Pécsi Tudományegyetem
Bölcsészettudományi Kar
Kommunikáció- és Médiatudományi Tanszék

Az információkeresés és a nyelv

Milyen szavakra kerestünk és keresünk

szigorlati dolgozat
a második kommunikációs szigorlathoz

Viszket Anita

Témavezető: Szijártó Zsolt PhD
egyetemi docens
Szak: Kommunikáció egyetemi szak, 1997 előtti változat, minor
Szakirány: Társadalmi kommunikáció szakirány

Benyújtás időpontja: 2008. november 20.

Tartalomjegyzék

Az előírt rövid összefoglalások

1. A dolgozatom témája

A dolgozatomban azt vizsgálom, hogy abban az esetben, ha egy könyvet vagy cikket akarunk megtalálni, milyen keresést preferálunk, és hogy ebben a preferenciában történt-e változás az utóbbi évtizedekben. Amellett érvelek, hogy a XXI. századra előtérbe került a kulcsszó alapú keresés a hagyományos könyvtári típusú keresésekkel, pl. tárgyszókereséssel szemben. Az állításom igazolására egyrészt összehasonlítom két tanulmány megállapításait és következtetéseit a jövő kereséseivel kapcsolatban, másrészt bemutatom néhány kereső működési különbségeit. A jelenségre magyarázatot is keresek, amelyet az információ-áramlás, a jelek, a zaj és a redundancia fogalomkörében vélek megtalálni.

A dolgozatom nagyobb részben a Társadalmi kommunikáció szakirányhoz, kisebb részben a Hálózati kommunikáció szakirányhoz kötődik.

2. A dolgozatom születésének körülményei[1]

A kultúrakutatásból mindig is elsősorban a változások vizsgálata érdekelt, illetve a változásokra való magyarázatkeresés. A megközelítésben preferáltam azokat a módszereket, amelyek a jelenségeket a szöveg-, tartalom- vagy egyenesen a mondatelemzés módszereivel közelítették meg. Az utóbbi években az időm nagy részét adatbázisokkal, az azokban való keresésekkel és nyelvtechnológiai problémákkal töltöttem. Mindezek együtt eredményezték a témaválasztásomat.

A témámhoz a következő kurzusaimat tartom a legrelevánsabbnak:

Ács Péter:	Dinamikus információs szolgáltatások (Web-programozás)
Viszket Anita:	Internetes minidiskurzusok elemzése
Horányi Özséb:	A kommunikációelmélet lehetőségei
Uherkovich Péter	Kommunikációs technológiák, informatika
Havasréti József:	Könyv, írás, szóbeliség, szöveg
Petőfi S. János:	A multimediális kommunikátumok
Nun Zsuzsa:	Az SQL (Structured Query Language) alapjai
Niedermüller Péter:	Társadalomtörténet és kultúrakutatás
Béres István:	A tartalomelemzés

[1] A dolgozatom lektorálását szeretném megköszönni Csatay Richárdnak és Viszket Zoltánnak. Az esetlegesen megmaradó hibákért természetesen nem övék a felelősség.

1. Bevezetés: a témaválasztásról és a célkitűzésről

1.1 A dolgozatom témája és célja

A dolgozatomban a számítógépes rendszerekben megvalósított könyvkereséssel foglalkozom. Ezen belül is a webes felületeken történő könyvkereséssel. Ezen belül pedig elsősorban a könyvtári adatbázisokban és az online könyvesboltok nyilvántartásában való kereséssel. Részben a feldolgozott tanulmányok megállapításai, részben az elérhető könyvkeresők működése alapján úgy látom, hogy az utóbbi évtizedben megváltoztak a keresési szokások ezen a területen. A dolgozatom célja részben ennek a változásnak a bemutatása és bizonyítása, részben a magyarázatkeresés arra, hogy mi okozhatta ezt a változást.

1.2 A dolgozatom felépítése

A *2. fejezetben* megvizsgálom azokat a fogalmakat, amelyekről a dolgozatom szól (nyelv, keresés, információ). A vizsgálatom nem terjed ki minden lehetséges aspektusra és megközelítési módra: csak azokat a szempontokat veszem figyelembe, amelyek a további elemzések során relevánsak lesznek.

A *3. fejezetben* a dolgozatomban felhasznált elméleti keretek legfontosabb jellemzőit mutatom be és bevezetem az általam is használt fogalmaikat.

A *4. fejezetben* ismertetem, hogy véleményem szerint miben áll a keresésben történt változás, a korábbi és a mai keresésnek mik a jellemzői, és melyek a legfontosabb különbségek a két típus között.

Az *5. fejezetben* két tanulmány feldolgozásával igazolom a 4. fejezetben megfogalmazott észrevételeimet.

A *6. és a 7. fejezetekben* különböző webes keresők összehasonlításával, tesztelésével mutatom meg a 4. fejezetben ismertetett különbségeket a kétféle keresési módban.

A *8. fejezetben* fogalmazom meg az általam adott magyarázatot a változásra.

A *9. fejezetben* összefoglalom a dolgozatom állításait és bemutatom az alkalmazhatóságukat egy webes könyv (vagyis dokumentum) -kereső rendszer javasolt módosításain keresztül.

2. A dolgozat alapfogalmainak vizsgálata

Ebben a fejezetben a dolgozatom fő fogalmait vizsgálom meg, és összefoglalom, hogy ezeket a fogalmakat milyen értelemben fogom használni, velük kapcsolatban milyen ismereteket feltételezek, és milyen megállapításokat fogadok el a továbbiakban külön magyarázat nélkül.

2.1 A nyelv

A „mi a nyelv?" kérdés nem a központi témája a nyelvészeti kutatásoknak. Ugyanúgy, ahogy a „mi az élet?" sem központi témája a biológiai kutatásoknak[2]. A nyelv (vagy az élet) alapanyagot szolgáltat ahhoz, hogy az egyes elméleti keretekben a kiválasztott összefüggéseket a nyelvész (vagy a biológus) vizsgálhassa. Ugyanakkor természetesen több definíció is született a nyelv fogalmára, noha a többségüket inkább kommunikáció-, kultúra- vagy beszédkutatók alkották, semmint nyelvészek. Ez nem is meglepő, ha arra gondolunk, hogy míg a nyelvész magát a nyelvet vizsgálja (ezért azt önmagában nem definiálja), az előbb említett diszciplínákban a nyelvhasználat segítségével vizsgálnak más jelenségeket.

A különböző definíciókban a közös részek[3]:

- a nyelv az interakció vagy a kommunikáció eszköze;
- a nyelv önkényes szimbólumokból áll, amelyeket a szokásszerű használat legitimál;
- a nyelv rendszer.

Egészítsük ki ezeket az állításokat a következő megállapításokkal:

- Chomsky[4]: A nyelv mondatok végtelen halmaza, ahol minden mondat véges, de korlátlan hosszúságú, és minden mondat az elemek véges halmazából véges számú szabállyal előállítható.
- Saussure[5]: A nyelvi jel jelölő (vagy hangkép) és jelölt (vagy fogalom) kapcsolata. A jel a kommunikáció alapegysége egy közösségen belül, a nyelv ezért a jelek rendszere.
- Hockett[6]: Az emberi nyelv, szemben az állati kommunikáció eszközeivel, kettős tagoltságú. Azaz a hangoknak nincs önálló jelentésük, de belőlük jelentéssel bíró egységek (szavak) épülnek fel, amelyek struktúrába szervezésével jönnek létre a nyelvi közlemények (pl. a mondatok).
- A nyelvészeti vizsgálatok nagy része, a számítógépes nyelvészeti vizsgálatoknak pedig még nagyobb része expliciten, vagy még gyakrabban impliciten a nyelv írott változatát vizsgálja. A beszélt nyelv sok tulajdonságában különbözhet az írott nyelvtől

[2] A párhuzamra rávilágít pl. Crystal (1998), 492. oldal

[3] Sapir, Hall, Trager definícióinak összevonásából. Forrás: Crystal (1998), 492. oldal

[4] Forrás pl. Alberti (2006), 136. oldal. Chomsky megállapításai közül nyilván nem csak ez a fontos és maradandó, hanem többek között a kompetencia és performancia megkülönböztetése is, de csak a dolgozatomban a későbbiekben fontos meghatározásokat összegzem.

[5] Forrás pl.: Crystal (1998), 506. oldal. Ahogy Chomskynál, úgy Saussure esetében is elmondhatjuk, hogy további fontos fogalompárokat is alkotott, de ezekre most nem térek ki.

[6] Források pl.: Crystal (1998), 494. oldal, Kenesei (2004), 12. oldal.

(beleértve a fonológiai változások írásban jelölt és jelöletlen alakjainak különbségét is)[7].

Összefoglalva a nyelvvel kapcsolatos tudnivalókat: a dolgozatomban egy olyan nyelvfogalmat használok a keresések vizsgálatánál, amely a nyelv írott nyelvi változatát jelenti; ahol megkülönböztethető a betűkből felépülő szavak és a szavakból felépülő megnyilatkozások szintje (mondatokkal valójában nem foglalkozom, csak szavakkal és szintagmákkal); ahol a hangkép és a fogalom közötti kapcsolat önkényes és a használat által legitimált; ahol a megnyilatkozások hossza korlátlan, így a lehetséges megnyilatkozások száma végtelen; ahol a (hangképből és fogalomból álló) nyelvi jelek rendszert alkotnak; és végül, ahol ez a nyelv a most vizsgált kommunikációs aktus, azaz a keresés eszköze lesz. A későbbiekben ezeket a megállapításokat mint adott ismereteket tételezem fel.

2.2 A keresés

A magyar nyelv szótára[8] és A magyar nyelv értelmező szótára[9] szócikkeinek alapján összeállított definíció szerint:

Jelentéslista:

Keres

Körüljárva, körültapogatódzva, szemeit körülhordozva kutat, vizsgál, megtalálni iparkodik valamit.
Két értelemben:

1. tárgyas <Meghatározott személyt, dolgot, aki, amely elveszett, ill. akiről, amelyről nem tudja (pontosan), hogy hol van> megtalálni igyekszik. Például: *keresi az Alberti-Medve-féle grammatikakönyvet*.
2. tárgyas (ritk. tárgy nélkül is) Találni igyekszik <ismeretlen személyt, dolgot, akiről, amiről még azt sem tudja, hogy létezik-e olyan minőségben, amilyenre szüksége van, amilyen neki megfelel>. Például: *keres egy pszichológiai tankönyvet*.

A keresésről megállapíthatjuk, hogy nem csak emberi sajátosság. Valaminek a megkeresése éppen olyan biológiailag motivált cselekvés, mint a mozgás, a táplálkozás stb. A helyváltoztatásra képes élőlények általában többféle dolog keresésére kényszerülnek: táplálék, pihenőhely, fajtársak stb. Amennyiben a *keresés* kifejezést a fizikai objektumok keresésére szűkítjük (vagyis eltekintünk a nyugalom, a biztonság stb. keresésétől), akkor alapvetően két fő típusát különböztethetjük meg a keresési célnak minden élőlény esetében:

1. egy már megismert dolog, helyszín, fajtárs (vagy egyéb élőlény) újra-megkeresése
2. még ismeretlen, de feltételezetten létező dolog, helyszín, más élőlény megkeresése

Ennek a felosztásnak felel meg a szótárakból kiemelt definíció is.

[7] Ez a megkülönböztetés ritkán kerül nyíltan kimondásra, de valójában a legtöbb kutatás mögött ott van, mint alapfeltételezés. Egy példa: Krippendorff (1995), 134. oldal
[8] Az irodalomlistában Czuczor-Fogarasi (1999)
[9] Az irodalomlistában: Bárczi-Országh (1984-1992).

A keresés tárgya mellett a keresés helye alapján is megkülönböztethetjük a kereséseket. Kereshetünk pl. a fizikai térben vagy valamilyen virtuális térben is. A dolgozatomban kizárólag az utóbbi esetekkel foglalkozom: azaz a számítógépes terekben végrehajtott, korábbról már ismert vagy még ismeretlen könyvekre vonatkozó keresésekkel.

Párhuzamba állítva a virtuális terekben való keresést a fizikai terekben való kereséssel, a sikeres keresési eseményhez a következő szükséges ismereteket, képességeket tételezhetjük fel a kereső ágens részéről:

1. Előzetes ismeretek a keresett objektumról (annak felismerhető mintázatairól).
2. Előzetes ismeretek a keresési tér célszerű kijelöléséről.
3. Ismeretek arról, hogy a térben hogyan lehet tájékozódni, útvonalakat követni (mentális térképek építése és használata).
4. Mintázat-felismerési képességek.
5. Az a feltételezés, hogy a keresett objektum nem változtatja a keresés szempontjából releváns mintázatát, és külön ráhatás nélkül nem tűnik el a keresési térből (tárgyállandóság[10]).

Két további megszorítást fogalmazok meg az általam vizsgált keresésekkel kapcsolatban:

- A vizsgálatomból kizárom azokat az eseteket, amikor a keresés célja maga a keresési tevékenység és nem a keresett objektum megtalálása.
- Továbbá kizárom a keresésnek azokat az eseteit, amikor a keresési útvonal ismert, azaz valójában felkeresésről és nem megkeresésről van szó.

A dolgozatomban a felsorolt ismereteket, képességeket tételezem fel a kereső ágensek részéről, illetve a fenti megszorításokat alkalmazom a vizsgált keresések leszűkítésében.

2.3 Az információ

Az információnak rengeteg definíciója van. A definíciók általában három típusba csoportosíthatók[11]:

- azok, amelyek az információt az anyag tulajdonságaként írják le;
- az alapvetően kibernetikai definíciók, amelyek az információ és vezérlés kapcsolatára összpontosítanak;
- illetve az alapvetően szemiotikai definíciók, amelyek az információt elválaszthatatlannak látják a jeltől, és alapvetően a jelentésre koncentrálnak.

Az információt szokás úgy értelmezni, mint hírt vagy ismeretet[12]. Ez a meghatározás is azt hangsúlyozza, hogy az információ leírásánál a kulcsfogalom a felhasználás, az értelmezés.

[10] Pl.: Cianciolo – Sternberg (2007), 28. oldal

[11] Lásd B. Hajdu – Babiczky (1998), 13-14. oldal. Azzal a témával, hogy az információ a harmadik vagy negyedik alapfogalom-e (az anyag, energia, illetve az energia, idő, tér mellett), illetve, hogy ugyanolyan létszükséglet-e, mint a levegő, víz, táplálék, hajlék stb., most nem foglalkozom. Csakis azokat az elemeit emelem ki az információnak, amelyek a dolgozatom szempontjából relevánsak.

[12] Lásd Fülöp (1984), 5-6. oldal

Ebből egyenesen következik, hogy nincs önmagában való információ, hanem csak egy jól meghatározott rendszeren belül beszélhetünk információról[13].

Kétféle módon dönthető el, hogy valami informál-e vagy sem. A megkülönböztetés alapja a halmazmegadás módszere, azaz az a módszer, ahogy megadhatom egy halmaz elemeit[14]. Egyrészt felsorolhatom azokat a jelenségeket, amelyek információhordozók. Ha a könyvkeresés témakörében maradunk, pl. felsorolhatom a lehetséges ETO-jelzeteket vagy a tárgyszavakat. Másrészt megadhatok egy szabályrendszert, amellyel eldönthető, hogy egy jelenség információhordozó vagy sem. Ismét a könyvkeresés példakörében maradva, ha megadom a tartalomindexelés szabályait, akkor az indexadatbázisban szabályszerűen létrejött index-kifejezések (kulcsszavak) információhordozók lesznek, míg más, véletlenszerűen választott szó, amelyet az indexelés szabályrendszere nem „állít elő", nem lesz az.

Ezekből a példákból is következik, hogy a dolgozatomban úgy tekintem, hogy a könyvek számítógépes rendszerekben való megkeresésének feltétele a könyvekről szóló információ elérhetősége. Az információ az, hogy melyik könyv milyen kereső-kifejezés megadásával érhető el. A könyvhöz rendelt adatokat jeleknek tekintem, amelyek egyik fajta jelentése nem más, mint a könyvre való rámutatás. Mindebből az is következik, hogy az információnak a szemiotikai megközelítését fogom alapul venni.

3. A használt elméleti keretek

Ebben a fejezetben az általam használt megközelítéseket mint elméleti kereteket mutatom be: a számomra most releváns alapfogalmaikat, megállapításaikat, illetve a dolgozatom tárgyának az ezekben a keretekben értelmezhető jelenségeit foglalom össze.

3.1 Nyelv és keresés – az információkereső nyelvek

Az információkereső nyelveket többféle szempontból lehet csoportosítani[15]. Lehetséges két fő csoportba osztani ezeket, annak alapján, hogy a „fogalomkészletüknek" mi az alapja: lehetnek természetes (tárgyszavak, tezauruszok) és mesterséges (ETO, Cutter-jelzet) nyelv alapúak[16]. Egy másik lehetséges osztályozás szerint vannak hierarchikus (ETO, Dewey-féle Tizedes Osztályozás), mellérendelő (tárgyszó) és szemantikus vagy relációs (tezaurusz) információkereső nyelvek[17]. A lényeges szempont a dolgozatomban az, hogy az

[13] Lásd Horányi (2006), 13-16. oldal. Stehr (2007) részletesebben vizsgálja az információ fogalmát, de számára a lényeges kérdés az, hogy mi a különbség információ és tudás között. Stehr összehasonlítja például Bell és Dosi információ-fogalmát, és úgy látja, hogy Dosi éppen úgy definiálja az információt, ahogy Bell a tudást: a tényekről szóló állítás szervezett rendszereként. Stehr saját információ-fogalma közel áll ahhoz az értelmezéshez, amit én használok a dolgozatomban. Ő az információt nem tekinti annyira szituációs jellegűnek, mint a tudást. A konkrét megfogalmazása az, hogy „Az információ csupán reflektál azokra a termékekre, amelyekből absztrahálták.".

[14] Az ezen a módon való megkülönböztetés ötletének forrása: Horányi (2006), 13-16. oldal

[15] Az említett osztályozási rendszereket összefoglalja B. Hajdu – Babiczky (1998), 16-20. oldal.

[16] Lásd Siró (2007), 77-84. oldal

[17] Lásd Ungváry – Vajda (2002), 24-34. oldal

információkereső nyelvek esetében mindig azt a módszerét választjuk az információ-hordozók definiálásának, hogy felsoroljuk azokat az elemeket, amelyeket információ-hordozóknak tekintünk. Vagyis azok a szavak, amelyek nem szerepelnek az előbb említett osztályokban (azaz az ETO-ban, a tezauruszban, a tárgyszavak között stb.), az információkereső nyelvek használata esetében a könyvkeresés szempontjából nem számítanak információhordozónak. Még pontosabban fogalmazva: ezeket a nem információhordozó szavakat nem használhatjuk a keresésnél, mert ezek használatával nem kaphatunk találatot.

3.2 Nyelv és információ – szemiotika

Az információval kapcsolatban leszögeztem, hogy az információt elsősorban a szemiotikai oldaláról fogom megközelíteni.

A jelek modern kutatását a legtöbben Saussure munkásságától eredeztetik[18]. Saussure-nek a jelekkel kapcsolatos egyik alapvető definíciója az, hogy a jelet a jelző (vagy jelölő) és a jelzett (vagy jelölt) együtt alkotja. Jelölő és jelölt kapcsolata Saussure szerint (aki elsősorban a nyelvi jelet vizsgálta), konvencionális és önkényes.

A jelentésnek is két aspektusát szokás megkülönböztetni: a denotációt és a konnotációt (a megkülönböztetés bevezetése Hjelmslev nevéhez fűződik[19]). A denotációt általában első (közvetlen) jelentésnek, míg a konnotációt második (közvetett) jelentésnek szokás tartani[20].

Már Saussure is hangsúlyozta, és azóta is minden szemiotikai kutató különböző szempontokból ismét megerősíti, hogy a jel soha nem magában áll, hanem egy rendszer eleme. Barthes több szemiotikai rendszert írt le[21], miközben a szemiotikai rendszerek közös tulajdonságait igyekezett feltárni. A szemiotikai rendszerek egyik általános, minden ilyen rendszerre igaz sajátosságának tartotta Barthes a konnotatív eltolódást: vagyis amikor az eredeti jelhez (jelölő és jelölt) egy új, másodrendű szemiotikai rendszer társul (konnotáció). Barthes ezeknek a konnotatív oldallépéseknek általában ideológiai, mítoszképző funkciót tulajdonít. Míg Barthes a jelek ilyen módon való lebegésére, bizonytalanságára hívja fel a figyelmet, addig Eco a kiterjesztés kérdésére[22]: vagyis arra, hogy a jelölt nem egyszer maga is jel, ami az egész szemiotikai rendszert egy enciklopédiához teszi hasonlatossá, ahol az egyik szócikk egy másikra hivatkozik.

Peirce modelljében kissé máshova esnek a hangsúlyok[23]. Peirce-ről mindenkinek a jelölő és jelölt kapcsolatának hármas csoportja jut eszébe: szimbólum, ikon, index. De ugyanennyire

[18] Pl. Griffin (2003), 335. oldal
[19] Lásd Gripsrud (2007), 105. oldal.
[20] A jelentésdimenziókat természetesen ennél sokkal finomabban is osztályozzák, lásd pl. Osgood faktoranalízise (Rosengren (2008), 77-78. oldal). De az egyes megemlített diszciplínáknak csak azokra a fogalmaira térek ki, amelyekre visszautalok további bekezdésekben vagy fejezetekben.
[21] Griffin (2003), 334-344. oldal
[22] Anderson (2005), 52-53. oldal. Eco „határtalan szemiózis”-áról lásd pl. Szőnyi (2004), 229-242. oldal
[23] Lásd Gripsrud (2007), 110-113. oldal

fontos, hogy Peirce-nél meghatározó szerepet kap az, aki a jelet értelmezi. Peirce vég nélküli szemiózis-modellje azt ábrázolja, ahogy állandóan új és új interpretációkat hozunk létre.

Ebben a rövid összefoglalásban a dolgozatom szempontjából a legfontosabb az, hogy amennyiben jeleknek a könyvekhez rendelt, a keresést lehetővé tevő adatokat tekintem (pl. a tárgyszavakat és a tartalomindexeléssel létrejövő kulcsszavakat), akkor milyen jelentést tudok ezeknek tulajdonítani?

Nos, azt feltételezem, hogy ezeknek a tárgy- és kulcsszavaknak az elsődleges jelentése (denotációja) az, amit a szó önmaga jelent. Vagyis a szónak az a jelentése, amit róla a szótárban megtalálnánk. De, ahogy korábban, az információról szólva már írtam, feltételezem, hogy ezeknek a jeleknek van egy másik jelentésük is, amit a könyvnyilvántartásban kaptak, ez pedig nem más, mint bizonyos könyvekre való rámutatásuk, illetve rá nem mutatásuk. Ez a jelentésük folyamatosan változik (úgy, ahogy az adatrögzítések változnak, illetve szaporodnak), és ebben az értelemben a tárgyszavaknak és kulcsszavaknak mint jeleknek a konnotációjuk bizonytalan, lebegő.

Visszatérve az információ fogalmához: az adatbázisokban tárolt könyvekről a rekordokhoz rendelt kulcs- és tárgyszavak mint jelek és az ő jelentésük (a könyvhöz való rendelésük) együttesen adnak információt. Így a könyvről szóló információ nem csak a jel, az adat, hanem a jel jelentése, értelmezése is, egy adott keretben, a könyvnyilvántartás keretében.

3.3 Információ és keresés – információelmélet

Mint azt az információ fogalmánál már láttuk, mindenekelőtt azt kell definiálnunk, hogy mi az a rendszer, amin belül az információt értelmezzük, illetve melyek azok a jelenségek, amelyeket információhordozónak fogunk tekinteni.

A számítógépes könyvkeresés mint információs rendszer valójában két alrendszerből épül fel.

Az egyik a könyvek nyilvántartása, és ilyen módon kereshetővé tétele. Ebben az alrendszerben az adatbázis (amiben a könyvek nyilván vannak tartva), illetve az ehhez hozzáférő keresőfelület és az a mögött működő keresőprogram képezik az információs rendszert. Információhordozónak azokat az egységeket tekintünk, amelyeket vagy mint egy véges készlet elemeit, vagy mint egy szabályrendszerrel előállított halmaz elemeit a könyv-adathoz rendelünk. Az elsőre példa a tárgyszavak rendszere, a másikra a tartalomindexeléssel előálló kulcsszó-nyilvántartás.

A másik információs alrendszer a találati lista kigenerálása, megmutatása és a kereső ágens által való értelmezése. Ebben információhordozónak a képernyőn megjelenő adatsorokat tekintjük, amelyeket egy előre definiált szabályrendszerrel jelenítünk meg. A szabályrendszerben definiáljuk a megjelenítendő adatokat, azok formáját és a megjelenítésük sorrendjét.

Az információk kezelésében két nagyon fontos fogalomra ki kell még térnem: az egyik a zaj, a másik a redundancia[24].

Zajnak azt tekintjük, ha az információáramlás során az információ valahogyan sérül. Ennek lehet szűkebb (technológiaibb) és tágabb (általában az információátvitel sikerességére vonatkozó) használata is. Én a zaj fogalmát a tágabb értelemben használom. Ebben az értelemben a zajnak oka lehet az, hogy irreleváns információk is bekerülnek a közlés folyamatába, vagy az, hogy bizonyos információk elvesznek, kimaradnak, esetleg hibásan jelennek meg. Mi lehet a zaj a számítógépes könyvkeresés kapcsán? A kérdésre akkor tudunk válaszolni, ha az előbb megkülönböztetett két információs alrendszert külön választjuk.

Az adattárolásban, illetve az adatoknak a keresőfelületen keresztül történő elérhetővé tételében a zaj lehet az, ha egy könyvhöz nem a megfelelő jel van társítva, vagyis vagy kevés, vagy túl sok az adott könyvhöz a tárgyszó: nem minden releváns tárgyszó szerepel a könyvnél, vagy éppen olyanok is szerepelnek, amelyek az adott könyvnél nem relevánsak. Zaj lehet még a rosszul begépelt tárgyszó vagy a rosszul begépelt kereső-kifejezés.

A találati listánál mint alrendszernél zaj lehet a túl sok találat, köztük irreleváns elemekkel, vagy a nem szerencsés megjelenítési forma, amiből a kereső ágens nehezen bogozza ki, hogy melyik könyvről van szó.

Redundanciának azt tekintjük, amikor ugyanazt az információt (jelentést) több jel is hordozza. Ennek két haszna is van. Az egyik az, hogy a redundancia segít a zaj elleni küzdelemben. Ha ugyanaz a jelentés több jelnél is tárolódik, akkor az egyik egység sérülése vagy el nem érhetősége esetén az információ még mindig eljuthat az azt keresőhöz. A másik haszna a redundanciának, hogy kevésbé követeli meg a maximális figyelmet attól, aki az információt kapja. Minél kevesebb jel hordozza ugyanazt az információt, annál nagyobb figyelem kell a megértéshez.

Ezt a két előnyt valójában egynek is tekinthetjük, ha a zaj fogalmába beleértjük annak a figyelmét, megértő- és koncentrálóképességét is (pontosabban annak csökkenését vagy hiányát), aki az információt kapja.

A redundancia a könyvkeresés esetében az adattárolás szintjén nyilvánvaló: minél több tárgyszóval vagy kulcsszóval közelíthető meg ugyanaz a könyv, minél több jel hordozza az ugyanarra a könyvre való rámutatást, annál redundánsabb a rendszer, és annál valószínűbb, hogy a keresést nehezítő tényezők ellenére a keresés sikeres lesz. A találati lista megjelenítésénél a redundancia két dolgot is jelenthet. Egyrészt azt, hogy ugyanabban a találati listában több találati sor mutat ugyanarra a könyvre: ez, mint tudjuk, nem jellemző a könyvkeresőknél. De az igen, hogy több, különböző keresés esetén is megjelenhet ugyanaz a könyv a találatok között – ez a találati lista redundanciájának másik lehetséges formája.

[24] A fogalmak meghatározásának forrása: Fülöp (1983), 16-20. oldal

Összefoglalom egy táblázatban az általam értelmezett zaj és redundancia fogalmakat a keresők rendszerében:

	információs rendszerek a keresésnél	
	Első információs alrendszer: az adattárolás és a keresés	Második információs alrendszer: a találati lista
zaj forrása	1) nem jó index-rekord (nem illik a könyvhöz vagy rosszul van begépelve vagy hiányzik a hozzárendelés) 2) rosszul megadott kereső-kifejezés	1) irreleváns elem a találati listában vagy hiányzó elem a listából 2) rossz, értelmezhetetlen adatmegjelenítés
redundancia forrása	több index-rekord mutat a könyvre	több keresésnél előjön ugyanaz a könyv

1. táblázat: a keresés mint információs rendszer; zaj és redundancia

A dolgozatomban arra mutatok rá, hogy az előbb definiált fogalmak értelmében a tárgyszólista és a kulcsszólista használata különböző méretű redundanciával és különböző méretű zajjal jár. Kérdés, hogy melyik használata éri meg jobban, melyik biztosít sikeresebb keresést. Természetesen a megvizsgált néhány keresőprogram esetében nem tudom számszerűsíteni a zaj és redundancia arányait: csak feltevéseim lehetnek néhány tesztelés, illetve a racionális meggondolások alapján. Mindenesetre a dolgozat végére levont következtetés (hogy melyik változatnak kellene a sikeresebbnek lennie) egybecseng a gyakorlattal (hogy melyik változat sikeresebb). Ezt valamilyenfajta bizonyítéknak is tekinthetjük a feltevésemre vonatkozóan.

3.4.1 Számítógépes keresés és nyelvtechnológia

A dolgozatomban a számítógépes könyvkeresésnek azokat az eseteit vizsgálom, amikor természetes nyelven adunk meg egy kereső-kifejezést, és erre keresünk a nyilvántartásban. Ez vonatkozhat a szerzőre, a címre, a tárgyszóra, a témakörre stb.

A természetes nyelvi kifejezést használó számítógépes kereséseknél egy kisebb sztringet (azaz karaktersort) igyekszünk megtalálni egy nagyobban: körülbelül ez az alapelv. Mivel elképzelhetetlen, hogy elég gyorsan lefusson a keresés, ha annak a kereső-kifejezés beírása után kell a teljes nyilvántartást (beleértve a tárolt könyvekhez tartozó tartalmakat is) végigböngésznie, ezért a keresés előre elkészített index-adatbázisokban történik.

Az kulcsszókeresés klasszikus esetében, az internetes kereséseknél például alapvetően három alkotóelem szükséges[25]:

- egy olyan program, ami a keresés előtt végigböngészi az Interneten található dokumentumokat, összeszedi belőlük az információt, és készít egy adatbázist arról, hogy milyen kulcsszóra milyen weboldalt érdemes megmutatni;
- ez az összeállított adatbázis (vagyis index – hasonlóan a könyvek végén található indexhez);
- az a program, amely a keresőprogramba beírt kereső-kifejezést összeveti az indexszel, és ennek alapján összegyűjti a megmutatandó weboldalakat, valamint azt is, hogy milyen sorrendben érdemes ezeket megmutatni.

A probléma, ami miatt a nyelvtechnológia említése releváns, az az, hogy nagyon nem mindegy, milyen szavak kerülnek az indexbe: ugyanis ezeket a szavakat lehet majd egybevetni a kereső felhasználó által megadott kereső-kifejezéssel.

Index kétféleképpen készülhet[26]:

- a dokumentum (pl. weboldal, vagy pedig a nyilvántartott könyv rekordja) alkotója maga megadja azokat a kulcsszavakat, amelyekre keresve szeretné, ha a weboldala, dokumentuma megtalálható lenne,
- vagy az indexkészítő segédprogram a dokumentum tartalma alapján próbálja kitalálni a megfelelő, indexben tárolandó kifejezéseket.

Mindkét esetben gondot okoz az, hogy pontosan milyen szavakat vagy szintagmákat tárolunk az indexben. Az előre adott tárgyszólisták esetében a szavak-szintagmák készlete adott: ezt könyvekkel foglalkozó szakemberek állítják össze, nem egy információkereső nyelv esetében nemzetközi szinten. Az automatikus (vagyis gépi) tartalomindexelés esetében a szólista nem adott előre, azaz az indexelő programnak kell eldöntenie, milyen szót (kifejezést) milyen formában fog az indexadatbázisba beemelni. Az egyik lehetséges megoldás, hogy szavakat indexelünk: azaz a dokumentum összes szavát egyenként elhelyezzük az index-adatbázisban.

A szavak indexelésénél két problémát kell megoldanunk. Az egyik az, hogy mit tegyünk azokkal a szavakkal, amelyek minden dokumentumban előfordulnak (például a névelők), a másik, hogy mi legyen a különböző toldalékolt alakokban előforduló szavakkal. Hiszen a szövegben a szavak toldalékolt alakja fordul elő, a kereső-kifejezésekben viszont az esetek nagyobb részében toldalékolatlan (vagy legfeljebb többes számú) főnévi alakokat használunk. A magyarban súlyosbítja ezt az utóbbi problémát az, hogy a toldalékoláskor sok szónak a töve is megváltozik. Ennek a legtipikusabb esete az *a*/*e* hangok írásban is jelölt rendszerszerű megnyúlása a szavak végén a toldalékok előtt (*alma* – *almák*, *kefe* – *kefét*). Mint korábban írtam, a számítógép sztringeket, vagyis konkrét karaktersorokat hasonlít össze, és számára az *a* és az *á* vagy az *e* és az *é* két különböző karakter.

[25] Rengeteg olyan szakkönyv van, amely a keresőprogramok működésével foglalkozik, és természetesen a legleegyszerűsítettebb formában ugyanezt a felosztást ismertetik valamennyien. Egy konkrét példa: Battelle (2006), 23-28. oldal. Már az oldalszámokból is látszik, hogy a lehető legegyszerűbb leírást én tovább egyszerűsítettem: csak annyi információt vezetek be, amennyi a mondanivalóm megértéséhez szükséges.

[26] Prószéky-Kis (1999), 178. oldal. Az indexelés nyelvészeti problémáinak listáját Prószéky-Kis (1999) alapján állítottam össze, jelentősen leegyszerűsítve az ő megállapításaikat.

Ami az első problémát illeti, nyilván érdemes azokat a szavakat kihagyni az indexelésből, amelyek gyakorlatilag minden dokumentumban szerepelnek. Ilyenek például a névelők, a kötőszavak, a névmások stb. Vagyis össze kell állítani a nemkívánatos szavak listáját[27].

A másodikként említett probléma megoldása nehezebb ennél. A kereső-kifejezésekben megadott és az indexben tárolt szavak összehasonlításakor szerencsés lenne, ha az összehasonlítandó sztringek mindig szótövek lennének. Ehhez mindkét esetben szótövesíteni (azaz lemmatizálni) kellene. Kérdés, hogy erre valóban szükség van-e. A kérdésre az indexelés egy újabb kérdésének megtárgyalása után visszatérek.

Az internetes keresések során nagyon fontos a szószint fölötti sztringek, azaz a kifejezések indexelése is. Ha csak szavakat indexelünk, akkor pl. a több szóból álló nevekre rákeresve nem fogunk jó találatokat kapni. A legtöbb keresőben az idézőjelek közé tett kereső-kifejezés nem lenne használható, hiszen csak az egyes szavak lennének eltárolva az index-adatbázisunkban, a szavak egymáshoz való közelsége, sorrendisége nem. Ezért szükség van a kifejezések indexelésére is. A számítógép nem csak a szó, de a szintagma, a tagmondat, a mondat fogalmakat sem ismeri. Vagyis a kifejezések indexelésénél egyszerűen csak azt mondhatjuk, hogy „több szó indexelése". Nyilván meg kell határoznunk a felső szóhatárt, amire indexelni fogunk.

Mivel nincs értelme teljesen összefüggéstelen szóláncokat indexelni, a legtöbb indexelő-eljárás korlátozott kifejezés-indexelést használ: vagyis nem automatikusan minden szóláncot vesz fel az indexfájlba, ami a maximális határon belül van, hanem igyekszik ezt bekezdésekre korlátozni, vagy címekre, nevekre stb. Ebben részben a dokumentum szerkesztésében megtalálható információkra támaszkodik az indexelőprogram, részben nyelvtechnológiai eszközökre (pl. a nevek megtalálásánál).

A kifejezés-indexelés az oka annak, hogy nem okoz problémát, ha több szóból álló kereső-kifejezésekben toldalékolt szavakat használunk: ezek a kifejezések ugyanis általában éppen így, kifejezésként kerülnek be az index-adatbázisba.

Visszatérve a szótövesítés problémájára: Prószéky-Kis (1999) amellett érvel, hogy a legelőnyösebb, ha a kereső-kifejezés feldolgozásakor is és az indexeléskor is lemmatizálunk. Ebben az esetben szótövek vannak az indexadatbázisban és szótövekre van lefordítva a kereső-kifejezés, könnyű tehát a kettőnek az egybevetése. Ugyan arról nincs információ, hogy az egyes nagy keresőprogramok pontosan hogyan működnek az indexelés és a kereső-kifejezés feldolgozásának szempontjából (ezek a megoldások mindig az üzleti titok részét képezik), de ha figyelembe vesszük azt, hogy toldalékolt szóra leginkább csak többszavas kifejezések esetében keresünk (aminek a visszakeresését a kifejezés-indexelés megoldja), akkor beláthatjuk, hogy erre valójában nincsen nagy szükség. A lemmatizálás (vagy stemming) amúgy sem kifejezetten grammatikai alapokon szokott megtörténni, inkább statisztikai alapon: ha a szót alkotó sztringnek egy nagy valószínűséggel előforduló részletét

[27] A Google-n itt lehet erről leírást találni: http://www.google.com/support/bin/answer.py?answer=981&ctx=sibling&topic=13912. Az Index.hu oldal fórumkeresőjének is van egy „nemkívánatos szavak" listája, itt: http://forum.index.hu/Custom/showStopwords (letöltések ideje: 2008. szept. 25.).

reguláris kifejezésként keressük[28], akkor ez által a tőváltakozást nem használó nyelvekben gyakorlatilag lemmatizáltunk. A magyarban leggyakoribb tőváltakozás, a szótövek végén álló *a*/*e* nyúlása pedig megoldható úgy, hogy a kereséskor nem a szótőig csonkolunk, hanem ha az *a*/*e*-re végződik, akkor egy karakterrel tovább[29]. Egy másik lehetséges megoldás az ékezetmentesítés: amikor a reguláris kifejezést úgy fogalmazzuk meg, hogy minden ékezetes karakter helyett keresse annak az ékezetmentes párját is. Így működik például az Acrobat Reader keresője[30].

Összefoglalva tehát a természetes nyelvi kifejezéseket használó számítógépes könyvkeresések fő jellemzőit, azt mondhatjuk, hogy a keresés előkészítéseként mindenképpen el kell látnunk a könyv-rekordunkat azokkal a szavakkal, amelyek jelekként a megfelelő információt fogják majd a könyvünkről hordozni, és ezen információk alapján a könyv kereshetővé válik.

Ez a jelekkel való ellátás gyakorlatilag az indexadatbázis feltöltésével történik. Ez történhet úgy, hogy egy előre megadott szókészletből válogatva jelöljük meg a könyveinket (ilyenkor valamilyen információkereső nyelv kifejezéseit használjuk), vagy úgy, hogy egy indexelő program találja ki, hogy az adott könyvhöz milyen kulcsszavak tartozzanak. Ez az indexelő program kiemelheti a könyv tartalmából a szavakat (ha tároljuk az egész könyvet), a könyv leírásából a szavakat, illetve a könyv címéből és szerzőjéből a megfelelő szavakat. Nyilván praktikus, ha nem csak szó-, hanem kifejezés-indexelést is végez az indexelő programunk, különösen a cím és szerző adatokra vonatkozóan.

3.4.2 Számítógépes keresés és a keresőfelület felismerhetősége (user interfész)

Amikor a keresések változását vizsgálom, noha elsősorban a kereső-kifejezések használatára, illetve ezeknek viszonyára koncentrálok a könyvekhez tárolt információkra vonatkozóan, mégis ki kell térnem, ha nagyon röviden is, a keresőfelületek vizsgálatára. A keresőfelületek tájékoztatják ugyanis arról a kereső ágenst, hogy mit és hogyan tud keresni a nyilvántartásban.

Kétféle alapvető keresési módot különböztethetünk meg: az egyszerű és az összetett keresést.

Az egyszerű keresésnél kettő vagy három keresés-aktust[31] lehet megvalósítani. Az egyik mindenképpen a természetes nyelvi fogalom vagy fogalmak beírása egy keresőmezőbe

[28] Pl. *dietary* helyett *diet**, vagy *múzeumokban* helyett *múzeum**.

[29] Például: *alma* helyett *alm** sztringgel hasonlítunk, és így az *alma*-nak az összes toldalékolt alakját megtalálhatjuk, azokat is, amelyekben az *almá* tő szerepel. A szótövet statisztikai alapon szokták megállapítani: a szavak elején leggyakrabban előforduló sztringekből. Természetesen fontos, hogy az *alma* szó esetében ne az *al* karaktersort tekintsük a szótőnek, noha az is gyakran előfordul a szavak elején, de ennek megakadályozására is vannak különböző statisztikai módszerek.

[30] Az ékezetmentesítés a bal- és jobboldali csonkolással ötvözve eredményezheti azt, hogy ha például a *keres* szót adjuk meg kereső-kifejezésként, akkor megtaláljuk a *keres*, *kérés*, *tekerés*, *keresés* stb. szavakat is.

[31] Keresés-aktus: hymesi értelemben, a beszédaktus mintájára, a keresési esemény legkisebb alapegysége (Hymes (1997)).

(editbox[32]). A beírt fogalomnak vagy fogalmaknak megtalálhatóknak kell lenniük az indexben (ami lehet szerző-, tárgyszó-, kulcsszó- stb. index), ami jó fogalom-kiválasztási képességet vár a kereső felhasználótól. A másik kötelező keresés-aktus a keresés megindítása, a legtöbb esetben egy Keresés feliratú gombbal, aminek a helye kötött a keresőmezőhöz képest. A harmadik opcionális keresés-aktus a „hol keres", illetve a „hogyan jelenít meg" paraméterezése. A szokásos formái: checkbox[33] vagy külön keresőmező (editbox) a természetes nyelvi fogalom beírásához, esetleg combobox[34] (a „hol keres" esetében).

Az összetett keresésnél annyira különbözőek a keresőfelületek, hogy nem lehet általánosításokat megfogalmazni sem a keresés-aktusokra, sem azok formájára vonatkozóan. Az ilyen felületeken külön kell meghatározni az adattároláskor releváns mezőkben keresendő adatot (pl. külön a szerzőt, külön a címre vonatkozó kereső-kifejezést stb.) Ebből is következik, hogy a felhasználók csak a legritkább esetekben használják ezeket a felületeket[35].

A keresőprogramok a webes programoknál használatos beviteli mezőkkel élnek, így ezek működése minden esetben azonos (editbox, combobox, checkbox, radiobutton[36], gomb stb.). Az egyszerű keresésnél ezeknek nem csak a formája, hanem a funkciója is mindig azonos, így nem lehet kétséges az értelmezésük. Az összetett kereséseknél azonban az adatstruktúra és az adatokat leíró nyelv ismerete szükséges lehet ahhoz, hogy a felület elemeit értelmezni tudjuk[37]. Ha nem tudjuk pontosan, hogy milyen adatokat tárolunk a kulcsszó mezőben, és ennek van-e összefüggése a dokumentum szerzőjével, akkor nem értelmezhető annak a keresés-aktusnak a jelentése, amikor a kulcsszó kategóriájú keresőmezőbe beírt kifejezés mellett megnyomjuk a keresőgombot. Ezért nem csak a felület mintázatának azonosításánál ütközik nehézségekbe a felhasználó, hanem a felület jelentésrétegeinek feltárásánál is. Az összetett keresők felületeinek igazán sikeres felhasználói az adatok rögzítői, akik pontosan tudják, hogy milyen adat milyen adatmezőbe kerül.

Vagyis összefoglalva a keresőfelület nyújtotta lehetőségeket: az egyszerű keresés amellett, hogy a felülete nagyon egyszerű (Krug (2008)[38] szerint a keresőfelület sémája mindössze ennyi: „egy keresőmező, egy gomb és egy 'keresés' szó"), nem kívánja meg a kereső ágenstől, hogy tisztában legyen azzal, hogy melyik kereső-kifejezést melyik beviteli mezőbe kell beírni a keresőfelületen. Ezzel szemben az összetett keresés döntést vár a kereső

[32] Egysoros beviteli mező.
[33] Jelölőnégyzet.
[34] Legördülő menüből választás.
[35] Battelle (a Google-ról és a keresésről szóló könyvében) hosszan méltatja az összetett keresés funkciót, hogy mennyivel biztosabban lehet vele elérni a jó találatokat, majd a fejezet végén megjegyzi: „őszintén szólva, a többséghez hasonlóan én is lusta vagyok használni ezeket". Ez azt jelenti, hogy az összetett keresések még egy olyan, a keresésben igazán képzett szakembernek is nehezére esnek, mint Battelle. Nem csodálkozhatunk hát azon, hogy (az általa idézett statisztikák szerint) a felhasználók több, mint 95%-a soha nem használ összetett keresés funkciót. (Battelle (2006), 28. oldal)
[36] Több alternatív lehetőségből egynek a kiválasztására lehetőség kör alakú jelölőmezőben.
[37] Ez az elvárás expliciten is megjelenik olyankor, amikor a kereső-kifejezés összeállítása után a felület megmutatja annak fordítását a keresőprogram számára értelmezhető formában. Gyakran lehet találkozni ezzel a funkcióval igazán összetett keresési felületeken. Előfordul, hogy nem csak a „lefordítás", hanem a kereső-kifejezés javítása is megjelenik. Ez a javítás valójában annak a funkciónak felel meg, amit más keresők az index láthatóvá tételével oldanak meg: hogy lehessen tudni, milyen kulcsszavakra érdemes keresni.
[38] Krug (2008), 77-78. oldal

ágenstől azzal kapcsolatban, hogy melyik kereső-kifejezést hova írja be, ráadásul az összetett keresés keresőfelületére vonatkozóan nem adható meg általános képi séma sem, ami a felület felismerését és használatát megkönnyítené.

Összefoglalva az eddig elmondottakat: mivel a kereső ágensek igen nagy százaléka az egyszerű keresés felületeket használja, ezért a dolgozatomban, amennyire lehetséges a könyvkereséssel kapcsolatban, az ilyen felületekre korlátozom a vizsgálatomat. Ahol ez nem lehetséges, mert a keresőfelület nem biztosít ilyen lehetőséget, ott a tárgyszó vagy a kulcsszó megadására alkalmas keresőmezőt igyekszem használni a vizsgálatomban.

4. A tárgyszókereséstől a kulcsszókeresésig

Az előző fejezetekben megállapítottam, hogy a számítógépes könyvkeresést lehetővé tévő előkészítő munka az, hogy a könyv-rekordot indexekkel látjuk el. Definiáltam, hogy ezeket az indexeket (szavakat és kifejezéseket az indexadatbázisban, amelyeket a könyv-rekordhoz rendelek) jeleknek tekintem. A vizsgált jelek jelentése egyrészt a szó jelentése, másrészt a könyvekre való rámutatásuk, amikhez hozzá lettek rendelve.

Az indexadatbázis kétféleképpen állhat elő: vagy a szavak előre definiálásával (ebben az esetben az információkereső nyelvek szavairól és kifejezéseiről beszélünk) vagy a könyv-rekord adatainak gépi indexelésével. Ez utóbbit következetesen kulcsszó-indexnek neveztem.

4.1 A különbség a tárgyszó és a kulcsszó között

A szakirodalomban nem egyértelmű a tárgyszó és a kulcsszó megkülönböztetése. Kulcsszón nagyon gyakran a könyv minden metaadatát értik (ez lehet a szerző, a cím vagy a tárgyszó, esetleg témakör stb.). Vagyis ebben a használatban a kulcsszó a lehetséges különböző indexek összefoglaló kategóriája.

Azért van néhány szakirodalom, ahol következetesen megkülönböztetik a tárgyszót a kulcsszótól: én ezek gyakorlatát követem a dolgozatomban. A megkülönböztetés néha nem egyéb egy újabb adatmező felvételénél[39], más esetekben viszont egyértelműen a megkülönböztetés alapja a szó forrása: azaz az, hogy míg a tárgyszót egy kötött halmazból választják ki, a kulcsszóra vonatkozóan nincsen ilyen zárt halmaz, hanem a forrásául a könyv szerző, cím, leírás mezőiben található szavak szolgálnak[40].

Vagyis a dolgozatomban a tárgyszót úgy értem, hogy egy információkereső nyelv egyik kifejezése, amelyet a könyv feltárását végző szakember a könyvhöz rendel, a kulcsszót pedig úgy, hogy egy olyan természetes nyelvi kifejezés, amit egy előre szabályozott módon működő indexelő program rendel a könyvhöz, a könyv adatleírásában található sztringekből válogatva. Mindkét esetben indexadatbázis készül, csak más-más tartalommal megtöltve.

[39] Lásd Drótos (1994)
[40] Lásd Drótos (1999)

4.2 A kétféle keresés összehasonlítása

A tárgyszó alapú és a kulcsszó alapú keresések között van legalább három nagy különbség. De van egy azonosság is, és első lépésként erre mutatok rá.

Ahogy már a számítógépes keresés és nyelvtechnológia ismertetésénél írtam, a keresés mindig sztringhasonlítást jelent. Vagyis a keresés úgy indul, hogy a kereső ágens begépel egy kereső-kifejezést (ami állhat egy vagy több szóból), megindítja a keresést, és a keresőprogram összehasonlítja a beírt kereső-kifejezést vagy annak elemeit (vagyis az abból lefordított keresőkérdést) az egy vagy több, a könyvekhez tartozó indextáblában. Az összehasonlítás szabályai természetesen függenek attól, hogy tárgyszóra vagy kulcsszóra kerestem-e, de az alapvető rendszer azonos. Ebből következően vannak azonos zaj-problémák is: hibásan gépelt szavak az indexben, hibásan gépelt kereső-kifejezések, téves (felesleges) szavak az indexben stb.

A fejezet további részében a különbségeket mutatom be a kétféle keresés között.

4.2.1 Hol keresünk

A két fogalom definiálásánál már leírtam, de hasznos lehet kiemelni, hogy nagy különbség a tárgy- és kulcsszó között, hogy míg a tárgyszó az egyik adata a könyvnek (a szerző, a cím stb. mellett), addig a kulcsszó (legalábbis az általam használt jelentésben) a könyv összes adatából létrehozott index, vagyis önmaga tartalmazza a szerzőt, a címet, a tárgyszót akár, és az egyéb adatok szavait. Ebből az is következik, hogy a kulcsszó alapú keresés szinte minden esetben egyszerű keresés, míg a tárgyszó alapú keresés szinte mindig összetett keresés kell, hogy legyen.

Vagyis, ha tárgyszóra keresünk, akkor nem érdemes beírni a (tárgyszó megadására alkalmas) keresőmezőbe, hogy ’agatha christie’, míg ha kulcsszóra keresünk, akkor nyugodtan beírhatjuk a szerző nevét (a feltételezhetően egyetlen keresőmezőbe).

4.2.2 Mennyi adatban keresünk

Egy másik nagy különbség a mennyiségben van. ETO-osztályból kb. 60000 (hatvanezer)[41] van, vagyis nem mondhatjuk azt, hogy a kötött, előre meghatározott szavakat tartalmazó index (az információkereső nyelvek szókészlete alapján előálló index, mint pl. a tárgyszóindex) mérete túl kicsi lenne. De a lenyűgöző szókincse ellenére a tárgyszóindex egy szempontból biztosan szegényesebb lesz, mint a kulcsszóindex, és ez a szinonimák használatát jelenti. A tárgyszóindexben nem cél a szinonimák párhuzamos használata. Sokkal elterjedtebb gyakorlat, hogy keresni lehet a tárgyszóindexben külön, így ki lehet választani a könyvkereséshez a legalkalmasabbnak tűnő tárgyszavakat. A kulcsszóindex azonban nagy eséllyel tartalmazza a szinonimákat, hiszen ha a tartalom vagy a leírás alapján indexelünk, akkor annak a szerzője valószínűleg egy-egy fogalomra több kifejezést is használt. Így egyszerűen a nyelvhasználatból adódóan a kulcsszóindex gazdag lesz szinonimákban, míg a tárgyszóindex éppen ellenkezőleg. Ez azt is jelenti, hogy a tárgyszóindexben jóval kisebb a

[41] Ungváry-Orbán (2001), 491-492. oldal

redundancia, mint a kulcsszóindexben, hiszen jóval kevesebb jel hordozza ugyanazt a jelentést, vagyis az erre a könyvre való rámutatást.

A kulcsszóindexnek természetesen megvannak a maga hátrányai. Ha például a dolgozatom tartalma indexelésre kerülne, akkor az egyik kulcsszó lehetne hozzá az, hogy 'agatha christie', hiszen ez a sztring is szerepelt a szövegben. Ugyanakkor annak, aki Agatha Christie-ről akar megtudni valamit, és ezért a nevére keres, az én dolgozatom semmi lényeges információt nem közöl, azaz ez a találat a zajt fogja erősíteni, nem a releváns találati listát.

4.2.3 Hogyan mutatjuk meg, amit találtunk

Az előbbiekben láttunk példát arra, hogy a kulcsszó alapú keresés több irreleváns találatot eredményezhet, mint a tárgyszó alapú. Van-e valamilyen lehetőség ennek a kivédésre?

A korábban definiált két információs alrendszer (adatnyilvántartás és az ehhez kapcsolódó keresőfelület, valamint a találati lista) közül az elsőben (vagyis az adatnyilvántartásban) nem sok lehetőségünk van a zaj csökkentésére. Az indexelés algoritmusa természetesen valamennyit segíthet ezen a problémán (pl. csak azokat a szavakat emeljük be az indexadatbázisba, amelyek többször is előfordulnak egy könyvvel kapcsolatban, nem csak egyszer stb.), de a megoldást a másik információs alrendszerben, a találati lista megjelenítésében találhatjuk meg.

A több száz tételből álló találati listát nem szokás végignézni. Vagyis minél előbbre vannak a releváns tételek a listában, a lista annál kevésbé tekinthető zajosnak. Ha az indexadatbázisunkban nem csak a megfelelő kifejezést tároljuk, hanem azt is, hogy hol találtuk (szerző, cím vagy leírás, esetleg tartalom), illetve hányszor, és hogy pontos találat volt-e vagy a találat csak a reguláris kifejezések alkalmazásával lett az, akkor egy megfelelő algoritmussal kiszámolhatjuk, hogy melyik könyv mennyire releváns a beírt kereső-kifejezés alapján. A relevancia alapján csökkenő sorrendbe rendezett könyvek jelentősen csökkentik a zajt, így a kulcsszóindexelés magasabb zaj-értéke már nem, vagy jóval kevésbé okozhat gondot, legalábbis jól megválasztott kereső-kifejezés esetén.

4.2.4 Összefoglalás

Az eddig elmondottak alapján a következőket állapíthatjuk meg a tárgyszó- és kulcsszókeresés hasonlóságáról és különbségeiről:

- Mindegyik esetben indexadatbázisban vagy -adatbázisokban keresünk.
- Mindegyik esetben természetes nyelvi kereső-kifejezést használ a kereső ágens.
- A keresés során a kereső ágens által beírt kereső-kifejezést hasonlítjuk az indexadatbázisban szereplő szavakkal.
- A hasonlítás sztringhasonlítás: azaz a nyelvnek csak azt a tagolását vesszük figyelembe, ahol jelentés nélküli hangokból a nyelvben létező szavak állnak össze. A szavakból felépülő szerkezeteket (szintagmák, mondatok) a keresés során nem vesszük figyelembe. Ha lehet is többszavas kifejezésre keresni, ezek egyeztetése

szintén sztringhasonlítással történik. Ilyen módon a szinonimákra valójában nem keresünk. A szinonimák használatát csak a kulcsszókeresés biztosítja, a nyelvhasználatnak arra a tulajdonságára támaszkodva, hogy a szövegeinkben használunk szinonimákat.

- Az indexben szereplő szavaknak két jelentést tulajdonítottam ebben a modellben. Az elsőt, a szavak szótári jelentését a keresés nem veszi figyelembe. A másik jelentésük, vagyis az, hogy melyik könyvre mutatnak rá, az egyetlen, a keresésben releváns jelentés. Az információ, amit ezek a szavak továbbítanak, így éppen ez: általuk melyik könyv található meg.
- A keresők általában nem végeznek nyelvi elemzést. A hasonlítást a biztosabb találat érdekében elvégezhetik a kereső-kifejezés csonkolt (vagyis reguláris kifejezéssé alakított) változatán, de ez a megoldás inkább a kulcsszavas keresésre jellemző, mert a tárgyszavas keresés esetében megmutatható a konkrétan adott indextábla, amiből a felhasználó kiválaszthatja, hogy melyik tárgyszóra fog keresni, karakterre pontosan.
- Mivel a kulcsszó alapú keresés egyszerre több adatmezőben is keres, míg a tárgyszó alapú keresés csak egyben, így a kereső ágensnek a kulcsszó alapú keresésnél elegendő a megfelelő kereső-kifejezés kiválasztására képesnek lennie, ugyanakkor a tárgyszó alapú keresésnél tisztában kell lennie a könyvadatbázis adattárolási struktúrájával, legalább nagyvonalakban. Ez két különböző bonyolultságú mintázat-felismerést követel meg a felhasználótól.
- A keresőfelületek (egyszerű, összetett) formája is különböző mintázat-felismerési képességet követel a felhasználótól: míg az egyszerű keresések mintázata jól meghatározható, ugyanez nem igaz az összetett keresésekre, így minden nyilvántartás összetett keresési felülete egyedi, külön-külön megismerendő felület.
- A tárgyszó alapú keresés esetében kisebb a zaj (noha nem nulla méretű), és kisebb a redundancia is. A kulcsszó alapú kereséseknél nagyobb a zaj, de nagyobb a redundancia is, ami segít a zaj leküzdésében. Ugyanakkor a túl nagy zaj a második információs alrendszerben, a találati listában csökkenthető.

A következő fejezetekben az eddigi megállapításokkal kapcsolatban keresek további információkat, részben kereső felhasználói szokásokról szóló tanulmányokban, részben keresőfelületek elemzésével.

5. Tanulmányok összehasonlítása

Ebben a fejezetben két tanulmány legalapvetőbb megállapításait veszem sorra. Ezeken keresztül kívánom megmutatni (saját, a felhasználói szokásokra vonatkozó empirikus kutatás helyett), hogy a felhasználók kifejezetten előnyben részesítik a kulcsszó alapú keresést a tárgyszó alapúval szemben.

5.1 Kereső-zombik

Ebben a fejezetben egy rövid cikket ismertetek[42]. Ennek nem a kiválasztott cikk különlegessége az oka, hanem az, hogy jól összefoglalja a mai webes keresésről szóló gondolkodás egyik irányzatát.

Több tanulmányban[43] előfordul, hogy a szerzők aggodalmukat fejezik ki a jövő nemzedék információszerzési tudása, szokásai miatt. Ezekben az írásokban a szerzők sürgetik a keresők használatának iskolai oktatását, mivel a kereső felhasználók szokásainak megfigyelésével arra az eredményre jutottak, hogy a keresőket használó hallgatók[44] nem használják ki a keresők lehetőségeit, és a tesztekkel gyűjtött információ alapján úgy tűnik, nem is ismerik ezeket.

Rethlefsen (2008) még radikálisabban fogalmaz. Egyenesen arról beszél, hogy „egyszerű-kereső zombikat" nevelünk a következő generációban. Cikkének az „Easy ≠ Right" címet adta, utalva ezzel arra, hogy az egyszerű keresés felületei könnyen és gyorsan használhatók, de ezekkel nem lehet eljutni az igazán jó, érdekes, különleges találatokig. Rethlefsen úgy véli, hogy az egyszerű keresők használata túlságosan egyszerűnek láttat egy nagyon összetett kérdést. Az újfajta kereső-zombik ész nélkül és stratégia nélkül gépelik a kereső-kifejezéseket az egyetlen keresőmezőbe, ahelyett, hogy megtanulnák, mi is az az adatbázis, mi lehet a struktúrája és hogyan kell benne keresni. Ahhoz, hogy valóban hatékonyan tudjunk keresni, komplex információkereső képességek kellenek. A sikeres keresésnek több része van: keresés az adatbázisokban, konzultálás szakemberekkel, releváns folyóiratok manuális átlapozása, bibliográfiákban való bányászkodás, és persze képesség arra, hogy felismerjük, ha megvan, amit kerestünk. Az egyszerű keresések úgy tüntetik fel ezt a folyamatot, mintha könnyű lenne, pedig nem lenne szabad, hogy az legyen.

Rethlefsen kételyének ad hangot a tekintetben, hogy a keresések fejlődésének iránya (egyre jobb kereső algoritmusok, szemantikai keresések, nyelvtechnológiai alkalmazások, helyesírás-ellenőrzés, szókincsegyezési vizsgálatok, egyre fejlődő indexelési és relevanciaszámítási-algoritmusok, metakeresők használata) valaha pótolhatja azt, amikor tudatosan, az adatstruktúra ismeretében keresünk az adatbázisban. Hiszen (érvel a szerző) minden adatbázis egyedi, és a keresési kapacitás és a jó keresési stratégia mindig a konkrét adatbázis struktúrájától és tartalmától, metaadataitól függ. Rethlefsen a jövő feladatai között nem az egyszerű keresés további fejlesztését jelölné meg, hanem a felhasználók képzését és az adatbázisok népszerűsítését.

5.2 Az OPAC és az Internet

Mi-Weng (2008) szerint az utóbbi években drasztikus változás történt a könyvtárhasználat terén. Egy 2005-ös tanulmányra hivatkozik, amely szerint a vizsgált felsőoktatási hallgatók közel 90%-a, de a kutatóknak is több, mint a fele az Interneten, nagyrészt a Google-ban kezdi

[42] Rethlefsen (2008)
[43] Jóföldi (2003), K. Princz-Nagy (2005), Pajor (2006)
[44] Az idézett tanulmányok hallgatói szokásokat vizsgáltak.

meg a kutatásához az információgyűjtést. A keresési folyamatban a kényelem az egyik legfontosabb szempont az Internethez szokott felhasználók számára. A Google, az Amazon és a hasonló felületek keresőinek használatához nem kell a keresést megelőzően információkeresési tanulmányokat folytatni, és nem kell különösebben odafigyelni a keresőkérdés formájára sem a begépeléskor. Ezek a felhasználók a könyvtári katalógusokban való keresést túlságosan nehéznek találják.

A könyvtári katalógusok a 19. századtól a szerző, cím, tárgyszó hármasára építenek elsősorban. Ez a felépítés a könyvtári adatbázisokban is megmaradt. Ez a nyilvántartás nagyon határozottan adat alapú: vagyis a keresésnél tudnom kell, hogy melyik adatot hol kell keresni, és mik a lehetséges kereső-kifejezések (pl. a tárgyszavak esetében). Ezzel szemben az internetes keresések sokkal inkább természetesnyelv alapúak: azon kívül, hogy nem kell a kereső-kifejezést pontosan eltalálnom, hiszen a szöveg minden szava gyakorlatilag tárgyszóként funkcionál, az internetes keresők nyelvtechnológiai és statisztikai módszerekkel is segítik a keresést (hasonló oldalak, relevancia alapú rangsorolás stb.). Ezek a segítségek a könyvtári keresésben nem állnak rendelkezésre.

Több, Mi-Weng (2008) által meghivatkozott, a felhasználói viselkedést elemző tanulmány is azt állítja, hogy az új, az Interneten felnövő generációnak a „beviteli mező (ahova szavakat lehet begépelni) + Keresés feliratú gomb + rangsorolt találati lista” = „A Kereső”, amelyben kulcsszavakkal lehet keresni. Ezért a könyvtári keresésben is a tetszőleges mezőre vonatkozó kulcsszókeresés lenne a preferált (vagyis az egyszerű keresés), de a legtöbb könyvtári rendszer csak a merev adatstruktúra alapú keresést támogatja.

Mi-Weng (2008) a következőképpen összegzi a mondanivalóját: ha az a kérdés, hogy hogyan tanítsuk meg a felhasználóinkat a nyilvántartásunkban való keresésre, a jó válasz az, hogy készítsünk olyan rendszert, aminek a használatát nem kell tanítani.

5.3 Összefoglalás

Az előző két tanulmány ugyanazt a jövőképet ábrázolta (noha nem ugyanazzal a hozzáállással): azt, hogy a könyvtári keresésekben terjed a kulcsszó alapú keresés és ezzel párhuzamosan az egyszerű keresés használata, vagy az igény ennek a használhatóságára. Ennek egyik oka lehet az egyszerűbb keresőfelület. De az is lehetséges, hogy (minden ellentétes állítás ellenére) egyszerűen jobb találatokat lehet elérni a kulcsszó alapú kereséssel. A következő fejezetekben néhány keresőfelületet tesztelek ebből a szempontból.

A 6. fejezetben két nyelvészeti portálrendszer keresőjét hasonlítom össze. Egyik esetben sem professzionális, üzleti alapon működő keresőről van szó, de mivel szélsőségesen különböznek, ezért a szemléletbeli eltéréseket jól lehet velük illusztrálni.

A 7. fejezetben már valóban releváns, népszerű, professzionális keresőket tesztelek, és azokon keresztül igyekszem megmutatni, hogy a kulcsszó alapú keresés nem csak könnyebb, hanem akár jobb találatot is eredményezhet. A 8. fejezetben erre a jelenségre keresem a magyarázatot.

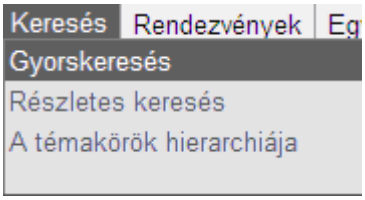
6. Két nyelvészeti portál keresőmódszereinek összehasonlítása

2001-ben a Széchenyi Terv támogatásával a magyar nyelvészeti közösségek összefogására több nyelvészeti portál is létrejött. A különböző portálok létrehozói más és más helyre helyezték a hangsúlyt a portáljuk megtervezésénél. Két szélsőségesen eltérő szemléletű portált mutatok be a következő két alfejezetben.

6.1 Nyelwww[45]

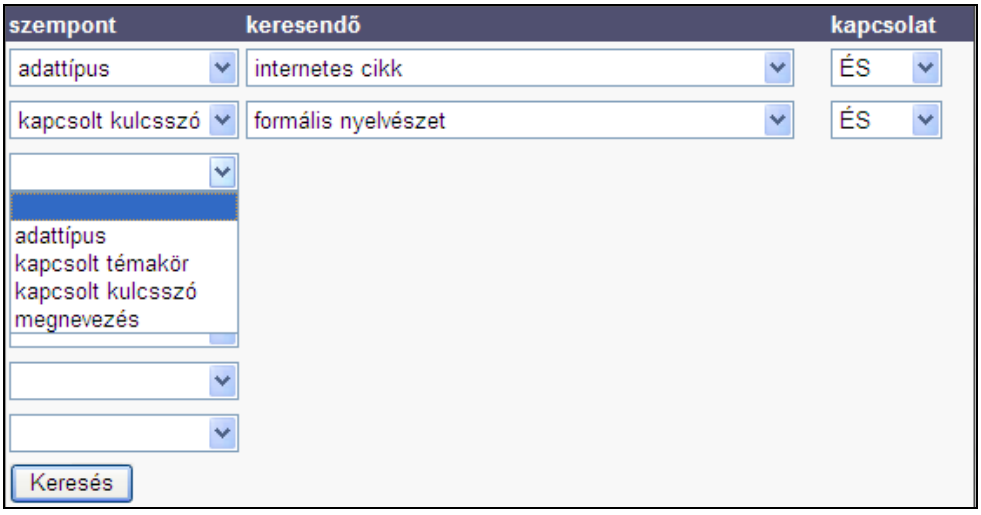
A Nyelwwwforum nevű portál a PTE BTK Nyelvtudományi Tanszéke által vezetett konzorcium munkájának eredményeként jött létre. A célja (a pályázati célkitűzéseknek megfelelően) az volt, hogy egy helyen nyújtson mindenkinek információt, aki nyelvészet, és azon belül is elsősorban nyelvtechnológia iránt érdeklődik.

A dolgozatom szempontjából a portál keresője a lényeges elem. Mint az 1. ábrán is látható, kétféle keresést definiáltunk a portál használatához: a gyorskeresést és a részletes keresést:



1. ábra: a Nyelwww portál keresői

A Gyorskeresésnél legfeljebb nyolc keresési szempontot lehetett beállítani, ÉS-sel vagy VAGY-gyal összekötve ezeket egymással. Mindegyik szempontnál egy négyelemű listából lehetett választani, attól függően, hogy adattípusra, témakörre vagy tárgyszóra akar-e a felhasználó keresni, esetleg a megnevezésben egy szóra:



2. ábra: a Nyelwww portál Gyorskereső felülete

[45] Az eredetileg Nyelwwwforumnak nevezett rendszer új címe: http://lingua.btk.pte.hu/nyelwww/main.asp?l=hu. Az eredeti szerkesztőbizottság tagjai voltak: Burányi Péter, Csatay Richárd és Viszket Anita. A letöltés ideje: 2008. november 15.

Az adattípus, a kulcsszó és a témakör esetében nem lehetett kereső-kifejezést begépelni, hanem a vonatkozó indexből lehetett választani. A témakörök struktúrája hierarchikus volt, és a magasabb hierarchiaszintre való keresés az alacsonyabbakhoz rendelt találatokat is megjelenítette. Így próbáltunk azon segíteni, ha a vonatkozó indexelemek rész-egész viszonyban vannak egymással. A témakörök hierarchiáját a keresés előtt meg lehetett nézni, lásd az 1. ábrán. A szinonimák problémájával nem tudtunk mit kezdeni.

Tegyük fel, hogy a szuperkurzusokra akarok keresni a Nyelwww portálon. Mivel nem tudom, hogy mi a kapcsolt adattípusuk, és egyelőre a témakörüket és a kulcsszavukat sem ismerem (nincs olyan kulcsszó, hogy „szuperkurzus"), ezért legegyszerűbb, ha a címükre keresek. Mivel a portál mögött lévő kereső nem használ reguláris kifejezéseket, ezért a kereső ágensnek kell tudnia, hogyan csonkolhatja a kereső-kifejezését:

3. ábra: keresem a szuperkurzusokat a Nyelwww portálon

4. ábra: találati lista a Nyelwww portálon

A kereső kialakításakor a következők voltak a céljaink:

- Minél dinamikusabb legyen a kereső: ezért nem kellett előre megmondani, hogy valaki kurzust, könyvet, tanfolyamot, pályázatot vagy egy embert keres, hanem lehetett például az összes rekordtípusban egyszerre keresni. A cél az volt, hogy ne kelljen a kereső felhasználónak átlátnia az adatstruktúrát.
- Minél jobban csökkentsük a kereséskor elkövetett hibák számát, vagyis a zajt: nem jól megválasztott kereső-kifejezések, gépelési hibák stb. Ennek érdekében a legtöbb esetben nem begépelni kell a kereső-kifejezést (sőt, általában nem is lehet), hanem ki kell választani a listából (indexből).

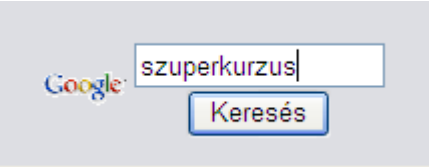
- A gondosan megalkotott tárgyszókészlettel és a hierarchikus témakörökkel csökkentsük a redundanciát.

Feltételezem, nem kell különösebben érvelni amellett, hogy a kereső, noha a kitűzött céloknak megfelelt, a felhasználók számára túl bonyolultnak bizonyult. Egy példa: adattípusból (mivel az összes adattábla típusai megjelentek az indexben) kb. hatvan féle közül lehetett választani, viszonylag alacsony feltöltöttség esetén is. Ez a szám nyilván tovább nőtt volna, ahogy növekszik az adatbázis mérete.

Az, hogy a felhasználónak ne kelljen ismernie az adatstruktúrát, a dolgozatomban leírt eddigi tapasztalatok alapján jó és fontos ötletnek tartható. De úgy tűnik, a zaj és a redundancia radikális csökkentése nem a megfelelő módszer a sikeres kereső építéséhez.

6.2 Nyelvészet.hu[46]

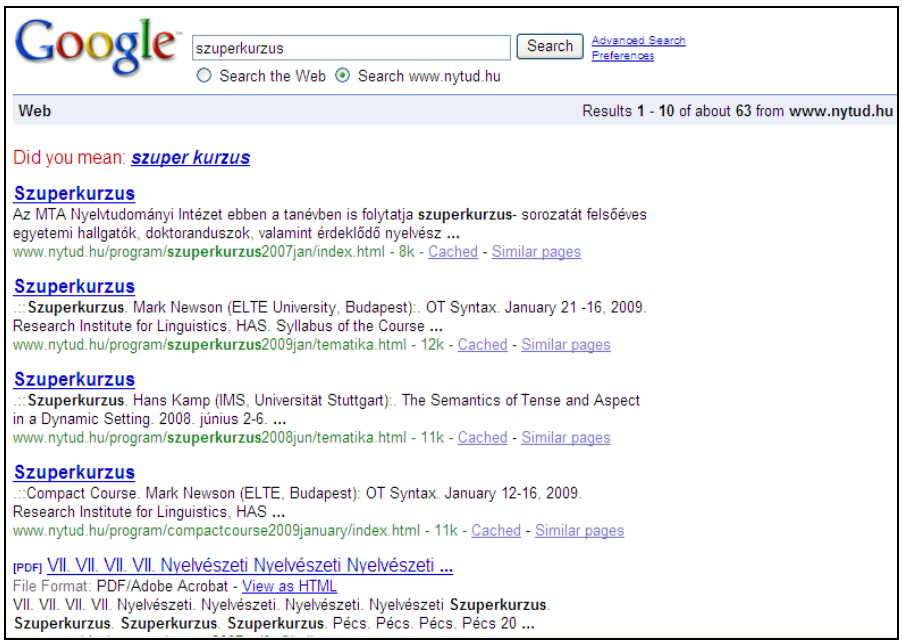
Ugyanebben az időszakban jött létre az MTA Nyelvtudományi Intézete által támogatott nyelvészet.hu portál. Az ő kiindulási pontjuk egészen más volt. Ők nem a kereső minél tökéletesebbre alkotásával foglalkoztak a projektidő alatt, helyette inkább igénybe vették a Google keresőjét. Sokkal több energiát szántak az adatbázisnak adatokkal való feltöltésére.



5. ábra: a nyelvészet.hu portál keresője

[46] A portál eredeti címe www.nyelveszet.hu volt, mostanra már nem üzemeltetik ezen a címen, hanem helyette a http://www.nytud.hu/portal/index.html a pontos cím. A letöltés ideje: 2008. november 15.

Ahogy a Google általában, úgy ez a kereső is csonkol magától, így egyszerűen elég beírni a kereső-kifejezést, és meg is jön a találati lista:



6. ábra: találati lista a nyelvészet.hu portálon

Annyit elsőre is láthatunk, hogy ezt a keresőt sokkal egyszerűbb használni. Több a találat is, de valójában az adat is több ezen a portálon. Viszont a több találat között nagyobb a zaj is: egy rövid vizsgálattal kideríthető, hogy a 4. találat nem teljesen, és a 20. találat szinte egyáltalán nem releváns. De szerencsére a lista elején a lehető legrelevánsabb találatok vannak, így a zaj ártalma jelentősen csökken.

A két portálrendszer keresőjének összehasonlításával megállapíthatjuk, hogy nem szerencsés stratégia túlságosan törekedni a zajtalan és redundancia-mentes keresésre. Nem ettől lesz sikeres a keresőprogram.

7. Keresőrendszerek összehasonlítása

Ebben a fejezetben konkrét könyvkereső rendszereket vizsgálok. Először összefoglalom, amit az Amazon.com-ról mondani szokás, vagyis mindazokat a tudnivalókat, amelyek miatt az elemzések szerint az Amazon.com sikeres lett. Ezután lefuttatom ugyanazt a keresést két magyar könyvkeresőn, a Bookline.hu-n és a Könyvkereső.hu-n, és összehasonlítom a működésüket, valamint a keresés sikerességét.

7.1 Amazon.com[47]

Mi az Amazon.com sikerének a titka?

Olyan terméket árul, amiről tudjuk, micsoda, ezért megértjük a találati listát, és vannak ötleteink a jó kereső-kifejezésre.

Spector (2003) az Amazon.com megalapításával kapcsolatban azt írja, hogy Jeffrey Bezos (az Amazon.com alapítója) eredendő célja a cégalapítással egy online üzlet létrehozása volt, támaszkodva az akkoriban (90-es évek első fele) rohamléptekben terjedő Internet-felhasználásra. Az, hogy ez az online üzlet mit áruljon, másodlagos kérdés volt Bezos számára. Azért döntött a könyvárusítás mellett, mert „mindenki tudja, mi az a könyv”[48]. Ez azért volt nagyon fontos, mert így nem kellett elmagyarázni a felhasználóknak, milyen adatai vannak egy könyvnek, és mit jelent a könyv keresése a beírt adatok alapján.

Nagyon könnyű benne keresni.

Az Amazon.com 1995. július 16-án indult, és ahogy Spector írja: „1995-ben még forradalminak számított soha nem látott egyszerűségével”[49]. 1994-ben a keresést még úgy tervezték, hogy a vásárló egy előre definiált formájú e-mailt küld a keresőgépnek, és válaszban kapja meg a keresett könyv adatait. A nyitáskor már nem volt szükség erre (a felhasználótól komoly technológiai felkészültséget feltételező) megoldásra, mert az Internet annyira általánossá vált.

Amazon.com sikerének egyik (Krug (2008)[50] szerint fő) titka az, hogy az Amazon.com (elsőként) megszüntette a könyvkeresésben a cím/szerző/kulcsszó hármast, és egyetlen keresőmezőben lehetett megfogalmazni a kereső-kifejezést. Ez nagyon sok felhasználó számára tette sokkal sikeresebbé a keresést. Krug megfigyelései szerint a felhasználók nem értik a „kulcsszó” meghatározást, és nagyon gyakran minden kereső-kifejezést (szerzőnevet is) kulcsszóként keresnek, és ilyenkor nem találják meg a keresett könyvet (hacsak az adatbázisban nincs minden szerző kulcsszóként is felvéve).

Olyasmit is meg lehet találni, amit sehol máshol (nagy adatbázis).

A kereső-kifejezéseket elemző cikkek tanúsága szerint nem csak az a jellemző, hogy általában több szóból álló kereső-kifejezéseket használunk az internetes keresésben, hanem az is, hogy sok olyan kereső-kifejezést használunk, amit soha senki más nem használt: a keresések kb. 50-60%-ában ilyen kereső-kifejezéseket gépelünk be[51]. Ez a tény egészen forradalmasította az üzleti világot. Ugyanis ez azt is jelenti, hogy olyan dolgokat (termékeket)

[47] http://www.amazon.com/, de valójában ezt a keresőt nem tesztelem, csak a róla szóló tanulmányból foglalom össze az állításokat: Spector (2003)
[48] Spector (2003), 52. oldal
[49] Spector (2003), az indulás dátuma: 93. oldal, az idézet helye: 102. oldal, további információk: 73-74. oldal
[50] Krug (2008), 77-78. oldal
[51] Jansen és tsai (2007) szerint az egyedi kereső-kifejezések aránya 58,9% (748. oldal). Battelle (2006) kb. 50%-ra teszi az egyedi kereső-kifejezések számát (31. oldal).

is megtalálunk és megveszünk, amelyek nagyon ritkák, és amelyekre az Internet nélkül nem találnánk rá.

A keresésnek és az eladásnak ez az összefüggése volt az egyik eleme az Amazon.com sikerének[52].

Elvileg egy online könyvesbolt kereső rendszere nem különbözik egy valódi, fizikai térben lévő könyvesbolt kereső rendszerétől (amit általában csak maguk az eladók használnak), és amiben szerzőre, címre, kiadóra, tárgyszavakra stb. lehet keresni. A nagy különbség az, hogy az online könyvesbolt olyan tételeket is felvehet az adatbázisába, amit aktuálisan nem raktároz, de valaha valahol kiadták, és van esély rá, hogy be tudja szerezni. Sőt, olyan tételeket is (ahogy ez az Amazon.com-nál is előfordult), amelyek soha nem kerültek könyvesbolti forgalomba, de a szerzők kis példányszámban megjelentették a művüket, és felvették a kapcsolatot az online könyvesbolttal értékesítésre[53].

Olyasmit is meg lehet találni, amit sehol máshol (gondosan fejlesztett kereső)

Spector (2003) leírja, hogy az Amazon.com sikeres üzletpolitikájának két kiemelkedő eleme volt: az egyik, hogy lehetővé tette, hogy a vásárló a bejutás után néhány perccel megtalálja amit keres, a lehető legegyszerűbb felületen; a másik, hogy a rendszer lelkének a keresési nyelvet tartotta, amely a lehető legjobb találatokat tudja a vásárló számára megjeleníteni[54]. Vagyis a keresőrendszer nagyon egyszerű keresési felületet, és mögötte nagyon körültekintően fejlesztett keresési rendszert tartalmazott, amelynek a működése nem volt ismert a felhasználó számára, de ennek ellenére elégedettséget okozott neki az, hogy a keresései könnyűek, sikeresek és gyorsak voltak.

Összefoglalás

Vagyis az Amazon.com egy jól időzített indulású, nagy adatbázissal rendelkező, könnyen használható kereső, ami mögött gondosan fejlesztett keresőmotor van (és persze számtalan vevőcsalogató ötlet, de ezek most nem relevánsak a keresés szempontjából).

A következő pontokban megvizsgálok két magyar könyvkeresőt. Azzal kapcsolatban nem állnak a rendelkezésemre statisztikák, hogy a kettő közül melyiket használják többen, tehát melyik a sikeresebb. Azért választottam ezt a két keresőt, mert nagyjából úgy viszonyulnak egymáshoz, mint a Nyelwww portál viszonyult a Nyelvészet.hu portálhoz (ami a keresési stratégiáikat illeti).

A feladat, amit mindkét keresőn magam elé tűztem, az volt, hogy meg kell találnom egy olyan krimit, aminek a címében valamilyen áfonyás süti van, meg a gyilkosság szó.

[52] Spector (2003), 154. oldal

[53] Spector (2003) leírása alapján az Amazon.com indulásakor valóban a nagy mennyiségű könyv-adat (úgy hirdették, hogy 1 millió könyv adata) volt a forrása a „Long Tail”-jelenségnek (amit természetesen csak 2004-től hívtak így, lásd Anderson (2004)). (Spector (2003), 179-180. oldal).

[54] Spector (2003), 94-95. oldal

Előtte természetesen ellenőriztem, hogy a kötet mindkét kereső által megtalálható legyen, ha pontos címre vagy pontos szerzőre keresek. A bibliográfiai pontosságú adatok a következők: *Fluke, Joanne: Áfonyás muffin és gyilkosság – Hannah Swensen titokzatos esetei 3. Illia & Co. Kiadó, 2007*. Ezt a címet azért választottam, mert van benne két jól kereshető kifejezés (ha a muffin-t is beleszámítom, akkor három, de azt túl ritkának találtam, és így a keresést túl könnyűnek), és a két kifejezésből az egyiknek a töve váltakozó tő a melléknévi képző előtt: *áfonya – áfonyá*. Így az is ellenőrizhető, hogy melyik kereső segíti a felhasználót az ékezetek „semlegesítésével".

7.2 Könyvkereső.hu[55]

A Könyvkereső.hu a korábban megmutatott Nyelwwwforumhoz hasonló szellemiségű (a keresés szempontjából). A lehető legprecízebben tájékoztatja a felhasználóit a keresőfelületen a csonkolási és egyéb lehetőségekről, de automatikusan semmit nem végez el a felhasználók helyett. Ez a kereső valójában nem tiszta tárgyszókereső rendszer: azaz a mellett, hogy lehet a címre és a szerzőre keresni, valójában van egy kulcsszókereső része is, ami ráadásul tartalomindexelés alapján is keres, de ezt csak a keresőfelület alapos végigolvasásával lehet megtudni, így aztán ez az információ az esetek nagy részében elsikkad. Először megpróbáltam a címben, aztán a kulcsszavak között keresni a két szóra, amire emlékeztem a címből, de sikertelenül (lásd 7. ábra):

7. ábra: akár a címben, akár a kulcsszóban keresem ezeket a kifejezéseket, nincs találat a Könyvkereső.hu-n

[55] http://www.konyvkereso.hu/, a letöltés ideje: 2008. nov. 15.

Ezután megpróbáltam (figyelembe véve a használati utasítást a csonkolással kapcsolatban) az **áfonya** kereső-kifejezésre keresni a címben. Kaptam három találatot, de a keresett könyv nem volt köztük. Nyilván azért, mert a keresőprogram nem csak nem csonkol, hanem nem is ékezetmentesít:

8. ábra: találati lista az *áfonya* kereső-kifejezésre a címben keresve a Könyvkereső.hu-n

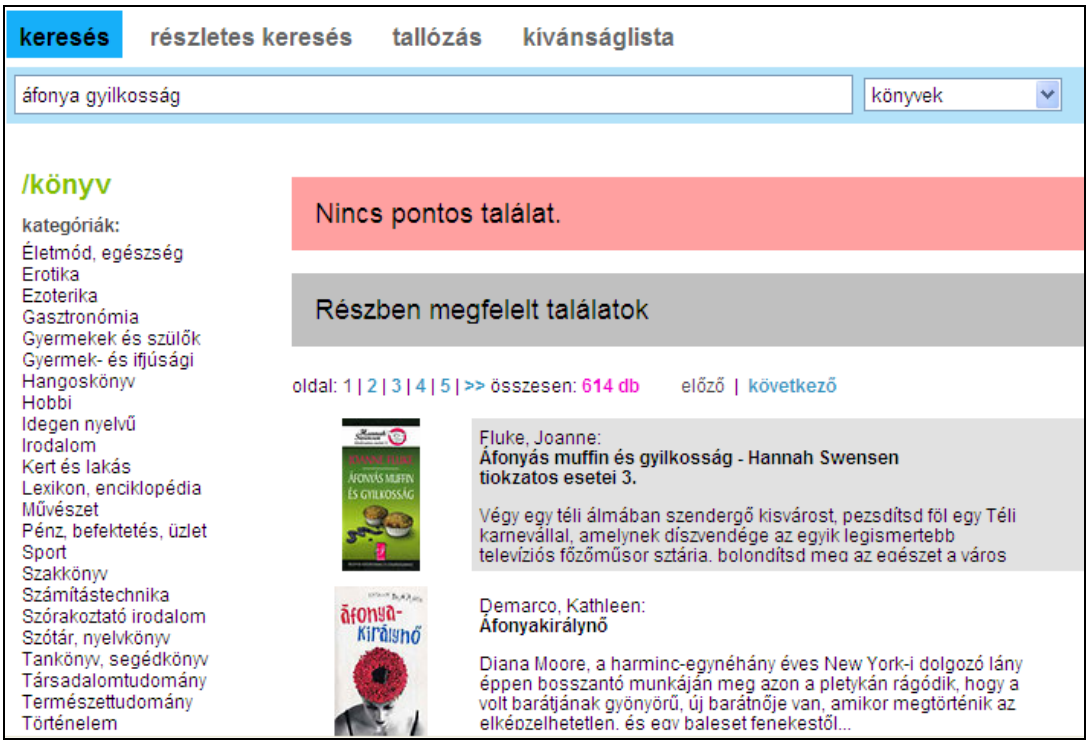
Utolsó próbálkozásként az *áfony* kereső-kifejezésre kerestem a címben, és így meg is lett a könyv, az utolsó helyen a találati listában (lásd 9. ábra):

9. ábra: találati lista az *áfony* kereső-kifejezésre a címben keresve a Könyvkereső.hu-n: az utolsó sorban megvan a keresett könyv

Azzal, hogy a Könyvkereső.hu keresője önmaga nem csonkol, nem ékezetmentesít, és nem keres automatikusan a tartalomleírásban, csak ha erre a megfelelő helyre beírt kereső-kifejezéssel utasítást adok, jelentősen csökkenti a zajt az információáramlásban. Annyira törekszik ez a kereső a zajcsökkentésre, hogy ha a kulcsszó szerinti keresésben azt a kereső-kifejezést adom meg, hogy **áfony* * gyilkosság**, akkor a válasz az, hogy „Túl általános keresőszó!”. De ezzel sajnos párhuzamosan csökkenti a zajártalom-megelőző redundanciát is. Az eredmény, úgy ítélem meg, nem a legsikeresebb könyvkereső.

7.3 Bookline.hu[56]

A Bookline.hu klasszikus kulcsszókereső rendszer. Ahogy a Nyelvészet.hu oldalon, így itt is csak egy keresőmező van, amibe bejegyezve a kereső-kifejezésemet, a lehető legteljesebb tartalomindexben fut a keresés, egyszerre kikeresve a pontos és a közelítő találatokat, és aztán relevancia szerinti sorrendbe rendezve azokat (vagyis a lehetséges csonkolásokat a keresőprogram elvégzi a kereső felhasználó helyett). Az eredmény a 10. ábrán látható: a keresett könyv az első a találati listában.



10. ábra: találati lista a Bookline.hu-n: az első sorban megvan a keresett könyv

A Bookline.hu nyilván nem aggódik a találati lista zajossága miatt, mert a Könyvkereső.hu öttalálatos listájával szemben akár a 614 találatos listát is megengedi magának., Viszont azzal, hogy tolerálja az olyan kereső-kifejezést, amit a Könyvkereső.hu nem engedélyez (**áfony* * gyilkosság** – hiszen feltehetően a háttérben a Bookline.hu is erre fordította le az általam ide beírt kereső-kifejezést), sokkal jobb relevanciát tud beállítani a találati lista elemeihez, így a keletkezett zaj jelentősen csökken, sőt, a tesztesetünkben a lehető legtökéletesebb lett az eredmény.

8. Magyarázatkeresés a változásokra

A dolgozatomban azt próbáltam megmutatni, hogy a hagyományos szerző, cím tárgyszó-típusú könyvkeresések helyett mind az online könyvesboltokban, mind a könyvtárakban egyre inkább az egy keresőmezőbe beírható kulcsszó alapú keresés az elvárás a felhasználók részéről. Ebben részben a felhasználói szokásokat elemző tanulmányok, részben a keresőfelületeken lefuttatott tesztek erősítettek meg.

[56] http://bookline.hu/, a letöltés ideje: 2008. nov. 15.

A kérdés az, hogy mi lehet ennek a változásnak az oka. Ebben a fejezetben kétféle magyarázatot adok a jelenségre. Mindegyik magyarázat elvileg lehetséges, de a bebizonyításukra ebben a dolgozatban nem tudok vállalkozni.

8.1 A felismerhető sémák a keresőfelületen

Az egyik lehetséges magyarázat, amit az 5. fejezetben feldolgozott tanulmányok is állítottak, az internetes keresés hatása a könyvkeresésre. Egészen pontosan a Google-szerű keresésről van szó, ahol egy keresőmezőbe beírva a természetes nyelvű kereső-kifejezést, megkapjuk az eredményt anélkül, hogy bármiféle adatstruktúráról vagy a kereső működéséről tudnunk kellene.

Ez azt jelenti, hogy kialakult egy keresési módszertanunk a Google (vagy más internetes kereső) alapján, és ebben a sémában próbáljuk értelmezni azokat a keresőprogramokat, amelyekkel találkozunk. Ezért kedveljük azokat, amelyek a Google-ra hasonlítanak, és nem kedveljük azokat, amelyek nem hasonlítanak rá.

Ez a magyarázat valóban megindokolhatja, hogy miért szeretjük jobban az egyszerű keresést, mint az összetett keresést. De arra nem ad magyarázatot, hogy miért jobb a keresés eredménye sok esetben (noha, el kell ismerni, nem minden esetben) a kulcsszó alapú keresésnél, mint a tárgyszó alapú keresésnél.

A kulcsszó alapú keresés többnyire egyszerű keresési felülettel, a tárgyszó alapú pedig többnyire összetettel társul, de ez természetesen legfeljebb a felhasználók hozzáállását magyarázhatná meg, a keresés sikerességét nem.

8.2 A redundancia és a zaj összefüggései

A dolgozatomban többször utaltam rá, hogy a kezelőfelület könnyű vagy nehéz kezelhetőségén túl van még egy nagy különbség a kulcsszó és a tárgyszó alapú keresések között, és ez a zaj, valamit a redundancia mérete.

Racionálisan beláthatjuk, hogy mind a redundanciának, mind a zajnak nagyobbnak kell lennie a kulcsszó alapú rendszereknél az első alrendszerben (vagyis az adatnyilvántartásban). Egy-egy könyvhöz biztosan több rámutató rekord létezik az indexadatbázisban kulcsszó alapú rendszereknél, ha azokat automatikusan hozzuk létre a szerzőből, címből, tárgyszóból, sőt leírásból. De ugyanakkor nem csak a redundancia (ami segítené a keresést), hanem a zaj is nagyobb lesz, mert sok könyvhöz olyan kulcsszó is hozzárendelődik, ami ugyan előfordult a könyvhöz tartozó szövegekben, de a könyv tartalma szempontjából mégsem releváns, vagy nem eléggé releváns.

Mi történik a második információs alrendszerben, a találati listában? Itt is nagyobb lesz a zaj a kulcsszó alapú rendszereknél, hiszen sokkal több lesz a találat, és a listából biztosan sok lesz irreleváns. Csakhogy ezt a zajt nagyon jól lehet kezelni, szinte megszüntetni a relevanciára támaszkodó megjelenítéssel, ami tárgyszó alapú rendszereknél nem lehetséges!

Úgy tűnik, hogy hiába nagyobb a zaj a kulcsszó alapú rendszerek adattárolásánál, ezt egyrészt ellensúlyozza az, hogy sokkal nagyobb a redundancia is ugyanott, másrészt sokat segít a zajt minimálisra csökkentő relevancia-rangsorú megjelenítés. Így a kulcsszó alapú rendszerekben, noha sokkal rendszertelenebbeknek, kevésbé kidolgozottaknak, véletlenszerűbbnek tűnnek, mégis könnyebben megtalálunk valamit, mint a szabályosan szervezett tárgyszó alapú rendszerekben.

9. Összefoglalás

Ebben a fejezetben összegzem a dolgozatom fő megállapításait és azután felvázolok egy továbbfejlesztési javaslatot a tanszéki digitális könyvtárakkal kapcsolatban a Coospace[57] rendszerében.

9.1 A megállapításaim összefoglalása

A dolgozatomban a számítógépes könyvkereséseket vizsgáltam. Az állításom az volt, hogy az utóbbi évtizedben megváltoztak a könyvkeresési szokásaink: ha a kereső lehetőséget ad rá, jobban szeretünk kulcsszó alapján keresni a korábbi tárgyszó alapú keresés helyett. A két fogalmat úgy definiáltam, hogy a tárgyszó alapú keresés egy kötött szókészletet használ, amit könyvekkel foglalkozó, az információkereső nyelvekben járatos szakemberek rendelnek a könyv-rekordokhoz, míg a kulcsszó alapú keresés szavait az indexelő program keresi ki a könyv-rekord elérhető összes adatából. Felhívtam rá a figyelmet, hogy az indexelés ekkor nélkülöz minden nyelvi intelligenciát, pusztán sztringek gyűjteményét jelenti. Azt is mondhatnánk, hogy a tárgyszavak elvileg nem nélkülözik a nyelvi intelligenciát, hiszen éppen a jelentésük alapján válogatjuk őket a könyvekhez.

Mégis az a tapasztalatunk, hogy noha a tárgyszavak jelentésükben pontosabban írják le a könyvet, és kevésbé redundánsak is, mint a kulcsszavak, mégis a kulcsszó alapú keresés a sikeresebb. Ezt részben a felhasználói szokások mások kutatásaira támaszkodó bemutatásával, részben a keresőfelületek tesztelésével illusztráltam.

A további észrevételem és kérdésfeltevésem az volt, hogy míg a felhasználói szokásokat magyarázhatjuk a kulcsszó alapú keresések esetén használt egyszerűbb keresőfelületekkel, a jobb találati listák nem magyarázhatók ugyanezzel. A jobb találati listák magyarázatára az információelméletben szokásos zaj és redundancia fogalmakat használtam. A redundanciának az a „feladata”, hogy csökkentse a zaj keltette ártalmakat. A keresőfelületek tesztelésével azt tapasztalhatjuk, hogy noha nő a zaj a kulcsszó alapú kereséseknél, az ugyancsak jelentősen növekvő redundancia és a találati lista jobb paraméterezési lehetőségei ezt bőven ellensúlyozzák, és összességében sikeresebb kereséseket eredményeznek, mint a tárgyszó alapú keresések.

[57] A kooperációs tér megnevezésből származó rövidítés. Tanulmányi és kutatási együttműködéseket támogató rendszer. Webcím: http://coospace.hu/portal/2.hu.page. Az általam bemutatott alkalmazások a PTE Coospace rendszeréből (https://coospace.tr.pte.hu/) vett minták. A letöltés ideje: 2008. nov. 16.

9.2 Következmények: hogyan tegyük még felhasználóbarátabbá a digitális könyvtárunkat?

A dolgozatom zárófejezetében az eddigi megállapításaim alapján igyekszem megmutatni egy optimális kereshetőségű digitális könyvtár létrehozásának alapjait a Coospace rendszerében.

A PTE Coospace-programban a BTK kezelésében jelenleg két ilyen digitális könyvtárról tudok: az egyik egy már adatokkal feltöltött, működő rendszer (Kommunikáció és Médiatudomány Digitális Könyvtár), a másik egy most induló új könyvtár, amely még nem is nyitott a látogatók előtt (Digitális Könyvtár – Nyelvtudományi Tanszék).

A Coospace lehetőséget ad arra, hogy színtereket hozzunk létre bizonyos együttműködési formákhoz (pl. egy digitális könyvtárhoz), ide dokumentumokat töltsünk fel és azokat keresni tudjuk (a felület még számtalan lehetőséget biztosít, de számomra most ezek a relevánsak).

A keresés egyszerű keresés felületen zajlik: azaz beírunk egy természetes nyelvi kifejezést, és a program keresni fog a könyvünk adataiban. A keresést a program a dokumentum címében, a szerzőben és a dokumentumhoz rendelt tárgyszókban hajtja végre. A keresés során a program használ csonkolást, de nem használ ékezetmentesítést, így a nagyon gyakori, a többes szám és a tárgyrag előtt tőváltó magyar szavak keresése nem egyszerű[58].

A másik probléma a tárgyszavak rendszere. Ez, ahogy a tárgyszórendszer lenni szokott, egy zárt rendszer. Az adatrögzítő szempontjából nem zárt, mert az adatrögzítéskor szabadon bővíthető, de a felhasználó (kereső) szempontjából zárt, mert praktikus a halmazzal megismerkednie a keresés előtt, hiszen nincs oka feltételezni (és valóban nincs is így), hogy tetszőleges szinonimák is előfordulhatnak tárgyszavakként (ahogy hosszabb leírásból indexelt kulcsszavakként nagy eséllyel előfordulnának). Nem kell nagyon sok könyvet rögzíteni ahhoz, hogy a tárgyszavak rendszere szinte áttekinthetetlenül nagy legyen, különösen, ha különböző időpontokban különböző emberek által felvitt „tudományos írásmű" témájú könyveknek mind az *esszéírás*, mind a *tudományos írásmű*, mind a *hogyan írjunk*, mind az *érvelési technikák*, mind a *hivatkozások kezelése* stb. tárgyszava lehet.

A tárgyszórendszerrel nem csak az a gond, hogy egy bizonyos méreten felül nem igazán átlátható, illetve, hogy a szinonimák szempontjából nem eléggé redundáns, hanem az is, hogy az összetett kifejezések egyetlen sztringként tárolódnak benne. Ha használtam azt a tárgyszót, hogy *hivatkozások kezelése*, akkor hiába keres valaki arra, hogy *hivatkozáskezelés* vagy *hivatkozás kezelés*, nem kap eredményt, mert a csonkolás csak a sztring elején vagy végén történik. Ugyanakkor az adatrögzítő még mindig úgy jár el legjobban, ha minél hosszabb szavakból álló és minél több szóból álló tárgyszavakat ad meg, mert úgy nagyobb az esély, hogy valaki eltalálja a megfelelő tárgyszórészletet, mint ha a tárgyszó rövid lenne. Például a *hivatkozások kezelése* tárgyszóval megjelölt könyv előjöhet találatként a *hivatkozás*, *hivatkozások*, *hivatkozások kezelése* kereső-kifejezésekre is, de persze a *hivatkozások kezelésének szabályai* kifejezésre már nem. Ha viszont a tárgyszó az lett volna,

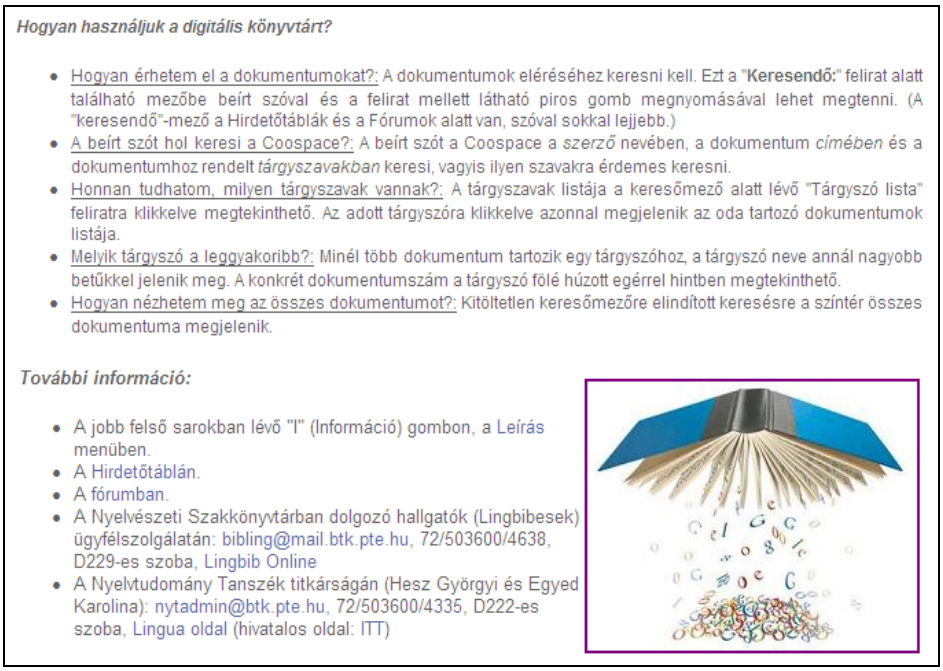
[58] Egy extrém példa: pl. ha a cím azt tartalmazza, hogy *logikát*, és bármilyen oknál fogva a *logika* tárgyszó nem tartozik a könyvhöz, akkor a *logika* kereső-kifejezésre ez a dokumentum nem lesz a találati listában.

hogy *hivatkozás*, akkor csakis erre az egy kereső-kifejezésre találnánk meg a könyvet, a másik felsorolt háromra nem.

Vagyis a dolgozatban leírt tapasztalatok alapján a javaslatom a gépi indexelés lenne: indexelés a címre, szerzőre, tárgyszóra (azaz a tárgyszó egyes szavaira is), illetve a leírásra, és ezek után a találati listának a relevancia szerint való rendezése. Saját tapasztalat alapján mondhatom (de lehetséges, hogy megoldható lenne a tevékenységek logolása is, és így ezt a feltételezést igazolni vagy cáfolni lehetne), hogy az ember (a kereső felhasználó) nem kéri le a tárgyszólistát, inkább újabb és újabb kereső-kifejezésekkel próbálkozik, és legfeljebb akkor rágja magát át a tárgyszólistán, ha nem sikerül találatot kicsikarnia a keresőprogramból.

Az indexelés bevezetésével nőne az egyes könyvekhez tartozó index-rekordok között a redundancia. Igaz, hogy nőne a zaj is, vagyis lennének nem jó indexelések, de a redundancia növelése és a találatok relevancia szerinti megmutatása az eddigi tapasztalatok alapján ezt ellensúlyozná[59].

Ma még tájékoztató szöveget kell írni a felhasználóknak a Coospace-es digitális könyvtár használatáról (lásd 11. ábra). Ron Tennant[60] megállapítását figyelembe véve nemsokára talán ez már nem lesz szükséges.



Hogyan használjuk a digitális könyvtárt?

- Hogyan érhetem el a dokumentumokat?: A dokumentumok eléréséhez keresni kell. Ezt a "**Keresendő:**" felirat alatt található mezőbe beírt szóval és a felirat mellett látható piros gomb megnyomásával lehet megtenni. (A "keresendő"-mező a Hirdetőtáblák és a Fórumok alatt van, szóval sokkal lejjebb.)
- A beírt szót hol keresi a Coospace?: A beírt szót a Coospace a *szerző* nevében, a dokumentum *címében* és a dokumentumhoz rendelt *tárgyszavakban* keresi, vagyis ilyen szavakra érdemes keresni.
- Honnan tudhatom, milyen tárgyszavak vannak?: A tárgyszavak listája a keresőmező alatt lévő "Tárgyszó lista" feliratra klikkelve megtekinthető. Az adott tárgyszóra klikkelve azonnal megjelenik az oda tartozó dokumentumok listája.
- Melyik tárgyszó a leggyakoribb?: Minél több dokumentum tartozik egy tárgyszóhoz, a tárgyszó neve annál nagyobb betűkkel jelenik meg. A konkrét dokumentumszám a tárgyszó fölé húzott egérrel hintben megtekinthető.
- Hogyan nézhetem meg az összes dokumentumot?: Kitöltetlen keresőmezőre elindított keresésre a színtér összes dokumentuma megjelenik.

További információ:

- A jobb felső sarokban lévő "I" (Információ) gombon, a Leírás menüben.
- A Hirdetőtáblán.
- A fórumban.
- A Nyelvészeti Szakkönyvtárban dolgozó hallgatók (Lingbibesek) ügyfélszolgálatán: bibling@mail.btk.pte.hu, 72/503600/4638, D229-es szoba, Lingbib Online
- A Nyelvtudomány Tanszék titkárságán (Hesz Györgyi és Egyed Karolina): nytadmin@btk.pte.hu, 72/503600/4335, D222-es szoba, Lingua oldal (hivatalos oldal: ITT)

11. ábra: tájékoztató a Digitális Könyvtár – Nyelvtudományi Tanszék Coospace-színtéren

[59] A zaj a tárgyszó alapú rendszereknél is elkerülhetetlen. Nem csak azért, mert a tárgyszavazó szakember döntése nem biztos, hogy mindig egybeesik a kereső felhasználók elképzeléseivel (lásd a korábban elemzett tárgyszó-problémát), hanem a mindig létező gépelési hibák és gépelési pontatlanságok miatt. A vizsgált digitális könyvtárakban találtam *inernet-hozzáférés*, *levelező szigorlat* és *levelezős szigorlat* tárgyszavakat. Ugyanezek a gépelési hibák vagy következetlenségek a kulcsszavas rendszerekben is előfordulnak. Például a Bookline oldalán a *maygar* kereső-kifejezésre is lehet találatot elérni. De mivel a kulcsszavas rendszerekben a kulcsszavak közé nagy eséllyel a jó gépelésű változat is bekerül (hiszen a kulcsszavak létrehozásánál az összes, a könyvhöz tartozó szöveget figyelembe vesszük), ezért az így előálló redundancia ellensúlyozza a zaj okozta problémát.

[60] Mi-Weng (2008) idézi Roy Tennant-tól (17. oldal): „I wish I had known that the solution for needing to teach our users how to search our catalog was to create a system that didn't need to be taught."

Felhasznált irodalom

1. ALBERTI Gábor (2006): *Matematika a természetes nyelvek leírásában*. Tinta Könyvkiadó, Bp.
2. Anderson, Chris (2004): The Long Tail. In: Wired Magazine, issue 12.10. Elérhetőség: http://www.wired.com/wired/archive/12.10/tail.html (a letöltés ideje: 2008. okt. 25.)
3. ANDERSON, James A. (2005): *A kommunikációelmélet ismeretelméleti alapjai*. Typotex, Bp.
4. B. HAJDU Ágnes – BABICZKY Béla (1998): *Bevezetés az információkereső nyelvek elméletébe és gyakorlatába*. Universitas Kiadó, Bp.
5. BÁRCZI Géza – ORSZÁGH László vezetőszerk. (1984-1992): *A magyar nyelv értelmező szótára*, 4. kiadás, Akadémiai Kiadó, Bp.
6. BATTELLE, John (2006): *Keress! Hogyan alakítja át kultúránkat, üzleti életünket a Google és az internetes keresés*. HVG Kiadó Zrt, Bp.
7. CIANCIOLO, Anna T. – Robert J. STERNBERG (2007): *IQ. Az intelligencia rövid története*. Corvina, Bp.
8. CRYSTAL, David (1998): *A nyelv enciklopédiája*. Osiris, Bp.
9. CZUCZOR Gergely – FOGARASI János (1999): *A magyar nyelv szótára*. MBE. Miskolc. Reprint. Eredeti kiadása: Budapest: Emich, 1862-1874
10. DRÓTOS László (1994): Hogyan keressünk az adatbázisokban? In: Bakonyi Géza – Drótos László – Kokas Károly (szerk.): *Navigáció a hálózaton. Információforrások a számítógépes hálózatokon*. Információs Infrastruktúra Fejlesztési Program, Bp. Lelőhely: Magyar Elektronikus Könyvtár. Link: http://mek.oszk.hu/01200/01290/html/navi8.htm#viii (letöltés ideje: 2008. szept. 24.)
11. DRÓTOS László (1999): Hálózati értelmező szótár. *NIIF Információs Füzetek, I/2*. N.I.I.F, Bp. Lelőhely: Magyar Elektronikus Könyvtár. Link: http://mek.oszk.hu/01200/01280/html/1.02/index.htm (letöltés ideje: 2008. szept. 24.)
12. FÜLÖP Géza (1984): *Ember és információ*. Második, átdolgozott kiadás (első kiadás: 1973). Múzsák Közművelődési Kiadó, Bp.
13. GRIFFIN, Em (2003): *Bevezetés a kommunikációelméletbe*, Harmat Kiadó, Bp.
14. GRIPSRUD, Jostein (2007): *Médiakultúra, médiatársadalom*. Új Mandátum Könyvkiadó, Bp.
15. HORÁNYI Özséb (2006): *Jel, jelentés, információ, kép*. (A jelen kötet alapjául szolgáló egyik eredeti mű kiadási adatai: Magvető Kiadó, 1975.) General Press Kiadó, Bp.
16. HYMES, Dell (1997): A nyelv és a társadalmi élet kölcsönhatásának vizsgálata. In: Pléh Csaba – Síklaki István – Terestyéni Tamás (szerk.): *Nyelv – kommunikáció – cselekvés*. Osiris, Bp. 458-495. oldal
17. JANSEN, Bernard J. – Amanda Spick – Sherry Kroshman (2007): Web Searcher Interaction With the Dogpile.com Metasearch Engine. In: *Journal of the American Society for Information Science and Technology*, 58(5): 744-755. oldal. Lelőhely: http://ist.psu.edu/faculty_pages/jjansen/academic/pubs/jansen_dogpile.pdf (letöltés ideje: 2008. szept. 18.)
18. JÓFÖLDI Hajnalka (2003): *Kultúrák hatása a weben való információkeresési szokásokra: Magyar felhasználók magatartása*. Szakdolgozat, BDF-TTK, Szombathely. Lelőhely: Magyar Elektronikus Könyvtár, http://mek.oszk.hu/01200/01288/ (letöltés ideje: 2008. szept. 25.)

19. K. Princz Mária – Nagy János (2005): A hallgatók keresési szokásai. In: Pethő Attila – Herdon Miklós (szerk.): *Informatika a felsőoktatásban 2005*, Debreceni Egyetem, Informatikai Kar, Debrecen (konferencia-kiadvány, Informatika a felsőoktatásban, Debrecen, 2005. aug. 24-26.). Lelőhely: http://agrinf.agr.unideb.hu/if2005/kiadvany/papers/G81.pdf (letöltés ideje: 2008. szept. 25.)
20. Kenesei István szerk. (2004): *A nyelv és a nyelvek*. Ötödik, átdolgozott kiadás, Akadémiai Kiadó, Bp.
21. Krippendorff, Klaus (1995): *A tartalomelemzés módszertanának alapjai*. Balassi Kiadó, Bp.
22. Krug, Steve (2008): *Ne törd a fejem! Felhasználóbarát webdizájn*. HVG Könyvek. HVG Kiadó Zrt, Bp. [Az eredeti mű címe: Don't Make Me Think! New Riders Publishing Berkely, California, 2006, 2. kiadás. Fordította: Nagy Marcell]
23. Mi, Jia – Cathy Weng (2008): Revitalizing the Library OPAC: Interface, Searching, and Display Challenges. In: *Information Technology & Libraries*, 2008. márc., Vol. 27. Issue 1, 5-22. oldal. (Elérhetőség: EBSCO adatbázis, http://search.ebscohost.com, 2008. dec. 31-ig elérhető a PTE hálózatából)
24. Pajor Enikő (2006): *A láthatatlan/mély web felhasználása a könyvtári tájékoztatásban*. Doktori disszertáció. ELTE BTK Irodalomtudományi Doktori Iskola – Könyvtártudományi Program, Bp. Lelőhely: http://szilleri.tvn.hu/disszertacio.pdf (letöltés ideje: 2008. okt. 17.)
25. Prószéky Gábor – Kis Ádám (1999): A szövegkeresés nyelvi támogatása. In: Uők: *Számítógéppel emberi nyelven. Intelligens szövegkezelés számítógéppel*. Szak Kiadó, Bicske. 177-226. oldal
26. Rethlefsen, Melissa L. (2008): Easy ≠ Right. In: *Library Journal*, 2008. nyár, Vol. 133. 12-14. oldal. (Elérhetőség: EBSCO adatbázis, http://search.ebscohost.com, 2008. dec. 31-ig elérhető a PTE hálózatából)
27. Rosengren, Karl Erik (2008): *Kommunikáció*. Ötödik kiadás, Typotex, Bp.
28. Siró Béla (2007): *Könyvtári ismeretek, könyvtárhasználat, információtechnika*. Pedellus Tankönyvkiadó Kft, Debrecen.
29. Spector, Robert (2003): *Amazon.com. Minél előbb – minél nagyobbra*. Alexandra Kiadó, Pécs
30. Stehr, Nico (2007): Tudás és információ. In: Uő: *A modern társadalmak törékenysége. Tudás és kockázat az információs korban*. Gondolat-Infonia, Budapest. 44-47. oldal
31. Szőnyi György Endre (2004): Szemiotika, ikonográfia, interpretáció. Eco és a kacsacsőrű emlős. In: Uő: *Pictura & scriptura. Hagyományalapú kulturális reprezentációk huszadik századi elméletei*. JatePress, Szeged. 229-242. oldal
32. Ungváry Rudolf – Vajda Erik (2002): *Könyvtári információkeresés*. Második, javított kiadás. Typotex, Bp.
33. Ungváry Tibor – Orbán Éva szerk. (2001): *Osztályozás és információkeresés. II. kötet: Az információkeresés és elmélete*. Országos Széchenyi könyvtár, Bp.