A gépi fordítás hetvenéves története

Prószéky Gábor[1]

[1] Nyelvtudományi Intézet
proszeky.gabor@nytud.hu

1. A hetven az hetven

Váradi Tamás hetvenedik születésnapja alkalmából arra teszek kísérletet, hogy a számítógépes nyelvészeti kutatások talán legismertebb, Tamással gyakorlatilag egyidős és az ő tevékenységei között többször is érintett területnek, a gépi fordításnak a hetvenéves történetét röviden összefoglaljam. Három nagy szakaszra szokás felosztani ezt az időszakot: a szabályalapú fordítás, a statisztikai fordítás és a neurális hálókkal történő gépi fordítás időszakára. Az első szakasz volt a leghosszabb, amely gyakorlatilag az ötvenes évek elejétől a géppel elérhető nagyméretű szövegkorpuszok megjelenéséig, a kilencvenes évekig tartott. Az ekkor megjelenő statisztikai közelítések egészen a neurális hálós módszerek megjelenéséig, a most befejeződött évtized elejéig tartottak. Napjainkban az uralkodó tudományos paradigma két fontos ismérve, hogy nem a nyelvészeti, de sokszor még nem is a programozási tudás az, ami a fordítási minőséget jelentős mértékben képes feljavítani, hanem a neurális rendszerek egyfajta „paraméter-beállítási" intuíciója. A tanítóanyagok gondos kiválasztásának és előfeldolgozásának megnőtt a jelentősége, ami ugyan igényli a tapasztalt nyelvész közreműködését, ám az új gépi környezetben sok, korábban jelentős eredményt elérő nyelvész a teljesen új szemlélet miatt kevésbé sikeres. Tamás azonban ebben a – hagyományos nyelvészeti orientációjú közelítéseknél sokkal több technikai ismeretet igénylő – világban, a neurális hálók világában is ígéretes első eredményeket tudhat magáénak.

2. A gép elkezd fordítani, a nyelvész szabályai alapján

A szélesebb értelemben vett *számítógépes nyelvészet* a számítógép és a nyelvészet számos lehetséges találkozási pontján kialakult szakterület. Ezen belül a *nyelvtechnológia* – azaz a mai tudományos világban hasz-

DOI: https://doi.org/10.18135/VT70.17

nált angol elnevezéssel: *human language technologies* – úgy definiálható, hogy ez az informatikának az az ága, ahol a nyelvészeti kutatásokon alapuló eredmények beépülnek a számítógépes rendszerekbe. Teszik ezt úgy, hogy a felhasználók számára a számítógéppel való kommunikáció folyamán az így kialakított szoftverrendszerek – bizonyos célhelyzetekben – a nyelvet jól használó emberéhez hasonló támogatást tudnak adni. Világosan kell látni, hogy eddig a nyelvet kizárólag az ember számára írta le a nyelvész, így bizonyos pontokon módja volt „összekacsintani" leendő olvasójával, építve arra, hogy az is ember, méghozzá nagy eséllyel hasonló kulturális háttérrel, így bizonyos alapvető fogalmak megmagyarázására nem volt szükség. A számítógép, amelynek számára leírjuk a nyelvet, nem rendelkezik azokkal a háttérismeretekkel, amivel egy nyelvtant értelmező ember, így minden olyan fogalmat, amelyre szükség lehet a rendszer működtetéséhez, a gépek számára részleteiben le kell írni. Egy egyszerű analógiával megvilágítva, ha a „vásárlást" mint tevékenységet leírnánk a gép számára, akkor azt a tényt, hogy a végén „oda kell menni a pénztárhoz", a gép csak úgy tudja értelmezni, ha az ehhez szükséges „menést" mint tevékenységet ismeri, különben kénytelenek vagyunk részletesen ezt is kibontani, azaz a „lépéseket" mint a „menés" alapelemeit is definiálni kell számára, és így tovább.

Napjainkra megjelent tehát egy új eszköz, mely az emberen kívül először képes a nyelvi leírás működtetésére: ez természetesen a számítógép, ami új nyelvészeti közelítések kialakítását is magával hozta. Talán a fentiekből az is világossá vált, hogy a 20. század közepétől kezdve ki kellett hogy alakuljon egy olyan nyelvleírási mód, amely csak részben azonos a nyelvészet addig meghatározónak tűnő elméleteivel, és sok olyan elemet tartalmaz, melyet az ember számára annak idején nem kellett leírni. A nyelvekkel kapcsolatban az általános tapasztalat ugyanis a 19. század közepéig az volt, hogy a nyelv változik. Ezért valójában a nyelvészet története a 20. századig elsősorban a történeti nyelvészet története volt. A 20. században megjelenő leíró, vagy más néven *deskriptív nyelvészet* viszont egyfajta „mechanikus" leírásnak is tekinthető, melyet már a számítógép létrejötte előtt egyfajta algoritmikus szemlélet jellemzett. A számítógépről ismeretes, hogy bizonyos értelemben a második világháború „hozadéka". Az eszköz neve igen sok nyelven a számolással, azaz a *comput-* latin tő valamely származékával kapcsolatos szóból alakult ki. Az egyik talán kevésbé ismeretes fő ok a számítógép létrejöttében a háborúkban oly fontos titkosírások mechanikus, sőt elektromechanikus kezelésének vágya, azaz a kódolás-dekódolás folyamatának gépesítése volt.

A világháború végén, a hidegháború kialakulásának hajnalán az Atlanti-óceán mindkét partján megjelent a gondolat, hogy a kódolás és dekódolás viszonya és az emberi nyelvek fordítása hasonló jellegű tevékenység, így egy ilyen eszköz létrejötte a gépi fordítás megvalósíthatóságának gondolatát is egyre erősítette (Hutchins, 1997). Ehhez nagy lökést adott, az MIT meghatározó hatású, kiváló matematikusának, Bar-Hillelnek az ötvenes évek elején tett kijelentése, miszerint idő kérdése csak, de a teljesen automatikus gépi fordítás megvalósítható (Bar-Hillel, 1951). Az Egyesült Államok kormánya nem kevés pénzt koncentrált erre az ígéretes kutatási területre, ami elsősorban az orosz műszaki-katonai szövegek fordításának automatizálását célozta meg. Az első működő gépi fordítást végző számítógép 1954-ben mutatkozott be az IBM georgetowni központjában (IBM, 1954). A fordításban részt vevő nyelvek leírása a gép számára azonban nem a nyelvészek által követett úton történt. Ennek egyik oka, hogy a gépi fordítást végző kutatók igazán nem is a nyelvet akarták leírni, hanem azt a módszert szerették volna megragadni, melynek segítségével az egyik nyelv szerkezeteit a másik nyelv szerkezeteivé tudja alakítani az ember. Az alapgondolat az volt, hogy ha ez a módszer megvan, akkor akár egy program is végre tudja hajtani a lépéseit. A fordítási egység a mondat volt, de nem abban a generatív értelemben, amelyről ebben az évtizedben már Chomsky egyre többet publikált (Chomsky, 1957). Ennek az egyik nyilvánvaló oka, hogy a Chomsky-modell az ideális beszélő nyelvi kompetenciáját volt hivatva megfogalmazni, a gépi fordításhoz pedig a bemenő mondatot egy nem feltétlenül ideális beszélő hozta létre, és a feldolgozás eredményeként sem egy absztrakt nyelvi szerkezetet, hanem egy másik nyelvi fordítást kellett a gépnek produkálnia. A világ akkori másik pólusán, a szovjet blokkban is folytak természetesen a kutatások, de a számítógépesítés alacsonyabb foka miatt a születendő nyelvi modelleket inkább matematikai nyelvészetinek nevezték (Papp, 1964). A Szovjetunió néhány neves nyelvészének hatására ezt követően az ún. szocialista országokban, így hazánkban is megindult a gépi fordítás kutatása. A gépi nyelvészet akkori amerikai eredményei – és nemcsak a „hivatalosan” publikáltak – ma is fellelhetők az Országos Műszaki Könyvtár által az ötvenes évek végén és a hatvanas évek elején beszerzett és félig-meddig titkos mikrofilmeken. A módszerek eleinte ugyan próbálták ötvözni az akkortájt születő generatív nyelvelméletek eredményeit a gépi feldolgozással, de egyre jellemzőbbé váltak nálunk is és máshol is a nyelvelméletmentes gépi kísérletek. Matematikai szempontból az volt az egyik probléma, hogy a Chomsky-féle

transzformációk nem invertálhatók. Ez számítógépes szempontból azt jelenti, hogy egy mondatátalakításkor kitörölt vagy elmozgatott elem helyét, az ún. nyomot a mondatelemző program nem találja meg, ui. a generatív levezetés végén ezek törlődnek. Az ilyen, a mondatban elvileg ott levő, de fizikailag nem megtalálható elemek visszaállítása az esetek jelentős részében nem, vagy csak nagyon hosszú idő alatt történhet meg. Már pedig komoly időbeli eltérés a mondatelemzés és a mondatlétrehozás között az emberi nyelvfeldolgozás esetén nem ismert, így furcsa volna egy olyan modell, mely egész máshogy működik generáláskor, mint elemzéskor. A számítógépes szakembereknek ugyanis elsősorban az emberek által létrehozott, és olykor nem pontosan megfogalmazott mondatokat kell elemezniük, és nem ideális mondatokat létrehozni. Így a számítógépes gyakorlatban egymás után jelentek meg olyan nyelvelméleti modellek, melyek nem a Chomsky-féle irányt követték, hanem például az őt megelőző strukturális leírást (Harris, 1951) vagy azt az alternatív elméletet, mely elsődlegesen a szavak közötti függőségi viszonyt szándékozott leírni (Tesnière, 1959). A teljesen automatikus gépi fordítás megvalósíthatóságát épp azok kezdték megkérdőjelezni az évtized végére, akik az évtized elején még az ügy élharcosai voltak, így a fordítással foglalkozó számítógépes kutatók elkezdtek a nyelv más, nemcsak fordítással kapcsolatos gépi feldolgozásával foglalkozni. Ekkor alakult ki az immár nemcsak a gépi fordítást magába foglaló számítógépes nyelvészet fogalma. Ehhez az Egyesült Államokban a gazdasági-politikai háttér is adott volt: a hidegháború eddig is a kutatási támogatás fő motiválója volt, de most már nemcsak a gépi fordításra koncentráltak. A Holdra szálláshoz például elkészült egy olyan számítógépes nyelvészeti program, amely a lehozott holdkőzetek adatbázisához angol nyelvű mondatok segítségével való hozzáférést biztosított (Woods, 1973). Ez bizonyos értelemben fordítóprogramnak volt tekinthető, bár a rendszer célnyelve nem emberi nyelv, hanem egy adatbázis-kezelő program nyelve volt. A számítógépes nyelvészet szempontjából lényeges, hogy Woods ennek a rendszernek a működtetéséhez létrehozta az Augmented Transition Network nevű leíró-működtető formalizmust (Woods, 1970). Az eljárás a véges állapotú automatáknak az emberi nyelvek rekurzív szerkezeteinek kezelésére is alkalmas kiterjesztésén alapult, és az ezt követő években, sőt évtizedekben a pszicholingvisztika és az amerikai számítógépes nyelvészet egyik alapmodelljévé vált, jóllehet visszalépéses

elven történő működése elsősorban a – gépi fordító rendszerek egyik állandó nyelve – az angol mondatainak feldolgozásakor volt csak evidens, a más típusú, például szabad szórendű nyelvek mondatainál nem.

A mesterségesintelligencia-kutatásból ekkortájt kinövőfélben levő nyelvvel kapcsolatos gépi alkalmazások másik legismertebbje Winograd nevéhez fűződik (Winograd, 1972). Ő a nyelv procedurális közelítésével kísérletezett. SHRDLU nevű rendszere egy olyan világot mozgat meg (angol) nyelvi instrukciók segítségével, melyben egy síklapon elrendezve háromdimenziós geometriai objektumok vannak csak, színükkel, méretükkel és alakjukkal. A nyelvi bemenet hatására a világ változásait reprezentálják, így az fizikai átrendezés ebben a virtuális világban a begépelt parancsok hatására megy végbe, amiről a gép „tud", és megfelelően reagál. Itt tehát a nyelv gépi reprezentációja procedurális, hiszen a nyelvi megnyilvánulások számítógép által végrehajtható műveletekbe való gépi fordításáról van szó.

Az eredeti értelemben vett gépi fordítás nagy túlélői viszont annak ellenére működtek, hogy az Egyesült Államok kormánya által a gépi fordítási eredmények – illetve egészen pontosan: az eredménytelenségek – vizsgálatára kijelölt bizottság szakvéleménye, az ALPAC-jelentés (Pierce et al., 1966) legtöbbjüket profilváltoztatásra kényszerítette. A korábban a georgetowni IBM-fordítókísérletet vezető magyar származású Toma Péter által alapított és az üzleti világban is sikeresnek mondható Systran rendszer az Európai Közösség érdeklődését is felkeltette, és hosszas tárgyalások után meg is vásárolták az egyre több nyelvet beszélő közösség fordítási gondjainak csökkentése céljából. A Logos fordítórendszer, melynek indulását a vietnami háború nyelvi nehézségei szolgáltatták, üzleti termékké vált, és a hetvenes évektől először a Wang, majd tőle az IBM, később pedig a Sun cég vásárolta meg, üzleti reményekkel. A Texas Egyetemen kifejlesztett angol–német fordítást végző Metal rendszer 1978-ban Európába került, a Siemenshez. A gépi fordítás az Egyesült Államokon kívül bizonyos értelemben érintetlenebb maradt az ALPAC-jelentés következményeitől. Így alakulhatott ki Kanadában az angol és francia időjárás-jelentéseket az egyik nyelvről a másikra fordító szolgáltatás, a METEO, vagy az egységes gazdaság irányába induló Európa néhány erődemonstrálási céllal indított K+F-projektje: az Eurotra és a DLT. Ez idő tájt jelentkezett az ötödik generációs számítógép gondolata is, és benne a japán álom, mely az akkor még két évtizednyi távolságban levő ezredfordulóra prognosztizálta a nyelvet intelligensen használó, beszélő és fordító számtógép megvalósítását. Mivel akkoriban

ettől még nagyon messze látszott lenni a világ, az amerikai oldalon megelégedtek az újonnan megjelenő fogalom, a *természetesnyelv-feldolgozás* (natural language processing: NLP) emlegetésével. Hazánkban egyébként az ötvenes évek végétől néhány évig szintén működött egy gépi nyelvészeti csoport, melynek kutatásait részben épp az ALPAC-jelentés közép-európai mellékhatásaként állították le (Prószéky, 2013).

A nyelvészet területén a gépi fordítás számára szóba jöhető újdonság csak a hetvenes évek végén jelentkezett, amikor Chomsky transzformációs nyelvtanának alapproblémáit egy új ügyes technikával kikerülve – bizonyos értelemben a strukturalista Harris és a generatív Chomsky közötti nyelvleírási különbségek újragondolásával – megjelent néhány új formalizmus: a GPSG, az LFG, majd a HPSG (Sells, 1985). Ezek az elképzelések azért voltak jelentősek, mert a számítógépes megvalósíthatóságot fontos szempontként maguk előtt tartva új lökést adtak a gépi nyelvészet művelőinek is. Azonban a lexikalizálódás, azaz a szótári információknak a szintaxis területén való hatékony térfoglalása meglehetősen komplex nyelvi struktúrákat és ebből következően (gép)időigényes műveletsorokat hozott. Így az ezeket működtetni szándékozó informatikai megoldások csak a fenti elméletek képességeinek demonstrálását szolgálták elsősorban, a gyakorlati életben, például a gépi fordítás területén nem játszottak meghatározó szerepet. Egy másik elméleti indíttatású gépi fordítási közelítés a modern formális logika egyik atyja, Gottlob Frege elmélete (Frege, 1923) egyfajta számítógépesítésének mondható Rosetta rendszer volt. Ez a „rule-to-rule" hipotézisen, azaz a szintaktikai és szemantikai szabályok párba állításán alapuló fordítási közelítés középpontba állításán alapult, de ennek sem lettek gyakorlati követői a gépi fordítás más művelői között. Időközben Chomsky folyamatosan megjelenő újabb generatív nyelvészeti elképzelései (Chomsky, 1981; 1993) meglehetősen átformálták a korábbi közelítést, de a generatív felfogás alapjai nem változtak, ezért a számítógépesek és különösen a gépi fordítók továbbra is jobban bíztak a hetvenes évek elején kialakult alapmodelljeikben. Ezek aktuális összefoglalását épp az a Winograd adta, aki a hetvenes évek elején bemutatott procedurális módszerével beírta magát a gépi nyelvészet történelmébe. Winograd nyelvi proceduralitásról szóló, összefoglaló, egyfajta „kvázi-formális" elméletről szóló könyve, a *Language as a Cognitive Process* 1983-ban jelent meg (Winograd, 1983). Ez idő tájt egyébként más kognitív grammatikák is megjelentek, melyek tudásalapú paradigmák formájában a gépi fordításon belül is fel-felbukkantak. Ezekben a világismeret és a nyelvi tudás keveredett, némiképp fittyet

hányva a nyelvészeti jelentéstan és a világismeret közötti falat szigorúan őrző nyelvészeti közelítéseknek. A Winograd-könyv egyik érdekessége egyébként, hogy bár összefoglalt szinte mindent, ami a számítógépes nyelvfeldolgozásban fontos lehetett a nyolcvanas évek elején, ám az a szó, hogy „morfológia", nem fordult elő benne. Itt is tetten érhető tehát, hogy a nyelv fogalma akkoriban többé-kevésbé az angol nyelvet jelentette. Ugyanebben az évben épp az említett területen történt egy fontos elméleti áttörés: a számítógépes nyelvészet morfológiai leírása egységes elméleti hátteret kapott, ugyanis megszületett egy új formalizmus, a reguláris nyelvtanok „újjászületésére" építkező kétszintes morfológia (Koskenniemi, 1983). Ettől kezdve a szabályalapú gépi fordító rendszerek legelső és legutolsó modulja, a szóalaktani elemzés és a szóalaktani generálás mostantól nem feltétlenül ad hoc karaktermanipulációkra, hanem ezekre a kétszintes rendszerekre épülhet.

A nyolcvanas évek első felében megjelentek az első személyi számítógépek, és hamarosan a számítógépes nyelvészet első piaci alkalmazásai is: a helyesírás-ellenőrző és az elválasztóprogramok (először Macintosh gépekre, majd IBM PC-re is). Nem sokkal később a gépi fordítás is megpróbált „leszállni" a személyi számítógépekre: kijött a PC Logos, majd a Siemens által megvásárolt Metal rendszer a szótárakat kiadó Langenscheidté lesz, és T1 néven – a sokak által jól ismert jellegzetes Langenscheidt-szótárak borítójához hasonló papírdobozban – a boltok kirakatába került. A Systrannak is kijött a PC-s változata, és létrejöttek az első, kimondottan a PC-s környezethez igazított képességű fordítórendszerek, mint pl. a finn Kielikone vagy az orosz ProMT. Magyarországon a nyolcvanas évek végén újra indult a számítógépes nyelvészet: megjelent az első magyar nyelvű összefoglaló az addigi eredményekről (Prószéky, 1989), majd 1991-ben létrejön az először csak nyelvhelyességi eszközöket, majd gépi fordító modulokat is létrehozó MorphoLogic cég (Mikolás, 2001).

3. A sok szöveg egyre jobbat tesz a gépi fordításnak

Miközben a PC-k hozták az első eladható gépi nyelvészeti megoldásokat, a tudomány újat lépett: beköszönt az internet és ezzel a számítógéppel távolról elérhető anyagok világa. Ráadásul egyre több anyag került ebben az időben már számítógépre, és előbb-utóbb a világhálóra is. A géppel feldolgozható szövegeknek egyfajta példatárként való használata mentén a nyelvtudománynak egy új, empirikus ága alakult ki: a korpusznyelvészet (részletesebben ld. McEnery és Hardie, 2013). Magának a

korpuszelméleti közelítésnek a gyökerei egyébként még a 19. század második felére mennek vissza, ahonnan még nagyon messze volt a számítógép. Az említett gondolatcsírák a kor egyik legnevesebb magyar nyelvészéhez, Simonyi Zsigmondhoz köthetők, akinek kis nyelvtanáról ezt olvashatjuk:

> „Simonyi új grammatikai módszert akar behozni, könyve inductive halad, azaz a példákból kiindulva tanítja a szabályt, nem pedig dogmatice. A grammaticát tehát valami olvasmány alapján akarja előadni, úgy hogy a szabályokat a tanár tanítványai közreműködésével vonhatja le ésszerű következtetések útján. Ilyenképp tehát ezen módszer véget vet a lelketlen magolásnak, és azt észfejlesztő inductióval pótolja. Eszerint a szabályok is mélyebben vésődnek be a gyermek emlékezetébe, mert amit magunk találunk, azt jobban tudjuk, mint amit más mond, vagy más tanultat velünk" (Riedl, 1882).

Erre az idézetre egyébként Sass Bálint, a Nyelvtudományi Intézet nyelvtechnológus kutatója, egykori doktoranduszom hívta fel a figyelmet, aminek lényegét mai világunkban úgy mondanánk, hogy egy új grammatikai módszer van megjelenőben, mely induktív módon halad, azaz a példákból kiindulva ismeri fel a szabályt. A grammatikát tehát az elolvasott, feldolgozott szövegek alapján építjük, úgy hogy a szabályokat a gép a példák segítségével állítja össze statisztikai következtetések útján. Ezáltal ez a módszer véget vet az előre megadott szabályok mechanikus alkalmazásának, és azt indukcióval pótolja. A szabályok így tárolódnak el a gép memóriájában, mert „amit magunk találunk, azt jobban tudjuk, mint amit más mond, vagy más tanultat velünk".

Ahol pedig megjelenik a mennyiség, ott megjelennek a valószínűségszámítási módszerek is. Így történt, hogy a kilencvenes években a statisztika „beszállt" a nyelvi modellezésbe is. A szövegek statisztikai feldolgozása ettől kezdve az IBM által kidolgozott algoritmusok alapján (Jelinek, 1997) elsősorban a beszédtechnológiából jól ismert zajoscsatorna-módszerrel történt. Ez olyan sikeresnek bizonyult, hogy rövid idő alatt kialakult a statisztikai módszerek nyelvészeti alkalmazásainak a világa. Ebben az időben jelent meg a világpiacon a belga Lernout és Hauspie, az akkoriban sikertörténetének a csúcsán járó PC-s hangkártya, a SoundBlaster két kifejlesztője. Cégük, az L&H a beszédtechnológia, sőt, a mesterséges intelligencia és a nyelvfeldolgozás rövid távú világméretű térhódítását prognosztizálta, és külső tőketámogatással elkezdték

felépíteni a terveik szerint az egész földgolyót átszövő technológiai hálózatukat, melyet SAIL-nek (= Speech, Artificial Intelligence, Language) kereszteltek el. A tervezett központok, az ún. kikötők, azaz „SAIL-portok" között még Budapest is szerepelt mint lehetséges kelet-európai központ, de az akkori magyar kormány idejében észlelte a szakmai figyelmeztetéseket, és végül nem állt be a SAIL rendszert anyagilag is támogató államok közé. A beszédfeldolgozás és a gépi fordítás L&H által ígért eredményei ugyan nagyon kecsegtetőek voltak, de az igazi és álüzletemberek hada komoly etikai, aztán jogi, majd anyagi nehézségekbe hozta az L&H vállalkozást, végül a börtönbe csukott két vállalkozó által összevásárolt nyelvtechnológiai és gépi fordító cégek hatalmas elegyét a ScanSoft, majd tőle a hazánkban a valahai Recognita karakterfelismerő cég mai tulajdonosaként ismert Nuance vásárolta meg. Érdembeli fejlesztés valójában nem sok történt az L&H környékén, de az események figyelmeztetésként hatottak sok, még éppen csak induló nyelvtechnológiai vállalkozás és az őket támogatók számára. Pozitív hozadéka volt az időszaknak, hogy a belga cég megjelenése a magyar politika legfelsőbb köreiben felhívta a figyelmet ennek az addig egyáltalán nem támogatott K+F terület létezésére. A 2000-es évek elejétől tehát hazánkban is megindultak a már központi forrásokból is támogatott nyelv- és beszédtechnológiai kutatások, és az addigra a MorphoLogic cég által kifejlesztett, angolról magyarra fordító MetaMorpho rendszer (Prószéky és Tihanyi, 2002) magyar–angol modulja már így jöhetett létre (Novák et al., 2008).

Ez volt az az időszak, amikor a világban kialakult a „human language technologies", azaz a *nyelvtechnológia* fogalma. Az IBM ezekben az években – átérezve az új kor üzenetét – komoly mesterségesintelligencia- és nyelvtechnológiai „erődemonstrációkat" tartott. Az első, a Deep Blue rendszerről szóló ugyan nem nyelvi megoldásokat, hanem a sakkozást népszerűsítette, de olyan szinten, hogy rendszerük megverte a regnáló sakkvilágbajnokot, Gari Kaszparovot (IBM, 1997). Ezzel a mesterségesintelligencia-technológiák bemutatták, hogy az alapismeretek (ez esetben a sakkfigurák lépéseinek szabályai) az eredmények szempontjából ugyan fontosak, de nem elsődlegesek, hiszen ezeket eddig is tudták a sakkprogramok, ezzel szemben rengeteg játszmát kell megfelelően elemezni és feldolgozni, mert akkor a program a sok-sok nemzetközi nagymester együttes tudásával le tud győzni gyakorlatilag akárkit, aki még ha nagyon okos is, de végül is csak egyetlen ember. A gépi fordításra alkal-

mazva ez a logika valahogy így hangzik: ha a nyelv mondatépítő szabályait ismerjük, az ugyan fontos, de ami igazán szükséges, az a rengeteg olyan minta, amit már emberek bizonyos szövegek fordításaként korábban létrehoztak. Ha a sok elérhető fordítást megtanítjuk a rendszernek, akkor a sakkprogramhoz hasonlóan fordítók ezreinek a tudását fogja tudni egyidejűleg alkalmazni (természetesen valamilyen statisztikai formában) egy adott, még le nem fordított szöveg célnyelvi megfelelőjének létrehozásához. A gépi fordításban ráadásul nem is valaki ellen kell használni ezt a tudást, mint a sakkban, hanem mindannyiunk javára. A gépi fordítás ezektől a matematikailag kifogástalan megoldásoktól tehát szárnyakra kapott, mindössze a kiinduló anyag mennyisége és minősége volt az, ami a géppel fordítandó szöveg más nyelven történő megfogalmazásának használhatóságát befolyásolta. Az új évezred első évtizedének a az IBM újabb, immár nyelvi csodarendszerként beharangozott alkalmazással állt elő, melyet a cég egyik legbefolyásosabb elnökéről, Thomas J. Watsonról neveztek el. A Watson rendszer ugyan nem a fordításban jeleskedett, hanem azt a tudást, amit az ezzel foglalkozó kutatók a rendszer számára elérhetővé tettek, viszonylag bonyolult kérdések megválaszolására használta fel. Ezt a tevékenységet természetesen fel lehet fogni úgy is, hogy a bemenő nyelvi adatot a belső keresőrendszer „nyelvére" kellett lefordítania. A rendszer demonstrációján egy népszerű kvízjáték győzteseit verte meg a televíziónézők millióinak szeme láttára (IBM, 2011).

4. A neurális hálók megjelennek a gépi fordítás területén is

Ezzel a *mesterséges intelligencia* fogalma ismét előtérbe került a nyelvfeldolgozással kapcsolatban. Nem sokkal ezután jött el az a pillanat, amikor a *mélytanulás* és a *neurális hálós módszerek* újra mesterséges intelligencia néven maguk alá gyűrték az addig kétkedő világot. Egy brnói hallgató PhD-disszertációjában kidolgozott egy olyan módszert, a *szóbeágyazást* (Mikolov, 2013), amellyel a nyelv szavait vektorokként tudta reprezentálni, méghozzá úgy, hogy a jelentésükben hasonló szavak a vektortérben közel kerültek egymáshoz, a távoliak pedig messze. Mindehhez semmilyen nyelven kívüli információt nem használt fel, mindössze a szavak különböző mondatokban talált előfordulásainak szókörnyezetét. Mivel megnyilatkozásainkban a szavak mindig mondatokban, nagyobb szövegegységekben fordulnak elő, és csak ott jelentik azt, amit, ha két szó környezete sokszor hasonló, akkor nagy eséllyel az adott szavaknak is hasonlítaniuk kell egymásra. Ez egy régóta ismert alapgondolat, hiszen tudományos megfogalmazásában ez eddig is valahogy úgy

hangzott, hogy a jel jelentése a jel használati szabálya (részletesebben ld. Wittgenstein, 1953). A jel itt a szó, és a használati szabályt a környező szavak közötti előfordulás jelenti. Mindössze az a különbség, hogy az eddigi meglehetősen absztrakt megfogalmazás helyett most Mikolov egy egzakt matematikai módszert mutatott, az ezt megvalósító programmal együtt. Ez a program a neurális hálók egyik első alkalmazása volt a nyelvtechnológiában, és alapvetően megváltoztatta a számítógépes nyelvészet világát. Az ilyen vektoros reprezentáción alapuló gépi fordító rendszerek nem a szavak betűalakját, hanem valójában ezeket a szemantikus térben megjelenő „jelentéscsomókat" fordítja, következésképp egy kicsit úgy tud viselkedni, mintha „értené" a szöveget, és nemcsak a betűit olvasná. Ez a közelítés a gépi fordítás azonnali minőségi javulását hozta. Például a gépi úton eddig nehezebben fordítható nyelvpárok minőségi ugrást mutattak, és közel kerültek azokhoz a nyelvpárokhoz, melyeket már korábban is sikeresen fordítottak a gépek. Örömünkre a magyart (és az EU más eddig nehezen kezelhető nyelveit, mint pl. a finnt vagy az észtet) tartalmazó nyelvpárok egyre használhatóbb minőségű fordításkat produkáltak. Ami viszont mind a korábbi statisztikai, mind ezeket a neurális fordítórendszereket illeti, van egy igen fontos probléma, ami a tanítóadatok mennyiségéből és minőségéből következik. Igen jelentős mennyiségű szöveg – ún. bitext, tehát forrásmondat-célmondat párokból álló kétnyelvű szövegkorpusz – szükséges a jó fordításhoz, viszont egy szűk szakterületnek még ha az összes valaha készített fordítását fel is tudnánk használni tanítóanyagnak, sokszor az is kevés a jó minőségű gépi fordításhoz. Ugyanez a probléma áll fenn azoknak a nyelvpároknak az esetében is, amelyeken az összes eddig készült fordítás együtt sem volna elég tanítóanyagnak. Gondoljuk el például a magyar–máltai gépi fordító rendszert, aminek a számára ha minden eddigi ember készített fordítást össze is szedünk, nem kapnánk megfelelő minőségű statisztikai/neurális gépi fordítást a gépi tanuláshoz a kis mennyiségűnek számító tanítóanyag miatt. Az, hogy bizonyos típusú fordítások (tehát ritka nyelvpárok, vagy gyakoribb nyelvpárok kevés fordítási mintával rendelkező szakterületei) esetében nincs megfelelő mennyiségű kiinduló anyag, a szakma „sparse data problem"-nak nevezi. Tehát mind a matematikai alapok, mind az informatikai megoldások elvileg tökéletesek, ám a nehézséget a gyakorlatban a nyelvi anyag hiánya vagy nem megfelelő minősége adja.

Ha nagyok a tanítókorpuszok, akkor viszont valószínűleg nagyon heterogének, mert mindenféle szövegtípus előfordul bennük (gondoljunk

csak az interneten fellelhető szövegek sokféleségére), így egy-egy kifejezésnek több lehetséges fordítása is előfordul bennük a különböző környezetekben. Hogy ezeket a lehetséges többértelműségeket szétválaszszuk egymástól, jó volna homogenizálni a korpuszokat, azaz szűkebb tematikus egységekre, doménekre bontani. Ezeken belül ugyanis már jóval kisebb lesz az egyes szavak többértelműsége, ám így a kiinduló korpusz mérete is kisebb lesz, ami egyfajta 22-es csapdájaként az említett „sparse data problem"-hoz vezethet. Előáll tehát a statisztikai/neurális rendszereknek egy nehezen feloldható kettőssége: ha kicsi a szövegkorpusz, bár a tanítóminta ilyenkor nagyrészt egyértelmű szavakat tartalmaz, sokszor nem lesz jó az erre épülő fordítás az egyes kifejezések relatíve kis előfordulási száma miatt. Ha növeljük a korpusz méretét, óhatatlanul megjelenik a többértelműség okozta „fordítási zaj", bár a korpusz mérete már más szempontból megfelelőnek tűnhet.

Egy másik nagy probléma napjaink neurális gépi fordításában, hogy az informatikai kutatóközpontokban ugyan készülnek nyelv(pár)független modellek, ám ezek minősége meg sem közelíti a nyelv(pár)specifikus modellekét. Nyilván nem minden kutatóhely rendelkezik minden nyelvre megfelelő mennyiségű olyan tanítóanyaggal, amiből jó minőségű fordítás volna várható. Ráadásul a neurális rendszerek nyelvmérnökei elsősorban nem a fordításban jók, de még csak nem is abban, hogy előkészítik a nyelvi anyagokat a programrendszerek számára, hanem abban, hogy a neurális megoldáshoz szükséges felfoghatatlan mennyiségű paramétert úgy állítják be, hogy a fordítóprogram jó minőségű eredményt adjon. A paraméterbeállítások mikéntje viszont jelenlegi tudásunk szerint nehezen hozható közvetlen logikai kapcsolatba az eredménnyel, tehát a gépi nyelvészet világában mindig is jelen lévő intuíciónak még jobban felértékelődik a szerepe a mai gépi fordító rendszerek létrehozásánál. Ha egy nagyobb cégnél sok intuitív ember jön össze, és ezeken a helyeken a géppark lehetőségei is komoly sebességelőnyt mutatnak egy kisvállalkozás gépeivel szemben, hamar megérthetjük, hogy ugyanolyan intuitív emberek kisebb kapacitású gépekkel nagyságrendekkel kevesebb kísérletet tudnak végezni ugyanannyi idő alatt a paraméterbeállítások világában, mint nagy céges társaik. Tehát a gépi fordítás területén a verseny ma elsősorban nem a nyelvi vagy programozási tudáson múlik, hanem a kísérletezésen, amelyben a gyorsabb környezet előbb jelzi vissza egy-egy kísérlet eredménytelenségét, mint a lassabbé. És ha mindezt több tízszer vagy százszor annyi kísérletező ember végzi, hamar belátható, hogy né-

hány világcég jelentős előnyt tud szerezni a mai gépi fordítási versenyben, mint bármikor korábban. Más szavakkal: egyre jobban nyílik az olló a kis és a nagy gépi fordító intézmények között. Egy dolog ugyanakkor egyre jobban látszik: az általános modellek általában nem elegendőek egy adott nyelvi közösség számára, hiszen az ilyen modellek azért készülnek, hogy relatíve kevés munkával lehessen összehasonlítható eredményeket felmutatni a bármely nyelvről bármely nyelvre való fordítás világában. Akiknek viszont az általános nyelvtechnológiai eszközök világában egy konkrét nyelvre, vagy a gépi fordítás esetében egy-egy konkrét nyelvpárra kell egyre jobb eredmény, azoknak a saját tanítóadataik egyre jobb minőségén és egyre nagyobb mennyiségén kell dolgozniuk, még ha az ezeket feldolgozó szoftverek mindössze néhány világcég műhelyéből jönnek is elő. És ebből következően talán mindenki számára érthető, hogy a neurális megoldások világában is van értelme támogatni a magyar nyelvtechnológiai fejlesztéseket, és ezáltal a magyarról és a magyarra történő gépi fordítást is, mert helyettünk ezt mások nem fogják jó minőségben megcsinálni.

5. A gépi fordítás és Váradi Tamás találkozásai

Ami Váradi Tamásnak a gépi fordítás világához való szakmai hozzájárulását illeti, személyes tapasztalataim is vannak, mert korábban több alkalommal is összetalálkoztunk a fenti kutatások egyikét-másikát megvalósító projektekben mint partnerek. Akkor még javában külön intézményekben dolgoztunk: Tamás az MTA Nyelvtudományi Intézet, én pedig a MorphoLogic képviseletében vettem részt gépi fordítás témájú munkálatokban. Az utóbbi négy évben azonban már intézményesen is egy hajóban ülünk, és – amint ezt mindjárt prognosztizálom is – könnyen elképzelhető egy közelgő, immár belső, újabb együttműködés is. Az első magyarországi gépi fordítási projekt egyébként a 2000 és 2002 között futó MATCHPAD (Machine Translation Systems for the Use of Hungarian and Polish Administrations) volt, amikor a korábbiakban már említett Systran rendszer magyarra és lengyelre való alkalmazhatóságának bizonyítása volt a cél. Ennek a kísérleti kutatásnak az idején már bontogatta szárnyait Tihanyi László szakmai vezetésével MorphoLogic cég MetaMorpho nevű gépi fordító rendszere (Novák et al., 2008), amelynek első publikus bemutatója 2001-ben, a MorphoLogic tizedik születésnapján volt. Néhány év múlva, miután a MetaMorpho angol–magyar modulja teljesen elkészült, a MorphoLogic megcélozta a magyar–angol változatot is, amihez immár külső partnerek is csatlakoztak: az MTA NYTI

és a SZTE (Tihanyi, 2007). Így újra együttműködhettünk egy Tamás vezette csapattal, amihez később még hozzákapcsolódott a MorphoLogic webforditas.hu portáljának alapgondolatából kinövő, több európai gépi-fordítás-szolgáltatót egy nagy nemzetközi fordítórendszerré összekapcsoló iTranslate4.eu projekt Tamás általi menedzselése is. Legutóbbi összetalálkozásunk területe, amit csak futtában említettem néhány sorral feljebb, a mai nyelvtechnológia javarészt mélytanulásos technológiákon alapuló megközelítése, vagyis amit ma a sajtó – némiképp összemosva a részleteket – mesterségesintelligencia-alapúnak nevez. Ebben, különösen a legújabb transzformer rendszerek létrehozásában Tamás már rövid idő alatt is sok eredményt ért el, így nem lehetetlen, hogy előbb-utóbb a vezetésével elkészült neurális modellek a gépi fordítás területén is bizonyíthatnak. Mindehhez jó tudni, hogy a saját intézeti eredményeinken túl azokból a szakmai közösségekből, melyek együttműködéséről az előbb szóltam, a MorphoLogic MetaMorpho projektjét vezető Tihanyi László és Tamás kutatócsoportjának egyik korábbi oszlopa, Oravecz Csaba ma az Európai Bizottság Fordítási Főigazgatóságának elismert kutatói, akik ezen a területen elért eredményeikkel, azaz neurális gépi fordítási megoldásaikkal bevezették a magyar nyelvet a világ fordítóprogramjainak használható forrás- és célnyelvei közé.

Egy kutatónak, aki régóta vezethet másokat, kétféle nagyszerű szakmai eredmény létezik: ha maga ér el új eredményeket, ahogy Tamás a legújabb típusú nyelvmodellek, a magyar BERT-Large létrehozásában (Feldmann et al., 2021), valamint ha az általa hosszú ideig menedzselt kutatócsapat valamely tagja maga is elér ilyeneket (Tihanyi és Oravecz, 2017). Kedves Tamás, kívánom, hogy mindkettőből a továbbiakban is sok jusson Neked!

Bibliográfia

Bar-Hillel, Y.: The Present State of Research on Mechanical Translation. American Documentation 2/4, 229–237 (1951)

Chomsky, N.: Syntactic Structures. Mouton (1957)

Chomsky, N.: Aspects of the Theory of Syntax. MIT Press (1965)

Chomsky, N.: Lectures on Government and Binding. Foris (1981)

Chomsky, N.: A Minimalist Program for Linguistic Theory. MIT Occasional Papers in Linguistics No. 1. Cambridge: MIT Press (1993)

Feldmann Á., Hajdu R., Indig B., Sass B., Makrai M., Mittelholcz I., Halász D., Yang Zijian Gy., Váradi T.: HILBERT, magyar nyelvű BERT-large modell tanítása felhő környezetben. In: Berend G., Gosztolya G., Vincze V.(szerk.) XVII. Magyar Számítógépes Nyelvészeti Konferencia. 29–36. Szegedi Tudományegyetem TTIK, Szeged (2021)
Frege, G.: Logische Untersuchungen. Dritter Teil: Gedankenfuge. In: Beiträge zur Philosophie des Deutschen Idealismus, Vol. III. pp. 36–51 (1923)
Harris, Z. S.: Methods in Structural Linguistics. University of Chicago Press, Chicago (1951)
Hutchins, J.: From First Conception to First Demonstration: the Nascent Years of Machine Translation, 1947–1954. In: A Chronology. Machine Translation 12/3, 195–252 (1997)
701 Translator, IBM Press Release, January 8, 1954. (1954)
Deep Blue Accepts Challenge to Compete in Ultimate Chess Match with Human Champ Kasparov. IBM Press Release, May 30, 1995. (1995)
Jeopardy! And IBM Announce Charities to Benefit from Watson Competition. IBM Press Release, Jan 13, 2011. (2011)
Jelinek, F.: Statistical Methods for Speech Recognition. MIT Press, Cambridge (1997)
Koskenniemi, K.: Two-level Morphology: A General Computational Model for Word-Form Recognition and Production. Publications, No. 11, University of Helsinki, Department of General Linguistics, Helsinki (1983)
McEnery, T.; Hardie, A.: Corpus Linguistics: Method, Theory and Practice. Cambridge University Press, Cambridge (2012)
Mikolás Z. (szerk.) MetaMorpho: A MorphoLogic tíz éve. Budapest: MorphoLogic (2001)
Mikolov, T.: Statistical Language Models Based On Neural Networks. Ph.D. Thesis: Masaryk University, Brno (2013)
Novák A., Tihanyi L., Prószéky G.: The MetaMorpho Translation System. In: Callison-Burch, C., Koehn, P., Monz, C., Schroeder, J. Shaw Fordyce, C. (eds.) Proceedings of the Third Workshop on Statistical Machine Translation at ACL. pp. 111–114. Association for Computational Linguistics, Stroudsburg, PA (2008)
Papp F.: Matematikai nyelvészet és gépi fordítás a Szovjetunióban. OMKDK, Budapest (1964)
Pierce, J. R., Carroll, J. B. et al.: Language and Machines – Computers in Translation and Linguistics. ALPAC Report, National Academy of Sciences, National Research Council, Washington, DC (1966)
Prószéky G.: Számítógépes nyelvészet: Természetes nyelvek használata számítógépes rendszerekben. Számalk, Budapest (1989)
Prószéky G., Tihanyi L.: MetaMorpho: A Pattern-Based Machine Translation System. In: Proceedings of the 24th 'Translating and the Computer' Conference. pp. 19–24. ASLIB, London, United Kingdom (2002)
Prószéky G.: A magyar számítógépes nyelvészet történeti áttekintése. In: Prószéky G., Váradi T. (szerk.) Általános Nyelvészeti Tanulmányok XXIV: Nyelvtechnológiai kutatások. pp. 17–45. Akadémiai Kiadó, Budapest (2012)
Riedl F.: Simonyi kis nyelvtana. Egyetemes Philologiai Közlöny 6/6, 573–590 (1882)
Simonyi Zs.: Kis magyar nyelvtan mondattani alapon. Negyedik átdolgozott s gyakorlatokkal bővített kiadás egy kötetben (1882)

Sells, Peter. Lectures on Contemporary Syntax Theories: An Introduction to Government-Binding Theory, Generalized Phrase Structure Grammar, and Lexical-Functional Grammar. CSLI, Stanford (1985)
Senellart, J., Dienes, P., Váradi, T.: New Generation Systran Translation System. In: Senellart, J., Yang, J., Rebollo, A. (eds.) Proceedings of MT Summit VIII. Santiago de Compostela, Spain (2001)
Tesnière, Lucien: Éléments de syntaxe structurale. Libraire C. Klincksieck, Paris (1959)
Tihanyi L.: A MetaMorpho projekt 2007-ben – a sorozat vége. In: Tanács A., Csendes D.(szerk.) V. Magyar Számítógépes Nyelvészeti Konferencia. pp. 179–186. SZTE, Szeged (2007)
Tihanyi L., Oravecz Cs.: First Experiments And Results in English-Hungarian Neural Machine Translation. In: Vincze V. (szerk.) XIII. Magyar Számítógépes Nyelvészeti Konferencia. pp. 275–286. SZTE, Szeged (2017)
Winograd, T.: Understanding Natural Language. Cognitive Psychology 3/1, 191 (1972)
Winograd, T.: Language as a Cognitive Process. Vol. 1. Syntax. Reading: Addison-Wesley (1983)
Wittgenstein, L.: Philosophical Investigations. Blackwell, Oxford (1953). [Magyar fordítás: Filozófiai vizsgálódások. Atlantisz, Budapest (1992), ford.: Neumer Katalin]
Woods, W. A.: Transition Network Grammars for Natural Language Analysis. Communications of the ACM 13/10, 591–606 (1970),
Woods, W. A.: Progress in Natural Language Understanding: An Application to LUNAR Geology. Proceedings of the National Computer Conference AFIPS pp. 441–450 (1973)