elő- és utótagok figyelembevételére sokkal hangsúlyosabb volt az „egyetem” szó keresésénél. A Morfixban önálló szóként jelent meg a szövegek 3%-ában, a többi esetben akár határozott névelővel, akár a birtokos eset részeként, viszonyszókkal vagy elő- és utótagok kombinációival. A Google-ban az alábbi logikai mondatra keresve: „egyetem VAGY az egyetem VAGY az egyetemen VAGY az egyetemre VAGY az egyetemen VAGY az egyetemről VAGY a ... egyetemre VAGY a ... egyetemről” 56 500 találatot eredményezett, majdnem nyolcszor többet, mint az „egyetem” szóra. Az „egyetem” szóra kereséskor a legtöbb felhasználó csak a szót gépeli be, nem gondolva az összes lehetséges kombinációra, amit emiatt elveszítene. Még akkor sem veszik a fáradságot a begépelésre, ha gondolnak arra, hogy a többi szóalakra is keresniük kellene. Egy másik probléma volt, hogy az önálló szóként történő keresésnél a kevésbé fontos találatok jelentek meg a lista elején. „Az egyetem” és az „... egyetem” sokkal jobb találatokat eredményezett mind a Google-ban, mind a Morfixban. Az „egyetem” szóra végzett morfológiai keresés első száz találatának elemzése során kiderült, hogy közülük nyolcvan nem volt releváns. A héber szempontjából az összes szóalak megtalálása kétségkívül elengedhetetlen, de ez a lehetőség önmagában nem elegendő a kiváló minőségű találatok elérésére.

Összefoglalás

A Word Wide Web nem csak az angol nyelvűeké: a felhasználók több mint 50%-a nem angol anyanyelvű, és a becslések azt mutatják, hogy a honlapok egyharmada nem angol nyelvű. Ha ezeket a keresőmotorokat nem teszik alkalmassá speciális

nyelvek keresésére, az adott honlapot talán soha nem fogják megtalálni, a nem angol nyelvű honlapok sokkal nagyobb eséllyel vesznek el a kibertérben.

A másik fontos probléma, hogy a Google túlságosan népszerű keresőeszköz a nem angol nyelvű felhasználók körében. Többnyire nincsenek is tisztában a Google hiányosságaival, ezért rengeteg releváns találatot veszítenek. A nem angol nyelvű beszélők hátrányos helyzetben vannak azokkal szemben, akik angolul keresnek információt, mert néhány szóalak annyira egyedi a nem angol nyelvekben, hogy morfológiai elemzés és felhasználói tapasztalat nélkül a keresőmotorok használata nem kielégítő. A lehetséges megoldás, hogy a felhasználóban tudatosítsuk a problémát, bár nem várható el tőle, hogy az összes lehetséges szóalakot begépélje, (különösen, hogy a Google 10 szóban maximálja a keresőkérdést, és a tapasztalatok szerint a VAGY operátor nem működik megfelelően).

Az információkeresést eddig főleg a morfológiai szempontból viszonylag egyszerű angol nyelvre dolgozták ki. A keresőképesség kibővítésekor a nem angol nyelvekre a morfológiai variánsokat figyelembe kell venni. Ha az internetet mindenki számára hozzáférhetővé akarjuk tenni, attól függetlenül, hogy a felhasználók milyen nyelven beszélnek, ki kell dolgozni a megfelelő eszközöket.

/BAR-ILAN, Judit – GUTMAN, Tatyana: How do search engines respond to some non-English queries? = Journal of Information Science, 31. köt. 1. sz. 2005. p. 13–28./

(Szalóki Gabriella)