2. SZÁMÍTÁSTECHNIKA ÉS NYELVTUDOMÁNY

2.1. Bevezetés

Az első fejezetben a korpusznyelvészet és a korpusz meghatározása után a korpusz általános jellemzőiről szóltunk. Az eddigiekből világosan kitűnik, hogy a korpusznyelvészet számítógépek nélkül nem létezhet, így ezt a fejezetet a technikai fejlődés rövid ismertetésével kezdjük. A korpuszok és a korpusznyelvészet fejlődése a technikai háttér minimális ismerete nélkül nem igazán értékelhető. Ezek után a szellemi, azaz nyelvészeti hátteret mutatjuk be, majd kapcsolódó tudományágakról és a számítógépes nyelvészet hazai fejlődéséről esik röviden szó.

2.2. A számítástechnika fejlődése

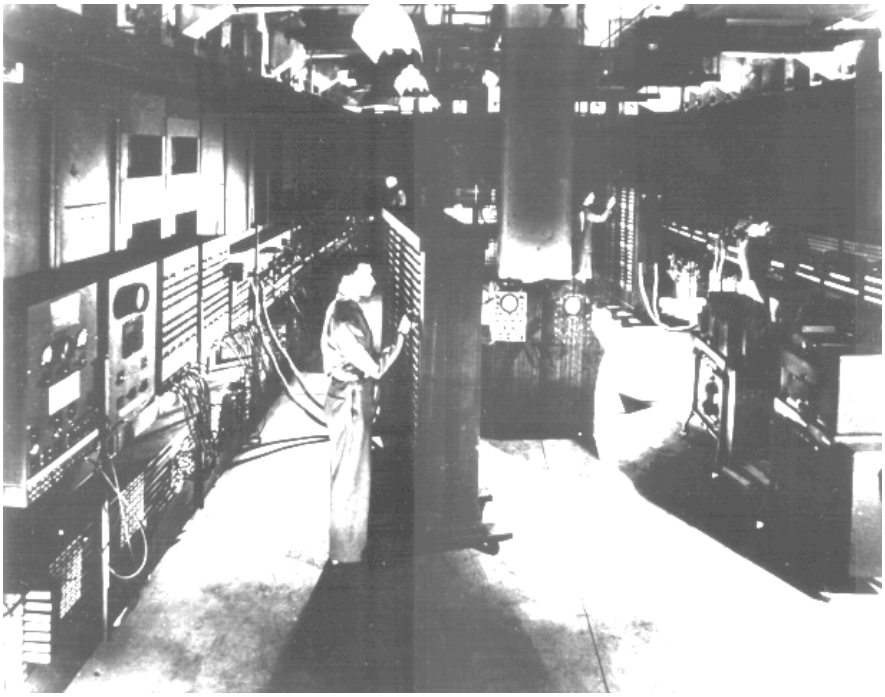
Jóllehet már az 1960-as évek előtt is készült nyelvészeti elemzés céljára néhány korpusz, ezeket azonban nem elektronikus formában tárolták, így az adatok elemzése és feldolgozása nem történhetett számítógépes programok segítségével, hanem csak kézzel és papíron végezték. Nem is lehet ezen csodálkozni, hiszen a számítógépek története sem nyúlik időben sokkal hátrább, mindössze az 1930-as, 40-es évekig. Mivel a korai korpusznyelvészet fejlődését meghatározta és behatárolta a számítógépek hardverének és szoftverének limitált teljesítőképessége, szükségesnek tűnik, hogy – ha nem is a teljesség igényével – egy pár szót szóljunk a számítógépek fejlődéséről. A korpusznyelvészet fejlődését és az úttörő munka nehézségeit is csak így lehet értékelni és megérteni.

Még számítástechnikai szakemberek számára is nehéz feladat eldönteni, hogy igazából mit is illethetünk „az első számítógép” névvel. A technika és tudomány történetében többször előfordult, hogy nagyjából azonos időpontban, de egymástól függetlenül két feltaláló kísérleteit is siker koronázta. A számítógép létrehozásához szükséges elméleti és gyakorlati ötletek közül számos Pascalig is visszavezethető. Ha a szabadalmak alapján szeretnénk ezt a vitát eldönteni, akkor sem járnánk sok sikerrel, mert vagy két évtizeden keresztül folytak perek e téren. Elégedjünk itt meg a köztudatba bekerült és elfogadott eseményekkel. A kíváncsiabbak figyelmébe ajánlom McCartney könyvét az ENIAC-ról (1999) és Goldstine (1993) könyvét a számítógép történetéről.

A legtöbb korai próbálkozással az volt a baj, hogy a gép emberi közreműködés nélkül csak előre meghatározott műveleteket tudott végrehajtani, sokszor egyszerre csak egyfélét, és a kapott eredményeket nem tudta újabb műveletre felhasználni, hanem azokat újra be kellett táplálni a gépbe. Az első elektronikus számítógép, az ABC (Atanasoff-Berry Computer), 1942-ben jelent meg, amit az ENIAC követett 1944–45-

ben. Az ENIAC sokkal többet „tudott” bármely más gépnél, hiszen másodpercenként 5000 összeadást vagy 300 szorzást volt képes elvégezni, míg más gépek csak egy szorzási műveletre voltak képesek ugyanennyi idő alatt. Manapság ez a szám valószínűleg több százmillió, attól függően, hogy milyen sebességű a központi processzor egység, azaz CPU (Jamsa, 1997). E mellett az ENIAC képes volt 20 darab 10 karakterből álló szó „gyors” memóriában és 450 szó ROM-ban való tárolására. Ennek ellenére egyik problémáról egy másikra való átállás így is órákat vagy napokat vett igénybe. Ezek után, a von Neumann, azaz Neumann János és H. H. Goldstine által az Institute for Advanced Study (IAS)-nál (Princeton, New Jersey) vezetett projekt eredményeképpen 1952-ben olyan gépet készítettek, amely a modern elektronikus számítógép prototípusa lett.

A kereskedelmi forgalomban vásárolható számítógépeket sokszor emlegetik „generációkra” utalva az alapján, hogy milyen technológiát használtak gyártásukhoz. Az első generációs számítógépek (1952–1958), elektroncsöves technológiával (vacuum-tube) készültek, elsősorban nagy, természettudományi tevékenységet folytató vevők, mint például a kormány által támogatott laboratóriumok számára. A legelső sorozatgyártásban készült gépek a UNIVAC mellett az IBM 701 és 702 voltak.



17. ábra: Az ENIAC 2

A második generációt tranzisztoros technológia jellemezte, 1959 és 1963 között került forgalomba. A harmadik generáció (1964-től 1975-ig) időszakában jelentek meg az integrált áramkörök. Az első számítógépes „család” is csak 1964-ben jelent meg (IBM

System/360), mely azt jelentette, hogy ugyanazt a programot a család minden egyes tagján futtatni lehetett. 1972-ben találták fel a csipeket, amelyek több integrált áramkör kombinációjából állnak.

A negyedik generációs számítógépekre (az 1970-es évek közepétől az 1980-as évek közepéig) a nagyon nagy méretű áramkörök, multiprocesszorok és hálózatok voltak jellemzők. A memóriacsipek olcsóbbak lettek és megjelentek a nagy felbontású képernyők, valamint a kép orientált programok. Az új áramkörök lehetővé tették a miniszámítógépek, a nagy teljesítményű személyi számítógépek és a munkaállomások kifejlesztését is. Ezzel a személyi számítógépek a nagyközönség számára is elérhetővé és megfizethetővé váltak.

Az Altair 8800, az első széles körben használt számítógép 1974-ben került piacra, majd ezt követte a Commodore PET 1977-ben és a Radio Shack által gyártott Tandy. Az Apple Computer céget 1976-ban alapították és 1980-ra az éves eladásuk már 100 millió dollár felett volt. Az IBM cég Intel 8086-os processzorral rendelkező személyi számítógépe 1981-ben jelent meg, és nem sokkal ezután a cég engedélyezte, hogy más cégek is gyártsák ennek olcsó „klónjait", valamint arra ösztönözték a szoftverfejlesztő mérnököket, hogy írjanak programokat az IBM személyi számítógépére.

Az ötödik generáció (1982–1994) figyelmét a nem numerikus adatfeldolgozásra, alkalmazott mesterséges intelligenciára, és az intelligens interfészekre irányította. Jóllehet a vizsgált problémákat nem sikerült ezen időszakban megoldani, számos technikai előrelépés történt a természetes nyelvek feldolgozása, a képfeldolgozás és a hálózatok terén. Az 1990-es évek elejétől kezdve a technikai fejlődés jelentősen felgyorsult. Szinte félévente egyre gyorsabb és gyorsabb processzorok kerülnek forgalomba. Már a notebook és a kézben tartható (handheld) számítógépek is gyorsabbak és nagyobb adattárolási kapacitással rendelkeznek, mint a korai korpuszelemzések idején használt hálózati (mainframe) számítógépek. Az Encyclopedia Americana Version 5.0 („Encyclopedia Americana", 1999) által közölt adatok (6. táblázat 2. és 3. oszlopa) az első kereskedelmi forgalomban beszerezhető UNIVAC I technikai paramétereit hasonlítja össze az első személyi számítógép paramétereivel. A technikai adatok mellett az ár is fel van tüntetve, mert ebből következtetni lehet arra, hogy kik lehetnek a potenciális vásárlók, és mennyire elterjedt vagy ritka lehet az adott technika. A laptopokra vonatkozó adatok tájékoztató jelleggel szerepelnek, hiszen a rengeteg különböző gyártmány közül nehéz lenne egyet kiválasztani. Az említett adatok közül manapság már nem használják az utasítás/másodpercet, az árak is igen változatosak, és a gép fogyasztása nagymértékben függ a használt eszközöktől (nyomtató, CD-író stb).

	UNIVAC I	IBM PC	Laptop
Megjelenés éve	1951	1981	2004
Sebesség (utasítás/másodperc)	20 000	400 000	Több 100 millió
Ár (dollárban)	250 000	5000	1000–2000
Energiafogyasztás (watt)	50 000	500	75–150

6. táblázat: A kereskedelmi forgalomban levő elektronikus számítógépek első 30 éve: technikai összehasonlítások

Ez a gyors fejlődés az utóbbi 5-10 évben még jobban érezhető. Hadd említsek itt egy személyes példát. Az első általam használt asztali számítógép, 256K RAM memóriával rendelkezett 1992-ben. Az első notebook számítógépem 1994-ben 120 Mb-os merevlemezzel rendelkezett. Manapság a kulcstartóra fűzhető flash memóriák, vagy pendriveok is sokkal több adatot tárolnak. Az 1996-ban vásárolt Toshiba számítógépem, amelyre oly büszke voltam akkor, ma már nemcsak hogy eladhatatlan, de még ajándékba sem nagyon akarná senki sem elfogadni. A jelenlegi – már háromévesen öregecske – notebook számítógépem most 256Mb RAM-mel, 1,19 GHz-es processzorral és 40Gb merevlemezzel rendelkezik. Változnak az idők!

A korpusznyelvészet szempontjából azonban nem csak az adattárolás és az adatfeldolgozás sebességének fejlődése volt meghatározó, hanem az adatbevitelé is. A korai korpuszok készítésekor jelentős akadályt jelentett az adatbevitel nehézsége. John Sinclair korai, Edinburghban, Manchesterben és Birminghamben végzett munkája során (J. Sinclair *et al.*, 1969) a beszélt nyelvi adatokat átírták, majd kézzel lyukkártyára vitték az adatokat és egy kártyaolvasó segítségével olvasták be a számítógépbe. Ez volt az egyik oka annak, hogy ez a korpusz mindössze 300 000 szóból állt (J. M. Sinclair, 2001). Az 1970-es években még mindig használtak lyukkártyákat a programok bevitelére, de az adatokat már kicsit gyorsabban, mágneses szalagról vagy papír lyukszalagról olvasták be. A birminghami Cobuild Korpusz esetében még 1980–83 között is gépelni kellett az adatokat, így 1984-ben is mindössze 7,3 millió szóból állt.

1984-ben a Birminghami Egyetem egy Kurzweill Data Entry Machine-t (KDEM), azaz optikai szkennert szerzett be, amellyel az adatbevitel sokkal gyorsabbá vált, és 1985-re a korpusz már kb. 18–20 millió szóra növekedett. Az 1980-as évek vége felé és az 1990-es évek elején egyre több anyagot kaptak mágneses szalagon, így például a Times és a BBC anyagát. Napjainkban a kézi adatbevitel, illetve a szkennerek használata háttérbe szorult, hiszen az adathordozók óriási kapacitása és az internetet gyorsító technológiák lehetővé teszik, hogy akár egyszerű fájl-átviteli eljárással (FTP) bővítsük az adatbázisokat. Az internetről tehát még az egyszerű érdeklődő is rengeteg adathoz juthat egy korpusz létrehozásához.

2.3. A korpuszok fejlődése

Az első kereskedelemben kapható számítógép 1951-ben jelent meg, és az első számítógépes korpusz munkálatait 10 évvel később kezdték meg. Szintén ekkorra készült el a világ leggyorsabb számítógépe, a Stretch, mely magas ára miatt nem válhatott sikeressé. A második generáció ideje ez, amikor az amerikai légitársaság, az American Airlines, az IBM-mel együttműködve létrehozza a Sabre nevű utazási helyfoglaló rendszert 1962-ben. Az Assembly programozási nyelv helyett magasabb szintű programnyelvek jelennek meg, mint például a FORTRAN vagy a COBOL (Nadar, 1998). Ebben az időben, és ilyen technikai feltételek között született meg az első elektronikus korpusz, a Brown University Standard Corpus of Present-Day American English, azaz a Brown Egyetem Mai Amerikai Angol Nyelvének Standard Korpusza, melyet mindenki csak Brown Korpuszként emleget.

2.3.1. A szellemi háttér

A technikai háttér lehetőséget ad a gondolatok megvalósítására, de mint oly sok más területen, a nyelvészetben is vannak kedvező és ellenséges szellemi erők és divatok, melyek erősen befolyásolják az új gondolatok fogadtatását. Noam Chomsky *Syntactic Structures* (1957) című műve megjelenésének eredményeképpen, különösen az Egyesült Államokban, a nyelvészek többsége évtizedeken keresztül a transzformációs-generatív nyelvészettel foglalkozott. Ezen irányzat teljesen figyelmen kívül hagyja a tényleges nyelvhasználat vizsgálatát, és céljának elsősorban a nyelv belső szabályrendszerének leírását tartja. Mivel a generativista irányzat a beszélő megnyilatkozását egy tökéletes nyelvtan tökéletlen, hibákkal tarkított megnyilvánulásának tekintette, tényleges megnyilvánulásokat, szövegeket nem vizsgált, és így az empirikus módszereket is megvetette. Az ilyenfajta vizsgálódást merő időpocsékolásnak, vagy az állami pénzek elherdálásának tekintették. Úgy tűnik, hogy Chomsky véleménye egyáltalán nem változott e tekintetben a majdnem 50 év alatt sem, hiszen egy Bas Aartsnak adott interjúban (2000) még a korpusznyelvészet puszta létét is tagadta. Az *Aspects of the Theory of Syntax* című művében ezt írja: *„A nyelv ismerete, mint a legtöbb fontos és érdekes tény, se nem figyelhető meg közvetlenül, se nem vonható ki adatokból semmilyen induktív módszer segítségével”*[16] (1965: 18). Andor József személyesen készített interjút Chomskyval 2004 januárjában és ekkor a korpusznyelvészetre vonatkozó kérdésre Chomsky így kezdte válaszát: *„A korpusznyelvészetnek nincs értelme”*[17] (2004: 97).

Egy ismert amerikai nyelvész, Charles Fillmore a generatív és korpusznyelvészt kicsit parodikusan a következőképpen írta le: A generatív vagy elméleti nyelvész egy kényelmes fotelban ülve, csukott szemmel elmélkedik, majd szemét kinyitva felkiált: Milyen érdekes tény! Felkapja a tollát, és jegyzetelni kezd. Öröme órák múltán sem csillapul, hiszen egy lépéssel közelebb került a nyelv megismeréséhez. A korpusznyelvész viszont hatalmas korpuszában valójában minden elsődleges tény birtokában van. Céljának azt tekinti, hogy az elsődleges tényekből másodlagos tényeket hozzon létre (1992: 35). Érdemes itt ismertetnünk egy másik összehasonlítást is, melyet Lager (1995: 3) doktori dolgozatából vettünk:

A KORPUSZNYELVÉSZ	AZ ELMÉLETI NYELVÉSZ
Figyelmét a **beszéd/performancia/folyamat**nak szenteli.	Figyelmét a **nyelv/kompetencia/rendszer**nek szenteli.
Célja a nyelv tényeinek és a nyelvhasználatnak a korpuszban való megjelenés szerinti leírása.	Célja az egyén tudatában élő nyelv tényeinek a magyarázatát adni.
A nyelvészeti kutatásai **adat-vezéreltek**. Imádja a hosszú listákat.	Nyelvészeti kutatásait **elméletek** vezérlik.
Megelégszik egyszerű módszerekkel, mert gyorsak, még hosszú szövegek esetében is. Úgy érzi, hogy a módszerei egyszerűsége által teremtett pontatlanságot el lehet viselni.	Az egyszerűségnél jobban kedveli a kifinomultságot, a sebességnél a pontosságot, még ha ez azzal is jár, hogy kevesebb példával is be kell érnie.

[16] Eredetiben: *“Knowledge of the language, like most facts of interest and importance, is neither presented for direct observation nor extractable from data by inductive procedures of any known sort.”*
[17] Eredetiben: *“Corpus linguistics doesn’t mean anything.”*

A KORPUSZNYELVÉSZ	AZ ELMÉLETI NYELVÉSZ
Kedveli a **kvantitatív** módszereket.	Kedveli a **kvalitatív** módszereket.
Önmagát az **empirikus** hagyományok követőjének tekinti.	Önmagát a **racionalista** hagyományok követőjének tekinti.
A szöveget **fizikai produktumként** kezeli.	A szöveget **absztrakt entitásként** kezeli.
Főként egyes nyelvek nyelvtana érdekli.	Az egyetemes nyelvtan érdekli.
Csak a **formára** figyel.	A formára **és** a jelentésre figyel.
A szövegeket **globálisan** vizsgálja. Át akarja tekinteni, hogy mi van a szövegben.	A szövegeket **lokálisan** vizsgálja, azaz a részletekre összpontosít.
Megkötésektől mentes szövegek széleskörű (bár valószínűleg felszínes) elemzésére összpontosít.	(Mesterségesen) megkötött területek mély elemzését végzi.
Kedveli a statisztikai és a valószínűség elméletén alapuló módszereket.	Kedveli a **szabályvezérelt** (tudásalapú, dedukciós, logikán alapuló) módszereket.
A nyílt nyelvi viselkedés (produktumainak) megfigyelésére támaszkodik.	Fő adatszerző módszere az intuícióra/önvizsgálatra támaszkodik.
Autentikus adatokkal és más beszédmegnyilvánulások kontextusában levő beszédmegnyilvánulásokkal dolgozik.	„Játék példákkal" dolgozik izolált, kontextus nélküli mondatok formájában.
Bacon gondonkodásmódját követi. Úgy gondolja, hogy a tudomány lényege az induktív módszer.	Úgy véli, hogy a feltételező-dedukciós módszer a tudomány lényege.
Szilárdan hisz a FELFEDEZŐ folyamatokban.	Az IGAZOLÓ/ÉRTÉKELŐ folyamatokban hisz.
Ha a számítógépet a szövegszerkesztésen kívül másra is használja, akkor azt konkordanciák készítése, lemmatizálás vagy statisztikai számlálás céljából teszi.	Ha érdekli egyáltalán a számítógép, akkor parserekkel, generátorokkal, és teoréma bizonyító programokkal dolgozik.
Ha egyáltalán foglalkozik programozással, akkor azt Pascal, C vagy Pearl nyelvben teszi.	A Prolog vagy Lisp az, ami számít, ha programozással is foglalkozik.

7. táblázat: A korpusz és az elméleti nyelvész paródiája (Lager 1995: 3)

Mint minden karikatúra, ez a leírás is alapjaiban igaz, de túlzó. A kétfajta nyelvészeti szemlélet ellentétének alapja tulajdonképpen a beszéd/performancia/folyamat és a nyelv/kompetencia/rendszerről vallott nézeteken alapszik, mely egyben a további különbségek kiváltó oka is. A performancia és kompetencia részletes vizsgálatát számos cikk taglalja (Fromkin, 1968; Milroy, 1985; Newmeyer, 1990; Taylor, 1988). Jóllehet ezt a fajta kettősséget a főbb modern nyelvészeti irányzatok elfogadták (Stubbs, 1996), elsősorban brit nyelvészek, mint Firth (1957), Halliday (1978) és Sinclair (1991) elvetették. A korpusznyelvészek a brit nyelvészeket követték.

A közmondás szerint is az „arany középutat" kell követni, melyet Fillmore úgy fogalmazott meg, hogy „két nyelvész egy testben" (1992: 35) kellene, hogy létezzen. Tehát, az elméleti nyelvésznek is szüksége van autentikus példákra és adatokra, míg a korpusznyelvész sem elégedhet meg puszta statisztikákkal, és az intuíciójára is támaszkodnia kell. Számos nyelvész fogadta meg Fillmore tanácsát mindkét táborból. Az elméleti nyelvészeti kutatási programok, mint az MIT Lexikon Projektje, a FrameNet, és a WordNet is használtak korpuszokat. A kutatók közül sokan kutatásaikhoz empirikus adatokat használtak, például Levin & Rappaport Hovav (1995), Miller & Fellbaum (1992), Jackendoff (1977), Pinker (1989, 2000), és Pustejovsky (1995). Az empirikus

adatok és korpuszok használata elsősorban a szemantika, különösen a lexikális szemantika területén vált elfogadottá. A szövegelemzés és textológia is sokat profitált a korpuszok használatából. Ezzel egyidőben a korpusznyelvészet is sokat változott. Véget vetettek a „szószámlálásnak", és egyre kvalitatívabb lett. A kétfajta szemlélet inkább kiegészíti egymást, semmint ellentmondana egymásnak.

2.3.2. A korpusznyelvészet és a kapcsolódó tudományágak

2.3.2.1. A számítógépes nyelvészet

A korpusznyelvészet a legszorosabb kapcsolatban a számítógépes nyelvészettel van, és sokszor igen nehéz a határvonalat meghúzni köztük. A számítógépes nyelvészetet az Encyclopedia Britannica Online úgy határozza meg, hogy az pusztán az „elektromos digitális számítógépek használata a nyelvészeti kutatásban"[18] (Linguistics, 2005). Két szempontból tűnik ez a meghatározás túl egyszerűnek. Először is azt az érzetet kelti, hogy ez semmiben nem különbözik a hagyományos nyelvészettől, pusztán a számítógép használatában. Ha ez igaz lenne, akkor más tudományok esetében is, ha nem is minden esetben, ezt külön jelzővel illetnék, és beszélhetnénk számítógépes fizikáról vagy számítógépes kémiáról is, hiszen ott is használnak számítógépeket a kutatásban. A meghatározás alapján tehát a hagyományos értelemben vett nyelvészet „elnyeli" a számítógépes nyelvészetet, és ezek alapján nem lehet önálló diszciplína. Ez egyben magában hordozza a másik problémát. Azt az érzetet is kelti, hogy a számítógépes nyelvészet kutatási problémái azonosak a hagyományos nyelvészetével, csak számítógépet használnak vizsgálatukhoz. Ezzel szemben a számítógépes nyelvészet kérdései mások, mint a hagyományos nyelvészeté, és ezek vizsgálalát nem lehetne a számítógép segítsége nélkül elvégezni, tehát a számítógép használata szerves része, feltétele a kutatásnak.

Az emberek igen kis százaléka rendelkezik nyelvészeti szakszótárral, így kíváncsiságukat legtöbbször enciklopédiák segítsével elégítik ki, ami tévedésekhez, pontatlanságokhoz vezethet. Lássuk tehát a számítógépes nyelvészetnek egy nyelvészeti szótárban megjelent meghatározását is: *„A nyelvészetet és az (alkalmazott) számítógépes tudományokat egyesítő tudományág, mely a természetes nyelvek számítógépes feldolgozásával foglalkozik (a nyelvi leírás minden szintjén)"*[19] (Bussmann, 1996: 91).

2.3.2.2. A mesterséges intelligencia

A számítástechnika területéről a számítógépes nyelvészek számára elsősorban a mesterséges intelligencia kutatásai fontosak, melyeknek célja „az emberi intelligencia és kognitív képességek megértése és szimulálása számítógépek segítségével"[20] (ibid. 38). Jóllehet a mesterséges intelligenciával foglalkozó kutatások már a számítógépek meg-

[18] Eredetiben: *"no more than the use of electronic digital computers in linguistic research"*.
[19] Eredetiben: *"Discipline straddling linguistics and (applied) computer science that is concerned with the computer processing of natural languages (on all levels of lingusitis description)."*
[20] Eredetiben: *"to simulate and understand human intelligence and cognitive abilities by using machines (i.e. computers)"*.

jelenése előtt megkezdődtek, igazából a számítógépek megjelenése után nőtt meg az ezzel foglalkozó kutatók száma. Az első mesterséges intelligenciáról szóló tudományos cikket 1950-ben Alan Turing angol tudós publikálta. Az első főállású alkalmazottakból álló kutatócsoport 1954-ben, a Carnegie Mellon Egyetemen alakult Allen Newell és Herbert Simon közreműködésével. Az első számítógépekről szóló konferenciát 1956-ban Dartmouth-ban tartották, melyen azt kívánták modellálni, hogy hogyan gondolkodik az ember. Így tehát a játékok, mint például a sakkozás, vagy fordítások készítésére próbálták a gépeket programozni. A technika fejletlensége miatt nem sikerült megfelelő eredményeket felmutatni.

Jóllehet nem igen vagyunk tudatában, de szinte minden nap használjuk a mesterséges intelligencia kutatásainak eredményeit, amint bekapcsoljuk a számítógépünket. Lawrence Tesler úgy határozta meg a mesterséges intelligenciát a számítástechnikához viszonyítva, hogy *„minden, amit még nem csináltunk meg”* (McEnery, 1992: 10). E szerint a mesterséges intelligencia mai problémái holnapra már közhelyekké válnak. Ha valaki ki szeretne próbálni egy korai, de mai napig jól ismert programot, annak Weitzenbaum (1976) Eliza nevű programját javasoljuk. Ez a program a Turing tesztet is kiállja, ami abból áll, hogy miközben egy személy a számítógéppel és egy másik személlyel kommunikál a számítógépen, meg kell mondania, hogy mikor „beszél” a gép és mikor a másik személy. Az interneten a következő két címen lehet elérni a program egy változatát, és angol nyelven „társalogni” a számítógéppel: http://www.manifestation.com/neurotoys/eliza.php3 és http://www-ai.ijs.si/eliza/eliza.html.

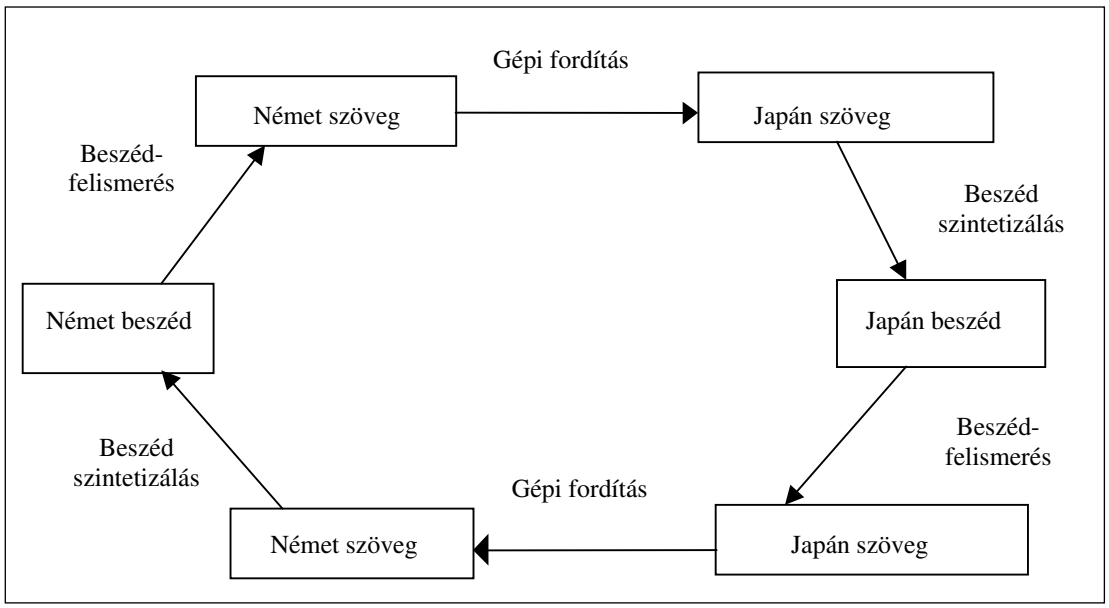
2.3.2.3. A számítógépes nyelvészet kutatási területe

Mint említettük, a számítógépes nyelvészet a nyelvészet és a mesterséges intelligencia interdiszciplináris kutatási területe, melynek műveléséhez mindkettő tudományág alapos ismerete szükséges. Ennek ellenére az 1950-es években a korai kutatásokat ezen a területen elsősorban mérnökök és számítógépes szakemberek végezték, akik nem sokat tudtak a nyelvről vagy nyelvészetről. Figyelmüket elsősorban a számítógépes fordítás kötötte le, melyről azt hitték, hogy viszonylag egyszerűen megoldható probléma. Azt képzelték, ami a mai nyelvtanulók körében sem ritka, hogy elegendő pusztán összepárosítani az egyik nyelv szavát a másik nyelv szavával, és kész a jó fordítás. Így fordulhatott elő, hogy a következő angol modatot: *“The spirit is willing but the flesh is weak.”* (ford. A szellem hajlandó, de a test gyenge, Máté 26, 41) oroszra fordították, majd ismét angolra és a következő mondat lett belőle: *“The vodka is excellent but the meat is rotten”* (ford. A vodka kiváló, de a hús rohadt.) (Pfaffenberger, 1993: 41). Nyilvánvaló, hogy ha egy adott szó több fordítási megfelelővel is rendelkezik, akkor nem lehet véletlenszerűen kiválasztani, mely fordítását használjuk. Ezt a fenti példa is bizonyítja.

A számítógépes nyelvészek ma is foglalkoznak a gépi fordítással, de nem tekintik azt kizárólagos feladatuknak. Kutatásaik többek között a következőket is vizsgálják: 1) természetes nyelvi interfész, 2) helyesírást ellenőrző programok és szókincstárak, 3) dokumentum-feldolgozás, 4) stílusellenőrzők és 5) lexikográfia.

A tágabban értelmezett számítógépes nyelvészet a beszédfelismeréssel is foglalkozik. A beszédfelismerő programok használatakor gondot okozhat a különböző hangok felis-

merése, például férfi-női, vagy felnőtt-gyerek hang, ezért e programok sokszor kiválóan működnek egy bizonyos beszélő vagy beszélő típus esetében, míg mások számára teljesen használhatatlanok. Így tehát azok a programok működnek megfelelően, amelyek vagy „taníthatók" a beszélőhöz való alkalmazkodásra, vagy eleve nem beszélőfüggők. 2000 augusztusában Hermann Maurer (személyes közlés, Tokyo) megemlítette, hogy a német Siemens cég szerint a közeljövőben egy német személy telefonbeszélgetést folytathat egy japánnal anélkül, hogy bármelyikük is beszélne a másik nyelvén. A következő ábra szemlélteti, hogy ez hogyan lehetséges.



18. ábra: A technológiával támogatott valós időben történő kommunikáció

Természetesen bármilyen nyelvvel helyettesíthető a japán és a német, csak megfelelő minőségű beszédfelismerő és szintetizáló programra van szükség az adott nyelven. A gépi fordítórendszernek gyorsnak és pontosnak kell lennie. Először is képesnek kell lennie arra, hogy az azonos hangzású szavakat helyesen értelmezze és írja le, vagy az azonos alakú, de több jelentésű szavak esetében a fordításhoz a megfelelő jelentést válassza ki. Ez a rendszer teljesen automatikusan, emberi közreműködés nélkül működne, és semmilyen közbeeső nyelv, mint például az angol nem szerepel. Nyilvánvaló, hogy egy közbülső természetes nyelv csak fokozná a hibalehetőségek számát. Hogy miért éppen a német–japán rendszer van már jelenleg kísérleti stádiumban? Nyilván gazdasági és technikai okok játszanak ebben döntő szerepet. Így hát ne is nagyon várjuk, hogy mi, magyarok is cseveghetünk majd másokkal nyelvtanulás nélkül.

2.3.3. A magyarországi számítógépes nyelvészetről

Mindenkiben felmerülhet a kérdés, hogy vajon Magyarországon mikor kezdtek el a számítógépes nyelvészettel foglalkozni? A válasz talán meglepő lesz. Már 1961-ben

megkezdődtek ilyen jellegű kutatások, és 1963-tól kezdve A számítógépes nyelvészet című szakkiadványt is kiadták évente egyszer vagy kétszer. A kiadvány 500 példányban jelent meg, a szerkesztőbizottság tagjai Papp Ferenc, Petőfi S. János, Szépe György és Varga Dénes voltak. 1975-re nem csak a szocialista blokk országaiban, hanem a John Benjamins B. V. kiadó révén az egész világon elérhetővé vált a *Computational Linguistics and Computer Languages* ('Számítógépes nyelvészet és számítógépes nyelvek') címmel. Tudomásom szerint az utolsó szám 1982-ben jelent meg.

Papp Ferenc (1930–2001) a Debreceni Egyetemen (amelyet akkor még Kossuth Lajos Tudományegyetemként ismertünk) állította be az első számítógépet a humán tudományok szolgálatára. Hunyadi László a magyar számítógépes nyelvészetről írt cikkében (1999) Papp Ferenc úttörő munkásságának egyik nagy eredményeként megemlíti *A magyar nyelv szóvégmutató szótárát* (1969). Papp Ferenc magyar, angol és orosz nyelven írt és szerkeszett könyvei, mint például a *Matematikai nyelvészet és gépi fordítás a Szovjetunióban* (Papp *et al.*, 1964), vagy a hasonló címmel angolul megjelent könyve (1966) mutatja, hogy milyen korán megindult Magyarországon az érdeklődés a számítógépes nyelvészet iránt.

A hazai számítógépes nyelvészet ismertetése során ki kell térnünk az MTA Nyelvtudományi Intézete Fonetikai Osztályán az 1980-as években nemzetközi szinten is úttörő beszédszintetizálási kutatásokra. E munkálatok keretében magyar és orosz nyelvű szöveg–beszéd (text to speech) rendszereket fejlesztett ki Bolla Kálmán, Kiss Gábor, Nikléczy Péter és Olaszy Gábor.

Kiefer Ferenc akadémikus a másik prominens személyisége a hazai számítógépes nyelvészetnek. 1992-től 2002-ig igazgatta a Magyar Tudományos Akadémia Nyelvtudományi Intézetét, amely mindig is kulcsszerepet játszott a hazai nyelvészeti kutatások fejlődésében. Kiefert már képzettsége is predesztinálta a számítógépes nyelvészettel való foglalkozásra, hiszen nyelvi diplomái mellett matematikusi diplomát is szerzett. Már 1964-es első publikációi is (Kiefer, 1964a, 1964b) azt jelezték, hogy elsősorban a nyelv és a nyelvészet érdekli, a matematikát pedig eszköznek tekinti, amely a nyelv megismerésében segíti. Számos írása jelent meg nemcsak itthon, hanem külföldön is, többek között a *Mathematical Linguistics in Eastern Europe* (1968) című könyve. Kiefer vezetése alatt az intézetben számos nemzetközi tudományos kutatás indult meg, melyek egy részét az Európai Unió támogatja. Kiefer Ferenc tiszteletére barátai és tanítványai könyvet adtak ki, melynek előszava (Szépe, 2001) bővebben ismerteti életútját és munkásságát.

Jóllehet a mai napig nincs önálló számítógépes nyelvészeti osztálya a Magyar Tudományos Akadémia Nyelvtudományi Intézetének, nemzetközileg is elismert szinten folyik az ilyen jellegű kutatómunka Váradi Tamás vezetésével a Korpusznyelvészeti Osztályon. Annak ellenére, hogy hivatalosan csak 1997 óta működik ez az osztály, a nyelvtechnológiai kutatások és fejlesztések már évekkel korábban megkezdődtek. A Nyelvtudományi Intézetben 1985-ben kezdődött el egy projekt, melynek az volt a célja, hogy a *Magyar nagyszótár* szerkesztését egy magyar nyelvű korpusz létrehozásával segítse. Pajzs Júlia a projekt előrehaladásáról és különböző aspektusairól számol be cikkeiben (pl., 1990, 1991, 1994; Pajzs *et al.*, 1992).

Az eddigiekből is kitűnik, hogy az e témával foglalkozó írások nagy része külföldön, vagy ha itthon is, de idegen nyelven jelent meg. Magyar nyelven három könyv jelent

meg (Bach, 1995; Prószéky, 1989; Prószéky & Kis, 1999). Bach írása az elektromérnöki kar hallgatói számára készült tankönyv, így nem széles körben terjesztik, és elsősorban a matematikai nyelvészet gyakorlati oldalával foglalkozik.

Prószéky Gábor könyve, *Számítógépes nyelvészet – Természetes nyelvek használata számítógépes rendszerekben* (1989), elsősorban a természetes nyelv feldolgozásával (NLP – natural language processing) és a természetes nyelv megértésével (NLU – natural language understanding) foglalkozik. A közel 600 oldalas könyvben négyoldalas (489–492) összefoglalást is találunk a számítógépes nyelvészet történetéről, mely a magyar nyelv feldolgozásáról szóló V. részt (489–550) vezeti be.

Prószéky és Kis *Számítógéppel emberi nyelven – Intelligens szövegkezelés számítógéppel* (1999) című könyve a szövegfeldolgozással kapcsolatos tudnivalókat írja le hétköznapi nyelven. E mellett azonban háttér információval is szolgál a gépi fordítás, lexikográfia és általában a számítógépek nyelvészetben való felhasználásának jobb megértéséhez. E könyvben számos olyan program működését is ismertetik, melyet a Prószéky által vezetett cég, a MorphoLogic készített, és a magyar nyelvű szövegfeldolgozó programok részeként működnek. A HuMor (Prószéky & Tihanyi, 1993) nevű program különösen figyelemre méltó, hiszen morfológiai elemző, amely elengedhetetlen része a magyar helyesírást ellenőrző programnak. Érdemes itt megemlíteni, hogy a cég honlapjáról elindulva http://www.morphologic.hu/h_pub.htm számos tudományos cikket találhatunk a linkeket követve. A címek már önmagunkban hasznos kis irodalomjegyzékként használhatók, hiszen Magyarországon nincs olyan jellegű szakfolyóirat, amely kizárólag a számítógépes nyelvészettel foglalkozna. Sajnálatos volt azonban, hogy a 2002-ben megjelent cikkek után semmilyen publikációt nem említettek a honlapon egy-két éven keresztül, még csak listaszerűen sem, és ez jelentősen megnehezítette írásaik figyelemmel kísérését. Örvendetes azonban, hogy Prószéky Gábor a közelmúltban felfrissítette publikációinak listáját, így már néhány 2005-ben megjelent írásának adatait is megtalálhatjuk itt.

A Nyelvtechnológiai Csoport 1998-ban a József Attila Tudományegyetem (2000 januárjától Szegedi Tudományegyetem) Informatikai Tanszékén alakult meg. A magyar nyelvre vonatkozó nyelvtechnológiai kutatások nagyon intenzíven folynak, és már szép eredmények is születtek. A különböző projektek sikere érdekében konzorciumot alkottak a fent említett MorphoLogic Kft. Budapest és az MTA Nyelvtudományi Intézet Korpusznyelvészeti Osztályával. A projektek ismertetését a csoport honlapján olvashatjuk http://www.inf.u-szeged.hu/projectdirs/hlt/. A honlap dicséretes részletességgel szól a kutatásokról és azok előzményiről is, így ezen a téren a legtöbb információt innen tudhatja meg az érdeklődő.

A Szegedi Tudományegyetem Informatikai Tanszékcsoportjának nevéhez fűződik a Magyar Számítógépes Nyelvészeti Konferencia megrendezése is. Az első konferencia 2003-ban, a második pedig 2004-ben tette lehetővé a hazai szakembereknek, hogy legfrissebb kutatásaik eredményeit bemutathassák és megvitathassák. Remélhetjük, hogy egy minden évben megrendezésre kerülő konferencia fogja a jövőben segíteni a hazai szakemberek munkáját és eszmecseréjét.

2.4. Folyóiratok

Mint említettük, a magyar nyelvű szakirodalom e téren könyvekben is igen szegényes, és a diszciplínát képviselő rendszeresen megjelenő folyóirat sem létezik, így néhány külföldi folyóiratot szeretnénk az érdeklődő olvasók figyelmébe ajánlani. A folyóiratok angol nyelvűek, de ha az egyes szerzők nevét egy internetes vagy könyvtári katalógusban megkeressük, más nyelven írott cikkekhez is eljuthatunk. Az első kettő után felsorolt folyóiratoknak nem minden száma tartalmaz feltétlen a számítógépes nyelvészethez kapcsolódó cikkeket. Esetleg több évfolyamot is át kell böngészni, mire az érdeklődésünknek megfelelőt megtaláljuk. Az itt felsorolt folyóiratok legtöbbjét a http://www.ingentaconnect.com/ honlapról elindulva elérhetjük, de a kiadóik honlapján is megtalálhatók. A folyóiratok tartalomjegyzéke és az összefoglalások ingyenesen megtekinthetők, csak a teljes cikkek letöltéséért kell fizetni, ha a könyvtárnak nincs előfizetése az adott folyóiratra.

- A **Computational Linguistics** című folyóirat (ISSN 0891-2017) az egyetlen, amelyet kizárólag a természetes nyelvfeldolgozásnak szenteltek. 1974-ben jelent meg az első szám, évente négyszer jelenik meg. Az 1994-től megjelent számok elektromos formában is megtekinthetők az előfizetők számára. A tartalomjegyzéket és a cikkek absztraktjait mindenki szabadon megtekintheti. A folyóirat honlapjáról http://mitpress.mit.edu/journals minden szám egy-egy cikkét ingyenesen is hozzáférhetővé tették. Egyes számokat kimondottan egy témának szenteltek, pl. a 29. évfolyam 3. száma (2003. szeptember) a „Web, mint korpusz” témával foglalkozik.
- **Journal of Machine Learning Research** (JMLR) on-line változata teljes egészében ingyenesen hozzáférhető a http://jmlr.org címről. Az első szám 2000 októberében jelent meg, egyes számok tematikusok, egy bizonyos témával foglalkoznak.
- **International Journal of Corpus Linguistics** (IJCL) a közelmúltig az egyetlen olyan folyóirat volt, amelyet teljes egészében a korpusznyelvészetnek szenteltek.
- **Corpus Linguistics and Linguistic Theory** teljesen új folyóirat, mely a http://www.degruyter.com/rs/384_7546_ENU_h.htm címen található meg. Első számának megjelenését 2005-re tervezik. A nyelvészet bármely területét érintő, korpusz alapú, de elméleti vonatkozású kutatások eredményeit kívánják itt publikálni.
- Applied Intelligence
- Artificial Intelligence
- Artificial Intelligence Review
- Computational Intelligence
- Computer Assisted Language Learning
- Computer Speech and Language
- Computers and Composition
- Computers and the Humanities
- Computers and Translation

- Hebrew Computational Linguistics
- International Journal of Computer Processing of Oriental Languages
- Language and Computation
- Literary & Linguistic Computing
- Machine Translation
- Research on Language and Computation
- Web Journal of Formal, Computational and Cognitive Linguistics http://fccl.ksu.ru

2.5. Összefoglalás

Ebben a fejezetben elsősorban a számítástechnika és maga a számítógép fejlődését tekintettük át, hiszen a korpusznyelvészet és a hozzá kapcsolódó társtudományok előrelépése szinte teljes egészében ezen alapult. A hardver fejlődésével (gyors memória és a merevlemez kapacitása, operációs rendszer stb.) e tudományterületek is „robosztusabbakká" váltak. A korpusznyelvészet és a számítógépes nyelvészet fejlődését a nyelvészeti elméleti háttér is igen befolyásolta. Chomsky és az őt követő, elsősorban amerikai (generatív) nyelvészet hatására az elméleti nyelvészet vált uralkodóvá. A nyelvhasználat vizsgálatát, és az empirikus kutatásokat sokan időpocsékolásnak tekintették.

A korpusznyelvészet tanulmányozása során elengedhetetlen, hogy a kapcsolódó tudományokról ne tájékozódjunk valamennyire. Sok cikkről igen nehéz megmondani, hogy vajon a korpusznyelvészet vagy a számítógépes nyelvészet címszó alá esik-e. A mesterséges intelligencia meghatározása után a számítógépes nyelvészet kutatási területét körvonalaztuk, amely az első korszakban szinte kizárólag a gépi fordítással foglalkozott.

A magyarországi számítógépes nyelvészet főbb képviselőinek – Papp Ferenc, Kiefer Ferenc, Prószéky Gábor, és az MTA Nyelvtudományi Intézete – munkásságát dióhéjban bemutattuk, majd a hazai szakirodalom hiányában néhány külföldi szakfolyóiratról és elérhetőségükről adtunk felvilágosítást a teljesség igénye nélkül.