# TEXT-PAC
## szabad szöveges információvisszakereső rendszer alkalmazása a KSH könyvtárában

SZABÓNÉ TÖRS HANNA

A KSH Könyvtár és Dokumentációs Szolgálat (a továbbiakban: KSH Könyvtár) jelenlegi állománya meghaladja a 600 ezer kötetet, ebből statisztikai állomány mintegy 200 ezer kötet (közel 90 ezer kötet hivatalos statisztikai adatközlő publikáció és több mint 100 ezer kötet módszertani, elemző és leíró dokumentum.) A KSH Könyvtár 900-nál több külföldi és 300-nál több magyar szakfolyóiratot járat. A könyvtári forgalomra jellemző, hogy a helyben olvasott, illetve kölcsönzött könyvek átlagos száma évenként 55–60 ezer, és ezt 30–35 ezer fő veszi igénybe.

A KSH Könyvtár mint az egyetlen magyarországi statisztikai szakkönyvtár kettős rendeltetésű: nyilvános szakkönyvtári minőségében ki kell elégítenie az igen széles spektrumú olvasói táborának még szélesebb skálán megfogalmazott igényeit. Hivatali könyvtárként pedig hatékonyan kell segítenie a KSH-ban dolgozó statisztikusokat, közgazdászokat, jogászokat, demográfusokat az idegen és magyar nyelvű szakirodalom gyors feltárásával, továbbá statisztikai adatforrásokat kell a számszerű információt kereső felhasználók rendelkezésére bocsátania.

A KSH alapítása óta a statisztika tárgykörei egyre szélesebbekké váltak. A statisztika elmélete, módszertana és a hozzá is kapcsolódó új tudományterületek pl. ökonometria, vezetéstudomány, operációkutatás, szociológia, regionális gazdaságtan stb. köre egyre terjeszkedik. Ahogyan az ismeretek bővülnek, úgy nőnek a követelmények és az igények a szakirodalmi tájékoztatás iránt, ezért a hagyományos feltárási eszközök és módszerek már nem kielégítőek. A manuális dokumentációs eszközök időigényesek, lassúak, az információk többoldalú feltárását korlátozzák. Ezért merült fel a KSH Könyvtárban a gépi adatfeldolgozás lehetősége: egy számítógépes információvisszakereső rendszer alkalmazásának terve.

Miután megfogalmazódott a gondolat, hogy az információszolgáltatást „új utakra" kell vezetni, fel kellett mérni az információszolgáltatás iránti igényeket. Az alkalmas információtároló és -visszakereső rendszer kiválasztása céljából elvégzett vizsgálat az alábbi szempontokat vette figyelembe:

- A KSH az éves tervben jelöli ki a Főosztályoknak a tárgyévi munkát. Ez az éves terv fontos útmutató a Könyvtár számára. Megmutatja, hogy mely szakterületeken milyen irányban, milyen időszakban, milyen mélységgel fog jelentkezni az információs igény.
- Az előbbihez hasonló jelzést adnak a kutatási tervek is.

– Mint a könyvtárakban általában, a KSH Könyvtárban is a referenszben csapódnak le a konkrét, rendszerint ad hoc olvasói kérdések, amelyekről a kijelölt figyelőnapok – évek óta rendszeresen vezetett – témalapjai nyújtanak bővebb felvilágosítást.
– A KSH Főosztályok a témafigyelési kéréseikben általában nagyobb témacsoportokat ölelnek fel, ezekről a tájékoztatást havonta igénylik.

A fenti szempontok szerint elvégzett előzetes igényfelmérés azt a felismerést eredményezte, hogy a feldolgozandó szövegben rejlő statisztikai adat, tábla, számítás avagy módszer olyan jellegű információ, amelyhez a felhasználó a hagyományos módszerekkel megnyugtató biztonsággal nem férhet hozzá. A megoldás egy olyan számítógépes szöveges feltáró és visszakereső rendszer, amellyel ki lehet elégíteni a rendkívül gazdag szakirodalomból a speciális igényeket.

1974–1975-ben az INFELOR Rendszertechnikai Vállalat többéves kutatómunka után a KSH Számítógépalkalmazási program keretében nyilvánosságra hozta az IBM cég nyolc információtároló és -visszakereső programcsomagjával (IRMS, TEXT-PAC, DPS, HYPHENATION/360, COMPOSITION/360, TEXT/360, KWIC/360, STAIRS) végzett kísérleteinek eredményeit. Ezek a programcsomagok szakkönyvtárakban, dokumentációs központokban alkalmazhatók dokumentumok nyilvántartására, kurrens és retrospektív témafigyelésre, bibliográfiák és katalógusok készítésére. Az említett információvisszakereső rendszereket összehasonlítva kitünt, hogy a KSH Könyvtár számára a TEXT-PAC rendszer a legmegfelelőbb.

A TEXT-PAC rendszer kiválasztását a következő előnyei indokolták:
– retrospektív és kurrens információvisszakeresési lehetőség;
– indexek készítése (a tárgyszavak, a szerzők neve, a dokumentum forrása, a kategóriák szerint; KWOC-index);
– a keresőkérdés eleme a szöveges feldolgozás bármely szava lehet;
– a keresőszavakat a szótő, illetve annak bármely képzett alakja alapján lehet kialakítani (maszkolás);
– a visszakeresést Boole-féle logikai kapcsolók (AND, OR, NOT) és szövegösszefüggést figyelembevevő kapcsolók (ADJ, WITH) segítik;
– lehetőség van meghatározott adatmezőkben (pl. címmező, szerző-mező stb.) való keresésre;
– a keresés során meg lehet határozni az adatbázisban a keresési tartományt;
– a számítógépes rendszer meg tudja különböztetni a kis- és nagybetűket;
– a szelektív információterjesztéshez a rendszer tárolja a keresőprofilokat, ezekre vissza lehet hivatkozni.

A TEXT-PAC az IBM cég 28 programból álló, kötegelt (batch) üzemmódban használható információvisszakereső rendszere. A KSH Könyvtár a rendszerrel való kísérleteit 1975-ben kezdte meg. A kísérletek először a feldolgozás munkameneteinek kialakítására, később az indexek készítésére, legvégül 1977 második felétől a természetes beszélt nyelv szavaiból összetevődő szabad szöveges keresőkérdések megfogalmazására terjedtek ki. A kísérletek elvégzői a hagyományos tájékoztatással szemben számos új problémával

találkoztak mind a feltárás, mind a visszakeresés során. Jelen cikk összefoglaló tájékoztatást ad a kísérletek tapasztalatairól és eredményeiről.

A TEXT-PAC rendszerrel – mint szabad szöveges információvisszakereső nyelvvel – történő tájékoztató munkának sok előnye van a hierarchikus és a fazettás szerkezetű indexelő nyelvekkel szemben. Mind a hierarchikus, mind a fazettás szerkezetű indexelő nyelvek esetében az indexkifejezések közötti alá- és fölérendeltségi viszonyokat (hierarchikus kapcsolatokat) az előkészítő munka során kell megadni. Ezek az indexelő nyelvek általában szinonímaszótárakat, vagy tezauruszokat igényelnek. A TEXT-PAC rendszer – mint szerkezet nélküli információvisszakereső nyelv – nem igényel sem szinonímaszótárt, sem tezauruszt. Használatához kategóriajegyzék és tárgyszójegyzék szükséges, amelyek sokszorosított formában állnak a dokumentátorok rendelkezésére. A kategóriák az érdeklődésre számot tartó anyagot szakterületek, ágazatok szerint csoportosítják (pl. mezőgazdaság, ipar, külkereskedelem stb.) és a visszakeresés hatékonyságát segítik elő. A tárgyszavak egy-egy kategórián belüli kulcsszavak egy-egy kérdéskör szűkebb lehatárolására. A dokumentátor minden feldolgozáshoz kategóriákat és tárgyszavakat rendel. A kurrens anyagot a számítógép meghatározható rendszeres időközökben az előbbiekben említett különféle szempontok szerinti rendezettségben kinyomtatja.

## A rendszer inputja

A TEXT-PAC rendszer egy-egy input tétele különféle típusú adatmezőkből épül fel. Az egyes adatmezők alfanumerikus karakterekből (a természetes nyelv szavaiból) tevődnek össze. A különféle típusú adatmezőket egy háromjegyű szám (a KSH Könyvtár csak kétjegyűt használ) azonosítja, amelyeket „print control"-nak nevezünk. A KSH Könyvtár által adaptált TEXT-PAC rendszer input adatlapja a következő adatmezőket tartalmazza:

| *print control* | *adatmező* |
|---|---|
| 00 | magyar cím |
| 01 | eredeti cím |
| 02 | nyelvi megjelölés |
| 10 | forrás (a folyóirat megjelenési adatai) |
| 20 | szerző(k) |
| 30 | forrás (raktári jelzet) |
| 40 | tartalmi feltárás (annotáció) |
| 41 | országkód |
| 42 | időhatárok |
| 43 | a feldolgozó, ill. az ellenőr nevének kezdőbetűi |
| 50 | statisztikai táblák |
| 60 | kategóriaszám |
| 61 | tárgyszavak |

A kategóriák, a tárgyszavak és a földrajzi kódok használatához segédletek készültek.

A tárgyszavak összeállítása körültekintő munkát igényelt, hiszen a statisztika társtudományai és társterületei (szociológia, demográfia, munkaügy stb.) ún. lágy (soft) tudományágak és minél „lágyabb" a tudományág, annál kevésbé rendezett az ismeretkészlete, logikailag annál ellentmondásosabb a fogalomrendszere.

Az egzaktabb társterületek (pl. közgazdaságtan, iparstatisztika stb.) fogalmi rendszere szintén nehezen közelíthető meg az egységesség igényével, ezért az azonos tartalmú, de eltérő formájú (szinonim) tárgyszavakat utalórendszerbe kellett foglalni (pl. nemzetgazdasági elszámolások → népgazdasági elszámolások). Komoly gondot okozott az eltérő gazdasági, ipari stb. ágazati nomenklaturák osztályozása (pl. az ipar felosztása az egyes országokban az ENSZ vagy a KGST ajánlása szerint minden esetben eltérő, közös nevezőre szinte nem is hozható).

A kidolgozott tárgyszógyűjtemény tárgyszavai között nincs létrehozva hierarchia, csupán a szinonímák és kvázi-szinonímák vannak utalórendszerbe foglalva. Bár a tárgyszavak mindegyike alá van rendelve egy kategóriának, önállóan, a kategórián kívül is használhatók (pl. az építőipar kategóriában található az építőiparra vonatkozó termelés, munkaügy, foglalkozásszociológia stb. tárgyszavak, de a munkaügy tárgyszó az egészségügyi, ipari, építőipari stb. kategóriába tartozó munkaüggyel foglalkozó cikkek indexeléséhez is használható).

## A rendszer outputjai

Jelen cikk írásakor az adatbázis nagysága kb. 9000 feldolgozott dokumentum-tétel. Ez az adatbázis havonta 6–700 tétellel növekszik.

A különböző szempontú indexek közül a KSH Könyvtár jelenleg a következőket állítja elő havonkénti rendszerességgel:

A *Bulletin* a kurrens dokumentumok feltárt részét tartalmazza. A Bulletin adatmezői megegyeznek az inputlap adatmezőivel, kiegészítő információ a hivatkozási azonosítószám, amely egyben a dokumentum sorszáma is.

A *kategóriaindexben* a kategóriák sorrendjében rendezve jelenik meg a dokumentum magyar címe, forrása és a Bulletinre való hivatkozás.

A *tárgyszóindex* a tárgyszavak betűrendjében tartalmazza a dokumentumok magyar címét, forrását és a Bulletinre való hivatkozást.

A *szerzői index* a szerzők nevének betűrendjében közli a hivatkozási adatokat.

A *forrásindex* a folyóiratok raktári számának növekvő sorrendjében közli a hivatkozási adatokat.

A TEXT-PAC rendszer outputjai közé tartoznak még a kurrens és a retrospektív kérdésekre kapott kinyomtatott válaszok is.

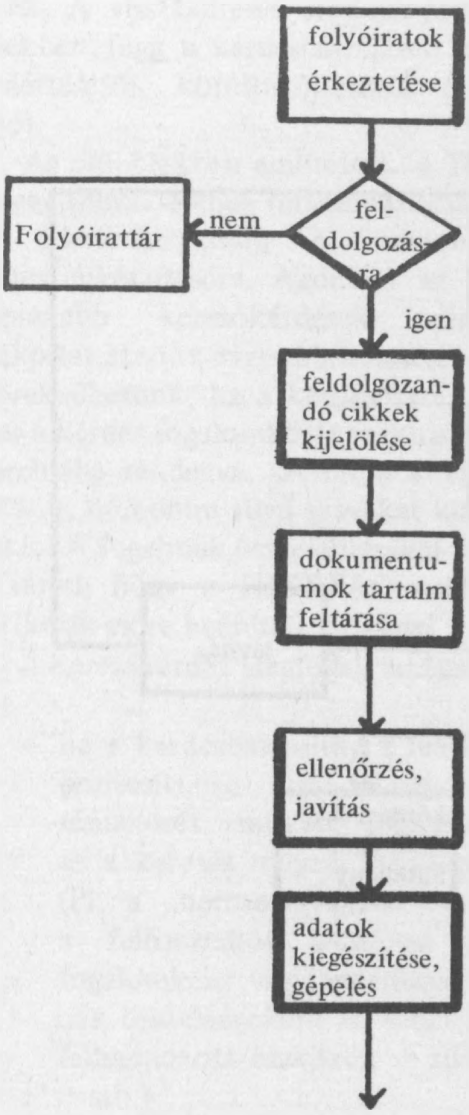
A *kurrens témafigyelés* lényegében az adott adatbázis legfrissebb részének vizsgálatát és lekérdezését jelenti. Az adatbázis havonta meghatározott tételszámmal növekszik és az információ ebből a friss adatbázis-részből sugárzódik szét a felhasználók felé. Az állandó keresőkérdések a felhasználói „profilok", amelyeket a folyamatos üzemelés során ki lehet egészíteni, át lehet alakítani vagy törölni.

A *retrospektív keresés* az adatbázis retrospektív (visszatekintő) feltárása, amely minden esetben a teljes adatbázisra vonatkozik. Míg a kurrens témafigyelésnél a kérdezés mindig ugyanazokkal a „profilokkal" történik, addig a retrospektív témakutatásnál a felhasználó ad hoc igényeire esetenként kell a keresőkérdéseket összeállítani.

## A TEXT-PAC rendszerrel való szöveges feldolgozás folyamata

(ld. 1. ábra)

A KSH Könyvtárba járó szakfolyóiratok közül jelenleg 355 idegen nyelvű és 111 magyar nyelvű van a tartalmi feltárásba bevonva. A feltárandó folyóiratcikkeket a következő szempontok szerint választjuk ki:




1. ábra

– a cikkek csak éves statisztikai adatokat tartalmazzanak;
– a feltárás iránya és mélysége feleljen meg mindenkor a felhasználói igényeknek;
– a feltárandó cikkek tartalma feleljen meg a Könyvtár gyűjtőkörének.

Az adatbázis tartalmazza a külföldi statisztikai hivatalok központi folyóiratainak összes cikkét címfordítás, kategória, tárgyszó hozzárendelésével, de ezek esetében nem minden cikkhez készül annotáció.

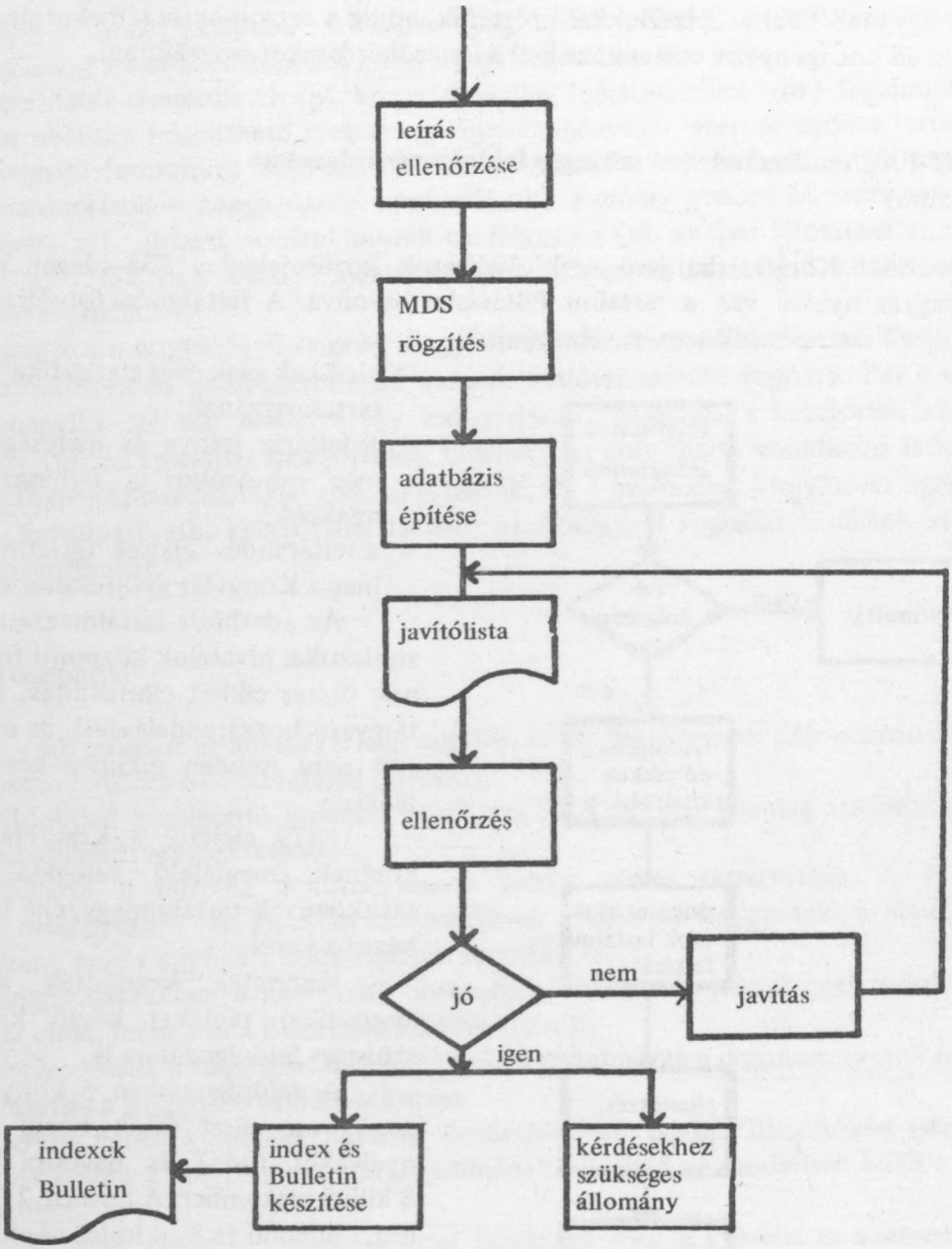
1979 elejétől a Könyvtár gyűjtőkörének megfelelő válogatás szerinti szakkönyvek tartalomjegyzéke is az adatbázisba kerül.

Kísérletek kezdődtek a csupán statisztikai táblákat közlő kiadványok szöveges feldolgozására is.

A feldolgozásban 5 könyvtári dolgozó vesz részt (akik közül 4 idegen nyelveket tudó) és havonta átlagosan 8 külső szakember. A javítást 2 fő, a gépelést 1 állandó és 3–4 külső gépíró végzi.

## A TEXT-PAC rendszerrel való visszakeresés folyamata

A feldolgozó szakemberek témák szerint vannak beosztva a nyelvi ismerete-




1/b ábra

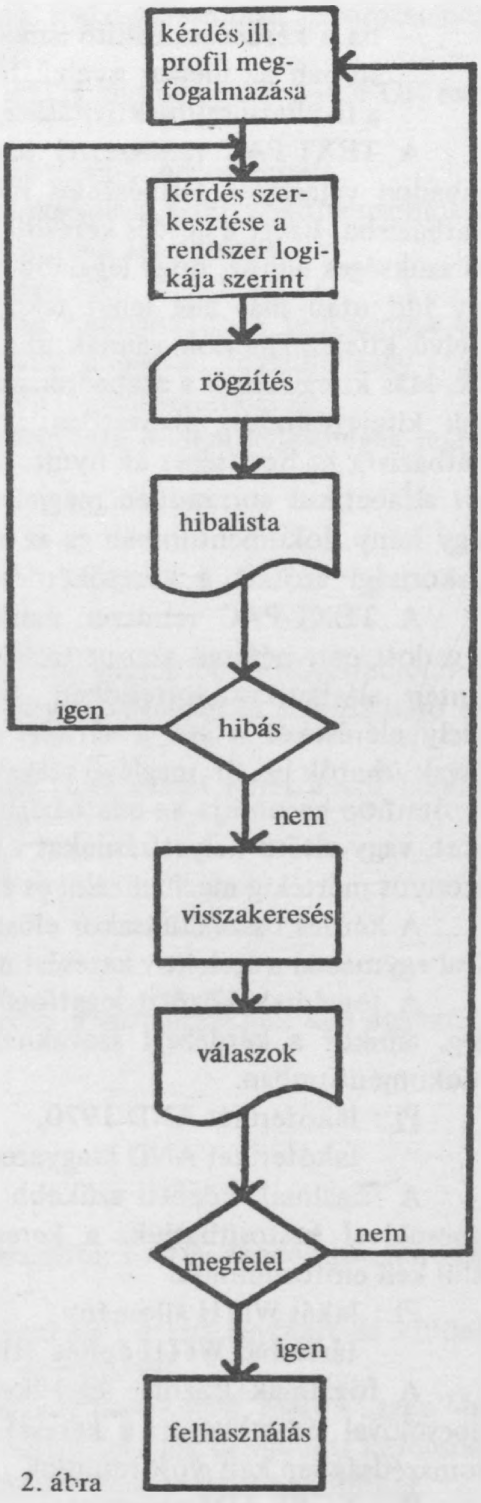
ket is figyelembe véve. Szakterületük folyóiratcikkeit feldolgozzák és később az adott témában a számítógépes rendszer számára a felhasználóval megbeszélt keresőkérdéseket is megfogalmazzák (ld. 2. ábra).

Az információvisszakereső rendszer akkor éri el célját, ha a felhasználó egyfelől a rendelkezésre álló adatbázisból az őt érdeklő valamennyi dokumentumról értesüléseket kap, másfelől ha nem kell foglalkoznia azokkal a dokumentumokkal, amelyek látszólag ugyan beletartoznak az érdeklődési körébe, de bizonyos kizáró okok miatt számára mégis érdektelenek. A visszakeresés eredményessége nagy mértékben függ a kérdésösszeállító szakember hozzáértésétől, körültekintésétől és gondosságától.

Az előbbiekben említettük: a TEXT-PAC rendszer működéséhez nincsen szükség a fogalmak hierarchizálására és szinonímaszótár előzetes elkészítésére. Azonban az átfogóbb, komplexebb keresőkérdések megoldásánál a gyakorlat szerint nagyobb keresési pontosságra törekedhetünk, ha a kérdésösszeállító szakember a kérdés fogalomkörét szakmailag pontos hierarchiába rendezve, szinoním kifejezésekkel kibővítve, homonim tövű szavakat kizárva állítja össze. A fogalmak összeállításánál figyelembe kell venni, hogy a TEXT-PAC rendszer nem rendelkezik előre beépített szótárral.

A keresőkérdés megfogalmazásánál előnyt jelent

– ha a kérdésösszeállító a felhasználóval pontosíttatni tudja, hogy a kérdés témakörét milyen fogalmak adják és a keresés milyen mélységű legyen. (Pl. a „nemzeti vagyon" érdekelheti a felhasználót általában generikus fogalomként vagy specifikus fogalmainak összességeként is, tehát nemzeti vagyon → tiszta vagyon → reáleszközök → felhalmozott eszközök → állóeszközök → befejezetlen beruházások → készletek → stb.)



2. ábra

– ha a kérdésösszeállító ismeri a szakterületet, annak szóhasználatát és nem utolsósorban az idegen nyelvű terminológiát. Ennek nagy jelentősége van elsősorban a fordítási szubjektivitásból adódó terminológiaeltérések felfedésénél.

A TEXT-PAC rendszerrel történő feldolgozás során megengedett, hogy minden szabadon választott természetes nyelvű kifejezésmód fokozatosan bekerüljön a kereső adatbázisba. Ezért a kérdés keresőfeltételeinek összeállításakor egyre nagyobb körültekintés szükséges ahhoz, hogy legalább megközelítően teljes legyen a visszakeresési eredmény. Egy idő után már alig lehet rekonstruálni, vagy áttekinteni, hogy milyen természetes nyelvű kifejezésmódok vannak az adatbázisban, amelyek keresőszavakként alkalmazhatók. Más kifejezéssel: a szabadon választott természetes nyelvek lényege az, hogy a fogalmak kifejezésmódja alapvetően előre nem kiszámítható és ennek hatása kiterjed az adatbázisra is. Segítséget az nyújt, hogy a TEXT-PAC rendszer egy *szógyakorisági szótárban* alfabetikus sorrendben megjeleníti az adatbázisban előforduló szavakat, jelezve azt, hogy hány dokumentumban és az adatbázisban összesen hányszor fordulnak elő. A szógyakorisági szótárt a keresőkérdések összeállításánál igen előnyösen lehet használni.

A TEXT-PAC rendszer másik szójegyzéke az ún. *sűrített szótár,* amely előre megadott paraméterek szerint tartalmazza az adatbázisban többször előforduló szavakat, szintén alfabetikus sorrendben. Paraméterrel az az előfordulási szám adható meg, amely elérésekor a szó a sűrített szótárba bekerül. A sűrített szótár módosítható: új szavak írhatók be, ill. meglévő szavak törölhetők. A számítógép a javítómenetekben ehhez a szótárhoz hasonlítja az adatbázisba újonnan bekerülő szavakat: a szótárban nem szereplőket, vagy eltérő helyesírásúakat a javítandó tételekhez írja ki. Ily módon a javítási fázist bizonyos mértékig mechanizálni és egyszerűsíteni lehet.

A kérdés összeállításakor először a kérdést leíró szavakat kell a kapcsolók használatával egymással a szelektív keresést meghatározó logikai kapcsolatba hozni:

A fogalmak közötti legátfogóbb kapcsolatot az [AND] kapcsolóval teremthetjük meg, amikor a kérdezett szavaknak, vagy kifejezéseknek együttesen kell előfordulniok a dokumentumban.

Pl.: lakóterület AND 1970,
lakóterület AND Magyarország stb.

A fogalmak közötti szűkebb kapcsolatot a keresés folyamán a [WITH] mondatkapcsolóval biztosíthatjuk: a keresett szavaknak vagy kifejezéseknek egy mondaton belül kell előfordulniok.

Pl.: lakás WITH állomány,
társasház WITH építés stb.

A fogalmak közötti legszűkebb kapcsolatot az [ADJ] (adjacent = szomszédos) kapcsolóval érhetjük el: a keresett szavaknak az adott sorrendben, egymással szoros szomszédságban kell előfordulniok.

Pl.: KGST ADJ országok,
Egyesült ADJ Királyság stb.

Az ADJ kapcsoló megfelelő használatával elkerülhetjük a hibás találatokat, így pl.: „a fejlődő országok gazdasága" különválasztható a „nyugati országok fejlődő gazdaságá"-tól.

A kérdést pontosító fogalmak, esetleg szinonímák, kvázi szinonímák felsorolásához, csoportosításához az [OR] (vagy) kapcsolót használjuk.

Pl.: lakás OR társasház OR ház OR öröklakás WITH állomány OR építés OR megszünés OR bontás stb.

A visszakeresés folyamán lehetnek különböző kikötéseink, nevezetesen:

– a [NOT] kapcsolót használjuk, ha nincs szükségünk azon dokumentumokra, amelyekben az általunk megjelölt bizonyos fogalmak szerepelnek.

Pl.: ház NOT házasság OR háztartás OR háztűznéző stb.

– az [ABS] (absolute) kapcsolót használjuk, ha mindenképpen szükségünk van azon dokumentumokra, amelyekben az ABS-sal megjelölt fogalom vagy fogalmak szerepelnek, függetlenül a lekérdezés egyéb logikai feltételeitől.

A logikai és pozicionális kapcsolókon kívül a következő logikai lehetőségek segítik a visszakeresés hatékonyságát:

**Maszkolás** (csonkolás)

A maszkolás lehetővé teszi, hogy a dokumentátor egy vagy több szótövet adjon meg a keresőképben és az eredményt a szótövekből bármilyen végződéssel kialakítható szó alapján szolgáltassa.

*Relatív maszkoláskor* adott számú, maximálisan 6 karaktert lehet a szó tövéhez csatolni, amely kijelöli a keresett szó maximális hosszát.

| | |
|---|---|
| Pl.: szén⌐⌐⌐⌐⌐⌐ | szén⌐⌐⌐ |
| szénfekete | széna |
| széna | szénről |
| | szénben |

*Abszolut maszkolás* esetében a kereséskor csupán a szótörzseknek kell egyezniök tekintet nélkül a szó hosszára.

Pl.: szén⌐*
széna
szénbányászati
széntüzelésű

Relatív maszkolást általában akkor célszerű használni, ha egy szónak csupán ragozott alakjaira van szükségünk, a szóösszetételeire nem.

Pl.: a szénexport, szénárak keresésénél számítani kell a szó egybeírt és különírt birtokos változatára is, tehát szénnek . . . . . exportjával
ebben az esetben nem alkalmazhatunk abszolut maszkolást, mert a „szén"-hez sok olyan összetétel kapcsolódhat, amely nem felel meg az eredeti kérdésnek, így pl.: a szénbányászati (termékek) exporthitele nem érdekel bennünket.

Abszolut maszkolást érdemes alkalmazni, ha a kérdésben olyan sok a szóösszetétel, hogy érdemesebb néhány irreleváns szóösszetételt megkapnunk, mint felsorolni az összes szükséges szóösszetételt.

Pl.: az „anyag- és energiahelyzet" kérdés elemzéséhez szükség van a
szénexport
import
kivitel
behozatal
ár
felhasználás
termelés
kitermelés
bányászat
szóösszetételekre, információs zajként ebben az esetben viszont megkaphatjuk a
szénfekete
széntüzelésű
széna
Szénási stb.
számunkra jelen esetben irreleváns szavakat. Ezeket természetesen a NOT kapcsoló használatával a keresésből letilthatjuk, ha figyelmünk erre kellően kiterjed és némi gyakorlatra is szert tettünk.

Abszolut maszkolást érdemes használni továbbá akkor is, ha a keresőszónknak feltehetően nincsen olyan további szóösszetétele, amely nem felel meg számunkra.

Pl.: gépkoncentráció¬*, kapacitástervezés¬*

A KSH Könyvtárban szerzett tapasztalatok azt bizonyítják, hogy amikor a különírás-egybeírás miatt kétféleképpen kell a szót megkérdezni, a különírt változatnál a relatív maszkolást, az egybeírt változatnál az abszolut maszkolást célszerű alkalmazni.

Pl.: kapacitások koncentrációja
kapacitásainak WITH koncentrációjával
koncentrációjának

(relatív maszkolás)
kapacitáskoncentráció¬*
(abszolut maszkolás)

Az elvégzett kísérletek szerint a 6 szóhosszkijelölő karakter elegendő. Ezt ugyan a TEXT-PAC rendszer angol változatánál állapították meg, azonban a kétféle maszkolási lehetőséget a fent említett elgondolások szerint alkalmazva a többes szám birtokos eset ragozott alakjai nem tesznek ki többet 6 karakternél.

Maszkoláskor ügyelni kell arra, hogy a képzett szavak (származékszavak) a keresésből le legyenek tiltva, amennyiben erre szükség van. Ellenkező esetben nagy lesz a hibás találatok száma.

Pl.: ha a „házak nagysága, megszünése" kérdés esetében – amennyiben katerógia-megkötés nélkül keresünk – nem tiltjuk le a „házasság", „háztartás" stb. szavakat, akkor az építőipari cikkek között megkapjuk a demográfiai (házasságok megszünése, háztartások nagysága) cikkeket is.

Ügyelni kell arra is, hogy léteznek olyan szavak, amelyek a kérdésben főnévként szerepelnek,és amelyeket a maszkolás alapján melléknévi igenévként megkapunk.

Pl.: kereset¬* keresett

Ezen segíthet a NOT letiltás.

Nem küszöbölhető ki azonban a kereséséből a műveltető igék egyes szám harmadik személyű alakja, amennyiben ez az alak azonos szóképű főnévnek felel meg:

Pl.: irat irat
(cselekvés) (tárgy),

valamint a homonimák többi formája (pl. ég, anyag stb.).

Ugyancsak maszkolási probléma, hogyha a felhasználó számára releváns két azonos tövű szó szerepel a kérdésben, akkor a maszkolást a közös tőnél kell kezdeni, nem pedig az illető szavak mindegyikénél.

Pl.: kapacitáskihasználásának
kapacitáskihasználtságából

Hangváltó töveknél (pl. utca – utcából, óra – órái stb.) a maszkolást az utolsó mássalhangzótól kell kezdeni.

A [CONTROL] kapcsolóval meghatározható a keresés végrehajtása szempontjából figyelembe veendő mező (print control).

Pl.: ha statisztikai számadatokra van szükség, akkor az 50-es statisztikai táblamezőre kell korlátozni a keresést.

A [NOT CONTROL] kapcsolóval meghatározott mezők kizárhatók a keresésből.

Pl.: a magyar „ón" szó a keresés folyamán az eredeti címmezőben az angol „on" szónak homonimája, ezért az irreleváns találatok kiküszöbölése érdekében a 01-es eredeti címmezőt a fenti kapcsolóval lehet a keresésből letiltani.

A magyar folyóiratokból történő feltáráskor a keresőszavak sok esetben azonosak lehetnek a folyóirat nevével, pl. Külgazdaság, Építőanyag, Magyar Alumínium stb. ilyenkor a forrásmezőt kell letiltani a keresésből.

**Összetett keresési módszerek**

Az előbbiekben bemutatott módszerekkel egyszerű keresőprofilok alakíthatók ki. A TEXT-PAC számítógépes rendszerben ezen egyszerű kérdéselemek felhasználásával tetszés szerinti bonyolultságú kérdéscsoportok alakíthatók ki. Így pl. a logikai és a pozicionális kapcsolók segítségével logikai kifejezések (logikai szintek) építhetők fel. A különféle logikai kifejezések egy magasabb szinten egymással logikai relációkba hozhatók. Ezeket a magasabb szintű logikai kifejezéseket a TEXT-PAC rendszerben CONCEPT-eknek nevezik. Ilyen módon kérdésstruktúrák jönnek létre, amelyekben a kérdések tetszés szerinti mélységű hierarchiája valósítható meg.

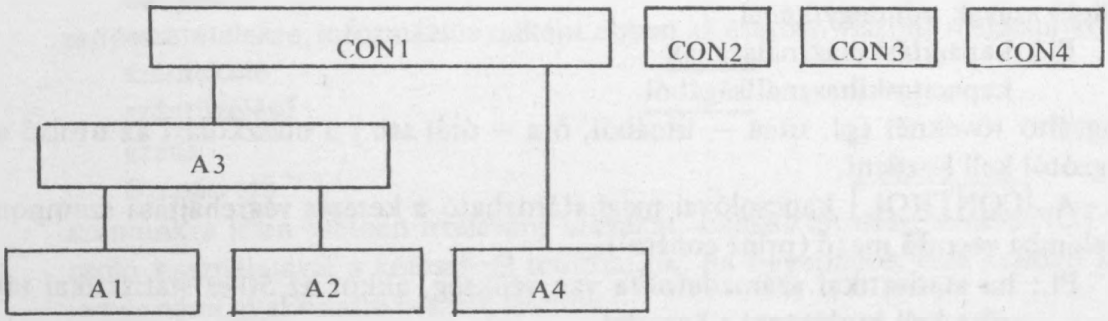
Példa:

Cím: Lakáshelyzet

A1 ház¬¬¬¬¬ OR lakás¬¬¬¬¬ OR társasház¬¬¬¬¬ OR panelház¬¬¬¬¬ OR öröklakás¬¬¬¬¬

A2 állomány¬* OR épít¬* OR megszün¬* OR bont¬* OR lerombol¬* OR felszerel¬* OR lebont¬* OR nagyság¬*

A3 A1 WITH A2
A4 lakásállomány¬* OR lakásépít¬* OR lakásmegszün¬* OR lakásbont¬* OR lakásfelszerel¬*
CON1 A3 OR A4
CON2 1977
CON3 Magyarország¬* OR Nagy-Britanni¬* OR Ausztri¬*
CON4 NOT házasfél OR házasok OR házaspár OR házasság OR házbér¬¬ OR házi¬¬¬¬




A keresőkérdés összeállításánál minden esetben meg kell adni, hogy hány CONCEPT-nek kell teljesülnie. (Jelen példában egyidejűleg négy CONCEPT-nek kell teljesülnie.)

A keresőkérdésekre a releváns válaszok általában a számítógépes program egy, két esetleg három futtatása során készülnek attól függően, hogy a kérdés mennyire vonatkozik konkrét, egyértelmű tárgyra. Ha a kérdés tárgyköre egyértelmű szakkifejezésekkel fedhető le, akkor elegendő egyetlen futtatás (pl. termelési függvény, kapacitáskihasználás az iparban stb.). Ha viszont túl átfogó, vagy túl speciális témát ölel fel a kérdés, és nehezen fogható meg a szakkifejezések halmaza, akkor több futtatásra van szükség, és ezek során a kérdéseket a kinyomtatott válaszok ismeretében javítani lehet. A felhasználókat a második futás után általában be kell vonni a javítási folyamatba, miután a szintaktikai, valamint a kérdés összeállításából származó esetleges hibák javítása már megtörtént.

Jelen cikk írásakor 91 állandó keresőprofil futtatása folyik a következő megoszlásban:

36 profil ipari témában,
23 profil demográfiai témában,
10 profil makroökonómiai témában,
4 profil társadalomstatisztikai témában,
7 profil beruházási és építőipari témában,
4 profil társadalmi szolgáltatások témában,
4 profil belkereskedelmi témában,
2 profil igazgatási témában,
1 profil mezőgazdasági témában.

Ezen kívül gépi feldolgozás adott választ kb. 20–30 egyszeri retrospektív kérdésre, különböző témákban.

A kísérleti eredmények ismeretében megállapítható, hogy a számítógép-segítette szabad szöveges információkeresés egy olyan hatékony eszköz, amelynek eredményei a hagyományos manuális kereséssel már nem valósíthatók meg, azonban mint minden modern műszaki segédeszköz az üzemeltetés során bizonyos új ismeretek elsajátítását igényli mind a feldolgozó, mind a felhasználó szakemberektől.

## IRODALOM

1. ROBSON, A. – LONGMAN, J. S.: Automatic aids to profile constructions. = Journal of the American Society for Information Science. 1976. No. 4. 213–223.p.
2. STANDERA, O.: Computerized information retrieval services. = Nachrichten für Dokumentation. 1972. No. 4. 154–157.p.
3. GORNOSZTAEV, Ju. – MOLIBZSENKO, Sz. – SZOSZIN, E.: Szravnitel'noe opiszanie treh paketov programm (IRMS, DPS i TEXT-PAC) firmü IBM, realizujuscsih IPSZ. = Naucsno-Tehnicseszkaja Informacija. 1974. No.2. 28–36.p.
4. BALOGH Z.: Egy IBM adatbáziskezelő rendszer és több szöveges információtároló és -visszakereső rendszer adaptálási és alkalmazási kísérlete. Programozási rendszerek '75. Szeged, 1975.
5. Az IBM TEXT-PAC nevű szöveges információfeldolgozási programcsomag alkalmazásának leírása. Nyersfordítás.
6. ANTAL L.: A tartalomelemzés alapjai. Bp. 1976. Magvető Kiadó. 150 p.
7. ANTAL L.: Egy új magyar nyelvtan felé. Bp. 1977. Magvető Kiadó. 188 p.