www.meta-net.eu
office@meta-net.eu
Tel: +49 30 3949 1833
Fax: +49 30 3949 1810

META-NET
Fehér könyvek sorozat

Nyelvek az európai információs társadalomban

Magyar

Ez a fehér könyv egy sorozat részét képezi, amelynek célja, hogy felhívja a figyelmet a nyelvtechnológiára és az abban rejlő lehetőségekre. Elsősorban oktatókat, újságírókat, politikusokat és nyelvi közösségeket szólít meg.

Az európai nyelvek nyelvtechnológiai feldolgozottsága és a nyelvtechnológia elterjedtsége meglehetősen eltérő. Ennek következményeként a nyelvtechnológia fejlődéséhez és a kutatás elősegítéséhez szükséges lépések szintén nyelvenként mások és mások, és olyan különböző tényezőkön múlnak, mint például az adott nyelv összetettsége, vagy a nyelvet használó közösség nagysága.

A META-NET, az Európai Közösség által alapított hálózat felmérést végzett a jelenleg rendelkezésünkre álló nyelvi erőforrásokról és technológiákról. Ez a felmérés a 23 hivatalos európai nyelv mellett egyéb nemzeti és regionális nyelvekre is kiterjed, és eredményei rámutatnak az egyes nyelvek terén fellelhető kutatási hiányosságokra. Egy a jelenlegi helyzetet bemutató részletes szakértői elemzés és értékelés segíthet maximalizálni a további kutatások hatását és minimalizálni a kockázatot.

A META-NET 31 ország 47 kutatóközpontjából áll, akik a területtel foglalkozó vállalkozásokkal, kormányzati szervekkel, kutatószervezetekkel, szoftvercégekkel, szolgáltatókkal és európai egyetemekkel dolgoznak együtt. Egységes technológiai víziót alkotva egy stratégiai kutatási terv létrehozásán dolgoznak, amelyben megfogalmazzák, hogyan tudnak a nyelvtechnológiai alkalmazások segíteni a kutatási hiányosságokon 2020-ig.

Szerző

Simon Eszter, Nyelvtudományi Intézet, Magyar Tudományos Akadémia

Szerkesztő

Lendvai Piroska, Nyelvtudományi Intézet, Magyar Tudományos Akadémia

Köszönetnyilvánítás

A kiadó hálás a német fehér könyv szerzőinek, hogy rendelkezésükre bocsátották saját anyagaikat.

Tartalomjegyzék

Vezetői összefoglaló

Számos európai nyelv szembesül a digitális világ jelentette kockázattal, mivel kevés erőforrással rendelkeznek, és alulreprezentáltak az interneten. A nyelvi korlátok miatt óriási piaci lehetőségek maradnak kihasználatlanul. Ha nem teszünk lépéseket most, sok európai polgár kerül társadalmilag és gazdaságilag hátrányos helyzetbe az anyanyelve miatt.

Az innovatív nyelvtechnológia közvetítőként működik, amely segít az európai polgároknak abban, hogy egy egyenlő, befogadó és gazdaságilag sikeres tudás- és információs társadalom részesei lehessenek. A többnyelvű nyelvtechnológia a nyelvi korlátok ledöntésével utat nyit a gyors, olcsó és egyszerű kommunikációnak.

Napjainkban a nyelvi szolgáltatások elsődlegesen amerikai szolgáltatókon keresztül érhetők el. A Google Translate ingyenes szolgáltatása csak egy példa a sok közül. Watson, az IBM számítógépes rendszere minden játékost maga mögé utasított a Jeopardy! nevű amerikai műveltségi vetélkedőn. Watson emberekkel szemben aratott sikere is mutatja a nyelvtechnológiában rejlő hatalmas lehetőségeket. Európaiként fel kell tennünk magunknak néhány égető kérdést:

- Hagyhatjuk-e, hogy a kommunikációnk és tudásinfrastruktúránk monopólium helyzetben levő vállalatoktól függjön?
- Támaszkodhatunk-e olyan nyelvi szolgáltatásokra, amelyek tőlünk függetlenül bármikor megszűnhetnek?
- Aktívan részt veszünk-e a globális piacon folyó nyelvtechnológiai kutatás és fejlesztés versenyében?
- Vajon meg fogja-e oldani egy másik kontinensről érkező harmadik fél az európai többnyelvűséget érintő fordítási és egyéb problémákat?
- Európai kulturális hátterünkből kiindulva vajon létre tudunk-e hozni egy tudásalapú társadalmat egy jobb, biztonságosabb, pontosabb, innovatívabb és robusztusabb csúcstechnológia segítségével?

Ez a magyar nyelvről szóló fehér könyv bemutatja, hogy Magyarországon aktív nyelvtechnológiai ágazat és kutatói hálózat létezik. Bár számos technológia és erőforrás áll rendelkezésre a magyar nyelvre, ez még mindig kevesebb és gyengébb minőségű, mint az angol nyelv esetében.

A jelentéshez kapcsolódó értékelésből kiderül, hogy sürgősen lépéseket kell tenni annak érdekében, hogy áttörést érjünk el a magyar nyelv területén is.

A META-NET hozzájárul egy erős, többnyelvű európai digitális információs tér kialakításához. Az országok között így létrejövő multikulturális unió egy békés és egyenlő nemzetközi együttműködés alapja lehet. Ha azonban ez nem sikerül, Európának választania kell kulturális sokszínűségének feláldozása és a gazdasági vereség között.

Bevezetés: Kockázat a nyelveinknek – kihívás a nyelvtechnológiának

Digitális forradalom szemtanúi vagyunk, amely drámaian befolyásolja a kommunikációt és a társadalmat.

Digitális forradalom szemtanúi vagyunk, amely drámaian befolyásolja a kommunikációt és a társadalmat. A digitális és hálózati kommunikációs technológia legújabb vívmányait sokszor Gutenberg nyomdájának feltalálásához hasonlítják. Mit mutat meg nekünk ez az analógia az európai információs társadalom és főleg nyelveink jövőjéről?

Gutenberg találmánya után a kommunikációban és tudáscserében a következő nagy áttörést Luther Biblia-fordítása jelentette. A következő századokban a különböző technikák fejlődése segítette a hatékonyabb nyelvi feldolgozást és tudáscserét:

- a fő nyelvek helyesírási és nyelvtani sztenderdizálása lehetővé tette az új tudományos és intellektuális ötletek gyors elterjesztését;
- a hivatalos nyelvek kialakulása lehetővé tette a polgárok számára a (gyakran politikai) határokon átívelő kommunikációt;
- a nyelvtanítás és fordítás lehetővé tette a nyelvek közötti cserét;
- az újságírói és bibliográfiai útmutatók biztosították a nyomtatott anyagok minőségét és elérhetőségét;
- a létrejövő médiumtípusok, mint az újság, a könyvkiadás, a rádió és a televízió különböző kommunikációs igényeket tudtak kielégíteni.

Az elmúlt húsz évben az információs technológia elősegítette számos folyamat automatizálását és könnyebb használatát:

- a kiadványszerkesztő szoftver felváltja a gépírást és formázást;
- a Microsoft PowerPoint szoftver felváltja az írásvetítő fóliákat;
- e-mailben gyorsabban küldünk és fogadunk dokumentumokat, mint faxszal;
- a Skype segítségével interneten keresztül telefonálhatunk és virtuális találkozókat szervezhetünk;
- a hang- és videókódolási formátumok segítségével könnyen cserélhetünk multimédia tartalmakat;
- a keresőprogramok kulcsszavas keresést biztosítanak;
- az online fordítóprogramok, mint a Google Translate gyors nyersfordítást adnak;
- a közösségi médiaplatformok elősegítik az együttműködést és az információmegosztást.

Bár ezek az eszközök és alkalmazások segítséget jelentenek, jelenleg még nem tudnak egy fenntartható, többnyelvű európai információs társadalmat kialakítani, ahol az információ és a javak szabadon áramolhatnak.

Az európai információs társadalom gátjai: a nyelvi határok

Nem tudjuk pontosan, hogyan fog kinézni a jövőbeli információs társadalom. Elképzelhető, hogy az európai külügyminiszterek saját nyelvükön fognak beszélni a közös európai energetikai és külpolitikai kérdésekről. Lehet, hogy lesz egy olyan platform, ahol különböző nyelveken különböző szinten beszélő emberek vitathatnak meg egy témát úgy, hogy közben a technológia automatikusan összegyűjti a véleményeket, és rövid összefoglalókat készít. Az is lehet, hogy minden nehézség nélkül tudunk egy külföldi egészségbiztosító társasággal beszélni.

Nyilvánvaló, hogy a kommunikációs igényeink sokban különböznek a néhány évvel ezelőttitől. A globális információs és gazdasági térben több nyelvvel, kommunikációs partnerrel és tartalommal kerülünk kapcsolatba, és mindez arra késztet minket, hogy gyorsan hasznosítsuk az új média típusait. A közösségi média (Wikipedia, Facebook, Twitter és YouTube) jelenlegi népszerűsége csak a jéghegy csúcsa.

A globális információs és gazdasági térben több nyelvvel, kommunikációs partnerrel és tartalommal kerülünk kapcsolatba.

Manapság több gigabájtnyi szöveget tudunk továbbítani a világ körül pár másodpercen belül anélkül, hogy észrevennénk, hogy a szöveget nem is értjük. Az Európai Bizottság felkérésére készített legutóbbi jelentésből kiderül, hogy az európai internethasználók 57%-a nem a saját anyanyelvén vásárol árukat és szolgáltatásokat. (Az angol a leggyakoribb idegen nyelv a francia, a német és a spanyol előtt.) A felhasználók 55%-a olvas idegen nyelvű szöveget az interneten, míg csak 35%-uk használ más nyelvet e-mailek vagy más üzenetek írásához a weben. Pár évvel ezelőtt az angol volt a web lingua franca-ja – az interneten megtalálható tartalom nagy része angolul volt –, de a helyzet mostanra jelentősen megváltozott. A más nyelvű (különösen az arab és más ázsiai nyelvű) online tartalom mennyisége robbanásszerűen emelkedett.

A nyelvi határoknak köszönhető mindenütt fellelhető digitális megosztottság eddig meglepően kevés figyelmet kapott a nyilvános vitákban; most azonban egy nagyon sürgető kérdést vet fel: „Mely európai nyelvek fognak boldogulni és kitartani a tudásalapú információs társadalomban?”

Mely európai nyelvek fognak boldogulni és kitartani a tudásalapú információs társadalomban?

Veszélyben a nyelveink

A nyomtatott sajtó megjelenése páratlan mértékű információcserét indított el Európában, viszont egyben sok európai nyelv pusztulását is magával hozta. A regionális és kisebbségi nyelvek alig kerültek nyomtatásba. Ennek eredményeként sok nyelv, mint a dalmát vagy a kelta, csak beszélt formában élt tovább, és ez korlátozta további fejlődésüket és használatukat.

Európa legfontosabb és leggazdagabb kulturális értékei közé tartozik a térségben használt csaknem 60 nyelv. Európa nyelvi sokszínűsége is hozzájárul társadalmi sikeréhez. Míg a népszerű nyelvek, mint az angol vagy a spanyol biztosan megmaradnak a feltörekvő digitális társadalomban és a piacon, sok más európai nyelv eltűnik a digitális kommunikációból és az internetes társadalom látóköréből. Ez biztosan

Európa legfontosabb és leggazdagabb kulturális értékei közé tartozik nyelvi sokszínűsége.

nem járható út. Egyrészt elveszne egy stratégiai lehetőség, ami gyengítené Európa globális helyzetét. Másrészt az ehhez hasonló változások szemben állnak azzal az elképzeléssel, hogy az európai polgárok azonos mértékben vehetnek részt az őket érintő ügyekben, nyelvtől függetlenül. Egy többnyelvűségről szóló UNESCO beszámoló szerint a nyelvek az alapvető jogok, mint például a politikai önkifejezés, oktatás vagy a társadalomban való részvétel fontos közvetítői.

Nyelvtechnológia, egy kulcsfontosságú technológia

A múltban a beruházások főleg a nyelvoktatásra és fordításra fókuszáltak. Példaként: becslések alapján Európának 2008-ban 8,4 milliárd eurós fordító, tolmács, szoftverlokalizációs és honlapglobalizációs piaca volt, és mindehhez még évi 10%-os növekedést vártak. De ez a kapacitás még mindig nem elég ahhoz, hogy kielégítse a jelenlegi és a jövőbeli igényeket.

A nyelvtechnológia kulcsfontosságú technológia, amely megvédheti és támogathatja az európai nyelveket. A nyelvtechnológia lehetővé teszi az együttműködést, a tanulást, az üzletkötést, a tudásmegosztást és a társadalmi és politikai vitákban való részvételt számítástechnikai tudástól és nyelvi határoktól függetlenül. A nyelvtechnológia már ma is segíti mindennapi munkánkat, például az e-mail írást, az online keresést vagy akár a repülőjegy-foglalást. A nyelvtechnológiát hasznosítjuk, amikor:

A nyelvtechnológia lehetővé teszi az együttműködést, a tanulást, az üzletkötést, a tudásmegosztást és a társadalmi és politikai vitákban való részvételt nyelvi határoktól függetlenül.

- információt keresünk egy internetes keresővel;
- helyesírást és nyelvtant ellenőrzünk egy szövegszerkesztőben;
- termékajánlókat nézünk egy online boltban;
- egy navigációs rendszer szóbeli utasításait hallgatjuk;
- online szolgáltatással fordítunk weboldalakat.

A jelen kiadványban bemutatott nyelvtechnológiai fejlesztések a jövő innovatív alkalmazásainak fontos részét képezik. A nyelvtechnológia tipikusan nagyobb alkalmazásokban, például navigációs rendszerekben vagy keresőprogramokban jelenik meg. A fehér könyvek az egyes nyelvekhez kötődő legfontosabb technológiák jelenlegi helyzetét mutatják be.

A közeljövőben minden európai nyelvre elérhető, olcsó és nagyobb szoftverkörnyezetbe integrált nyelvtechnológiára van szükségünk. Az interaktív, multimédiás és többnyelvű internethasználat nyelvtechnológia nélkül nem képzelhető el.

A nyelvtechnológia lehetőségei

A nyelvtechnológia segítségével elérhetővé válik az automatikus fordítás és tartalom-előállítás, az információfeldolgozás és a tudásmenedzsment minden európai nyelven. Emellett elősegíti az intuitív nyelvalapú interfészek fejlesztését a háztartási elektronika, a gépészet, a járműgyártás, a számítástechnika és a robotika területén is. Bár már sok prototípus létezik, a kereskedelmi és ipari alkalmazások még mindig a fejlesztés kezdetleges fázisában vannak. A kutatásban és fejlesztésben elért eredmények lehetőségek egész tárházát nyitották meg. Például a gépi fordítás adott témákon belül kellő pontossággal

működik, a kísérleti alkalmazások pedig számos európai nyelven nyújtanak többnyelvű információ- és tudásmenedzsment szolgáltatásokat.

Nyelvi alkalmazásokat, hangvezérelt felhasználói interfészeket és dialógusrendszereket általában speciális területeken találunk, ám ezek gyakran korlátozott teljesítményt mutatnak. A kutatás aktív része a nyelvtechnológia katasztrófa sújtotta helyeken, mentési munkálatoknál való felhasználása. Ilyen magas kockázatú környezetekben a fordítás pontossága élet-halál kérdés lehet. Hasonlóan fontos a pontosság a nyelvtechnológia egészségügyben való felhasználásában is. Az intelligens robotok nyelvi képességeikkel életeket menthetnek.

Nagy piaci lehetőségek rejlenek a nyelvtechnológiának az oktatásba és a szórakoztatóiparba, például játékokba, oktatóprogramokba, szimulációs környezetekbe való integrálásában is. A mobilinformációs szolgáltatások, a számítógéppel támogatott nyelvtanulás, az e-learning, az önértékelő eszközök és a plágiumszűrő szoftverek csak kiragadott példák arra, hogy hány helyen játszik fontos szerepet a nyelvtechnológia. A közösségi oldalak, mint a Twitter vagy a Facebook népszerűsége szintén arra utal, hogy igény van a kifinomultabb nyelvtechnológiai alkalmazásokra, amelyek figyelemmel követik a bejegyzéseket, összegzik a vitákat, ajánlásokat tesznek, kiszűrik az érzelmi válaszokat, szerzői jogi szabálytalanságokat vagy visszaéléseket.

A nyelvtechnológia hatalmas lehetőséget jelent az Európai Unió számára mind gazdasági, mind kulturális szempontból. Európában törvényszerű a többnyelvűség; az európai cégek, szervezetek és iskolák multinacionálisak és sokfélék. Az EU polgárait azonban még ma is hátráltatják az Európai Közös Piac nyelvi határai. A nyelvtechnológia segíthet a nyelvi gátak ledöntésében, támogatva a szabad és nyitott nyelvhasználatot. Emellett az innovatív, többnyelvű nyelvtechnológia segít a nemzetközi partnerekkel és a többnyelvű szervezetekkel való kommunikációban is. A nyelvtechnológia hozzájárul a nemzetközi gazdasági lehetőségek sikeréhez is.

A többnyelvűség a szabály, nem a kivétel.

A nyelvtechnológia kihívásai

Bár a nyelvtechnológia nagy fejlődésen ment keresztül az utóbbi években, a termékinnováció és -fejlesztés még mindig túl lassú. Nem várhatunk tíz vagy húsz évet azokra a számottevő változásokra, amelyek elősegítik a kommunikáció és a termelékenység fejlődését a többnyelvű környezetünkben.

A termékinnováció és -fejlesztés még mindig túl lassú. Nem várhatunk tíz-húsz évet a számottevő változásokra.

A széles körben használt nyelvtechnológiai alkalmazások, mint például a szövegszerkesztők helyesírás-ellenőrzői, tipikusan egynyelvűek, és csak néhány nyelvre elérhetőek. A többnyelvű alkalmazások bizonyos szintű kifinomultságot igényelnek. Az olyan gépi fordítók, mint a Google Translate vagy a Bing Translator kitűnőek arra, hogy nyersfordítást adjanak a dokumentum tartalmáról. De az ilyen és ehhez hasonló online szolgáltatások és professzionális gépi fordító alkalmazások nem alkalmasak pontos fordításra. Mindannyian ismerünk vicces félrefordításokat, mint például a *Bush* vagy a *Kohl*

nevek szó szerinti lefordítása, amelyek mutatják, hogy a nyelvtechnológia számára még akadnak kihívások.

Nyelvelsajátítás

Ahhoz, hogy bemutassuk, hogyan birkóznak meg a számítógépek a nyelvvel, és miért olyan nehéz a nyelvelsajátítás, először egy kis kitekintést adunk arra, hogyan sajátítja el az ember az anyanyelvét, valamint idegen nyelveket, majd felvázoljuk, hogy hogyan működik a gépi fordítás. Látni fogjuk, hogy nem véletlen, hogy a nyelvtechnológia olyan közel áll a mesterséges intelligenciához.

Két különböző módon sajátíthatunk el egy nyelvet. Először a gyermek a környezetében folyó beszédet hallgatva tanul beszélni. A nyelvhasználók, vagyis a szülők, testvérek és más családtagok által használt konkrét nyelvi példák segítik a gyerekeket abban, hogy kétéves koruk körül kiejtsék első szavaikat és rövid mondataikat. Ez egy speciális, genetikailag adott nyelvi képességnek köszönhető, amely lehetővé teszi, hogy elsajátítsunk egy nyelvet.

A második nyelv elsajátítása már ennél sokkal nagyobb erőfeszítésbe kerül, ha a gyermek nem anyanyelvi társaságban van. Iskolás korban az idegen nyelv elsajátítása a nyelv nyelvtani szerkezetének, szókincsének és helyesírásának könyvekből és oktató anyagokból való megtanulásával zajlik, amelyek a nyelvtudást szabályokon, táblázatokon és példaszövegeken keresztül mutatják be. Egy idegen nyelv megtanulása sok erőfeszítést és időt igényel, és ez csak nehezedik az évek múlásával.

A nyelvtechnológiai rendszereknek is két fő típusát különítjük el, hasonlóan az emberi nyelvelsajátításhoz. A statisztikai megközelítést követő rendszerek a nyelvtudást nagy mennyiségű, egy- vagy többnyelvű (párhuzamos) szövegből nyerik. A gépi tanuló algoritmusok a nyelvi képességet modellezik, amely képes meghatározni, hogy a szavakat, rövid kifejezéseket és mondatokat hogyan használjuk egy adott nyelvben, illetve hogyan fordítjuk le őket egyik nyelvről egy másikra. A statisztikai módszerek hatalmas mennyiségű szöveget igényelnek; teljesítményük az elemzett szöveg mennyiségével növekszik. Nem ritka, hogy az ilyen rendszereket több millió mondaton tanítják. Ez az egyik oka annak, amiért a kereső programok szolgáltatói lehetőség szerint minél több írott anyagot akarnak összegyűjteni. A szövegszerkesztőkben található helyesírás-ellenőrzők, a webes keresők és gépi fordító szolgáltatások, mint a Google keresője és fordítója egyaránt statisztikai (adatvezérelt) megközelítésen alapulnak.

A nyelvtechnológiai rendszereknek is két fő típusát különítjük el, hasonlóan az emberi nyelvelsajátításhoz.

A nyelvtechnológia másik nagy típusát a szabályalapú rendszerek alkotják. Ebben az esetben nyelvészek, számítógépes nyelvészek és számítástechnikusok dolgozzák ki a nyelvtani szabályokat és építik meg a lexikont. Egy szabályalapú rendszer megalkotása nagyon idő- és munkaigényes feladat, emellett magasan kvalifikált szakembereket igényel. Vannak olyan szabályalapú gépi fordító rendszerek, amelyek több mint húsz éve folyamatos fejlesztés alatt állnak. A szabályalapú rendszerek előnyei közé tartozik viszont, hogy a szakértők jobban tudják irányítani a nyelvfeldolgozás folyamatát, vagyis könnyebben tudják javítani a szisztematikus hibákat, illetve vissza tudnak jelezni a

felhasználónak. Ez utóbbi abban az esetben lehet különösen hasznos, ha a szabályalapú rendszert nyelvtanulásra használják. Pénzügyi szempontból viszont a szabályalapú technológia csak a nagy nyelvekre kifizetődő.

A magyar nyelv az európai információs társadalomban

Általános tények

A magyar nyelv a legtöbb ember által beszélt nem indoeurópai nyelv Európában. A Magyar Köztársaság államnyelve, itt a 10 milliós lakosságnak kb. 97%-a magyar anyanyelvű. A szomszédos hét országban is találunk magyar nyelvű közösségeket, amelyek közül a legnagyobb a romániai diaszpóra, megközelítőleg másfél millió nyelvhasználóval. Becslések szerint a magyar nyelvet összesen 13 millióan beszélik, ezzel a 12. a legtöbb beszélővel rendelkező nyelvek listáján Európában. A magyar nyelv hivatalos nyelv még a Vajdaságban, továbbá három szlovéniai községben. Regionális vagy kisebbségi nyelvként beszélik még Ausztriában, Horvátországban, Ukrajnában, Szlovákiában és a már említett Romániában. Ezen felül emigráns közösségek használják világszerte, elsősorban az Amerikai Egyesült Államokban, Kanadában és Izraelben.

Érdekes, hogy a magyarnak alig vannak érdemleges változatai: a nyelvjárások egymástól és a köznyelvtől kevéssé térnek el, megértési nehézségeket alig okoznak. Ez talán a hosszú szomszédsági lét miatt van, mely – a más nyelvekkel folyamatosan ütközve – egységességre indíthatta a beszélőket. A hagyományos felosztás szerint a magyar nyelvnek hét dialektusát különböztetik meg Magyarország mai területén. Ezen felül két magyar dialektus létezik Romániában: a székely és a csángó.

A Magyar Köztársaságban és a szomszéd országokban használt magyar között ugyancsak kevés különbség van, különösen a művelt nyelvhasználat és a helyesírás egységes. Apró, de jellemző különbségek persze adódnak. Míg a magyarországi magyar döntően német hatás alatt fejlődött, addig a romániai magyar inkább román hatás alatt él. A csángó közösség viszonylag szeparáltan élt a többi magyartól, ezért ők egy, a középkori magyarhoz közelebb álló nyelvváltozatot őriztek meg.

A magyar nyelv különlegességei

A legtöbb európai nyelv az indoeurópai nyelvcsaládba tartozik, s így egymásnak rokona az orosz, a spanyol, a görög, a norvég, az angol, az albán – de a magyarnak nem! A magyar az Urál heygségből származik, Európa és Ázsia határvidékéről. Az uráli nyelvcsaládnak két ága van: szamojéd és finnugor. A magyar az utóbbiba tartozik, ezért szoktuk finnugor nyelvnek is nevezni. Rokonai a finn, az észt, a lapp és néhány más nyelv a mai Oroszországban.

Az uráli nyelvek néhány közös, ősi jellemzője:

- nincsenek nemek: ugyanaz a szó – a magyarban *ő* – fejezi ki a "he" és a "she" fogalmát.
- csak két igeidő van: jelen és múlt; ezek árnyalatait meg a jövő időt körülírással lehet kifejezni.
- az irányhármasság: a helyet kifejező ragokból 3x3 van, mint az alábbi tábla mutatja a *doboz* szó példáján (a névelő változatlan, nincs egyeztetve a főnévvel):

	Hova? *'Where to?'*	Hol? *'Where?'*	Honnan? *'Where from?'*
belül *'inside'*	a dobozba *into the box*	a dobozban *inside the box*	a dobozból *out of the box*
rajta *'on'*	a dobozra *onto the box*	a dobozon *on the box*	a dobozról *off the box*
közelében *'near'*	a dobozhoz *to the box*	a doboznál *at the box*	a doboztól *from near the box*

A magyart latin betűkkel írják, de a magyar szöveg mégsem hasonlít semelyik európai nyelvre. Íme egy klasszikus vers két sora, egyszerű fordításban (Kölcsey Ferenc 1823-as *Hymnus* című verséből, amely ma a magyarok nemzeti himnusza):

Isten, áldd meg a magyart	"God bless the Hungarians
Jókedvvel, bőséggel.	With merriment and plenty."

Egyetlen szót sem lehet felismerni az átlagos európai nyelvkincs alapján; a magyarok nemcsak "God"-ot hívják *Isten*-nek, de saját magukat sem hívják "Hungarian"-nek, hanem *magyar*-nak. De többről van szó, mint a szavak különbözéséről:

Isten	*áldd*	*meg*	*a*	*magyart*
God	bless	?	the	Hungarian

A kérdőjeles szó nem létezik a legtöbb nyelvben: a neve igekötő ("verb-binder", szakszóval "Verbal Prefix"). Szerepe igen sokféle: itt a befejezettséget fejezi ki. A magyar nyelv egyik szépsége (és nehézsége) éppen az igekötők használatában van.

De nézzük a második sort:

	jókedv-	*-vel*		*bőség-*	*-gel*
with	merriment		with	plenty	

Ahol az angolban "with" áll (elöljáró), ott a magyarban végződések vannak. A magyarban nincsenek elöljárók, példánkban a *-vel*, *-gel* ragok fejezik ki azt, amit az angol "with".

Még egy fontos magyar sajátságot említünk: a birtokviszonyt (possession) fordítva fejezik ki, mint az indoeurópai nyelvek. Például a "Paul's radio" megfelelőjében a magyar nem a birtokoshoz, Pálhoz teszi a ragot, hanem a birtokhoz, a rádióhoz: *Pál rádió*-ja, ami olyan, mintha azt mondanám: "Paul radio-his".

Inkább kultúrtörténeti, mint nyelvészeti érdekesség, hogy a magyarban a családnév áll elöl, az "utónév" ("given name, Christian name") hátul, tehát Liszt Ferenc (=Franz Liszt), Bem József (=Józef Bem), Bartók Béla, Márai Sándor a megszokott sorrend.

A magyar ún. "szintetikus" nyelv: a nyelvtani elemeket többnyire egyetlen szóban, toldalékokkal fejezi ki, szemben az "analitikus" nyelvekkel, melyek inkább külön szavakat – elöljárókat, névmásokat, segédigéket – használnak. Például az angol *can* megfelelője a -hat/-het rag.

*Leó-**val***	*a kocsi-**ból***	*utaz-**hat***	*jár-**ogat***
with Leo	**from** the car	**can** travel	**usually** goes

A végződéseket szigorú sorrend szerint kell a szóhoz ragasztani, gyakran többet is egymás után, s így a szavak jó hosszúra nőhetnek. A szintetikus szóépítésnek ezt a módját agglutinációnak (azaz "szóragasztásnak") nevezzük. Például:

bolondozhattunk “we could fool [around]” (=’fool-verb-can-past-we’)
ösztönözhettünk “we could stimulate” (=’stimulus-verb-can-past-we’)

A két szó felépítése azonos – a látszólagos különbséget a magánhangzók okozzák, az ún. magánhangzó-harmónia (más néven illeszkedés) miatt. A magánhangzók két osztályba sorolódnak: "mélyek" (deep): *a o u*, és "magasak" (high): *e i ö ü*. A végződésekben a magánhangzó az alapszónak megfelelően jelenik meg: a *bolond* mély, így a többi magánhangzó is mély: *o - o + o - a - u*.

Modernkori fejlődés

A magyar nyelv bizonyos szempontból mindig kisebbségi nyelv volt, és más nyelvekből folyamatosan vett át szavakat. Bár a magyar a térség legnépesebb nyelve volt, sosem került abszolút többségbe: összességében mindig több másnyelvű élt a Kárpát-medencében: szláv (elsősorban szlovák, szerb, horvát), később pedig német, román, zsidó és cigány népesség. Hivatalos nyelvként a latin volt használatos egészen a 19. század elejéig, ez volt a közigazgatás és a tudomány nyelve. A magyar országgyűlés csak 1844-től vezette be a magyarul való ülésezést, addig latinul folyt a vita.

A magyar nyelv mindig inkább importőr volt, mint exportőr. A mai magyar szókincs számos szláv, latin, román és olasz eredetű szót tartalmaz. A legerősebb a német hatás volt, hiszen Magyarország 400 évig volt a Habsburg-birodalom része. Rengeteg német eredetű szó van, ilyen például a *tánc* és a *hering*. A más nyelvekből való szóátvétel napjainkban is folytatódik: francia *fritőz*, *bagett*; olasz *maffiózó*, *paparazzi*; angol *fitnesz*, *szerver* stb. A mostanában átvett szavak nagy része anglicizmus, köszönhetően az amerikai filmipar, zene és technológia erős hatásának.

Nyelvművelés Magyarországon

A magyarországi nyelvművelés egyik központja a Balassi Intézet, amely a határon túli magyar kultúra magyarországi és az egyetemes magyar kultúra külföldi bemutatásáért felel, hasonlóan, mint a német Goethe Institut vagy az angol British Council. Az egységes és egyetemes magyar kultúrát terjeszti és népszerűsíti a nagyvilágban, úgy hogy ezzel párhuzamosan segíti a külföldön vagy határon kívül létező magyar hagyományok és kultúra megismertetését Magyarországon. A Balassi

Intézet központi szerepet tölt be a magyar nyelv tanulása, tanítása, a képzés módszertani központjának kialakítása vonatkozásában is.

A magyar nyelv kutatásának vezető magyarországi központja a Magyar Tudományos Akadémia Nyelvtudományi Intézete. A Nyelvtudományi Intézet 1949-ben jött létre, a Közoktatási Minisztérium felügyelete alatt, majd 1951-ben került az MTA felügyelete alá. Alapfeladata a magyar nyelvészet, az általános és alkalmazott nyelvészet, az uráli nyelvészet és a fonetika területén tudományos kutatásokat végezni, a magyar irodalmi és köznyelv nagyszótárát elkészíteni, archív anyagát gondozni, valamint a magyar nyelv különböző változatait és az országon belül és kívül beszélt kisebbségi nyelveket vizsgálni, beleértve az európai integráción belüli nyelvpolitikai kérdéseket is. Kiegészítő feladatként nyelvi korpuszok és adatbázisok létrehozásával, számítógépes alkalmazások nyelvészeti alapjainak megalkotásával, valamint közönségszolgálati tevékenységgel, szakértői vélemények készítésével is foglalkozik. Mindemellett a felsőoktatásban is részt vesz, az itt működő MTA-ELTE Elméleti Nyelvészet Szakcsoport révén.

A magyar helyesírási kérdések szigorú akadémiai szabályozás alá tartoznak: a magyar helyesírást a Nyelvtudományi Intézet Helyesírási Bizottsága szabályozza úgynevezett helyesírási szabályzatok kiadásával. A szabályok alkalmazása nem kötelező, de Magyarországon a helyesírásnak presztízsértéke van.

Manapság sok lelkes hagyományőrző érvel amellett, hogy az elsősorban az angolból származó neologizmusok nem erősítik, hanem inkább gyengítik a magyar nyelvet. "Nyelvvédő" tevékenységüknek köszönhetően 2002-ben bevezették az ún. nyelvtörvényt, ami kötelezővé teszi az összes angol nyelvű hirdetés és szlogen magyarra cserélését. Emellett egyéb nyelvművelő és -védő lépések is történtek: például 2011 elején lépett életbe az új médiatörvény, amely megszabja a televízióban és a rádióban sugárzott magyar és külföldi zenék arányát.

A magyar nyelv az oktatásban

A magyar nyelv 1844-ben lett a közigazgatás, a tudomány és az oktatás hivatalos nyelve – azóta lehet magyarul tanulni az általános iskolákban is. Az 1868-as oktatási reform után pedig a felsőbb szintű oktatási intézmények nyelve is a magyar lett. Ma már a Kárpát-medence számos felsőoktatási intézményében lehet magyar nyelvű diplomát szerezni, Nyitrától (Nitra, Szlovákia) a magyarországi egyetemeken, főiskolákon át Újvidékig (Novi Sad, Szerbia) vagy Kolozsvárig (Cluj-Napoca, Románia).

A 19. század óta a magyar nyelv és irodalom meghatározó szerepet tölt be az oktatásban. A magyar tantárgy 6-tól 18 éves korig kötelező az iskolákban. Az általános iskola alsó évfolyamaiban, 6 és 10 éves kor között a tananyag írás, olvasás és fogalmazás területekre oszlik. 10 éves kor után a magyar nyelvtant és irodalmat külön tanítják.

A 2009-es PISA felmérés szerint, amely a tanulók szövegértési képességeit mérte, a magyar tanulók átlageredménye emelkedett 2000-hez képest, ezzel elérte az OECD-átlagot. Így Magyarország olyan or-

országokkal került egy csoportba, mint Franciaország, Németország vagy az Egyesült Királyság.

Nemzetközi vonatkozások

Magyarország számos világhíres fizikust (Teller Ede, Wigner Jenő és Szilárd Leó, a Manhattan terv résztvevői), matematikust (Rényi Alfréd, Erdős Pál, az Erdős-szám névadója) és zenészt (Liszt Ferenc, Bartók Béla) adott a világnak. A magyar tudósok számos Nobel-díjat nyertek a fizika, a kémia és az orvostudomány terén.

Ahogy mindenhol máshol a tudományos világban, a magyar kutatók is szembesülnek az állandó publikációs nyomással. Mivel a vezető nemzetközi folyóiratok jelentős része angol nyelvű, tovább nő az angol nyelv szerepe. A helyzet hasonló az üzleti világban is: a nagy multinacionális vállalatoknál az angol lett a lingua franca a szóbeli és az írott kommunikációban is. Ám egy 2005-ös felmérés szerint Magyarországon a valamilyen idegen nyelvet beszélő emberek száma még mindig az európai átlag alatt van: a magyar embereknek csak 35%-a beszél legalább egy idegen nyelvet.

A nyelvtechnológia erre a kihívásra egy más nézőpontból tud megoldást nyújtani: olyan szolgáltatásokkal, mint a gépi fordítás vagy a nyelvközi információ-visszakeresés, így csökkentve a nem angol anyanyelvűek személyes és gazdasági hátrányait.

Magyar az interneten

2009-ben a magyarországi lakosság 61,6%-a volt internethasználó. A fiatal generáció körében, 14-17 éves korban, ez az arány magasabb. Az internetpenetráció az európai átlag alatt van, de folyamatosan emelkedik a rendszerváltás óta. Egy 2010-es európai felmérés szerint a közösségi oldalak használata az európai átlag fölött van, ami talán annak köszönhető, hogy Magyarországon a Facebook megjelenése előtt már létezett egy népszerű közösségi oldal, az iWiw. Egy meglehetősen aktív magyar nyelvű webes közösség létezéséről tanúskodik az is, hogy a magyar Wikipédia a 19. legnagyobb, megelőzve olyan több beszélővel rendelkező európai nyelveket, mint a török, a román vagy a dán, és olyan világnyelveket, mint az arab vagy a koreai.

A magyar nyelvtechnológia számára az internet növekvő jelentősége két szempontból is fontos. Egyrészt a digitálisan elérhető nyelvi adatok mennyisége gazdag forrást nyújt a nyelvhasználat statisztikai elemzéséhez. Másrészt az internet adja a nyelvtechnológiai alkalmazások elsődleges felhasználási helyét.

A leggyakrabban használt alkalmazás a webes keresés, ami feltételezi a nyelv többszintű automatikus feldolgozását, ahogy majd részleteiben látni fogjuk fehér könyvünk második felében. A webes keresés minden nyelvre különböző, szofisztikált nyelvtechnológiát igényel. Például a magyarra nézve ez magában foglalja azt is, hogy a főnevek, melléknevek és igék különböző végződésekkel ellátott alakjait, illetve az eltérő tőváltozatokat is meg kell találnunk, mint például a *ló-lovak* esetében.

Az internethasználók és szolgáltatók azért ennél kevésbé transzparens módon is profitálhatnak a nyelvtechnológiából, például abban az esetben, amikor webes tartalmakat fordítanak egyik nyelvről egy másikra. Tekintve az emberi fordítás magas költségeit, ebben az esetben még az olyan nyelvtechnológiai eszközök fejlesztése is megéri, amelyek az elvártnál kevésbé teljesítenek jól. Ez utóbbi helyzet előállhat amiatt is, mert a magyar nyelv meglehetősen komplex, továbbá mert egy tipikus nyelvtechnológiai alkalmazás kifejlesztésében nagyszámú más technológia is érintve van. A következő fejezetekben bevezetést adunk a nyelvtechnológiába és annak főbb alkalmazási területeibe, valamint értékeljük a magyarországi nyelvtechnológia jelenlegi állapotát.

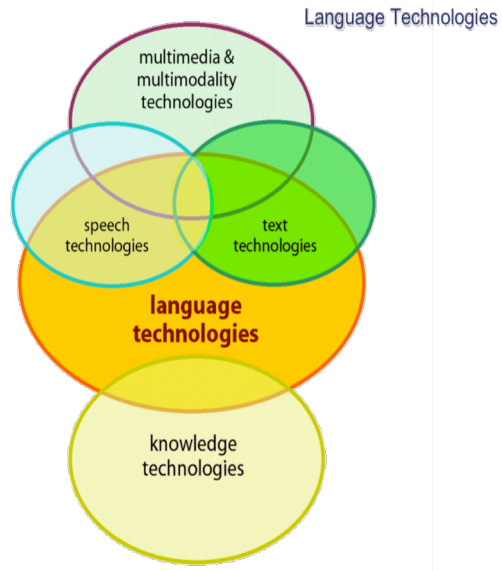
További olvasnivalók

- Ismeretterjesztő kiadvány a magyar nyelvről. Szöveg: Nádasdy Ádám, Kiadja: Balassi Intézet, Nemzeti Évfordulók Titkársága
- http://hu.wikipedia.org/wiki/Magyar_nyelv
- Péter Rebrus, Anna Babarczy: Hungarian descriptive grammar. In: S. Nagy Katalin, Szakadát István (szerk.): Média és társadalom: válogatás a Szociológia és Kommunikáció Tanszék Média Oktató és Kutató Központ munkatársainak legújabb munkáiból. Budapest, 2006. 331-381. o.

Nyelvtechnológia magyarul

Nyelvi technológiák

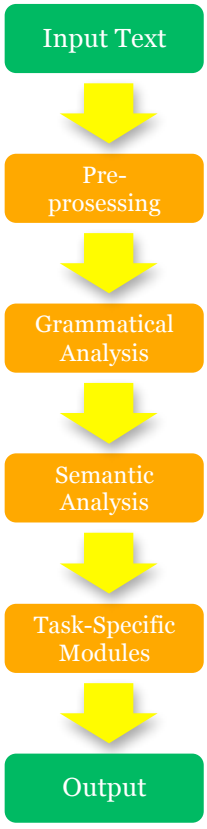
A nyelvi technológiák olyan információtechnológiai eszközök, amelyek kifejezetten a természetes emberi nyelv feldolgozására lettek specializálva. Ezért ezeket a technológiákat összefoglaló névvel természetesnyelv-feldolgozásnak szokták nevezni. Az emberi nyelv előfordul beszélt és írott változatban. Míg a beszéd a legősibb és legtermészetesebb módja az emberi kommunikációnak, a komplex információk és az emberi tudás nagy része írott formában létezik. A beszéd- és a nyelvtechnológia az emberi kommunikációnak ezt a két módját dolgozza fel, illetve állítja elő. De persze a nyelvnek vannak olyan aspektusai, amelyek a beszédet és a szöveget egyaránt érintik: ilyenek a szótárak, a nyelvtanok és a szemantika. Ezek a nyelvfeldolgozás olyan területei, amelyeket nem lehet besorolni kizárólag a beszéd- vagy a nyelvtechnológia alá. Ezek között a technológiák között találunk olyanokat is, amelyek összekötik a nyelvet a világról való tudásunkkal. A jobb oldalon látható ábra a természetesnyelv-feldolgozás egészét illusztrálja. Kommunikációnkban vegyítjük a nyelvet és a kommunikáció más módjait és csatornáit. A beszédet gesztusokkal és arckifejezésekkel kísérjük. A digitális szövegek képekkel és hangzó anyagokkal együtt jelennek meg. A filmek a nyelvet beszélt és írott formában is megjelenítik. Vagyis a beszéd- és nyelvtechnológia átfed és együttműködik más technológiákkal, amelyek így együtt erősítik a multimodális kommunikáció és a multimédiás tartalmak feldolgozását.



A nyelvtechnológiai alkalmazások felépítése

A tipikus nyelvtechnológiai alkalmazások több komponensből állnak össze, anelyek a nyelv és az alkalmazási terület egyes szintjeit tükrözik. A jobb oldali ábra egy szövegfeldolgozó rendszer egyszerűsített felépítését mutatja. Az első három modul a bemenő szöveg szerkezetét és jelentését dolgozza fel:

- Előfeldolgozás: adattisztítás, a formázás eltávolítása, a bemenő szöveg nyelvének megállapítása, a speciális karakterek kezelése (pl. a magyar ékezetes betűk esetében) stb.
- Nyelvtani elemzés: az ige és argumentumainak megkeresése, a mondat szerkezetének feltárása.
- Szemantikai elemzés: egyértelműsítés (egy adott szónak az adott kontextusban mi a jelentése), az anaforák feloldása (pl. az *ő* kire vonatkozik), a mondat jelentésének reprezentálása valamilyen gép által olvasható formában.



Ezután következnek a különféle feladatspecifikus modulok, mint például a bemenő szöveg automatikus tömörítése, az adatbázisokban való keresés és ehhez hasonlók.

A következő fejezetekben a legfontosabb alkalmazási területeket és a hozzájuk kapcsolódó modulokat fogjuk bemutatni. A különböző alkalmazások felépítése a fentihez hasonlóan egyszerűsített formában történik – célunk a nyelvtechnológiai alkalmazások komplexitásának könnyen érthető módon való illusztrálása.

A legfontosabb eszközök és erőforrások a szövegben aláhúzással vannak jelölve, és mindegyik megtalálható a fejezet végén levő táblázatban. A legfontosabb alkalmazási területeket tárgyaló alfejezetek a magyarországi kutatás-fejlesztésről is áttekintést adnak.

A legfontosabb alkalmazási területek bemutatása után egy rövid kitekintésben beszámolunk a nyelvtechnológiai kutatási és oktatási helyzetről, különös tekintettel a már lezárult és a folyó kutatási programokra. A fejezet végén egy szakértői értékelést adunk a legfontosabb nyelvtechnológiai eszközökről és erőforrásokról olyan dimenziók mentén, mint az elérhetőség, a fejlettség vagy a minőség. Ez a táblázat jó áttekintést ad a magyar nyelvtechnológia állapotáról.

A legfontosabb alkalmazási területek

Nyelvi ellenőrzés

Mindenki, aki használt már a Microsoft Word-höz hasonló szövegszerkesztőt, találkozott már helyesírás-ellenőrző programmal, ami jelzi a helyesírási hibákat, és javítási javaslatokat tesz. 40 évvel az első helyesírás-ellenőrző program után, amely Ralph Gorin nevéhez fűződik, a nyelvi ellenőrzők már nem szimplán összehasonlítják az ellenőrizendő szavakat a jól írt szavak listájával, hanem ennél sokkal kifinomultabb eszközökkel dolgoznak. A nyelvfüggő algoritmusokon felül, amelyek a morfológiát (pl. a többes számú alakokat) tudják kezelni, ma már némelyik program a mondatszintű hibákat is detektálja, például ha hiányzik a ragozott ige a mondatból, vagy ha az ige és az alany nincsenek számban-személyben egyeztetve, pl.: "Én *írsz levelet." Azonban a legtöbb elérhető nyelvi ellenőrző (beleértve a Microsoft Word-öt is) nem találna hibát Jerrold H. Zar versének következő versszakában:

> *Eye have a spelling chequer,*
> *It came with my Pea Sea.*
> *It plane lee marks four my revue*
> *Miss Steaks I can knot sea.*

Az ilyen típusú hibák kezeléséhez az esetek nagy részében a kontextus elemzését is el kell végezni. A magyarban például vannak olyan ragozott szavak, amelyek különböző jelentésekkel bírhatnak, pl.:

> *várunk*$_1$ ('we are waiting')
> *várunk*$_2$ ('our castle')

A jelenség kezeléséhez vagy nyelvspecifikus nyelvtani szabályok előállítására, vagyis magas szintű szakértői munkára, vagy statisztikai alapú nyelvmodellekre van szükség. Az ilyen modellek alapján egy bizonyos szó adott környezetben való előfordulásának valószínűségét tudjuk kiszámolni. Például a *várunk* valószínűleg nem ige, ha a mondatban már szerepel egy másik ragozott ige. Statisztikai alapú nyelvmodellek automatikusan előállíthatók nagy méretű helyes adatot tartalmazó szöveghalmazokból, más néven korpuszokból. Ez a megközelítés elsősorban angol nyelvű adatokra lett kifejlesztve, de a magyarra is al-

kalmazható. Azonban azt is figyelembe kell venni, hogy a módszerek nem ültethetők át egy az egyben a magyar nyelv agglutináló jellege és szabad szórendje miatt.

A nyelvi ellenőrzők használata nem csak a szövegszerkesztőkre korlátozódik, alkalmazzák még az ún. szerzői támogatási rendszerekben (authoring support systems) is. A technikai termékek számának növekedésével a hozzájuk tartozó dokumentáció mennyisége is nagyon megnőtt az elmúlt évtizedekben. A hibás vagy nehezen érthető használati útmutatók miatt bekövetkező károkról szóló vásárlói panaszoktól tartva a vállalatok egyre nagyobb hangsúlyt fektetnek a technikai dokumentáció minőségére, nemzetközi viszonylatokban is. A természetesnyelv-feldolgozás eredményei a szerzői támogatási rendszerekben is fejlődést hoztak: a technikai dokumentáció szerzőit szótárak, terminológiai adatbázisok és mondattani szabályok segítik.

A helyesírás-ellenőrzés és a szerzői támogatás mellett a nyelvi ellenőrzés a gép által támogatott nyelvtanulás terén is fontos szerepet tölt be, továbbá a webes keresőkben is alkalmazzák a lekérdezések automatikus javítására, pl. a Google keresési javaslatai esetében.

Tekintettel a magyar nyelv erősen agglutináló jellegére egy magyar nyelvű helyesírás-ellenőrzőnek tartalmaznia kell egy morfológiai elemző komponenst, hogy kezelni tudja a ragozott és összetett szavakat is. Az első magyar helyesírás-ellenőrzőt a MorphoLogic Kft. fejlesztette ki a 80-as években, amely egy helyesírás-ellenőrző modul és egy morfológiai modell kombinációjából állt elő. A *Helyes-e?* programcsomag a Microsoft Office, a QuarkXPress, az Adobe InDesign és más szöveg- és kiadványszerkesztővel is használható. A MorphoLogic nyelvhelyesség-ellenőrző programokat is fejlesztett, amelyek felismernek olyan helyesírási hibákat, amelyeket a szóellenőrző programok nem tudnak megtalálni, mert a szöveget nem összefüggéseiben, hanem szavanként vizsgálják. A program nem feltétlenül hibákat jelez, hanem csak figyelmeztet. A jelzések nagy része tényleges hibára utal, mások csak felhívják a figyelmet egy-egy lehetséges hibára. Az utóbbi esetben a felhasználónak kell eldöntenie, hogy tényleges hibáról van-e szó.

Nyílt forráskódú helyesírás-ellenőrző is létezik a magyarra. A Hunspell a MySpellen alapul, és integrálva lett az OpenOffice-ba, a Mozilla Firefox 3-ba, a Mozilla Thunderbirdbe és a Google Chrome-ba.

Webes keresés

A weben, intraneten vagy digitális könyvtárakban való keresés valószínűleg a legtöbbet használt és a legkevésbé fejlett nyelvtechnológiai alkalmazás jelenleg. A Google kereső 1998-ban indult, és napjainkban a világ összes lekérdezésének 80%-át végzi. Már a magyar nyelvben is elterjedt a *guglizni* szó, bár a nyomtatott szótárakba még nem került bele. Sem a Google lekérdező felülete, sem a találati lista prezentációja nem változott szignifikánsan az első verzió óta. A jelenlegi változatban van viszont ellenőrző program, amely az elgépeléseket javítja, továbbá 2009-ben alapszintű szemantikai kereső alkalmazást építettek be, amely növeli a találati pontosságot azzal, hogy kontextusban vizsgálja a keresőkifejezést. A Google sikersztorija azt mutatja,

hogy nagy mennyiségű adattal és hatékony indexelési technológiával a statisztikai alapú megközelítés kielégítő eredményt tud hozni.

Azonban ha részletesebb és szofisztikáltabb információhoz akarunk jutni, mélyebb nyelvi tudásra van szükségünk. A kutatóműhelyekben a gép által olvasható tezauruszokkal és a WordNethez hasonló ontológiákkal kísérleteznek, amelyek javítják a keresés hatékonyságát azáltal, hogy a keresőkifejezés szinonimáit (pl. *atomenergia*, *magenergia*, *nukleáris energia*) és a hozzá kapcsolódó szavakat is figyelembe veszik.

A keresőmotorok új generációjának már sokkal több specifikus nyelvtechnológiát kell tartalmaznia. Ha a lekérdezés kérdést vagy más típusú mondatot tartalmaz, nem csak szavak listáját, a releváns válasz megtalálásához szükség van a mondat szintaktikai és szemantikai szintű elemzésére, valamint a releváns dokumentumok gyors elérését lehetővé tevő indexelésre is. Például képzeljünk el egy olyan lekérdezést, hogy "Adj egy listát azokról a cégekről, amelyeket felvásároltak az elmúlt öt évben". Ahhoz, hogy erre kielégítő választ kapjunk, a mondat teljes szintaktikai elemzését el kell végezni, és rá kell jönni, hogy a felhasználó azokra a cégekre kíváncsi, amelyeket felvásároltak, és nem azokra, amelyek felvásároltak cégeket. Ezen felül az időt jelölő kifejezést is fel kell dolgozni ahhoz, hogy kiderüljön, hogy mely évekről van szó.

Végül a feldolgozott keresőkifejezést össze kell vetni nagy mennyiségű strukturálatlan adattal, hogy megtaláljuk azt az információt, amelyet a felhasználó keres. Ezt, vagyis a keresést és a releváns találatok sorrendezését hívják információ-visszakeresésnek. Továbbá ahhoz, hogy cégek listáját kapjuk, ki kell nyernünk azt az információt a dokumentumokból, hogy szavak egy adott sorozata egy cégre utal. Ezt a fajta információkinyerést végzik az automatikus tulajdonnév-felismerők.

Még több nyelvtechnológiát igényel egy keresőkifejezés megtalálása más nyelvű dokumentumokban. A nyelvközi információ-visszakereséshez először le kell fordítani a keresőkifejezést az összes lehetséges forrásnyelvre, majd a találatokat vissza kell fordítani a célnyelvre. A nem szöveges formában levő adatok növekvő aránya életre hívta az igényt a multimédiás információ-visszakereső szolgáltatásokra, vagyis a képekben, hangzó anyagokban, videókban való keresésre. Az audió- és videófájlok esetében szükség van egy beszédfelismerő modulra is, amely a beszédet szöveggé alakítja át, amelyben így már lehet keresni.

A ragozó nyelvek esetében, amilyen a magyar is, fontos, hogy a keresés során egy adott szó minden ragozott alakját is megtaláljuk. Erre a célra több magyar morfológiai elemző is létezik. A főnévi csoportok automatikus azonosítása már egy magasabb szintű elemzést tesz lehetővé: a magyarra egy szabályalapú és egy statisztikai alkalmazás is működik.

Mivel a magyar nem olyan kötött szórendű, mint például az angol, a magyar mondatelemzők fejlesztése során nem tudunk csak a mondat lineáris szerkezetére támaszkodni. Viszont az esetragok és névutók fogódzót jelentenek, mivel ezek határozzák meg a mondatrészek szerepét. Az igék és a hozzájuk tartozó vonzatok alkotják a mondat

szerkezetének alapját, ezért fontosak az ún. vonzatkerettárak. Egy ilyen adatbázist fejlesztettek a Nyelvtudományi Intézet munkatársai, amely magasabb szintű elemző alkalmazásokba, például szabályalapú szintaktikai elemzőbe is beépíthető. Ez utóbbiból több is létezik a magyarra – egyik a Szeged Treebankbe, egy másik pedig a MetaMorpho nevű szabályalapú gépi fordítóba lett beépítve.

A nyelvtechnológiával foglalkozó cégek és kutatóműhelyek fő kutatási irányai között szerepel az olyan trend- és szövegelemző eszközök fejlesztése, amelyek természetesnyelv-feldolgozó alkalmazásokat integrálnak annak érdekében, hogy a strukturálatlan szövegben megtalálják a releváns információkat. Erre a célra magyar nyelvű morfológiai elemzők és egyértelműsítők, valamint tulajdonnév-felismerők állnak rendelkezésre, melyek nagyrészt statisztikai tanuló algoritmusokon alapulnak.

Létezik egy magyar nyelvű általános célú metakereső, a PolyMeta, amely lehetőséget nyújt tetszőleges számú, interneten keresztül elérhető kereső (adatbázis, forrás) egyidejű keresésére. Az eredményekből közös találati lista készül, amelyben az elemek fontossági sorrendbe rendezettek. A metakereső természetesnyelv-feldolgozási és információvisszakeresési algoritmusokat használ a keresőkifejezések elemzéséhez és a találatok sorrendezéséhez.

De nemcsak kis- és középvállalatok fejlesztenek információkinyerő eszközöket Magyarországon. Számos olyan projekt fut különböző egyetemeken és kutatóintézetekben, melyek célja szemantikai alapú keresőrendszerek fejlesztése, vagy magyar nyelvű ontológiák (pl. Magyar WordNet, Magyar Egységes Ontológia) építése.

Beszédtechnológia

A beszédtechnológia szolgáltatja az alapot olyan interfészek előállításához, amelyek lehetővé teszik, hogy a felhasználók a gépekkel természetes emberi nyelven, és ne csak grafikus felület, billentyűzet vagy egér segítségével kommunikáljanak. Napjainkban ilyen beszédinterfészeket alkalmaznak bizonyos szolgáltatások részleges vagy teljes automatizálására. Az üzleti szférában elsősorban a bankok, a logisztikával, a szállítással és a telekommunikációval foglalkozó cégek használják. A beszédtechnológiát alkalmazzák még autós navigációs rendszerekben és az okostelefonokban a grafikus felület alternatívájaként.

A beszédtechnológia az alábbi négy fő technológiai területet foglalja magában:

- Az automatikus beszédfelismerés határozza meg, hogy mely szavakat mondta ki a felhasználó.
- A szintaktikai elemzés és a szemantikai interpretáció segítségével elemezhető a felhasználó közlésének szintaktikai szerkezete, valamint leképezhető annak szemantikai interpretációja az adott rendszer céljainak megfelelően.
- A dialógusvezérlés az input nyelvi jellemzői, az adott felhasználó és feladat egyéni beállításai alapján valósítja meg a rendszer megfelelő lépését, az adatbázis-lekérdezést.

- A beszédszintézis technológiáját alkalmazzák arra, hogy a gép előállítsa a megfelelő beszédkimenetet.

Az egyik legnagyobb kihívást az automatikus beszédfelismerés jelenti, vagyis hogy a rendszer minél pontosabban felismerje a felhasználó által kiejtett szavakat. Ez kétféleképpen történhet: a felhasználó által használható kifejezéseket csökkentjük kulcsszavak egy limitált nagyságú halmazára, vagy nyelvmodelleket állítunk elő, amelyek a természetes nyelvi kifejezések egy nagyobb hányadát fedik le. Míg az első módszer merev és nehezen használható beszédinterfészt, valamint kevésbé elfogadható kimenetet eredményez, addig a nyelvmodellek előállítása és finomhangolása a költségeket emeli meg erőteljesen. Azonban a nyelvmodelleket alkalmazó beszédinterfész nagyobb elfogadottsággal rendelkezik a felhasználók körében, előnyösebb, mint a kevésbé rugalmas rendszervezérelt megközelítés.

Ami a beszédinterfész kimeneti oldalát illeti, a vállalatok egyre inkább profik által előre felmondott kifejezéseket használnak. A statikus kifejezések esetében, amikor a beszéd nem függ adott kontextustól vagy a felhasználó adataitól, ez a módszer kellő mértékű felhasználói elégedettséget eredményez. Viszont minél dinamikusabb a tartalom, annál rosszabb lesz a prozódia az audiófájlok összevágása miatt. Ennek ellenére a mai beszédszintetizáló rendszerek egyre jobban teljesítenek, köszönhetően a természetes prozódiának.

A beszédtechnológia piacán az elmúlt évtizedekben erős sztenderdizálási lépések történtek a különböző technológiai komponensek közötti interfészek, valamint az egyes alkalmazásokra épülő termékek esetében is. Nagyon erős piaci konszolidáció zajlott le az elmúlt tíz évben, főként a beszédfelismerés és -szintézis terén. A G20 országok nemzeti piacait kevesebb mint 5 cég dominálja, mint a Nuance és a Loquendo, csak hogy a legprominensebbeket említsük az európai piacról.

A magyar nyelv speciális jellege miatt a világszerte széles körben alkalmazott módszerek vagy egyáltalán nem, vagy csak nehezen adaptálhatók a magyarra. Viszont a kifejezetten a magyarra kifejlesztett módszerek könnyen alkalmazhatók lehetnek a hasonlóan agglutináló nyelvekre, mint a finnre, a törökre vagy az arabra.

A magyar beszédszintézis piacát a Budapesti Műszaki és Gazdaságtudományi Egyetemen (BME) dolgozó kutatócsoportok dominálják. A legszélesebb körben használt magyar beszédszintetizátor a Profivox, ami több alkalmazásba is be lett építve: SMS- és emailfelolvasó szoftverbe, autós és mobiltelefonos GPS rendszerbe, valamint e-book- és képernyőolvasó szolgáltatásba, amelyek segíthetik a látássérült emberek integrációját az információs társadalomba.

Beszédfelismeréssel Magyarországon az egyetemi kutatóműhelyek (pl. a Szegedi Tudományegyetem Informatikai Tanszékcsoportja) mellett kisebb vállalkozások is foglalkoznak, mint az Alkalmazott Logikai Laboratórium, az Aitia vagy a Digital Natives. A már említett nyelvi nehézségek ellenére több magyar nyelvű gépi beszédfelismerő alkalmazást is kifejlesztettek az elmúlt években. A BME Távközlési és Médiainformatikai Tanszékén kifejlesztettek egy statisztikai alapú folyam-

atosbeszéd-felismerő motort és fejlesztői környezetet, továbbá egy kötött hangsúlyozáson alapuló szóhatár-detektáló alkalmazást magyar és finn nyelvekre, melyet egy nyelvközi vizsgálat előzött meg. Továbbá a már említett kutatóműhelyek közös munkájának eredményeképpen különféle orvosi leletező beszédfelismerők készültek el, melyek az orvosi vizsgálatokat közvetlen beszéd-szöveg átalakítással segítik. A beszédfelismerő rendszerek számos további gyakorlati alkalmazást segíthetnek. Ilyen például a telefonos hívások kezelése vagy a telefonközpont-irányítás.

Gépi fordítás

A digitális számítógépek alkalmazásának ötlete természetes nyelvek lefordítására 1946-ban merült fel először, és A. D. Booth-tól származik. Az ötletet az ötvenes években anyagi támogatás is követte, ami azonban csak a nyolcvanas években folytatódott. Mindemellett a gépi fordítás még mindig nem váltotta be a kezdeti nagy reményeket.

A legalacsonyabb szinten a gépi fordítás szimplán behelyettesítés: az egyik természetes nyelvű szót lecseréljük egy másik nyelvű szóra. Ez az eljárás csak nagyon szűk szókincsű, formalizált nyelvű szövegek (pl. időjárás-jelentések) esetében működik. Azonban a kevésbé sztenderdizált szövegek jó fordításához nagyobb szövegegységeket (frázisokat, mondatokat vagy teljes bekezdéseket) kell illeszteni a másik nyelvű megfelelőikhez. A legfőbb problémát az a tény okozza, hogy az emberi nyelv sokszor többértelmű, ami kihívások elé állítja a nyelvfeldolgozókat minden szinten. A lexikai szinten a szójelentés-egyértelműsítés (pl. a *nyúl* lehet ige és állat is), míg a mondat szintjén akár az esetragos főnévi csoportok is okozhatnak nehézségeket, mint ezekben a mondatokban:

> *A rendőr látta az embert a távcsővel.*
> [The policeman observed a man with a telescope.]
>
> *A rendőr látta az embert a revolverrel.*
> [The policeman observed a man with a revolver.]

A feladat egyik megközelítési módja nyelvtani szabályokon alapul. Közeli rokon nyelvek esetében a közvetlen fordítás kivitelezhető lehet a fenti példákra is. De általában a szabályalapú (vagy tudásvezérelt) rendszerek úgy működnek, hogy először elemzik a bemenő szöveget, majd egy közvetítő, szimbolikus reprezentációt alkotnak, amiből legenerálják a célnyelvi kimenetet. Ezeknek a rendszereknek a teljesítménye morfológiai, szintaktikai és szemantikai információt egyaránt tartalmazó lexikonok, valamint nyelvész szakértők által aprólékosan kidolgozott nyelvtani szabályok meglététől egyaránt erősen függ.

A nyolcvanas évek elejétől kezdve, ahogy a számítógépek egyre olcsóbbak lettek, és teljesítményük nőtt, egyre nagyobb érdeklődés mutatkozott a statisztikai modellek iránt a gépi fordítás terén. Ezeknek a statisztikai modelleknek a paramétereit kétnyelvű korpuszokból lehet kiszámítani, mint amilyen az Europarl párhuzamos korpusz, amely az Európai Parlament jegyzőkönyveit tartalmazza 11 európai nyelven. Kellő mennyiségű adat birtokában a statisztikai alapú gépi fordítás elég jó becslést tud adni egy idegen nyelvű szöveg jelentéséről. Azonban a

szabályalapú rendszerekkel ellentétben a statisztikai (más néven adatvezérelt) gépi fordítók gyakran agrammatikus kimenetet produkálnak. Másrészről az adatvezérelt rendszereknek több előnyük is van: amellett, hogy kevesebb emberi munkát igényelnek, a nyelv olyan különlegességeit is tudják kezelni (pl. az idiomatikus kifejezéseket), amilyeneket a szabályalapúak nem.

Mivel a szabályalapú és a statisztikai alapú módszerek erősségei és gyengeségei kiegészítik egymást, a kutatók manapság már inkább a két megközelítést ötvöző hibrid rendszereken dolgoznak. Ezt többféleképpen lehet megvalósítani. Az egyik út, amikor mindkét módszert használjuk, és minden mondatra kiválasztjuk a legjobb kimenetet. Ennél jobb megoldást ad, ha a különböző kimenetekből összeválogatjuk a legjobb mondatrészeket, ami meglehetősen komplex feladat lehet, hiszen nincs mindig egyértelmű megfelelés a mondatrészek között.

A gépi fordítás magyar piacán is léteznek szabályalapú és adatvezérelt megoldások. A MorphoLogic Kft. számítógépes gépi fordító programcsomagokat és online fordító szolgáltatást egyaránt kínál. A programjaik angol és magyar nyelv között fordítanak, mindkét irányba. Rendszerük szabályalapú és statisztikai módszereket ötvöz, de a fő komponens a fordítandó szöveghez egy belső reprezentációt rendel, majd ezt alakítja célnyelvű szöveggé.

A nagyméretű kétnyelvű szövegek kulcsfontosságúak a statisztikai alapú gépi fordításhoz. A Hunglish korpusz egy szabadon elérhető, mondatszinten párhuzamosított magyar-angol párhuzamos korpusz, amely 2,07 millió mondatban 54,2 millió szót tartalmaz. Jelenleg ez a legnagyobb magyar-angol párhuzamos korpusz. A mondatillesztés a hunalign nevű eszközzel történt, amelyet a BME kutatói fejlesztettek ki, és az egyik leggyakrabban használt mondatszintű illesztőprogram. A Hunglish mondattár egy online felületen keresztül kereshető, amely így nyersfordítóként vagy kétnyelvű szótárként is használható.

2010 márciusában indult az iTranslate4.eu projekt, melynek célja egy olyan gépi fordító szolgáltatás nyújtása, amely nemcsak lefedi az Európai Unió összes nyelvét, hanem az összes nyelvpár esetében a mindenkori legjobb minőségű fordítást is adja. Mindemellett a hivatalos fordítók felé is közvetít. Ezt a tervet az Európa legjobb gépifordító-rendszereinek működtetői által létrehozott konzorcium valósítja meg, amelynek tagjai legalább egy nyelvpár legjobb fordítását biztosítják. A projektnek két magyar résztvevője is van: a konzorciumvezető a Nyelvtudományi Intézet, míg a szolgáltatások közös interfészét a MorhoLogic nyújtja.

Nyelvtechnológia a színfalak mögött

A nyelvtechnológiai alkalmazások mögött egy sor olyan alfeladat van, amely általában nem jelenik meg a felhasználó szintjén, de fontos szerepet tölt be a rendszerben. Ezek jelentős kutatási irányokat alkotnak, és saját tudományos területet követelnek maguknak a számítógépes nyelvészeten belül.

Az egyik aktívan kutatott terület a kérdésmegválaszolás (question answering), amelyhez annotált (nyelvi információval ellátott) korpuszo-

kat építenek, és tudományos versenyeket rendeznek. A lényege, hogy a kulcsszóalapú kereséstől elmozdulva (amelynek során a keresőmotor a potenciálisan releváns dokumentumok teljes listájával tér vissza) egy olyan rendszert hozzanak létre, amelyben a felhasználó egy konkrét kérdést tehet fel, amire egy konkrét választ kap, pl.: "Hány éves volt Neil Armstrong, amikor a Holdra lépett?" – "38". Ez a kutatási terület nagyon hasonló, mint a fentebb már említett webes keresés, de a kérdésmegválaszolás inkább gyűjtőfogalma az olyan kutatási kérdéseknek, mint hogy milyen típusú kérdéseket lehet megkülönböztetni, és ezeket hogyan lehet kezelni; hogy a választ potenciálisan tartalmazó dokumentumhalmazokat hogyan lehet elemezni és összehasonlítani (mi van, ha egymásnak ellentmondó válaszokat tartalmaznak?); valamint hogy a választ hogyan lehet megbízhatóan kinyerni egy dokumentumból a kontextus figyelembevételével.

Ez pedig kapcsolódik az információkinyeréshez, amely nagyon népszerű feladat volt a számítógépes nyelvészet statisztikai fordulata idején, a kilencvenes évek elején. Az információkinyerő rendszerek célja, hogy speciális információkat hordozó egységeket azonosítsanak különböző típusú szövegekben, pl. cégfelvásárlások kulcsszereplőit felismerjék újságcikkekben. Egy másik tipikus felhasználási terület a terroristatámadásokról szóló riportokból való információkinyerés: ki volt a támadó, mi volt a támadás célpontja, ideje és helye, mi volt a következménye. A területspecifikus információkinyerés egy másik kiváló példája a háttérben működő nyelvtechnológiának, ami egy jól körülhatárolt kutatási terület, de igazán csak más alkalmazásokba építve használható.

A szövegtömörítés és a szöveggenerálás két olyan határterület, amelyek időnként önálló alkalmazásként jelennek meg, időnként viszont támogató háttérkomponensei valamely nagyobb rendszernek. A tömörítés során hosszabb szövegből készítünk rövidebb változatot. Ez a funkció már be lett építve például a Microsoft Wordbe. Nagyrészt statisztikai alapon működik: a rendszer először a fontos szavakat azonosítja a szövegben (jellemzően azok számítanak fontos szavaknak, amelyek a szövegben gyakoriak, míg általában nem), majd kiválasztja azokat a mondatokat, amelyekben sok fontos szó van. Ezekből a mondatokból épül fel a tömörített szöveg. Ebben az esetben, ami egyébként a legnépszerűbb, a tömörítés igazából mondatok kinyerésével egyenlő: a szöveg mondatainak egy részhalmazára csökken. Minden kereskedelmi forgalomban kapható tömörítő program ezen az ötleten alapul. Egy másik módszer új mondatokat hoz létre, vagyis olyanokat, amelyek ugyanebben a formában nem szerepelnek a forrásszövegben. Ez a szöveg mélyebb megértését követeli, és ezáltal nem is olyan robusztus. Mindent egybevetve egy szövegtömörítő- és generáló alkalmazás az esetek túlnyomó részében egy nagyobb szoftverkörnyezetbe építve jelenik meg, például klinikai információs rendszerekben, amelyben betegek adatait gyűjtik össze, tárolják és dolgozzák fel, és amelynek a jelentésgenerálás csak egy a sok funkciója közül.

A tulajdonnevek automatikus felismerése és klasszifikálása az egyik alapvető alkalmazás az információkinyerés területén belül. A magyarra is létezik kézzel tulajdonnév-annotált korpusz, amely több magyar nyelvű tulajdonnév-felismerő rendszer fejlesztéséhez szolgált be-

menetként. Ennek a területnek az egyik alkalmazási lehetősége az orvosi témájú szövegfeldolgozás. Az orvosi kórlapok folyó szöveges részeiből számos rejtett információ nyerhető ki, amelyekből például a gyógyszerkutatók statisztikákat, elemzéseket készíthetnek a betegekről. Ehhez viszont elengedhetetlen az orvosi adatbázisban szereplő személyes adatok anonimizálása. Ezt és ehhez hasonló feladatokat egy általános célú tulajdonnév-felismerő rendszer adaptálásával lehet megoldani.

A világban folyó biológiai célú kutatások eredményei exponenciálisan növekvő mennyiségű publikációkban öltenek testet, ennek eredményeképpen a biológiai témájú információkinyerés is egyre fontosabb. A Szegedi Tudományegyetem munkatársai kifejlesztettek egy olyan rendszert, amely biológiai kifejezéseket egyértelműsít a szövegben, valamint könnyen olvasható kimenetben rendszerezi azokat.

A magyar nyelvű kérdésmegválaszolás és szöveggenerálás sokkal kevésbé fejlett, mint az angol nyelv esetében, ahol ezeken a kutatási területeken a kilencvenes évek óta rendszeresen tudományos versenyeket rendeznek, elsősorban az amerikai DARPA/NIST támogatásával. Ezek a versenyek nagy mértékben elősegítették a fejlődést, de mindig csak az angolra koncentráltak. Néhány versenyen többnyelvű feladatok is voltak, de a magyar ezekben soha nem szerepelt. A szövegtömörítő rendszerek általában tisztán statisztikai alapon működnek, amelyek nyelvfüggetlenek, de ezekből csak néhány prototípus érhető el. A szöveggeneráló modulok viszont nyelvfüggők, és szintén leginkább csak az angolra működnek.

Nyelvtechnológia az oktatásban

A nyelvtechnológia tipikus interdiszciplináris terület: nyelvészeti, számítástechnikai, matematikai, filozófiai, pszicholingvisztikai és idegtudományi szakértelmet egyaránt kíván. Valószínűleg emiatt még nem találta meg a helyét a magyar oktatási rendszerben – Magyarországon egy egyetemen sincs számítógépes nyelvészeti tanszék. Nyelvtechnológiai oktatás azért folyik néhány kapcsolódó tanszéken. Pár egyetemen az alap- vagy a mesterképzés szintjén tartanak kurzusokat a témában, máshol nyelvtechnológia modulokat is kialakítottak egyéb szakokon, főként a nyelvészeten belül. Ám ezek a kurzusok és modulok sem rendelkeznek nagy múlttal: csak az elmúlt néhány évben indultak. Jelenleg hat magyarországi egyetemen folyik valamilyen formában nyelvtechnológia-oktatás.

A hazai felsőoktatásban, az utóbbi évek jelentős erőfeszítéseinek ellenére, a jövendő nemzedékek nyelvtechnológusainak oktatása ma még közel sem áll a megfelelő szinten. A magyarországi nyelvtechnológus közösség célja egy az általános európai rendszerbe illeszkedő BA/BSc-MA/MSc-PhD szekvencia tantervének kidolgozása. További problémát jelent a fiatal kutatók alacsony fizetése, és részben emiatti elvándorlása a szakmából. Számukra versenyképes ösztöndíjakat kéne létesíteni, valamint az ipar és az oktatási intézmények közötti kapcsolat megerősítésének keretében lehetővé kellene tenni képzésük egy részének kihelyezését ipari szereplőkhöz.

Nyelvtechnológiai programok

Más országokhoz hasonlóan a magyarországi természetesnyelv-feldolgozás kezdetei is a gépi fordításhoz kapcsolódnak. Az első próbálkozások a hatvanas években zajlottak – akkor még oroszról magyarra fordítottak. A hetvenes-nyolcvanas években a lexikográfusi munka adta a következő lökést: ez vezetett az első magyar morfológiai rendszer kifejlesztéséhez. Ezekben az években nem voltak szervezett nemzeti keretprogramok, továbbá Magyarország az európai támogatási lehetőségektől is el volt zárva.

A rendszerváltás után, a kilencvenes években egymás után alakultak a szakterületen egyetemi tanszékek (pl. a Szegedi Tudományegyetem Nyelvtechnológiai Csoportja), illetve kutatóintézeti osztályok (pl. a Nyelvtudományi Intézet Korpusznyelvészeti Osztálya). Az elmúlt tíz évben az európai és a nemzeti finanszírozású projektek száma nagy mértékben megemelkedett, ez utóbbiakat elsősorban a Nemzeti Kutatási és Technológiai Hivatal (NKTH) és az Oktatási Minisztérium támogatta.

Ezek következményeképpen az elmúlt évtizedben a magyar kutatók szép számú adatbázist (korpuszokat, szótárakat, lexikai adatbázisokat) és szövegfeldolgozó eszközt (helyesírás-ellenőrzőket, morfológiai elemzőket stb.) fejlesztettek ki. A különböző műhelyek sokáig elszigetelten működtek, ezért fordulhatott elő, hogy egymástól függetlenül hasonló eszközöket hoztak létre (pl. magyar morfológiai elemzőből legalább három van). Ezek általában össze nem egyeztethető formátumúak, nem követnek egy sztenderdet, továbbá hiányos a dokumentációjuk, és sok esetben a jogi státuszuk is tisztázatlan. Mindezek ellenére – vagy éppen ezért – az elmúlt pár évben Magyarországot is elérte a sztenderdizálás és egységesítés nemzetközi trendje. Több, az integrációt és interoperabilitást célul tűző projekt is indult, például a magyar egységes ontológia megépítését, vagy a morfológiai elemzők különböző kódolási rendszerének harmonizálását célzó projektek.

2008-ban élenjáró magyarországi kutató-fejlesztő közösségek létrehozták a Nyelv- és Beszédtechnológiai Platformot azzal a céllal, hogy összehangolt munkával erősítsék és elősegítsék az innovációt a nyelv- és beszédtechnológia területén. A Platform hivatalos keretet nyújtva összefogja a jelentősebb hazai nyelv- és beszédtechnológiai kutatás-fejlesztést végző központokat, és ezáltal elősegíti az eddig viszonylagos elszigeteltségben működő központokban felhalmozódott magas szintű tudás megosztását, illetve integrációját; részletes stratégiai és arra épülő megvalósítási terveket dolgoz ki; közvetíti az informatikai szektor érdekelt résztvevői felé a Platform elemzéseit, stratégiáit, javaslatait; megjeleníti és képviseli a magyar szempontokat és érdekeket a nemzetközi színtéren; és elősegíti a Platform eredményeinek tudatosítását a magyar gazdaság potenciális felhasználói felé, különös tekintettel a kis- és középvállalkozásokra. A Platform egyes résztvevői részt vesznek a CLARIN projektben is.

A magyar eszközök és erőforrások állapota

A következő táblázat összegzi a magyar nyelvtechnológia jelenlegi állapotát. Az egyes technológiák és erőforrások értékelése vezető sza-

kértők becslése alapján készült, az alábbi kritériumok alapján (minden értékelés 0-tól 6-ig terjed).

1. **Mennyiség**: Létezik az adott eszköz/erőforrás a nyelvre? Minél több van, annál magasabb az értékelés.
   - 0: egyáltalán nincs eszköz/erőforrás
   - 6: sok eszköz/erőforrás létezik, nagy változatosság
2. **Hozzáférhetőség**: Elérhetőek az eszközök/erőforrások, vagyis nyílt forráskódúak, ingyenesen használhatóak bármely platformon, vagy kizárólag magas áron vagy szigorú körülmények között érhetőek el?
   - 0: lényegében minden eszköz/erőforrás csak magas áron érhető el
   - 6: a technológiák/erőforrások nagy része ingyenesen, szabadon hozzáférhető Open Source vagy Creative Commons licenszek alatt, melyek engedélyezik az újrafelhasználást
3. **Min□ség**: Az adott technológiák/erőforrások mennyire közelítik meg az elérhető legjobb technológiák/erőforrások teljesítményét? Aktív fejlesztés alatt állnak ezek a technológiák/erőforrások?
   - 0: játékrendszer/-erőforrás
   - 6: kiváló minőségű technológia, emberi minőségű annotáció az erőforrásban
4. **Lefedettség**: Milyen szinten valósítják meg a legjobb technológiák a legújabb lefedettségi követelményeket (stílus, műfaj, nyelvészeti jelenségek, bemeneti/kimeneti formátumok, gépi fordító esetében a támogatott nyelvek száma stb.)? Az erőforrások mennyire reprezentatívak a célnyelvek tekintetében?
   - 0: speciális célú erőforrás vagy technológia, specifikus célokra, nagyon kicsi lefedettséggel, csak specifikus, nem általános célra használható
   - 6: széles lefedettségű erőforrás, nagyon robusztus technológia, széles körben alkalmazható, több nyelvet támogat
5. **Fejlettség**: Mennyire fejlett, stabil, piackész a technológia/erőforrás? Az elérhető legjobb technológiák/erőforrások termékkészek, vagy adaptációt igényelnek? Készek ezek a technológiák/erőforrások a gyártásra, vagy csak prototípusszinten vannak? Ez az indikátor azt is jelzi, hogy az adott technológiák/erőforrások mennyire elismertek a közösségen belül, és mennyire használhatóak más nyelvtechnológiai rendszerekben.
   - 0: előzetes prototípus, fejlesztés alatt, játékrendszer, csak szemelvények léteznek az erőforrásból
   - 6: azonnal alkalmazható, integrálható komponens
6. **Fenntarthatóság**: Mennyire jól integrálható a technológia/erőforrás a már meglévő rendszerekbe? Elér a technológia/erőforrás egy bizonyos fenntarthatósági szintet a

dokumentáció, használati utasítás, grafikus felhasználói felület stb. terén? Használja az általánosan elfogadott/sztenderd programozói környezeteket (pl. Java EE)? Léteznek ipari/kutatói sztenderdek, és ha igen, a technológia/erőforrás mennyire követi azokat (pl. adatformátum tekintetében)?

- 0: teljesen esetleges, ad hoc adatformátum és API
- 6: teljesen sztenderdkövető, jól dokumentált

7. **Alkalmazhatóság**: Mennyire alkalmazható/adaptálható/terjeszthető ki új feladatokra/területekre/műfajokra/szövegtípusokra/esetekre stb. a technológia/erőforrás?
   - 0: gyakorlatilag lehetetlen a technológia/erőforrás adaptálása más feladatra, lehetetlen/rengeteg emberi munkát igényel újabb erőforrások hozzáadása
   - 6: az alkalmazhatóság magas foka, nagyon könnyen és hatékonyan adaptálható

Az eszközök és erőforrások táblázata

	Mennyiség	Hozzáférhetőség	Minőség	Lefedettség	Fejlettség	Fenntarthatóság	Alkalmazhatóság
Nyelvtechnológia (Eszközök, Technológiák, Alkalmazások)							
Tokenizálás, Morfológia (tokenizálás, szófaji címkézés, morfológiai elemzés/generálás)	6	2	4	5	5	1	5
Elemzés (felszíni vagy mély szintaktikai elemzés)	3	2	4	4	3	5	4
Mondatszemantika (szójelentés-egyértelműsítés, argumentumstruktúra, szemantikai szerepek)	1	3	3	1	0	0	3
Szövegszemantika (koreferenciafeloldás, kontextus, pragmatika, következtetés)	1	3	2	0	0	0	3
Diskurzuselemzés (szövegstruktúra, koherencia, retorikai szerkezet, retorikai elemzés, érvelési szerkezet, szövegmintázatok, szövegtípusok stb.)	0	0	0	0	0	0	0
Információ-visszakeresés (szövegindexelés, multimédiás és nyelvközi információ-visszakeresés)	1	0	2	1	0	1	1
Információkinyerés (tulajdonnév-felismerés, esemény-/relációkinyerés, vélemény-/érzelemdetekció, szövegbányászat)	6	6	6	6	5	5	5
Nyelvgenerálás (mondatgenerálás, jelentésgenerálás, szöveggenerálás)	0	0	0	0	0	0	0
Tömörítés, Kérdésmegválaszolás, magasabb szintű információelérési technológiák	0	0	0	0	0	0	0
Gépi fordítás	6	1	5	5	6	5	6
Beszédfelismerés	3	0	4	2	4	3	3
Beszédszintézis	4	3	4	4	5	3	3
Dialógusvezérlés (dialóguslehetőségek és felhasználómodellezés)	0	0	0	0	0	0	0
Nyelvi Erőforrások (Erőforrások, Adatok, Tudásbázisok)							
Referenciakorpuszok	6	6	6	6	6	6	4
Szintaktikai korpuszok (treebankek, függőségi fák)	1	6	6	5	6	6	4
Szemantikai korpuszok	3	6	6	1	3	5	5
Diskurzuskorpuszok	0	0	0	0	0	0	0
Párhuzamos korpuszok, Fordítómemóriák	6	4	6	6	6	6	6
Beszédkorpuszok (nyers beszélt nyelvi adatok, címkézett/annotált beszélt nyelvi adatok, beszélt nyelvi dialóguskorpuszok)	2	2	4	2	4	4	0
Multimédia- és multimodális adatok (audió/videó adatokkal kombinált szöveges adatok)	1	0	1	1	1	0	0
Nyelvmodellek	6	3	4	3	6	6	5
Lexikonok, Terminológia-adatbázisok	5	1	6	6	2	2	6
Nyelvtanok	3	3	6	5	6	4	3
Tezauruszok, WordNetek	1	1	6	3	5	5	3
Ontológiai erőforrások (pl. felső szintű ontológiák, linkelt adatok)	2	6	1	1	1	4	2

Összegzés

A magyarországi nyelvtechnológiai helyzet az elmúlt pár év alapján óvatos optimizmusra ad okot. Nagyrészt állami támogatással, de létezik nyelvtechnológiai kutatás Magyarországon. A magyar természetesnyelv-feldolgozás piacát elsősorban egyetemi kutatócsoportok és akadémiai intézetek uralják, de mellettük kisebb cégek is vannak a piacon.

Számos technológia és erőforrás áll rendelkezésre a magyar nyelvre, bár közel sem annyi, mint az angolra. A magyar nyelvtechnológia különleges helyzetben van: egyrészt a nemzetközi, angolközpontú trendeket követi, másrészt a magyar nyelv speciális jellege miatt új módszereket kell kifejlesztenie.

A fehér könyvek sorozatával megtörtént az első lépés afelé, hogy a hiányosságokat és az igényeket egyaránt feltáró átfogó felmérést készítsünk az európai nyelvek helyzetéről, különös tekintettel a nyelvtechnológiára.

A magyar nyelvet illetően a technológiákat és erőforrásokat érintő kulcsfontosságú eredmények a következők:

- Létezik ugyan néhány kiváló minőségű specifikus korpusz, de nagyon nagy méretű szintaktikailag annotált korpusz nincs.
- Létezik egy manuálisan szintaktikailag annotált korpusz a magyarra, ami ingyenesen hozzáférhető, és ami számos alkalmazás alapjául szolgált már, de ennek nem elég nagy a mérete.
- Az erőforrások nagy része nem sztenderdizált; létrehozásukkor a fenntarthatóság nem szerepelt a tervek között. Szervezett programok keretében, a megfelelő előírásokat követve sztenderdizálni kellene a meglévő adatbázisokat.
- Minél magasabb szintű a nyelvfeldolgozás, annál nehezebb: a szemantikai feladatok bonyolultabbak, mint a szintaktikaiak; a szövegszintű szemantika bonyolultabb, mint a szó- vagy mondatszintű.
- Minél több szemantikát alkalmaz egy eszköz, annál nehezebb a fejlesztéséhez megfelelő adatokat találni, továbbá a mély elemzés több erőfeszítést igényel.
- A világról való tudásunkat leképező tudásbázisokhoz szükséges szemantikai sztenderdek (RDF, OWL stb.) léteznek ugyan, de nehezen alkalmazhatók a természetesnyelv-feldolgozási feladatokra.
- A sztenderd előfeldolgozó lépések (tokenizálás, morfológiai elemzés, felszíni szintaktikai elemzés stb.) már megoldottnak tekinthetők a magyarra, de a bonyolultabb szemantikai feldolgozás és diskurzuselemzés fejlesztése még folyamatban van.
- A magyarországi kutatás sikeresnek mondható a jó minőségű egyedi szoftverek fejlesztésében, de a jelenlegi kutatásfinanszírozási helyzetben szinte lehetetlen sztenderdizált és fenntartható megoldásokkal előállni.

- A magyar nyelvű eszközök és erőforrások értékelésében elég nagy szórás tapasztalható: bizonyos területeken (pl. morfológia, információkinyerés, gépi fordítás, párhuzamos korpuszok) magasabbak a pontszámok, másokon viszont (pl. diskurzusfeldolgozás, dialóguselemzés) a nullához közelítenek.
- Magyarországon beszédfelismeréssel és gépi fordítással számos kutatóműhelyben foglalkoznak, mégis alig van szabadon használható eszköz és erőforrás. Ez a jelenség elég tipikus a magyar nyelvtechnológiában: a nyílt forráskódú programok és a szabadon felhasználható adatbázisok száma – néhány üdítő kivételtől eltekintve – meglehetősen alacsony.

A fentiekből világossá válik, hogy a magyarországi kutatásfejlesztéshez, innovációhoz, a magyar nyelvű eszközök és erőforrások előállításához még több támogatás szükséges. A nagy mennyiségű adatra való igény és a nyelvtechnológiai rendszerek magasfokú komplexitása kötelezővé teszi az együttműködéshez szükséges közös infrastruktúra megteremtését.

A META-NET-ről

A META-NET az Európai Bizottság által alapított hálózat, amelynek jelenleg 47 tagja van 31 országból. A META-NET támogatja a META-t (Multilingual Europe Technology Alliance), amely az európai nyelvtechnológiával foglalkozó szakértők és intézmények egyre növekvő közössége.

META – The Multilingual Europe Technology Alliance

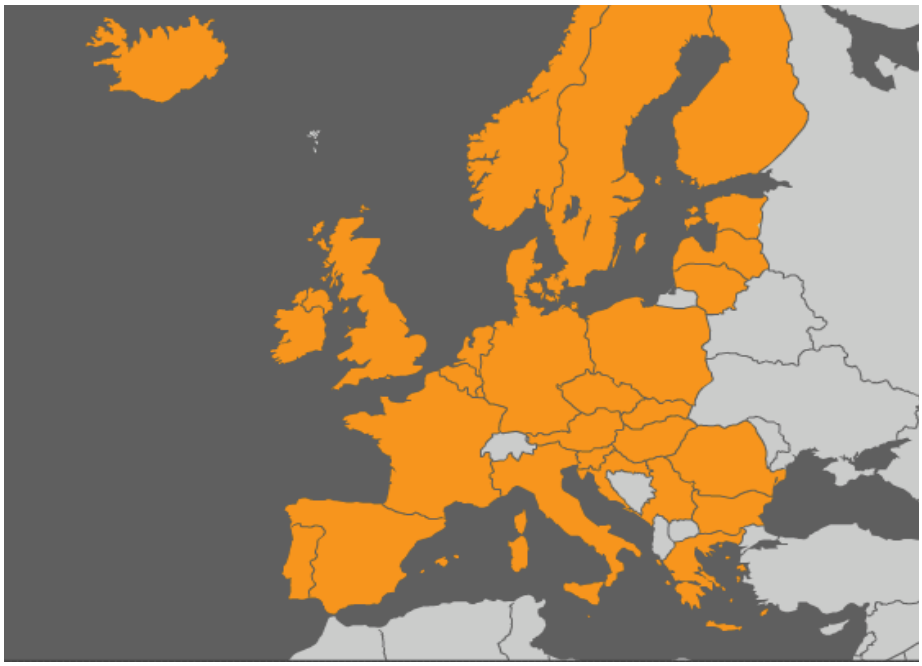


Figure 1: Countries Represented in META-NET

A META-NET olyan kezdeményezésekkel is együtt dolgozik, mint a Common Language Resources and Technology Infrastructure (CLARIN), amely segít megalapozni a digitális bölcsészeti kutatást Európában. A META-NET elősegíti a technológiai alapok létrehozását és fenntartását a többnyelvű európai információs társadalom számára, amely:

- megvalósítja a különböző nyelveken történő kommunikációt és együttműködést;
- minden nyelvhasználó számára biztosítja az információhoz és tudáshoz való hozzáférést;
- felhasználja, valamint fejleszti a hálózati információs technológiát.

A META-NET támogatja a többnyelvű technológiákat minden európai nyelvre. Ezek a technológiák az alkalmazások széles körében elérhetővé teszik az automatikus fordítást, az információfeldolgozást és a tudásmenedzsmentet. A hálózat célja, hogy fejlessze a jelenlegi módszereket, ezáltal jobb kommunikáció és együttműködés alakulhat ki a nyelvek között. Az európaiaknak a nyelvtől függetlenül egyenlő joguk van az információhoz és tudáshoz való hozzáféréshez.

A tevékenység irányvonalai

A META-NET 2010. február 1-én alakult azzal a céllal, hogy fejlessze a kutatást a nyelvtechnológia területén. A hálózat egy olyan Európát támogat, amely egységes digitális piacként és információs térként működik. A META-NET számos tevékenységet végez céljai elérése érdekében. A META-VISION, a META-SHARE és a META-RESEARCH alkotják a hálózat tevékenységének három fő irányvonalát.

A **META-VISION** támogatja egy olyan dinamikus és befolyásos döntéshozó közösség létrejöttét, amely egy közös vízió és az arra épülő stratégiai kutatási terv köré szerveződik. Ezen tevékenység fő célja, hogy összetartó és egységes nyelvtechnológiai közösséget alakítson ki Európában, azáltal, hogy elősegíti a döntéshozók széttöredezett és elszigetelt csoportjainak találkozását. A META-NET első évében szervezett találkozók elsősorban a partnerkeresésre, terjeszkedésre szolgáltak: FLaReNet Forum (Spanyolország), Language Technology Days (Luxembourg), JIAMCATT 2010 (Luxembourg), LREC 2010 (Málta), EAMT 2010 (Franciaország) és ICT 2010 (Belgium). Előzetes becslések alapján a META-NET eddig több mint 2500 nyelvtechnológus szakértővel vette fel a kapcsolatot a közös célok és víziók kifejlesztése érdekében. A META-FORUM 2010 eseményen Brüsszelben a META-NET több mint 250 résztvevő előtt publikálta jövőtervezési munkájának első eredményeit, melyre a résztvevőktől visszajelzést is kaptak az interaktív szekciók során.

A **META-SHARE** célja egy nyílt rendszer létrehozása, amely lehetővé teszi a nyelvi erőforrások megosztását. Az ún. peer-to-peer hálózat nyelvi adatokat, eszközöket és webes szolgáltatásokat fog tartalmazni, amelyek metaadatokkal lesznek ellátva, és sztenderdizált kategóriákba lesznek rendezve. Az erőforrások könnyen hozzáférhetőek és egységesen kereshetőek lesznek. Az elérhető erőforrások között találunk ingyenes, nyílt forráskódú eszközöket és kereskedelmi forgalomban kapható, fizetős szolgáltatásokat is. A META-SHARE a már meglévő nyelvi adatok, eszközök és rendszerek mellett olyan új, fejlesztés alatt álló termékeket is megcéloz, amelyek új technológiák, termékek és szolgáltatások kifejlesztéséhez vagy kiértékeléséhez szükségesek. A nyelvi adatok és eszközök újrafelhasználhatósága, kombinációja és újratervezése különösen fontos szerepet játszanak. A META-SHARE végül a fejlesztők, a lokalizálási szakemberek, a kutatók, a fordítók és a nyelvi szakértők számára egyaránt kritikus szerepet fog betölteni a nyelvtechnológiai piacon, a kis- és középvállalkozásoktól egészen a nagyvállalatokig. A META-SHARE felöleli az egész nyelvtechnológiai fejlesztési kört – a kutatástól egészen az innovatív termékekig és szolgáltatásokig. Ezen tevékenység

kulcsfontosságú szempontja, hogy a META-SHARE az európai és globális nyelvtechnológiai közösség fontos és értékes részévé váljon.

A **META-RESEARCH** hidakat épít a kapcsolódó technológiai területek között. Ez az irányvonal más területek fejlesztési eredményeit próbálja meg átemelni a nyelvtechnológiába. A gépi fordítás esetében például ezáltal több szemantikát lehetne belevinni a rendszerbe, optimizálni lehetne a munkamegosztást a szabályalapú és a statisztikai komponensek között, valamint ki lehetne terjeszteni a kontextust a célnyelvi megfelelő előállításához. De a META-RESEARCH más területekkel és tudományágakkal is foglalkozik, mint például a gépi tanulás és a szemantikus web. A META-RESEARCH az adatgyűjtésre, adatelőkészítésre fókuszál, valamint nyelvi erőforrásokat állít elő az eszközök kiértékeléséhez; elkészíti az eszközök és módszerek leltárját; valamint workshopokat és tréningeket szervez a közösség tagjainak. Továbbá ajánlások készültek arról, hogy hogyan lehet a szemantikai információt integrálni a gépi fordításba. A META-RESEARCH egy új nyelvi erőforrást is épített, az Annotated Hybrid Sample MT Corpust, amely angol-német, angol-spanyol és angol-cseh nyelvpárokra szolgáltat adatot. A META-RESEARCH emellett kifejlesztett egy szoftvert, amely többnyelvű korpuszt tud gyűjteni a webről.

Tagszervezetek

Country	Member (Affiliation)	Contacts
Austria	Universität Wien	Gerhard Budin
Belgium	University of Antwerp	Walter Daelemans
	University of Leuven	Dirk van Compernolle
Bulgaria	Bulgarian Academy of Sciences	Svetla Koeva
Croatia	Zagreb University	Marko Tadic
Cyprus	University of Cyprus	Jack Burston
Czech Rep.	Charles University in Prague	Jan Hajic
Denmark	University of Copenhagen	Bente Maegaard
Estonia	University of Tartu	Tiit Roosmaa
Finland	Aalto University	Timo Honkela
	University of Helsinki	Kimmo Koskenniemi, Krister Linden
France	CNRS, LIMSI	Joseph Mariani
	ELDA	Khalid Choukri
Germany	DFKI	Hans Uszkoreit, Georg Rehm
	RWTH Aachen	Hermann Ney
Greece	ILSP, R.C. “Athena”	Stelios Piperidis
Hungary	Hungarian Academy of Sciences	Tamás Váradi
	Budapest Technical University	Géza Németh, Gábor Olaszy
Iceland	University of Iceland	Eirikur Rögnvaldsson
Ireland	Dublin City University	Josef van Genabith
Italy	Consiglio Nazionale Ricerche	Nicoletta Calzolari
	Fondazione Bruno Kessler	Bernardo Magnini
Latvia	Tilde	Andrejs Vasiljevs
	University of Latvia	Inguna Skadina
Lithuania	Institute of the Lithuanian Language	Jolanta Zabarskaitė
Luxembourg	Arax Ltd.	Vartkes Goetcherian
Malta	University of Malta	Mike Rosner
Netherlands	Universiteit Utrecht	Jan Odijk
Norway	University of Bergen	Koenraad De Smedt

Poland	Polish Academy of Sciences	Adam Przepiórkowski
	University of Lódz	Barbara L.-Tomaszczyk
Portugal	University of Lisbon	Antonio Branco
	Inst. for Systems Engineering and Computers	Isabel Trancoso
Romania	Romanian Academy of Sciences	Dan Tufis
	University Alexandru Ioan Cuza	Dan Cristea
Serbia	Belgrade University	Dusko Vitas, Cvetana Krstev
	Pupin Institute	Sanja Vranes
Slovakia	Slovak Academy of Sciences	Radovan Garabik
Slovenia	Jozef Stefan Institute	Marko Grobelnik
Spain	Barcelona Media	Toni Badia
	Technical University of Catalonia	Asunción Moreno
	University Pompeu Fabra	Núria Bel
Sweden	University of Gothenburg	Lars Borin
UK	University of Manchester	Sophia Ananiandou