# SZÁMÍTÓGÉPES SZÓTÁRAK

Noha az első számítógépes szótárak a hatvanas években készültek, a számítógépes szótárírás a nyolcvanas években kezd önálló tudományággá válni. Míg az első számítógépes szótárak főként a fényszedés melléktermékeként készültek, a hetvenes évek végén, nyolcvanas évek elején mind több olyan projektum indult meg, amelynek fő célja egy lehetőleg sokoldalúan felhasználható szótári adatbázis, avagy nyelvi korpusz előállítása, és csak mintegy melléktermékeként áll elő ebből a nyomtatott szótár. A számítógépes lexikográfia mint önálló tudományág jelen pillanatban van születőben; jól jelzik ezt a sűrűsödő számítógépes lexikográfiai konferenciák, szimpóziumok és „workshop"-ok, illetőleg a lexikográfiai konferenciák számítógépes szekciói. E tudományág kialakulását elsősorban a számítógépek technikai fejlődése tette lehetővé, és a korábbi számítógépes szótárak kedvező és kedvezőtlen tapasztalatai tették sürgetővé.

Mivel 1984-ben az MTA Elnökségének határozata alapján megindult A magyar irodalmi és köznyelv nagyszótára elnevezésű projektum, amelynek célja a magyar nyelv történeti nagyszótárának elkészítése számítógép felhasználásával (a továbbiakban NSz.), szükségessé vált a téma nemzetközi szakirodalmának gondos áttanulmányozása. A meglepően bőséges angol nyelvű szakirodalom és a magyar publikációk csekély száma arra ösztönzött, hogy magyar nyelvű összefoglalást készítsek a témakörben az utóbbi években megjelent angol nyelvű publikációkról.

## I. Mire jó a számítógépes szótár (és mire nem)?

A számítógépes szótárak legfőbb előnye a nyomtatottakkal szemben az, hogy míg a nyomtatott szótárakban kizárólag a címszavak ábécérendjében tudunk keresni, a számítógépes szótári adatbázisban — az adatbázishoz használt szoftver minőségétől függően — számtalan különböző szempont szerint kereshetünk. Az Oxford English Dictionary számítógépes változatához eddig elkészült program segítségével például választ kaphatunk arra a kérdésre, hogy melyek a magyar eredetű szavak az OED-ben, és mely forrásokban találták meg az első előfordulásaikat.[1] Kiírathatjuk az összes Shakespeare-idézetet, vagy az összes olyan idézet szövegét, amely tartalmaz egy bizonyos állandósult szókapcsolatot, így kikutathatjuk egy idióma használatának történetét stb. A szótár gépesített változatából tehát olyan kérdésekre kaphatunk választ, amelyekre a nyomtatottból nem, vagy csak óriási, lélekölő és pontatlan munkával válaszolhatunk.

Le kell szögeznünk azonban, hogy a számítógépes szótárak általában nem képesek helyettesíteni a nyomtatott változatot (l. Landau 1984. 290; Benbow 1986a. 10; Merkin 1983b. 383). Amikor például egy címszót akarunk megkeresni, és egyaránt rendelkezésünkre áll a nyomtatott szótár meg a számítógépesített, sokkal egyszerűbb levenni a polcról a könyvet, és fellapozni a címszót, mint bekapcsolni a gépet, beindítani a megfelelő szoftvert, beírni a keresett címszót, megvárni a választ, lapozgatni a képernyőn, keresgélve a minket érdeklő részt. Kétségtelen viszont, hogy a kompakt lemezeken való adattárolás olyan ütemben fejlődik, hogy 10-20 éven belül minden valószínűség szerint sokkal olcsóbb lesz megvenni egy nagyszótárat optikai lemezen, mint nyomtatott formában, így feltehetően sokan fogják a számítógépes változatot használni a nyomtatott helyett is. Ettől azonban — különösen Magyarországon — egyelőre nagyon messze vagyunk, ezért is célszerű olyan szótári adatbázist készítenünk, amely nyomtatott formában is kiadható.

A számítógépes szótárak további nagy előnye, hogy viszonylag könnyen javíthatók, aktualizálhatók, pótkötetek készítése helyett a kiegészítések, javítások az egységes anyagba beágyazhatók. Ha egyszer létrehozunk egy szótárat számítógéppel olvas-

ható formában, utána a szótár újabb, módosított kiadásai könnyen elkészíthetők, ráadásul minimális az esélye annak, hogy a javítás során újabb hibákat vigyünk fel az adatbázisba. Kellő körültekintéssel készíthető olyan számítógépes szótári adatbázis is, amely egyszerre több nyomtatott szótár anyagát tartalmazza, például egy szótári család három tagja, a nagyszótár, a kéziszótár és a zsebszótár szócikkei tárolhatók közös adatbázisban, ha speciális jelekkel megkülönböztetjük, hogy melyik címszó szerepel mindháromban, melyik csak a nagyszótárban. A szócikkekben szintén elkülönítjük a közös, és csak az egyik vagy másik változatban publikálandó részeket, jelentéseket stb. Erre az egyik legszebb példa a New Van Dale projektum (erről részletesebben l. Sterkenburg 1981), amelyben összesen 21 szótár alapanyagát tárolják egy közös adatbázisban: először is létrehoznak a meglévő Van Dale nagyszótár alapján egy alapszótárat, majd ebből kiindulva elkészítik a módosított Van Dale egynyelvű nagyszótárat, a holland-francia, francia-holland, holland-német, német-holland, és a holland-angol, angol-holland nagyszótárakat, u g y a n a z t a címszóállományt felhasználva; végül pedig elkészítik az összes felsorolt szótár kéziszótár és zsebszótár változatát. Az ilyen projektumok teremtik meg a számítógépes lexikográfiát mint önálló, alkalmazott tudományágat, mivel effajta bonyolult munkára számítógép nélkül még csak gondolni sem lehetett.

Mindezen túl, bármely számítógépes szótár tekinthető úgy, mint az adott nyelv egy korpusza, amely sok szempontból érdekesebb információkat tartalmaz, mint egy folyó szövegekből összeállított korpusz, hiszen számos grammatikai, esetleg nyelvtörténeti adat is található benne. Különösen izgalmas kutatásokat folytathatunk akkor, ha egy nyelvről folyó szövegekből álló korpuszunk és szótárunk is van számítógépesített formában: összehasonlíthatjuk a kétféle adatbázisból kapott adatokat. Sőt, a szótárból eleve legalább háromféle adatot kaphatunk ugyanarra a nyelvi jelenségre, attól függően, hogy a címszavak között, az értelmezések között vagy az idézetek között keresgélünk, hiszen a címszóállományból például a szókincsre vonatkozó információkat kaphatunk, az idézetekből korok szerint csoportosított szó/szókapcsolat előfordulásokat kerestethetünk, míg az értel-

mezésekből a szótárírók által használt nyelvről nyerhetünk adatokat.

## II. A számítógépes szótárak fajtái és használatuk

A számítógépes szótárakat többféleképpen csoportosíthatjuk. Megkülönböztethetjük a számítógéppel olvasható szótárakat (machine-readable dictionaries) és a számítógépesített szótárakat (computerized dictionaries) (Kazman 1986. 10.). A különbség a kettő között az, hogy a számítógéppel olvasható szótár rendszerint csupán a fényszedés melléktermékeként jön létre, és ezért csak a fényszedésnél használt speciális jeleket tartalmazza, nem tükrözi megfelelően a szótár belső struktúráját, és így nem támogatja a szótárban való több szempontú keresést. A számítógépesített szótár ezzel szemben olyan adatbázis, amely jól tükrözi a szótár belső szerkezetét, speciális adatbáziskezelő szoftverrel van ellátva, amelynek segítségével hatékonyan kereshetünk a szótár bármely részében. A számítógéppel olvasható szótárak egy részét fokozatosan átalakítják számítógépesített formára. Mivel a strukturálatlan, csupán számítógéppel olvasható szótárak meglehetősen érdektelenek, a továbbiakban csak a számítógépesített szótárakkal foglalkozom és az egyszerűség kedvéért ezentúl csak ezeket nevezem számítógépes szótárnak.

A számítógépes szótárak feloszthatók használati körük szerint is. Némelyik a gépi fordítás vagy a természetes nyelvű számítógépes interfész segédeszközéül szolgál, némelyik elsősorban korpuszként használatos különböző nyelvészeti-lexikológiai kutatáshoz, mások pedig elsősorban a nyomtatott szótár számítógéppel támogatott előállítására vannak hivatva. Mivel a NSz. elsődleges célja egy nyomtatott szótár előállítása párhuzamosan a szótári számítógépes adatbázis kialakításával, elsősorban olyan szótári projektumokat fogok ismertetni, amelyeknek céljai hasonlók. Kimerítő ismertetésre azonban nem törekszem, hiszen egyrészt angolul elérhetőek teljességre törekvő beszámolók (Amsler 1984, Kipfer 1984, Keitz 1982, Warwick 1986)[2], másrészt a NSz. szempontjából hasznosabbnak tűnik néhány szótári projektum valamivel részletesebb ismertetése, mint az összes vázlatos felsorolása. Nem foglalkozom gyakorisági szótárakkal

és a természetes nyelvi interfészekhez, fordítórendszerekhez készített szótárakkal. Igyekszem tehát azokra a projektumokra korlátozni az ismertetést, amelyek a NSz. szempontjából relevánsnak tűnnek. Nem célom itt a magyar számítógépes szótárak ismertetése, noha az Értelmező Szótár számítógépre vitele világviszonylatban is úttörő jellegű vállalkozás volt, erről azonban már számos publikáció látott napvilágot (Papp 1969, Kornai 1986).

## 1. Néhány szótári adatbázis, amely nyomtatott szótárból készült

### 1.1. Merriam-Webster

Az első számítógépen is tárolt angol szótár a Merriam-Webster Seventh New Collegiate Dictionary (W7), és a Merriam-Webster New Pocket Dictionary volt (MPD) (Amsler 1984). Miután elkészültek a nyomtatott szótárak, kutatást indítottak a számítógépes változat kidolgozására. Az erre irányuló projektum 1966-68-ig tartott Olney és Ziff vezetésével. Először is lyukkártyára gépelték a két szótár teljes anyagát, ami nem volt könnyű, mivel a lyukkártyák karakterkészlete jóval kisebb, mint a nyomtatott szótáré. A felvitel után programokat írtak a szócikkek egy részének elemzésére, konkordanciát készítettek az értelmezésekre, morfológiai elemző és generáló programokat fejlesztettek és leválogatták a szócikkek egy részét további kutatás céljára. Ez a projektum a számítógépes lexikográfia „hőskorának" terméke. Az adatbázis jelenleg is a kutatók rendelkezésére áll, számtalan — elsősorban nyelvészeti — kutatás alapanyagául használták.

Az egyik ilyen, a W7-et alapanyagául használó kutatás során (Byrd 1984, 1985) a szótárhoz egy olyan adatbáziskezelőt készítettek, amellyel 3 „dimenzióban" lehet kérdéseket feltenni. A Wordsmith elnevezésű szótári adatbáziskezelő használatakor a képernyő közepén állandóan az aktuális szócikket látjuk egy ablakban, az ablak körül csillag alakban kiírathatók a kapcsolódó címszavak. A dimenzió itt az anyag háromféle szempontból való rendezettségét jelenti. Eszerint az első dimenzió a címszavak ábécérendje, a második dimenzió a rímelő szavak listája, a harmadik dimenzió az atergo rendezés. Azaz minden aktuális szócikkhez kiírathatók a szótárban előtte-utána lévő címszavak,

az ezzel rímelő szavak, a megegyező végződésű szavak. Mivel az adatbázisba beleépítették Roget Thesaurusának anyagát is, kérhetjük egy további dimenzióban az egyes szavak szinonimáit is. Később a Longman Dictionary of Contemporary English szócikkállományát szintén hozzáillesztették az adatbázishoz, és folyamatosan újabb és újabb kérdés-dimenziókat adnak a rendszerhez. Ilyen pl. a kiejtés-dimenzió, amely az aktuális címszó összes lehetséges kiejtését listázza ki. Ez a projektum mindent összevéve jó példa arra, hogyan lehet és kell fokozatosan továbbfejleszteni egy szótári adatbázist, részben újabb és újabb szótárak hozzáadásával, részben pedig a szoftver eszközök állandó fejlesztésével.

### 1.2. Az új Van Dale szótár

A Van Dale szótárat (Nieuw Woordeboek der Nederlandsche Taal) először 1872-ben adták ki, ezt újabb és újabb átdolgozott kiadások követték (l. Sterkenburg 1981). A szócikkek a Woordenboek der Nederlandsche Taal (A holland nyelv szótára) alapján készültek, amely az 1500-1920 közötti szókincsállományt öleli fel.

1976-ban a Van Dale korábbi kiadóvállalata (Kluwer) létrehozta a Van Dale Projektum leányvállalatot, abból a célból, hogy elkészítsék a Van Dale-nek egy olyan módosított változatát, amelynek címszóállománya többnyelvű szótárak alapanyagául szolgálhat.

Első lépésként a jelenlegi holland szókincs feltérképezését tűzték ki célul. Mivel a kiadó viszonylag gyors eredményt várt, nem törekedhettek arra, hogy sok nyomtatott szöveget vigyenek számítógépre, és ebből próbálják előállítani az aktuális szókincset. Inkább a Van Dale címszóállományát vitték gépre, csupán néhány népszerű folyóirat és magazin anyagával egészítették ki. Ugyanakkor egy csoport hagyományos módon gyűjtötte az új szavakat/jelentésárnyalatokat, az ezekre talált idézeteket szintén hozzáadták az adatbázishoz.

A Van Dale mintegy 220 000 címszavát körülbelül 90 000-re kellett csökkenteni, részben azért, hogy csak a szinkrón szókincset tartalmazza, részben azért, hogy kevesebb helyet foglaljon el. Ezért felvitték a Van Dale utolsó kiadását teljes terjedelmében mágnesszalagra, majd a lexikográfusok minden címszó minden jelentésárnyalatát ellátták egy „szinkrónia kóddal",

amely azt jelölte, felvegyék-e a címszót/jelentést az új címszóállományba. (A kód 0-5 terjedő érték volt.) A kód felhasználásával a gép válogatta le az új címszóállományt.

A címszóállomány összeállítása után megtervezték az új szótárak szerkezetét. Ehhez segítségképpen felhasználták egy felmérés eredményeit. A felmérést 1979-ben piackutatással együtt végezték, megkérdezték a várható szótárhasználókat, szerintük milyen adatoknak kell a szótárakban lenniük. Az általános vélemény szerint az egynyelvű értelmező szótárban elsősorban a szó helyesírásának és értelmezésének kell szerepelni, de szeretnék, ha a régies szavak és a szinonimák is benne lennének. Az etimológiát, tabu szavakat, antonimákat kevésbé tartják fontosnak.

Elhatározták, hogy a nyelvtant belefoglalják a szótárba, a szócikkekben pedig erre a nyelvtanra fognak hivatkozni. A kétnyelvű szótárakban összehasonlító nyelvtan lesz, példákkal.

Mivel a teljes szótári adatbázist számítógépen fogják tárolni, az általános szócikkszerkezetet és az alkalmazott kódokat úgy alakították ki, hogy az egyértelmű gépi tárolást és keresést biztosítsák. Külön kód jelzi a címszó elején, melyik szótárban legyen benne (nagy-, közép-, zseb-), utána nyelvtani szempontból szétbontva következnek a jelentések, példákkal.

Az így kialakított adatbázis legfőbb előnye az, hogy a három nyelv (francia, német, angol) kétnyelvű szótárai ugyanazt a szókincset fogják tartalmazni, a szerkezetük szintén egyforma lesz, így aki az egyik szótárat megtanulja kezelni, az összes többit is tudja használni. Az egységes grammatikai-szemantikai kódolás biztosítja a számítógépes változatban való keresést.

### 1.3. Az új Oxford English Dictionary

Az Oxford English Dictionary (OED) a legnagyobb egynyelvű történeti angol szótár (Murray, 1884-1928). A 19. század közepén kezdődött meg az anyaggyűjtés, az első füzet 1884-ben jelent meg, az utolsó 1928-ban, eredetileg 125 papírkötésű vékony füzetben. Ezt egy pótkötet kiadása követte 1933-ban, amit a szótár első és utolsó kötetének megjelenése közt eltelt 44 év tett szükségessé (Országh 1966. 341).

Az ötvenes évek végén az Oxford University Press (a továbbiakban OUP) Bob Burchfield irányításával hozzákezdett egy

újabb pótkötet készítéséhez, amely az 1933 óta keletkezett új szavakat, illetve új jelentésárnyalatokat tartalmazta, s amely végül is négykötetes lett: az első 1972-ben, az utolsó 1986-ban jelent meg. Ebben az időszakban vált nyilvánvalóvá, hogy az OED-t legközelebb csak fényszedéssel lehet kiadni, mivel az eredeti nyomólemezek alapján több utánnyomást már nem lehet készíteni. Ezért elhatározták, hogy a teljes OED-t és a négy pótkötetet számítógépre viszik, a pótkötetek anyagát — amennyire lehet, automatizálva — integrálják az OED eredeti szócikkállományával, és egyúttal a nyolcvanas években keletkezett új szavakat is hozzáadják a szótárhoz. Az így elkészítendő szótári adatbázis anyagát teljes egészében kiadják, a továbbiakban pedig a szótári anyag kiegészítését mindig a számítógépes adatbázison végzik el.

A New OED számítógépesítése két fő munkafázisból áll: az elsőben számítógépre viszik a szótár teljes szövegét a pótkötetekkel együtt, összeolvasztva őket a szótárral, és némi javítás után kiadják a teljes szótárat. Az új szótár várhatóan 1989-ben fog megjelenni. A második fázis a szótári adatbázis kialakítása oly módon, hogy az mind a lexikográfusok munkáját segítse, mind pedig az átlagos kutatói igényeket kielégítse.

A program megvalósítására több intézmény fogott össze. A rögzítéssel kapcsolatos legnagyobb probléma az volt, hogyan őrizzék meg a szótár eredeti belső struktúráját, amit elsősorban a tipográfiával fejeztek ki a szerkesztők. Végül is kidolgoztak egy speciális kódrendszert, amely átmenet a tipográfiai kódolás és a strukturális kódolás között. Elsősorban az volt a cél, hogy a rögzítők számára könnyen érthetők és alkalmazhatók legyenek a jelek, de ugyanakkor semmilyen információ ne vesszen el az eredeti szótárból.

Ahhoz azonban, hogy az ily módon rögzített szövegből létre tudjanak hozni egy olyan adatbázist, amelynek struktúrája tükrözi az eredeti szótár struktúráját, és ugyanakkor biztosítja a különböző szempontok szerinti keresést, egy külön elemző programot kellett írniuk. A program külön erre a célra fejlesztett mesterséges nyelv használatával működik. A nyelv szerepe elsősorban az volt, hogy explicitté tegye azt a struktúrát, amelyet a kódolt változat csak implicite tartalmazott (vö.

Kazman 1986). A pótkötetek anyagát csak részben lehet automatikusan összefésülni az eredeti szótár anyagával, ezért egy speciális interaktív programot készítettek ennek a munkafolyamatnak a támogatására. A pótkötetekből származó címszavak egyértelmű információt tartalmaznak arról, milyen műveletet kell velük végezni (helyettesíteni egy régi definíciót egy újjal, vagy csak idézeteket kell hozzáadni a megfelelő jelentéshez, esetleg egy új címszót felvenni a megfelelő helyre).

A pótkötetek szövegének integrálása után a korábbi hivatkozások jelentős része helytelen lesz, ezért ezeket — amennyire lehet automatikusan — módosítani kell az OED teljes szövegében. Az OED-ben speciális, Murray által kifejlesztett fonetikai jeleket alkalmaztak, ezeket most helyettesítik az International Phonetic Alphabet jeleivel.

Az új kiadás előtt a címszavak egy részét is átdolgozzák, ill. kijavítják, felvesznek egy-két új jelentést is, mindez azonban csak néhány címszót fog érinteni. Az új címszókhoz hagyományosan, kézzel gyűjtik az idézeteket, majd a kiválasztott anyagot az e célra fejlesztett szócikk-szerkesztést támogató programrendszer segítségével viszik az adatbázisba (l. Simpson 1985a). A szótári adatbázist jelenleg egy standard IBM adatbáziskezelő nyelv, az SQL segítségével érik el, de a szótári projektum második fázisában kifejlesztendő adatbázis nem az SQL-t fogja használni, hanem külön szótárra orientált adatbáziskezelő rendszert írnak.

Az adatbázisból fényszedéssel fogják előállítani a nyomtatott szótárat. A kiadás várható időpontja 1989 tavasza (Benbow 1986b). A nyomtatott formában való kiadáson túl az optikai lemezen való terjesztést is tervezik. Már 1987 végén ki akarják adni az eredeti OED szövegét a pótkötetek nélkül optikai lemezen, főként abból a célból, hogy tapasztalatokat szerezzenek a számítógépes változat használatával kapcsolatban. Ezen tapasztalatok felhasználásával akarják kifejleszteni azt a szoftvert, amelyet majd a teljes szótárhoz fognak árusítani.

Az ún. második fázis — a szótári adatbázis kiadáson túli, kutatási célokra való továbbfejlesztése — a valóságban párhuzamosan kezdődött az elsővel, és valószínűleg több egymást követő fázisból fog állni. A szótári adatbázisokkal kapcsolatos

kutatói és szoftver-fejlesztői tevékenységet a waterlooi egyetem (Kanada) számítógépes lexikográfiai kutatócsoportja végzi. A kutatóközpont — szoros együttműködésben az OUP-vel — részt vett az elemző program kidolgozásában, és hozzákezdett a szöveges adatbázisok számára leghatékonyabb adatbáziskezelő rendszer kialakításához. Ezért először is felmérést készítettek a számítógépes szótár várható felhasználói körében arról, milyen kérdésekre akarnak majd választ kapni (Benbow 1986a).

Kérdőív segítségével összehasonlították a nyomtatott és a számítógépes változat legjellemzőbb gyakorlati hasznosítását, melyből kiderült, hogy a felhasználók egészen másfajta keresésekre akarják használni a számítógépes változatot, mint a nyomtatott szótárat. Míg a nyomtatott szótárban elsősorban az egyes címszavakhoz tartozó értelmezéseket akarják kikeresni, a számítógépes változatban az idézeteket, ill. a hozzájuk tartozó címszavakat akarják a leggyakrabban kiíratni. Részben ebből a felmérésből, részben a gyakorlati tapasztalatokból azt a következtetést vonták le, hogy a szótári adatbáziskezelő szoftver fejlesztésekor nem arra kell törekedniük, hogy a nyomtatott szótár helyett használják a számítógépes változatot, hanem arra, hogy olyan kérdésekre tudjanak választ adni, amelyekre számítógép nélkül nem, vagy csak igen nehezen lehetne válaszolni.

Az adatbevitellel párhuzamosan megkezdődött a hatékony szövegkezelő programok fejlesztése. A PAT (Gonnet 1986) olyan szövegkereső program, amely hatalmas szövegállományokban képes szavakat, szópárokat, karaktersorozatokat keresni rendkívüli gyorsasággal. Pillanatok alatt megjeleníthetjük a képernyőn bármely szó összes, vagy valamilyen szempont szerint válogatott előfordulását a szótár teljes szövegében. Egy új programnyelvet is fejlesztettek, abból a célból, hogy a New OED adatbázisában a szótári struktúrát használva kereshessünk adatokat. Ennek segítségével megadhatjuk, hogy egy szót vagy betűsorozatot a címszavak közül, vagy csak az idézetek közül írjon ki a program. Így például táblázatot készíthetünk vele az összes, képzőt tartalmazó címszóról képzők szerint csoportosítva, feltüntetve a címszó első előfordulását; így kimutatható, melyik képző mennyire volt produktív a különböző évszázadokban.

## 1.4. A Longman szótár

A Longman Dictionary of Contemporary English (LDOCE) (Procter 1978) az első olyan számítógép segítségével készült nyomtatott szótár, amelynél a számítógépet nem pusztán a fényszedéshez használták, hanem a szócikkek kidolgozásához is. Az LDOCE legfőbb újítása, hogy az értelmezéseket egy 2000 szónyi alapszókincsállományra vezeti vissza. A definíciókban csak olyan szavak szerepelhetnek, amelyek vagy ennek az alapszókincsnek elemei, vagy az alapszavak segítségével definiálva vannak. Ez többek közt az angolt idegen nyelvként tanulók számára lehet nagyon hasznos: ha először megtanulják az első 2000 szó jelentését, utána eligazodnak a szótárban. Lexikográfiai szempontból is logikus törekvésnek tűnik a körkörös definíciók kiszűrése. Noha egyes lexikográfusok szerint az így készült értelmezések nem feltétlenül egyszerűbbek és használhatóbbak, mint az e megszorítás nélkül készült egynyelvű szótárakéi (Herbst 1986), a számítógépes szótárkészítésben mégis óriási jelentősége van ennek a műnek.

A korlátozott szókincsállomány használata lehetővé tette a körkörös definíciók kiküszöbölését. Így a címszavak definíciói jól formalizálhatók, a számítógépes változatot kiegészítették egy szemantikai kódrendszerrel (MICHIELS 1981). A szemantikai kódok egy része hierarchikus (a főneveknek pl. ilyen kódjai lehetnek: +Animal, +Female stb.), más része a címszó használatára vonatkozó információ (az igénél például a lehetséges alany és tárgy sajátosságait jellemzik). Grammatikai kódrendszerrel is ellátták a címszavakat, ill. a különböző jelentésárnyalatokat (pl. a "claim" egyik jelentésében C5 kódot kap, annak jelzésére, hogy megszámlálható, és "that"-tel kezdődő mellékmondatot vonz, a másik jelentésében C3-at kap, megszámlálható és "to"-val kezdődő infinitívuszt vonz stb.). Az egyes jelentésárnyalatokat a számítógépes verzióban tárgyszavakkal is ellátták (pl. a "hammer" szónak a tőzsdei csőd jelentésárnyalatához a "Gazdaság, "Tőzsde" tárgyszavakat rendelték).

A Longman-Liege projektum az LDOCE és a Longman Dictionary of English Idioms (LONG 1979) (LDOEI) számítógépes változatát felhasználva alakít ki egy olyan szótári adatbázist, amelyet gépi fordítórendszerek szintaktikai-szemantikai elemző-generá-

ló moduljaihoz akarnak majd felhasználni. Tekintettel arra, hogy az LDOCE korlátozott definiáló szókincset használ, és a fent röviden jellemzett kódrendszerrel van ellátva, ideális alapanyagnak tűnik a gépi fordítás számára. A kutatócsoport 81-ben részben a kódrendszer továbbfejlesztésén dolgozott, részben pedig automatikus szintaktikai-szemantikai elemzőt írt az angolra. Az elemző fő újdonsága, hogy elsősorban a lexikonban lévő információk irányítják az elemzést. A grammatikai kódok alapján az elemző először is megjósolja, milyen szintaktikai szerkezet következik; ha azonosítani tudja a keresett szerkezetet, sikeres az elemzés. Az elemző segítségével megvizsgálják a címszavak értelmezéseit, és feljegyzik, hogy a grammatikai kódok és a ténylegesen előfordult szövegkörnyezetek összhangban vannak-e. Ennek segítségével elkészítenek egy annotált szótári adatbázist, amiben szerepel az LDOCE-beli homográfkód, az értelmezés száma, a választott grammatikai kód, és a fenti elemzés eredménye. Egy interaktív programmal feljegyzik az anaforákat és az operátorok hatókörét is.

Az adatbázis felépítésére és lekérdezésére az IBM STAIRS adatbáziskezelőt használják.

## 2. Új szótárak készítése számítógép segítségével

Kevés az olyan projektum, amelyben teljes mértékben számítógépes gyűjtésre támaszkodva próbálnának új szótárat készíteni. Úgy tűnik, a NSz. tervezésekor nem sok idevágó nemzetközi tapasztalatot hasznosíthatunk.

### 2.1. A Trésor de la langue française

A francia Trésor talán a legambiciózusabb számítógépes történeti szótári projektum. A szótári adatgyűjtés 1960-ban kezdődött meg, Paul Imbs vezetésével. Folyamatos szövegekből számítógépes archívumot készítettek. 1800 szépirodalmi szövegből vették a mintát, a feldolgozandó források listáját az egyes korszakok szakértői állították össze. A szövegek mintegy 20%-a nem szépirodalom. A szövegeket 1789-1960 között megjelent művekből válogatták. A szótárszerkesztéshez felhasználtak mintegy 6 millió cédulából álló, 1936-1969 között gyűjtött anyagot is, lerögzítették továbbá valamennyi eddig megjelent egynyelvű

francia szótár anyagát is, ezek között számos szakszótárét is. Az összesen mintegy 150 millió szövegszó terjedelmű korpusz három logikai részből áll: a szépirodalmi anyagból, a szótárakból és a felhasznált források bibliográfiai adatbázisából.

Ezt a forrásanyagot elsősorban egy publikálandó történeti szótár alapanyagának szánták, ugyanakkor azonban ezt a hatalmas nyelvi korpuszt a kutatók számára is hozzáférhetővé tették hatékony szövegkereső programok segítségével. Az archívumból a fantasztikusabbnál fantasztikusabb kérdésekre tudnak választ adni a számítógép segítségével, például ilyenekre: „Melyek a halállal kapcsolatos kifejezések Zolánál?", „Milyen szavak írják le a *kellemetlenséget* a 18. század utolsó 10 évében?", „Hogyan fejlődött az alkoholizmus szó, milyen jelentésváltozáson ment keresztül?" (Martin 1984).

Ahhoz, hogy ilyen, s ezeknél jóval bonyolultabb kérdésekre választ adhassanak gépi úton, a puszta szövegrögzítésen túl számos kiegészítő információt is fel kellett vinniük. A bibliográfiai adatokat nyilvántartó file-ban például a kiadás adatain túl rögzítették a műfajt, azt, hogy a mű valós vagy képzelt világban játszódik-e, a benne szereplők nevét, azt, hogy milyen olvasóközönségnek szánták, milyen célból írták, és milyen a mű hangvétele. A szövegrészlet által felölelt tudományterületeket is bejelölik. A felvitt szótárak anyagát is többféleképpen használják fel az archívumban; a szinonimaszótár és egy tezaurusz használata teszi lehetővé, hogy szemantikailag összefüggő szavakat részben automatizálva kilistázhassanak. Az archívum melléktermékeként gyakorisági szótárat is készítettek az első fázisban felvitt 90 millió szövegszónyi adatállományból, ezt 1971-ben publikálták.

Azt remélték, hogy a számítógépes archívumból néhány év alatt a teljes szótárat ki tudják adni, ezzel szemben 1971 és 1986 között 12 kötet jelent meg; ez körülbelül fele-kétharmada lehet a teljes szótárnak. A viszonylagos lassúságnak az az oka, hogy — bár a számítógépes archívum sokat lendít a lexikográfusok munkáján, megóva őket a cédulák rendezésének és különböző szempontok szerinti válogatásának lélekölő munkájától — a szócikkeket továbbra is az embereknek kell megtervezniük, nekik kell kiválogatniuk a sokszor sok ezer konkordancia-sornyi,

egy-egy címszóra vonatkozó adatból azt a néhányat, amely meggyőződésük szerint legpregnánsabban tükrözi az egyik vagy a másik jelentést, el kell dönteniük, hány jelentésárnyalatot különböztessenek meg minden egyes szócikknél stb. Azaz az érdemi lexikográfiai munkát ugyanúgy, és minden valószínűség szerint hasonlóan elmélyülten, körültekintően kell elvégezniük, mint cédulákon dolgozó elődeiknek. Ráadásul még ilyen elképesztő mennyiségű forrásanyag esetén sem biztos, hogy minden szó minden jelentésárnyalatára akad példa a teljes számítógépes archívumban, minden bizonnyal ezért kellett hagyományosan, cédulákon gyűjtött anyaggal is kiegészíteniük a forrásanyagot. A teljes szótár talán a század végére készülhet el ebben a tempóban.

Az elkészült kötetek szócikkeinek szerkezeti felépítésén jól látszik, hogy számítógép segítségével készítették, feltűnően áttekinthető a szócikk struktúrája. Míg az OED-ben a jelentésekre bontás bizonyos esetekben rendkívül bonyolult, itt jól elkülönülnek az összefüggő és távolabb eső aljelentések. Történeti szótárban újdonság, hogy a címszók után az előfordulás gyakorisága is szerepel, ez a melléktermékként előállított gyakorisági szótárnak köszönhető. Az értelmezések rendkívül rövidek, csak a legfontosabb információkat tartalmazzák, az idézetek hivatottak a jelentés pontosabb megvilágítására. Az idézeteket jelentésárnyalatonként mindig szigorúan a forrás időrendjében közlik, egy-egy jelentéshez 5-10 idézetet válogattak ki az archívumból. A szó stilisztikai értékét, használati körét is megjelölik. A szótár normatív. Külön tárgyalják a címszó részletesebb történetét, és külön a szorosan vett etimológiát. A szócikkek végén felsorolják azoknak a munkáknak a bibliográfiai adatait, amelyek az adott címszóval kapcsolatban készültek.

### 2.2. Dictionary of Old English

A Dictionary of Old English (DOE) adatgyűjtését 1970-ben kezdték meg a torontói egyetemen. A szótár első főszerkesztője Angus Cameron volt, számítástechnikai tanácsadója Richard Venezky, akinek már számottevő számítógépes lexikográfiai tapasztalatai voltak, mivel részt vett az amerikai tájszótár gépesítésében is.

A DOE alapanyaga az összes 750 és 1200 között keletkezett és fennmaradt angol kézirat. Első lépésben fénymásolat formájá-

ban összegyűjtötték az összes forrásanyagot, az eredeti és publikált verziókat egyaránt. (Az anyaggyűjtés 5 évet vett igénybe.) A teljes korpusz mintegy 3 millió szövegszó terjedelmű, az egészet számítógépre vitték (Amos 1984).

A számítógéprevitelt, megfelelő hardver hiányában, elképesztően bonyolultan oldották meg. Először egy gépírónő legépelte az összes szöveget olyan speciális írógépen, amelynek karaktereit a rendelkezésükre álló optikai olvasó le tudta olvasni; az ellenőrzés és javítás után olvasta be az optikai olvasó, természetesen újabb hibákat téve a szövegbe. Mivel akkoriban a munkacsoportnak nem volt saját számítógépe, a mágnesszalagon tárolt szöveget kinyomtatták, ellenőrizték, visszaküldték javításra Madisonba, újra kinyomtatták stb.

Az adatgyűjtéssel párhuzamosan készültek a feldolgozó programok (Venezky 1971). A LEXICO elnevezésű programrendszer a konkordanciakészítést támogatta. Ehhez először is lemmatizálni kellett a teljes szöveget; a program lehetővé tette az interaktív, nem automatikus lemmatizálást. A lemmatizálással párhuzamosan az idézeteket besorolták az egyes címszavak alá, és a géppel cédulákat készítettek. (Erre azért vol szükségük, mert hardver hiányában a lexikográfusok nem tudtak közvetlenül a számítógépen dolgozni.)

Amint elkészültek a forrásanyag felvitelével, kiadták a teljes konkordanciát mikrofichen. Egy-egy szó konkordanciája két soros, lexémára rendezett. Külön táblázat tartalmazza a címszavak listáját és indexét, továbbá a gyakoriságra vonatkozó adatokat. A konkordancia és a gép által készített cédulák felhasználásával készítik a szócikkeket. Eddig az A és C betű anyagát tudták kiadni, 1987 tavaszán kezdték meg a B betű szerkesztését.

Az utóbbi években saját számítógépet kaptak, egy VAX 11/730 minigépen dolgoznak, két 121 Mbyte-os lemezen tárolják a szótár anyagát, és XEROX Dandelion munkaállomásokat csatlakoztattak hozzá, amelynek nagy grafikus képernyője, saját memóriája (1,5 Mbyte), és 43 Mbyte-os fix lemeze van. Jelenleg fejlesztik azokat a programokat, amelyek az on-line (azaz közvetlenül a gépen történő) szócikkírást támogatják; a program kikeresi a konkordanciaállományból a kijelölt azonosítójú idézeteket, és bemásolja a szócikkek megfelelő részébe, segít a nyomtatási formátum kialakításában. Mivel a képernyő grafikus, a teljes karak-

terkészlet megjeleníthető rajta, úgy látjuk a szócikket, ahogy a nyomtatott szótárban meg fog jelenni. A szócikkeket aztán laser printeren nyomtatják ki, a szerkesztők ellenőrzik és javítják.

Csak most kísérleteznek azzal, hogy a lexikográfusok számítógépen szerkesszék a szócikkeket (jelenleg papírra írnak mindent, a gépírónő rögzíti a szócikkeket), olyan programot készítenek, amellyel a címszóhoz tartozó összes konkordancia kikereshető és a képernyőre íratható, vagy válogatások, minták készíthetők az idézetekből. Becslésük szerint a szócikkírás várhatóan újabb 15 évet fog igénybe venni. Tapasztalataik szerint még a számítógép segítségével is nagyon nehéz azoknak a szócikkeknek a kidolgozása, amelyekre túl sok vagy túl kevés adat van.

### 2.3. New York Times Everyday Dictionary

A New York Times Everyday Dictionaryt 1982-ben adták ki. A szótár címszóanyagát a New York Times fényszedőszalagjairól készített konkordanciákból állították össze (Paikeday 1983). A konkordanciákat a mikroszámítógépek hőskorában készítették, egy icipici számítógépen. (TRS-80 Model I. A gépnek 48K memóriája volt, óra frekvenciája 2MHz!)

A Lexicon elnevezésű program először is összefűzte a szövegállományokat egy logikai állományba. Az összefűzött szövegeken különféle kereséseket lehet végrehajtani, a FIND parancs hatására a program egy szó teljes képernyőnyi szövegkörnyezetét, a PHRASE hatására szavak, karakterek együttes előfordulását írja ki. (Lényegileg csupán egy átlagos szövegszerkesztő képességeivel rendelkezik.) Ezenkívül lehetőség van a szövegben talált szövegszavak mennyiségének kiíratására, továbbá az egyes szavak egysoros konkordanciáinak előállítására.

Hősi erőfeszítésnek tűnik ilyen kis számítógépen szótári programokat fejleszteni, a szerző azonban, úgy tűnik, sosem alkalmazta 200 000 szövegszónál nagyobb anyagra a programot, ami egy szótár készítésénél meglehetősen kis adatmennyiségnek tűnik. Mindenesetre úgy látszik, hasznos volt a számítógép a címszóállomány kiválasztásában és a jelentések definiálásában, de azért ne higgyünk a szerzőnek, aki szerint a Merriam-Webster vagy az OED sok milliós cédulaállománya mikroszámítógépre tehető és azon hatékonyan kezelhető. (Ezen a gépen a 200 000 szónyi anyag-

ban egy szó kikeresése átlagosan 8 perc volt, ugyanez a New OED SUN gépén néhány század másodperc.)

2.4. Német lexikográfiai adatbázis

A mannheimi Institut für deutsche Sprache számítógépes nyelvészeti osztálya elhatározta, hogy létrehoznak egy olyan német szótári adatbázist, amely folyamatos szövegmintákból készül, de az ezekből kapott adatokat összevetik a korábbi szótárak szócikkállományával (Teubert 1984).

Első lépésként összegyűjtötték az összes számítógépes adathordozón lévő német nyelvű korpuszt, amelyek természetesen mind különböző konvenció szerint voltak lerögzítve, ezeket programmal egyformára alakították. Mintegy 7 millió szövegszónyi korpuszt sikerült így összeállítaniuk. Optikai olvasó segítségével további szöveges korpusszal fogják kiegészíteni.

Ahhoz, hogy a meglévő szótárakkal összehasonlíthassák a szövegekből kapott címszóállományt, egy szótári adatbankot is létre kell hozniuk, amit elsősorban a BONNLEX „Kumuliertes Lexikon" szótári adatbázis felhasználásával remélnek megoldani.

A BONNLEX (l. Brustkern 1981, 1982) szótári adatbázist a bonni egyetem Kommunikationsforschung und Phonetik intézetében hozták létre abból a célból, hogy az összes német számítógépes szótárat egy helyen, azonos formátumban tárolják. Tizenegy, más-más célra készült számítógépes szótárat fésültek össze egy adatbázisba, ezért a bennük található információ tartalma és struktúrája sokféle. Miután felmérték, melyek azok az információk, amelyek minden szótárban szükségesek, megtervezték az új adatbázis szerkezetét és az előállításához szükséges szoftvereket. Az így létrejövő szótár elsődleges funkciója a természetes nyelvű interfészek és fordítórendszerek kiszolgálása lesz, de azt remélik, hogy az itt tárolt adatokat a lexikográfusok is fel tudják majd használni. A központi szótár lehetővé teszi, hogy ne kelljen minden projektumnak elölről kezdenie a szótárkészítést, elég, ha kiegészítik a meglévő alapszótárat azokkal a specifikus információkkal, amelyekre az adott kutatásnak szüksége lehet. Az egyes szócikkek általános szerkezete formálisan:

A szótár

$W = LE_1, LE_2 \ldots LE_m$

szótári címszavak halmaza, (ahol $LE_i$ szótári címszó). Minden szótári címszó n információt tartalmaz:

$$LE_i = (I_{i,1}, I_{i,2}, I_{i,3}, \ldots I_{i,n}); \; I_{i,j} \; K_j$$

Minden $I_{i,j}$ információ egy lexikológiai információ osztályba tartozik, ezeket az osztályokat így definiálták:

$K_1$: a címszó grafikus reprezentációja, az írásváltozatokkal együtt
$K_2$: fonetika, fonológia, szótagolás, hangsúly
$K_3$: szófaj
$K_4$: morfológia: információ a ragozásról, képzésről stb., a szótő megadása
$K_5$: szintaktikai információk: vonzatok, lehetséges szintaktikai szerkezetek, mélyszerkezetek stb.
$K_6$: szemantikai információk: szemantikai primitívek, szemantikai mezők, releváns szövegkörnyezetek, jelentés-definíció
$K_7$: pragmatikai megjegyzések: stílusérték, használat stb.

Az így előállított szótári és szöveges adatbázisokat felhasználva, speciális szótárszerkesztői munkaállomások segítségével írják az új szócikkeket a lexikográfusok. Minden szócikket külön állományban helyeznek el, ezeket módosítják, javítják stb. Külön szoftver segíti az egységes stílusú szócikkírást.

## 3. A történeti szótárak és a számítógép

Amikor az első számítógéppel támogatott szótári projektumok megindultak a hatvanas évek végén, hetvenes évek elején (a francia Trésor, a DOE), a főszerkesztők rendkívül optimisták voltak, úgy gondolták, a számítógépes gyűjtés töredékére csökkentheti a szótárkészítésre fordítandó időt. A DOE-t 15 év alatt remélték befejezni — ezzel szemben 15 év után kezdődött meg a tényleges szócikkírás, és most úgy gondolják, jó ha újabb 15 év elég lesz a szótár elkészítéséhez. A Trésor főszerkesztője 1971-ben úgy nyilatkozott, hogy várhatóan 6-7 éven belül a teljes szótárat kiadják — jelen pillanatban, 16 évvel később kb. a szótár felét-kétharmadát sikerült kiadniuk.

Tekintélyes lexikográfusok is azt állították, hogy a számítógép forradalmasítani fogja a történeti szótárkészítést. Zgusta szerint: „It is also quite possible that large academic dictionaries will not be published anymore. The point is that even the academic dictionaries which consist of ten, twenty, or any number of volumes, do not and cannot present the whole material contained in the archieve... then why publish a twenty-volume reduction of the material if a one, two or four-volume reduction could suffice for the first information, which must be eventually be followed by the archive search, in any case?[3] (Zgusta 1971. 354-355). Aitken is ugyanerre a következtetésre jutott: „the existence of computer archives would often seem to remove the need to burden library shelves with still larger dictionaries filled with still more detailed information of interest to only a few people"[4] (Aitken 1971. 16).

A gyakorlat azonban nem igazolja ezt a derűlátást. Egyrészt bármilyen hasznos is egy nagy számítógépen tárolt szövegarchívum, nem helyettesíti a lexikográfusok által kiválogatott, jelentésárnyalatok szerint rendezett szócikkeket. A lexikográfusok több száz idézetből válogatják ki azt a néhányat, ami véleményük szerint legjobban reprezentálja az adott jelentésárnyalatot, s az átlagos szótárfelhasználó számára sokkal hasznosabb az így készített összefoglalás, mint az archívumból kiíratható, esetleg több ezer sorból is álló konkordancia. Amsler is azt írja: „Simply obtaining a multimegabyte set of words by grabbing all the machine-readable sources available and merging them together hardly provides the basis for scientific observation of the nature of the language as a whole. Thus ten million words of newspaper stories can be less useful than one million words of carefully sampled text taken from variety of sources.[5] (Amsler 1982. 661) Ha pedig a több ezer konkordanciából véletlenszerűen íratunk ki néhányat, semmi biztosíték sem lesz arra, hogy az összes jelentésre kapunk példát, a jelentésárnyalatok kialakulása pedig végképp nem követhető a puszta konkordanciákból. Legcélszerűbbnek az látszik, hogy párhuzamosan hozzáférhetővé tegyük a szótárkészítésre használt archívumot és a szótári adatbázist egyaránt.

Ami a számítógéppel segített adatgyűjtést illeti, mint a Trésor és a DOE példáján láthatjuk, nem számíthatunk arra, hogy

a számítógép alkalmazása lényegesen gyorsabbá teszi a szótárírást. Ha lerögzítettük a korpuszt, a gép nagyon gyorsan tud szólistát készíteni, ábécébe rendezni, konkordanciát listázni. A baj éppen az, hogy „túl jól" csinálja (Oakman 1980. 16.). Ha nagy a korpuszunk, bizonyos szavakról kétségbeejtő mennyiségű idézetünk lesz, ezeket egyenként megvizsgálni, eldönteni, melyiket érdemes felhasználni, időnként csaknem lehetetlen. Ez a művelet gyorsítható ugyan azzal, ha a gyakori szavakról csak bizonyos számú véletlenszerűen kiválasztott mintát kérünk, ilyenkor azonban nem biztos, hogy a kapott mintában még mindig benne lesz a szó összes jelentésárnyalata. Lehetséges ugyan, hogy lassan kialakul egy új, kifejezetten számítógépes szótáríró módszer, ami talán valóban gyorsítani fogja a szócikkírást, a valószínűbb azonban az, hogy változatlanul kb. 50-60 évig fog készülni egy-egy történeti szótár, feltéve, hogy a történeti szótártól ugyanazt várjuk el ezentúl is, amit Grimm vagy Murray nyújtott az első igazi történeti szótárakban. Ahhoz, hogy minden szó összes jelentésárnyalatának legkorábbi előfordulását megadjuk, a nyelv teljes írott korpuszát számítógépre kellene vinni, ami lehetetlen, vagy kombinálni kell a régi gyűjtési módszert és a számítógépes gyűjtést.

Mivel teljes korpuszt csak a holt nyelvek esetében vihetünk számítógépre, az élő nyelvek szótárainak készítésekor törekedhetünk a teljesség helyett arra is, hogy egy jó mintánk legyen az adott nyelv szókincséről. Megtehetjük, hogy egy valamilyen elv szerint kiválogatott zárt korpuszt gépre viszünk, ennek szinkrón vagy diakrón leírását elkészítjük és publikáljuk, leszögezve, hogy az így készült szótár milyen zárt korpusz leírására vállalkozik. Az így készülő diakrón szótár ugyan nem azonos a hagyományos történeti szótárral, de többé-kevésbé pótolhatja azt.

A történetiszótár-készítés ugyan, mint láttuk, nem gyorsul fel a remélt mértékben a számítógép alkalmazásától, mégis célszerűbbnek látszik a jelenleg induló projektumok esetén a gépi gyűjtés használata. Egyrészt számos rendkívül unalmas műveletet takarítunk meg (ábécébe rendezés, kor szerinti rendezés stb.), másrészt a számítógépen tárolt adatok sokkal többféleképpen elérhetők, könnyen módosíthatók, kiegészíthetők. Számolnunk kell azonban azzal, hogy a számítógépes szótár sokkal

alaposabb tervezést igényel, mint a hagyományosan készülő, különösen akkor, ha olyan szótári adatbázist akarunk készíteni, ami nem csak egy célra használható fel; márpedig számítógépen éppen sokcélú adatbázist célszerű létrehozni. A tervezéskor tehát gondolnunk kell a későbbi revíziókra, rövidített kiadásokra és a szótár egyes részeinek leválaszthatóságára is, ami a tervezés időszakában még nagyon távolinak tűnik. A DOE és a Trésor meglepő lassúsága valószínűleg annak is köszönhető, hogy még csak most ismerkednek a számítógép felhasználási lehetőségeivel a szótárírásban, számos olyan problémát kell megoldani, amellyel nagy lexikográfus elődeinknek nem kellett szembenézni. (Pl. a számítógépes szótárban megsérülhet a lemez, elveszhet rengeteg adat; külön tudomány, hogy lehet visszaállítani az elveszett adatokat. Murray céduláinak egy részét viszont az egerekették meg.) Röviden szólva, a számítógépes lexikográfus dolga nem könnyebb és nem nehezebb, mint elődeié volt: más.

Lényegesen meggyorsíthatja azonban a számítógép használata az új, egynyelvű, hétköznapi használatra készülő értelmező szótárak írását. Erre a New York Times Everyday Dictionary nagyon jó példa, hiszen ha kész fényszedő szalagokat használhatunk fel a konkordanciakészítéshez, valóban töredékére csökkenhet az adatgyűjtésre fordítandó idő.

## III. A számítógépes szótárak előállítása

### 1. Adatgyűjtés és rögzítés

A szótárkészítés első munkafázisa a gyűjtendő anyag kijelölése. A számítógépes tervezés és gyűjtés csak a forrásanyag kijelölése után, annak ismeretében kezdődhet meg. A számítógépre vitelnek számos módja van. A régebbi szótárak egy részét lyukkártyára, ill. lyukszalagra rögzítették, ma már ezt a technikát (szerencsére) nemigen alkalmazzák. Felhasználhatók korpuszkészítésre a nyomdai fényszedő szalagok is, így gyűjtötte alapanyagát a New York Times Everyday Dictionary. A legfejlettebb adatbeviteli mód az intelligens optikai karakterolvasó használata, ez azonban egyelőre még a fejlett országokban sem terjedt el. Az OED szövegét ugyan megpróbálták a Kurzweil géppel beolvastatni, amely számos betűtípus felismerésére megta-

nítható, ez a gép azonban csak akkor használható hatékonyan, ha folyamatosan egy betűtípust olvasunk be vele. Így például hatékonyan használható írói szótár készítésekor, amikor teljes köteteket folyamatosan akarunk felvinni.

Nem nevezhető hatékonynak az optikai olvasásnak az a módja, amelyet a DOE munkatársai voltak kénytelenek alkalmazni. Ha az optikai olvasó csak egyfajta betűtípust ismer fel, és ezért külön le kell gépelni a szöveget azzal a betűtípussal, jobb nem használni, inkább egyből valamilyen számítógépes adathordozóra kell gépelni az anyagot.

A legelterjedtebb rögzítési mód a terminálokon vagy személyi számítógépeken való adatrögzítés. Legfőbb előnye, hogy könnyen és gyorsan javítható, jobbnál jobb szövegszerkesztő programok vannak, amelyek megkönnyítik az adatbevitelt. A rögzített adatokat mágneses adathordozón tárolják, innen bármikor lemásolható, kinyomtatható, lekérdezhető.

A legtöbb szótári projektumnak problémát okoz a számítógépek szűkös karakterkészlete. Rendszerint valamilyen kódkombinációval vagy helyettesítéssel oldják meg az összes szükséges karakter ábrázolását. Az Old English Dictionarynél elegendő volt a 256 karakter, így a behelyettesítést választották. Külön szoftverjük gondoskodik arról, hogy a szócikkíráskor visszahelyettesítsék a megfelelő karaktereket. Mivel grafikus képernyőjük van, az összes „old English" karaktert meg tudják jeleníteni a képernyőn és a laser printeren egyaránt. A New OED-ben speciális tagoló kódokkal jelölték a különböző betűtípusokat.

## 2. Konkordanciák

A számítógépes szótáríráshoz általában csak az egyszerű konkordanciákat használják. A konkordanciák rendszerint minden szó 1-2 sornyi szövegkörnyezetét és az előfordulás pontos helyét tartalmazzák. A konkordancia-programok túlnyomó része lehetőséget ad arra, hogy a leggyakoribb szavakat — rendszerint a formaszavakat — kizárjuk a konkordanciából, mivel ezek általában az anyag 30-40 %-át teszik ki. Néhány közismert, készen megvásárolható konkordancia-program:

Oxford Concordance Program (OCP): Az OCP parancsnyelve egyszerű angol szavakból áll, könnyen elsajátítható. A programot

ANSI FORTRANban írták, így számos géptípuson futtatható (IBM, CDC, Digital, ICL, Univac, Burroughs, Honeywell, Prime). Meghatározhatjuk azokat a szavakat, amelyeket nem akarunk a konkordanciába kiíratni, vagy ha csak néhány szóról akarunk listát kapni, ezeket is megadhatjuk. A kulcsszavakat rendeztethetjük jobbról vagy balról ábécébe, gyakoriság vagy hosszúság szerint, csökkenő vagy növekvő sorrendbe. A listák főbb típusai:

- szóalak lista, gyakorisággal
- index: a szavak előfordulási helyei
- konkordancia: szólista szövegkörnyezettel, előfordulási hellyel és gyakorisággal
- statisztikák

A konkordancia-sorok hosszúsága szükség szerint változtatható.

A pisai konkordancia-program (Zampolli 1983) szintén lehetővé teszi, hogy felsoroljuk azokat a szavakat, amelyeknek a konkordanciájára nem vagyunk kíváncsiak. A nagyon gyakori szavak esetén a program automatikusan csak az előfordulás számát írja ki, bizonyos esetekben pedig az előfordulás száma mellett mintát ír ki a teljes konkordancia helyett.

Változtatható a kiírandó szövegkörnyezet mérete, a keresett szó mindig a konkordancia közepén van. A felhasználó megadhatja a szöveghatárok jeleit (versszak vége, bekezdés vége stb.), továbbá a szóhatárt jelölő írásjeleket. Grammatikai címkékkel ellátott szöveg esetén a keresett szó grammatikai tulajdonságai meghatározhatják a kontextus méretét.

A konkordanciát többféleképpen rendeztethetjük: a címszóra, és ezen belül a morfológiai kódokra, a szó után vagy előtt lévő szövegre, a szöveg eredeti sorrendjére, vagy időrendben.

A Lexicon nevű ún. konkordancia program (Paikeday 1983) alig több, mint egy átlagos szövegszerkesztő. Mindazonáltal ezzel is kiírathatjuk kisebb szövegfile-ok konkordanciáit, de ez csak egy-egy szóalakot tud keresni, szókapcsolatok konkordanciáját nem tudja előállítani. A konkordancia-sor mérete nem változtatható rugalmasan.

A Lexico, amelyet Venezky készített a DOE-hez, nem pusztán konkordancia-program, hanem egy összetett, szótárírást segítő programrendszer. Fő funkciói: a szöveg tárolása, szerkesztése, konkordancia készítése és lemmatizálás. A szerkesztés során a

programmal kikerestethetjük az összes olyan idézetet, amelyben az éppen szerkesztés alatt álló címszó előfordult. Az idézeteket azonosító szerint is kereshetjük, és besorolhatjuk őket a megfelelő címszavak alá, azaz lemmatizálhatjuk. A lemmatizált idézetállomány konkordanciáját ezután újra elkészíttethetjük a programmal; ekkor már címszavakra rendezett konkordanciát kapunk.

A Wordcruncher a konkordancia-készítésen túl, számos szövegkeresési műveletre alkalmazható (Hughes 1987). A program nem szótári projektum keretében készült, IBM XT-n és ezzel kompatibilis gépeken futtatható, felhasználó-orientált szövegkezelő rendszer.

Mivel mikroszámítógépre készült, elsősorban kisebb szövegállományokban való hatékony keresésre alkalmazható, lehetőség van azonban arra, hogy a kisebb szövegeket összefűzzük, és a keresést az összefűzött állományon végeztessük. Egy-egy kis szövegállományban 13-15 000 szó lehet; 50 kis szöveget fűzhetünk össze; kb. 300 000 szó lehet maximálisan az összefűzött állományban. Egy szó hossza legfeljebb 31 karakter lehet. A program futtatásához legalább 512K memória, 2 floppy disk vagy 1 hard disk szükséges, DOS 2.1 vagy 3.2-es operációs rendszer. A programot az Electronic Text Corporation terjeszti.

Három fő programból áll a rendszer: az IndexETC hozza létre az indexállományt a kötetlen formátumú szövegekből, a ViewETC-vel kereshetünk az indexelt állományban, a BYU concordance-szal készíthetjük el a konkordanciákat. Az index-készítés előtt megadhatjuk a nem indexelendő szavak listáját, a szóvéget jelölő írásjeleket, a kívánt rendezési szempontot, és a szövegekben használt elhatároló jeleket (bekezdés, versszak stb.). A ViewETC-vel kiválaszthatunk egy szövegrészt, kiírathatjuk ábécérendben a szóalakokat gyakoriságukkal együtt. Kereshetünk szavakat vagy szókapcsolatokat, logikai műveleteket használva. Változtathatjuk a kiírandó szövegkörnyezet méretét, a kiírás formátumát. Kérhetjük azt is, hogy mindig csak egy előfordulást, vagy hogy csak bizonyos kóddal rendelkező szavakat írjon ki.

A nem angol nyelvű szövegekben való keresést is támogatja; német, francia és spanyol szöveghez vannak kész interfészek; grafikus képernyő használata esetén görög, orosz és héber karaktereket is tud kezelni.

Felhasználó interfésze rendkívül jó, könnyen elsajátítható. A funkcióbillentyűk használata és a Help-menük nagyon megkönnyítik a rendszerrel való ismerkedést.

## 3. Lemmatizálás

A Lexico, mint láttuk, nem vállalkozik automatikus lemmatizálásra. Vannak azonban olyan szótári projektumok is, amelyek megkíséreltek automatikus morfológiai elemzőt készíteni. A pisai Istituto di linguistica computazionale-ban többféle morfológiai elemző eljárást is kidolgoztak (Zampolli 1983).

A legegyszerűbb automatikus morfológiai elemző tulajdonképp egy morfológiai szintetizáló programot használ fel a lemmatizáláshoz. Először egy szótárat felhasználva létrehozzák az összes lehetséges toldalékolt alakot, mindegyiket ellátják a megfelelő morfológiai kódokkal. Az elemzéskor a szövegben talált szóalakokat ezekhez illesztik; ha megtalálják a szövegszót a szótárban, kész az „elemzés".

Egy másik elemző program, amelyet a spanyol nyelvre fejlesztettek (Catarzi 1982), már valóban szegmentálja a szövegszavakat. Mielőtt hozzákezdenének az elemzéshez, egy előfeldolgozó program kiszűri a szövegből a leggyakoribb szavakat és állandósult szókapcsolatokat, a Juillard Gyakorisági szótár első 750 szavának felhasználásával. Így a szöveg 50%-át kiszűrik az első lépésben. A tulajdonképpeni elemző egy tőtárat és egy toldaléktárat felhasználva működik; minden tő- és toldalékmorféma morfológiai kódokkal van ellátva. Amennyiben nem sikerül elemezni a szóalakot, a program jelzi a lexikográfusnak a hibát. A homográf alakokat a szövegkörnyezet vizsgálatával próbálja azonosítani a program. Ezzel a módszerrel mintegy 80%-ban helyes elemzéshez jutnak.

A Siemens cég hasonló elven működő morfológiai elemzőt fejlesztett a német nyelvre (Thurmair 1982; Meya 1982; Gehrke 1986). Mivel ezt az elemzőt elsősorban nem szótárkészítésre, hanem természetes nyelvű interfészhez szeretnék felhasználni, rendkívül részletesen elemzik a szavakat. (Az összes képzőt is levágják.) Az elemző lényegileg ugyanúgy működik, ahogy a magyar elemző fog (Prószéky 1985). Arra törekedtek, hogy az elem-

zőt a tőtár és a toldaléktár felépítése vezérelje és az elemző algoritmus minél inkább gép- és felhasználás-független legyen. A tőtárban lévő szótövek mellett morfológiai kódokat találhatunk, a toldaléktárban található morfémák szintén kódolva vannak. A program először illeszti a szövegszót a tövekhez, a leghosszabb tőhöz tartozó toldalékot elemzi, azután, ha a maradékot megtalálja a toldaléktárban, ellenőrzi, hogy helyes-e az elemzés; ha nem, próbálkozik más lehetséges felbontásokkal.

## 4. Szócikkírás

A szócikkírásban a számítógép csak kifejezetten e célra fejlesztett szerkesztő-programokkal tud valamelyes segítséget nyújtani. A Lexico talán a legjobb ilyen program jelenleg. A szócikkszerkesztő program ablakokat használ: a képernyőn az egyik ablakban láthatjuk a billentyűzetet, rajta a speciális karakterek képével, a fő ablakban megjelenik a kitöltetlen szócikk. A megfelelő helyre bemásolhatjuk azokat az adatokat, amelyek a gépen is tárolva vannak (gyakoriság, legkorábbi előfordulás), ezt kiegészíthetjük új adatokkal (értelmezés). A lexikográfusok által bejelölt azonosítójú idézeteket lehívhatjuk a konkordanciaállományból, a program beírja a kívánt helyre. Végül a képernyőn lévő billentyűzetet felhasználva kicserélhetjük azokat a speciális karaktereket a szócikkfile-ban, amelyeket a konkordanciában valamilyen más karakterrel helyettesítettek. A képernyőn úgy látjuk a szócikket, ahogy majd a szótárban fog megjelenni. (Dőlt betűk, vastag betűk stb.) Most fejlesztenek ehhez olyan programot, amely szócikkírás közben a képernyőre írja a konkordanciát, és a lexikográfus rögtön bejelölheti, ill. átmásolhatja a jónak tartott idézeteket.

*

A számítógépek bizonyos fokig valóban forradalmasíthatják a szótárkészítést, ha nem is oly mértékben, ahogy Zgusta vagy Aitken remélte. Az igazi szótárakat továbbra is csak lexikográfusok írhatják, a gép csak a mechanikusan ismétlődő munkafolyamatok jelentős részét tudja átvállalni tőlük. A számítógép

azonban csak akkor lesz igazán hasznos segítőtársa a szótáríróknak, ha a szerkesztők már az anyaggyűjtés megkezdésekor számításba veszik a gép nyújtotta lehetőségeket, és már a forrásanyag kijelölésekor pontosan tudják, mit akarnak a géppel csináltatni, és mi az, amit továbbra is csak emberek tudnak elvégezni.

**Pajzs Júlia**

**Jegyzetek**

[1]Így megtudtam, hogy nemcsak a közismert „coach" és „hussar" szerepel benne, de a „ban" szó, 'bán, a szlavóniai tartomány vezetője' jelentésben szintén, továbbá számos magyar eredetű ásványnév.

[2]Amsler tanulmánya a számítógépes szótárakról jó kiindulópont a téma tanulmányozásához, elsősorban azért, mert kimerítő bibliográfia egészíti ki. A tanulmány egyébként rengeteg adatot tartalmaz meglehetősen eklektikus elrendezésben, sok, a témához nem szorosan kapcsolódó részt és kevés eredeti gondolatot. Viszont a témában nem járatos érdeklődőnek képet ad a számítógépes lexikográfia jelenlegi állásáról.

Kipfer tanulmánya lényegesen értékesebb, jól szerkesztett, ennélfogva több hasznos információt tartalmaz, bibliográfiája kevésbé részletes. Nem törekszik arra, hogy az összes számítógépes szótárat ismertesse, de nagyon használható felbontásban ismerteti egyes kutatások részeredményeit.

Keitz csupán felsorolja a European Science Foundation által támogatott összes számítógépes szótári projektumot, országonként csoportosítva. Ez az összeállítás azért nagyon hasznos, mert a teljesség igényével készült, megadja az egyes témák legfontosabb adatait, a munkálatot végző intézmény nevét és címét.

Warwick arra törekszik, hogy egy általános bevezető után ismertesse és értékelje az összes európai, nem angol projektumot. Hiánypótló munka.

[3]Lehetséges, hogy a nagy akadémiai szótárakat nem is fogják ezentúl kiadni. Mivel még a 10-20 vagy akárhány kötetes szótárak sem tudják a teljes archívumot magukba foglalni, érdemes-e egy 20 kötetes sűrítményt publikálni, amikor egy 1-2 vagy 4 kötetes kivonat tartalmazhatná a legfontosabb információkat, amelyet aztán az archívumban való keresés követhetne időnként?

[4]A számítógépes archívumok, úgy tűnik, szükségtelenné teszik, hogy a könyvtárak polcait egyre részletesebb szótárakkal rakjuk meg, amelyek csak néhány ember számára érdekesek.

[5]Ha csupán létrehozunk sok megabyte-nyi szóállományt úgy, hogy vesszük az összes számítógéppel olvasható forrásanyagot és összefésüljük őket egy állományba, aligha nyújtunk lehetőséget

a nyelv természetének tudományos megfigyelésére. Így 10 millió szónyi újságszöveg kevésbé lehet hasznos, mint 1 millió szónyi gondosan kiválogatott szövegminta, sokféle forrásanyagból.

## Irodalom


AITKEN, A.J. (1971), Historical dictionaries and the computer. In: R.A. WISBEY (szerk.), The computer in Literary and Linguistic Research, Cambridge.

ALLÉN, S. (1983), Sprakdata Lexibase System. An Integrated View of a Lexical Project. In: ZAMPOLLI 1983a. 51-63.

AMSLER, R.A. (1984), Machine-Readable Dictionaries. In: M.E. William (szerk.), Annual Review of Information Science. Vol. 19. 161-209.

AMOS, A. (1984), Computers and Lexicography: The Dictionary of Old English. Status Report on the Dictionary of Old English Project. Kézirat.

BENBOW, T. — WEINER, E. (1986a), Machine-readable Dictionaries for the General Public. Workshop on Automating the Lexicon. Second. Grosseto, Italy: European Community, University of Pisa and Instituto di Linguistica Computazionale del CNR. May 19-23, 1986. 10.

BENBOW, T. (1986b), Status Report on the New OED Project. Second Annual Conference of the UW Centre for the New OED. University of Waterloo, 9-11 Nov 1986. Kézirat.

BERG, D.L. (1986), NED to OED to New OED: a Bibliographic Essay. (Preliminary version) Centre for the New OED. Kézirat.

BRATLEY, P. (1983), Computers and Lexicography: Advances and Tends. In: ZAMPOLLI 1983a. 83-95.

BRUSTKERN, J. — HESS, K.D. (1982), Machine-Readable German Dictionaries: from a Comparative Study to an Integration. In: CIGNONI 1982. 77-181.

BRUSTKERN, J. — HESS, K.D. (1983), The Bonnlex Lexicon System. GOETSCHALCKX 1982. 33-40.

BUTLER, S. (1982), Problems with Headwords in Old English. In: ZGUSTA 1980. 105-114.

BYRD, R.J. (1984), DAM - A Dictionary Access Method. Computer Science Dept. I.B.M. Thomas J. Watson Research Centre, Yorktown Heights, New York. Kézirat.

BYRD, R.J. (1985), WordSmith. User Guide. Computer Science Dept. I.B.M. Thomas J. Watson Research Center, Yorktown Heights, New York. Kézirat.

CALZOARI, N — CECOTTI, M.L. — ROVENTINI, A. (1984), Computational Tools for an Analysis of Terminological Data in a General Dictionary. In: HARTMANN 1984. 328-332.

CASSIDY, F. (1980), Computer Maping of Lexical Variants for DARE. In: ZGUSTA 1980. 147-162.

CATARZI, M.N. — CAPELLI, G. — RATTI, D. — SABA, A. (1982), A morphosyntactic analyzer for Spanish. In: CIGNONI 1982. 115-122.

CIGNONI, L. — PETERS, C. (1982), Computers in Literary and Linguistic Research. Proceedings of the VII International Symposium of the Association for Literary and Linguistic Computing. Pisa 1982. Linguistica Computazionale vol. III. Supplement.

DOMENIG, M. — SHANN, P. (1986), Towards a Dedicated Database Management System for Dictionaries. Proceedings of COLING '86. Bonn. 91-96.
ENGEL, G. — MADSEN, B. N. (1983), From Dictionary to Data-Base. In: HARTMANN 1983. 339-344.
GEHRKE, M. — BLOCK, H.U. (1986), Morpheme-based Lexical Analysis. In: JOHANNESEN 1986. 1-15.
GOETSCHALCKX, J. — ROLLING, L. (szerk.) (1982), Lexicography in the Electronic Age. Proceedings of a Symposium held in Luxembourg, 7-9 July, 1981. North-Holland.
GONNET, G.H. (1987a), PAT — An efficient text searching system. UW Centre for the New OED. Kézirat.
GONNET, G.H. — TOMPA, F. WM. (1987b), Mind Your Grammar: a New Approach to Modelling Text. UW Centre for the New OED. Kézirat.
GUCKLER, G. (1983), A Computer Based Monolingual Dictionary: a Case study. In: R.R.K. HARTMANN (szerk.), Lexicography: Principles and Practice. Academic Press. 198-201.
HARTMANN, R.R.K. (szerk.) (1983), LEXeter '83 Proceedings. Papers from the International Conference on Lexicography at Exeter, 9-12 September 1983. Lexicographica Series Maior 1. Max Niemeyer Verlag, Tübingen.
HERBST, T. (1986), Defining With a Controlled Defining Vocabulary in Foreign Learners' Dictionaries. A. KUCERA, A. REY, H.E. WIEGAND, L. ZGUSTA (szerk.), Lexicographica. International Annual for Lexicography. 2/1986. Niemeyer, Tübingen. 101-119.
HOFLAND, K. — HAUGE, J. H. (1983), A Lemmatized Ibsen Concordance and Potential Consequences for Dictionary Production. In: ZAMPOLLI 1983a. 145-151.
HUGHES, J.J. (1987), Wordcruncher: High Powered Text-Retrieval Program. Bits & Bytes Rewiev. 1/3. Whitefish, Montana.
JOHANESSEN, G. (szerk.) (1985), Information in Data. First Annual Conference of the UW Centre for the New OED. November 6-7. 1985.
JOHANESSEN, G. (szerk.) (1986), Advances in Lexicology. Second Annual Conference of the UW Centre for the New OED. November 9-11. 1986.
KAY, M. (1983), The Dictionary of the Future and the Future of the Dictionary. In: ZAMPOLLI 1983a. 161-174.
KAZMAN, R. (1986), Structuring the Text of the Oxford English Dictionary Through Finite State Transduction. Master of Mathematics in Computer Science Thesis. University of Waterloo.
KEITZ, W. V. (1982), Projekte zur maschinellen Lexicographie. Sprach and Patenverb. 6/1-2:11-22.
KIPFER, B.A. (1984), The Dictionary of the Future: Computer Applications. Workbook on Lexicography. A Course for Dictionary Users with a Glossary of English Lexicographical Terms. Exeter Linguistic Studies 8. University of Exeter. 161-172.
KNOWLES, F.E. (1983), Dictionaries and Computers. In: HARTMANN 1983. 301-314.
KORNAI A. (1986), Szótári adatbázis az akadémiai nagyszámítógépen. Műhelymunkák 2:65-79.
KUCERA, H. és FRANCIS, W.N. (1967), Computational Analysis of Present-Day American English. Brown University Press.

LANDAU, S. I. (1984), Computer Use and the Future of Dictionary Making. Dictionaries. The Art and Craft of Lexicography. Charles Scribners's sons. New York. 272-294.
LONG, T.H. (1979), Longman Dictionary of English Idioms. Longman.
MAKKAI, A. (1980), Theoretical and Practical Aspects of an Associative Lexicon for the 20th Century English. In: ZGUSTA 1980. 125-145.
MARINONE, N. (1982), A project for a Latin Lexical Database. In: CIGNONI 1982. 175-178.
MARTIN, E. (1984), Une banque de données sur la langue francaise. BRISES. Bulletin de recherches sur l'information en sciences économiques, humaines et sociales. La Linguistique dans les systemes documentaires 4.
MATHIAS, J. (1983), Computer-Aided Processing of Chinese Lexicographic Materials. In: HARTMANN 1983. 371-376.
MERKIN, R. (1983a), The historical/academic dictionary. In: R.R.K. HARTMANN (szerk.), Lexicography: Principles and Practice. Academic Press, London - New York - Paris. 123-133.
MERKIN, R. (1983b), Historical dictionaries and the Computer - Another View. In: HARTMANN 1983. 377-384.
MEYA, M. (1982), Treatment of Suffixes in Automatic Morphological Analysis. In: CIGNONI 1982. 199-207.
MICHIELS, A. — MULLENDERS, J. — NOEL, J (1981), The Longman - Liege Project. In: GOETSCHALCKX 1982. 201-210.
MURRAY, J.H. — BRADLEY, H. — RAIGIE, W.A. — ONIONS, C.T. (1882-1928), The Oxford English Dictionary. 1 - 12 kötet. Clarendon Press, Oxford.
NAGAD, M. — TSUJI, J. — UEDA Y. — TAKIYAMA M. (1981), An Attempt to Computerize Dictionary Data Bases. In: GOETSCHALCKX 1982. 51-73.
OAKMAN, R. (1980), Computer Methods for Literary Research. University of South Carolina Press.
ORSZÁGH L. (1966), A mai angol szótárirodalom. In: Országh L. (szerk.), Szótártani tanulmányok. Tankönyvkiadó, Budapest.
PAIKEDAY, T.M. (1983), The Joy of Lex. Creative Computing. 240-245.
PAPP F. (1969), A magyar nyelv szóvégmutató szótára. Akadémiai Kiadó, Budapest.
PETERSEN, P.R. (1983), New Words in Danish 1955-75. A Dictionary Compiled and Worked out in a Traditional Way and Managed and Typed via Computer. In: ZAMPOLLI 1983a. 179-186.
PROCTER, P. (1978), Longman Dictionary of Contemporary English. Longman, London.
PRÓSZÉKY G. (1985), Magyar szövegek számítógépes morfológiai elemzése. MTA Nyelvtudományi Intézet. Kézirat.
SIMPSON, J. (1985), Opening Address: The New OED Project. In: JOHANNESEN 1985. 1-6.
STERKENBURG, P.V. — MARTIN, W. — AL, B. (1981), A New Van Dale Project: Bilingual Dictionaries on one and the same Monolingual Basis. In: GOETSCHALCKX 1982. 221-238.
SVENSEN B. (1983), A Computerized Concordance Based on a Bilingual Dictionary: a Case Study. In: R.R.K. HARTMANN (szerk.), Lexicography: Principles and Practice. Academic Press 202-205.

TEUBERT (1983), Setting up a Lexicographical Data-Base for German. In: HARTMANN 1983. 426-429.
THALLER, M. (1982), Recycling the Drudgery. On the Integration of Software Supporting Secondary Analysis of Machine-Readable Texts in a DBMS. In: CIGNONI 1982. 253-268.
THURMAIR, G. (1982), An Integrated Algorithm for Morphological Decomposition. In: CIGNONI 1982. 269-277.
TOMPA, F. WM. (1985), Database Design for the Dictionary of the Future. UW Centre for the New OED. Kézirat.
VENEZKY, R. — RELLES, N. — PRICE, L. (1976), Man-Machine Integration in a Lexical Processing System. University of Wisconsin Computer Center, Madison.
VOLLNHALS, O. (1983), Utilization of a Commercial Linguistic Data-Base System for Electronic Storage and Automated Production of Dictionaries. In: HARTMANN 1983. 430-434.
WARWICK, S. (1986), A Survey of Lexical Resources in Europe. Workshop on Automating the Lexicon. Second. Grosseto, Italy European Committee, University of Pisa. May 19-23, 1986. Kézirat.
WEBSTER, J.J. (1983), The 'PROLEX' Project. In: HARTMANN 1983. 435-440.
WEINER, E. (1984), The New Oxford English Dictionary. EURALEX Bulletin 1(2): 31-33.
WEINER, E. (1985a), Computerizing the Oxford English Dictionary. Scholarly Publishing. 240-253.
WEINER, E. (1985b), The New OED: Problems in the Computerization of the Dictionary. University Computing 7:66-71.
ZAMPOLLI, A. — CAPELLI, A. (szerk.) (1983a), Linguistica Computazionale III. The Possibilities and Limits of the Computer in Producing and Publishing Dictionaries. Proceedings of the European Science Foundation Workshop. Pisa 1981.
ZAMPOLLI, A. (1983b), Lexicological and Lexicographical Activities at the Instituto di Linguistica Computazionale. In: ZAMPOLLI 1983a. 237-278.
ZGUSTA, L. (1971), Manual of Lexicography. Mouton, The Hague.
ZGUSTA, L. (szerk.) (1980), Theory and Method in Lexicography. Hornbeam Press, South Carolina.
ZIMMERMANN, H.H. (1983), Multifunctional Dictionaries. In: ZAMPOLLI 1983a. 279-288.

## MACHINE READABLE DICTIONARIES

by Júlia Pajzs

Machine readable dictionaries has been made since the sixtieths. We selected some of the undergoing projects to show the state of the art in computational lexicography, with special emphasis on compiling historical dictionaries. A critical overview of the softwares used for computerization of dictionaries is also presented.