META NET

White Paper Series Fehér könyvek sorozat

THE HUNGARIAN LANGUAGE IN THE DIGITAL AGE

A MAGYAR NYELV A DIGITÁLIS KORBAN

Simon Eszter
Lendvai Piroska
Németh Géza
Olaszy Gábor
Vicsi Klára

Springer

White Paper Series Fehér könyvek sorozat

THE HUNGARIAN LANGUAGE IN THE DIGITAL AGE

A MAGYAR NYELV A DIGITÁLIS KORBAN

Simon Eszter MTA Nyelvtudományi Intézet
Lendvai Piroska MTA Nyelvtudományi Intézet
Németh Géza BME
Olaszy Gábor BME
Vicsi Klára BME

Georg Rehm, Hans Uszkoreit
(szerkesztők, editors)

ELŐSZÓ PREFACE

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. Ezért a nyelvtechnológia fejlődéséhez és a kutatás elősegítéséhez szükséges lépések is nyelvenként mások és mások, és olyan különféle tényezőkön múlnak, mint az adott nyelv összetettsége, vagy a nyelvet használó közösség nagysága.

A META-NET, az Európai Bizottság által alapított hálózat felmérést végzett a rendelkezésre álló nyelvi erőforrásokról és technológiákról (lásd a 73. oldalt). 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 a további kutatások hatásának maximalizálásában.

A META-NET 33 ország 54 kutatóközpontjából áll (2011. novemberi helyzet szerint, lásd a 69. oldalt), akik a területtel foglalkozó vállalkozásokkal, kormányzati szervekkel, kutatószervezetekkel, szoftvercégekkel, szolgáltatókkal és európai egyetemekkel dolgoznak együtt. Egységes technológiai víziót alkotva egy olyan stratégiai kutatási terv létrehozásán dolgoznak, amelyben megfogalmazzák, hogyan tudnak a nyelvtechnológiai alkalmazások a kutatási hiányosságokon enyhíteni a 2020-ig terjedő időszakban.

This white paper is part of a series that promotes knowledge about language technology and its potential. It addresses journalists, politicians, language communities, educators and others. The availability and use of language technology in Europe varies between languages. Consequently, the actions that are required to further support research and development of language technologies also differ. The required actions depend on many factors, such as the complexity of a given language and the size of its community.

META-NET, a Network of Excellence funded by the European Commission, has conducted an analysis of current language resources and technologies in this white paper series (p. 73). The analysis focused on the 23 official European languages as well as other important national and regional languages in Europe. The results of this analysis suggest that there are tremendous deficits in technology support and significant research gaps for each language. The given detailed expert analysis and assessment of the current situation will help maximise the impact of additional research.

As of November 2011, META-NET consists of 54 research centres from 33 European countries (p. 69). META-NET is working with stakeholders from economy (software companies, technology providers, users), government agencies, research organisations, non-governmental organisations, language communities and European universities. Together with these communities, META-NET is creating a common technology vision and strategic research agenda for multilingual Europe 2020.

A dokumentum szerzői köszönettel tartoznak a német fehér könyv szerzőinek azért, hogy engedélyezték a német változat egyes nyelvfüggetlen részeinek újrafelhasználását [1].

A fehér könyv megírását az Európai Bizottság 7. keretprogramja és ICT PSP programja támogatta a T4ME (szerződésszám: 249 119), a CESAR (szerződésszám: 271 022), a METANET4U (szerződésszám: 270 893) és a META-NORD (szerződésszám: 270 899) projekteken keresztül.

The authors of this document are grateful to the authors of the white paper on German for permission to re-use selected language-independent materials from their document [1].

The development of this white paper has been funded by the Seventh Framework Programme and the ICT Policy Support Programme of the European Commission under the contracts T4ME (Grant Agreement 249 119), CESAR (Grant Agreement 271 022), METANET4U (Grant Agreement 270 893) and META-NORD (Grant Agreement 270 899).

TARTALOMJEGYZÉK TABLE OF CONTENTS

A MAGYAR NYELV A DIGITÁLIS KORBAN

THE HUNGARIAN LANGUAGE IN THE DIGITAL AGE

1

VEZETŐI ÖSSZEFOGLALÓ

Az információs technológia jelentősen megváltoztatta mindennapi életünket. Jellemzően számítógépet használunk írásra, szerkesztésre, számolásra és információkeresésre, továbbá egyre inkább olvasásra, zenehallgatásra, fotó- és filmnézésre. Kis számítógépeket hordunk a zsebünkben, és használjuk őket telefonálásra, e-mailírásra, információszerzésre és szórakozásra, bármerre is járunk. Milyen hatással van az információnak, a tudásnak és a mindennapi kommunikációnak ez a masszív digitalizálódása a nyelvünkre? Megváltozik a nyelvünk, vagy eltűnik? Számítógépeink egy nagy és erőteljes globális hálózat részeit képezik. A lány Ipanemában, a hivatalnok Budapesten és a mérnök Delhiben ugyanúgy tudnak csetelni a barátaikkal a Facebookon, de nem valószínű, hogy valaha is találkoznak egymással online közösségekben vagy fórumokon. Ha azon aggódnak, hogy hogy lehet kezelni a fülfájást, akkor mindannyian a Wikipédiát fogják megnézni, de nem ugyanazt a cikket fogják olvasni. Amikor európai netezők a fukushimai atomkatasztrófának az európai energetikai piacra gyakorolt hatásairól beszélgetnek a fórumokon, akkor ezt jellemzően nyelvileg szeparált közösségekben teszik. Amit az internet összeköt, azt a használók nyelvi korlátai még mindig szétválasztják. Ez vajon mindig így lesz? A világ 6000 nyelve közül sok nem fog életben maradni egy globalizált digitális információs társadalomban. Becslések szerint legalább 2000 nyelv kihalásra van ítélve a következő évszázadokban. Mások szerepet fognak ugyan játszani a családi és ismerősi körben, de a szélesebb üzleti és tudományos szférában nem. Milyenek a magyar nyelv esélyei a túlélésre?

Becslések szerint a magyar nyelvet összesen 13 millióan beszélik, ezzel a 12. helyen áll a legtöbb beszélővel rendelkező nyelvek listáján Európában. A Magyar Köztársaság államnyelve, ahol a 10 milliós lakosságnak kb. 97%-a magyar anyanyelvű. A szomszédos hét országban is találunk magyar nyelvű közösségeket, amelyek közül a legnagyobb a romániai, megközelítőleg másfél millió nyelvhasználóval. Ezen felül emigráns közösségek használják világszerte, elsősorban az Amerikai Egyesült Államokban, Kanadában és Izraelben.

A magyar nyelv szigetet alkot Európában, ugyanis a legtöbb európai nyelv az indoeurópai nyelvcsaládba tartozik, de a magyar nem. A magyar finnugor nyelv, rokonai a finn, az észt és néhány más, Oroszországban élő népek által beszélt nyelv. A magyar nyelv a legtöbb beszélővel rendelkező nem indoeurópai nyelv Európában, de – ellentétben olyan világnyelvekkel, mint az angol vagy a kínai, és az olyan gyakran használt európai nyelvekkel, mint a német vagy a francia – a magyar nem játszik prominens szerepet nemzetközi szinten.

Sokan panaszkodnak Magyarországon az anglicizmusok egyre erősödő használata miatt, és attól tartanak, hogy a magyar nyelvet elárasztják az angol szavak és kifejezések. Ez a megközelítés félrevezető. A magyar nyelv már túlélte az új szavak hatását, melyeket különböző török nyelvekből vettünk át a honfoglalás előtti korban, és túlélte az erős szláv hatást is a Kárpát-medencében. Később az Oszmán Birodalom része volt az ország 150 évig, majd a Habsburg Birodalom ideje alatt a latin és a német nyelv hatása volt nagyon erős. Kedves kis magyar szavaink elvesztésének egy jó ellenszere, ha használjuk

őket – gyakran és tudatosan. Nyelvészeti polémiák az idegen befolyásról és kormányzati rendelkezések nem segítenek. Nem a nyelvünk fokozatos anglicizálódása miatt kellene aggódnunk, inkább amiatt, hogy nyelvünk a személyes élet főbb területeiről eltűnhet. Nem a tudományra, a légiirányításra vagy a globális üzleti piacra gondolunk, hanem az élet olyan területeire, ahol sokkal fontosabb, hogy a nyelv közel álljon az ország lakóihoz, mint nemzetközi partnerekhez, ilyenek például a helyi szokások, az ügyintézés, a törvényalkotás, a kultúra és a vásárlás.

Egy nyelv státusza nem csak beszélőinek számától függ, hanem attól is, hogy mennyire van jelen az információs térben és a szoftveralkalmazásokban. Egy meglehetősen aktív magyar nyelvű webes közösség létezéséről tanúskodik az, 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. Néhány fontos nemzetközi szoftver magyar változatban is elérhető, azonban a magyar nyelv specialitása megnehezíti az angolalapú alkalmazások adaptálását. A költséges magyar nyelvű technológiák fejlesztését az is hátráltatja, hogy a magyar piac meglehetősen kicsi.

Ami a nyelvtechnológiát illeti, a magyarországi helyzet óvatos optimizmusra ad okot. Nagyrészt állami, az utóbbi időben európai támogatással, de létezik nyelvtechnológiai kutatás Magyarországon. A jelenleg futó két európai, Magyarország által koordinált ICT projekt közül mindkettő nyelvtechnológiai témájú. Számos technológia és erőforrás áll rendelkezésre a magyar nyelvre, bár közel sem annyi, mint az angolra, és ezek nem elégségesek egy valódi többnyelvű tudásalapú társadalom igényeinek kielégítésére.

Az információs és kommunikációs technológia már a következő forradalomra készül. A személyi számítógépek, hálózatok, miniatürizáció, multimédia, mobileszközök és a *cloud* technológia után a következő generációba olyan szoftverek fognak tartozni, amelyek nem csak a kiejtett hangokat vagy a leírt betűket, hanem a szavakat és a mondatokat is értik, és a felhasználókat jobban támogatják azáltal, hogy beszélik és értik a nyelvüket. Ezeknek a fejlesztéseknek az előfutárai az olyan szabadon elérhető online szolgáltatások, mint a Google Translate, amely 57 nyelv között fordít, vagy európai versenytársa, az itranslate4.eu (egy Magyarország által vezetett konzorcium terméke), továbbá az IBM szuperszámítógépe, Watson, amely megverte a Jeopardy nevű játék amerikai bajnokát, és az Apple mobilalkalmazása, a Siri, amely reagál a hangvezérlésre, és válaszol angol, német, francia és japán nyelven.

Az információs technológia következő generációja olyan szintű nyelvi képességgel fog rendelkezni, hogy az emberi felhasználók saját nyelvükön tudnak majd kommunikálni ezt a technológiát használva. Az eszközök képesek lesznek automatikusan megtalálni a legfontosabb híreket és információkat a világ digitális tudásbázisából – mindezt könnyen használható hangvezérléssel. A nyelvtechnológia képes lesz automatikusan fordítani, vagy segíteni a tolmácsok munkáját; beszélgetéseket és dokumentumokat összefoglalni; és támogatni a felhasználókat a tanulásban.

Az információs és kommunikációs technológia következő generációja eléri az ipari robotokat is (jelenleg fejlesztés alatt áll a kutatólaboratóriumokban), melynek következményeképpen érteni fogják, hogy a felhasználó mit akar, és még be is számolnak az eredményekről.

A teljesítménynek ez a szintje azt jelenti, hogy túl kell lépni az egyszerű karakterhalmazokon, szótárakon, helyesírás-ellenőrzőkön és kiejtési szabályokon. A technológiának meg kell haladnia az egyszerűsítő megközelítéseket, és el kell kezdenie olyan nyelvmodelleket gyártani, amelyek figyelembe veszik a szintaxissal együtt a szemantikát is ahhoz, hogy megértsék a kérdések sorát, és releváns válaszokat adjanak rájuk.

Azonban az angol és a magyar nyelv között hatalmas technológiai szakadék tátong, és egyre mélyül. A kutatás-fejlesztési támogatások folytonossága nem megfelelő. Rövid távú programok váltják egymást alacsony támogatású időszakokkal, és az EU-s országok és az Európai Bizottság programjainak koordinációjában is általános hiányosságok mutatkoznak. Ennek eredményeképpen Magyarország (és az EU általában) több nagyon ígéretes innovációt vesztett az Amerikai Egyesült Államokkal szemben, ahol a stratégiai kutatások tervezésében nagyobb kontinuitás tapasztalható, és ahol jobban támogatják az új technológiák piacra kerülését. A technológiai innováció versenyében csak egy jövőbe mutató koncepcióval rendelkező korai start biztosíthat versenyelőnyt, persze csak akkor, ha valóban eléri a célt. Különben minden amit elérünk, csak egy tiszteletbeli említés a Wikipédiában.

Mindezek ellenére még mindig nagy kutatói potenciál rejlik Magyarországban és az EU-ban. A nemzetközileg is elismert kutatóközpontokon és egyetemeken kívül számos innovatív nyelvtechnológiai kis- és középvállalkozás működik, melyek nagy kreativitással és hatalmas erőfeszítésekkel próbálnak túlélni, stabil támogatás és kockázati tőke hiányában is.

Habár Magyarország fontos fejlesztéseket támogatott a korpuszépítés és a nyelvi erőforrások létrehozása terén, a magyar nyelvtechnológiai erőforrások és eszközök még mindig nem érik el minőségben és lefedettségben angol nyelvű megfelelőik színvonalát, amelyek majd minden nyelvtechnológiai területen az élvonalat képviselik. Minden nemzetközi nyelvtechnológiai verseny azt mutatja, hogy az eredmények az angol nyelv automatikus elemzése terén sokkal jobbak, mint ugyanezek a magyarra. Ez igaz az információkinyerésre, a nyelvi ellenőrzésre, a gépi fordításra és sok más alkalmazásra is. Sok kutató szerint ez annak köszönhető, hogy az elmúlt ötven évben a számítógépes nyelvészeti módszerek és algoritmusok, valamint a nyelvtechnológiai alkalmazások fejlesztése elsősorban az angolra fókuszált. Más kutatók azt gondolják, hogy az angol jellegénél fogva alkalmasabb a számítógépes feldolgozásra, továbbá az olyan nyelvek, mint a spanyol vagy a francia, szintén könnyebben kezelhetők a jelenlegi módszerekkel, mint a magyar. Ez azt jelenti, hogy következetes és fenntartható erőfeszítéseket kell tennünk a kutatás terén, ha a következő generációs infokommunikációs technológiákat magyarul akarjuk használni privát és hivatalos életünkben is.

Összefoglalva: a magyar nyelv romlásáról szóló próféciák ellenére nyelvünk nincs veszélyben, még az angol nyelv erejével szemben sem. Viszont a helyzet drámaian megváltozhat akkor, amikor a technológiák új generációja elkezdi valóban hatékonyan kezelni az emberi nyelvet. A gépi fordítás tökéletesítése által a nyelvtechnológia segít a nyelvi korlátok ledöntésében, de csak azon nyelvek esetében, amelyek képesek fennmaradni a digitális világban. Ha létezik használható nyelvtechnológia, akkor még a kevés beszélővel rendelkező nyelvek is biztosíthatják túlélésüket. Ha nem, akkor még a nagyobb nyelvek is erőteljes nyomás alá kerülhetnek.

A fogorvos tréfás intése: „Csak azt a fogát mossa, amelyet meg akar tartani!" Ez a mondás a kutatási politikára is igaz – egy kikötéssel: megtanulhatsz minden nyelvet, ami csak létezik a nap alatt, de költséges technológiákat csak azokra fejlessz, amelyeket igazán életben akarsz tartani.

2

VESZÉLYBEN A NYELVEINK: KIHÍVÁS A NYELVTECHNOLÓGIÁNAK

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

> Digitális forradalom szemtanúi vagyunk, amelyet Gutenberg invenciójához, a nyomtatás feltalálásához hasonlítanak.

Gutenberg találmánya után a kommunikációban és tudáscserében a következő nagy áttörést Luther bibliafordítása jelentette. Az ezt követő századokban a különböző technikák fejlődése segítette a hatékonyabb nyelvi feldolgozást és tudáscserét:

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

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

- A kiadványszerkesztő szoftver felváltotta a gépírást és a nyomdai formázást.
- A Microsoft PowerPoint szoftver felváltotta az írásvetítő fóliákat.
- E-mailben gyorsabban küldünk és fogadunk dokumentumokat, mint faxszal.
- A Skype segítségével interneten keresztül telefonálhatunk és szervezhetünk virtuális találkozókat.
- A hang- és videókódolási formátumok segítségével könnyen cserélhetünk multimédiás tartalmakat.
- A keresőprogramok kulcsszavas keresést tesznek lehetővé.
- Az online fordítóprogramok, mint a Google Translate, gyors nyersfordítást adnak.
- A közösségi médiaplatformok, mint a Facebook, a Twitter és a Google+ elősegítik az együttműködést és az információmegosztást.

Bár ezek az eszközök és alkalmazások fontos segítséget jelentenek, továbbra sem tudnak olyan fenntartható, többnyelvű európai információs társadalmat kialakítani, amelyben az információ és a javak szabadon áramolhatnak.

2.1 AZ EURÓPAI INFORMÁCIÓS TÁRSADALOM GÁTJAI: A NYELVI HATÁROK

Nem tudjuk pontosan, hogyan fog kinézni a jövőbeli információs társadalom. Azonban igen valószínű, hogy a kommunikációs technológia forradalma a különböző nyelveket beszélő embereket összehozza. Ez a folyamat az embereket új nyelvek tanulására, míg a fejlesztőket új alkalmazások létrehozására készteti, ami erősíti a kölcsönös megértést és elérhetővé teszi a közös tudást.

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

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, YouTube és legújabban Google+) jelenlegi népszerűsége csak a jéghegy csúcsa.

Manapság több gigabájtnyi szöveget tudunk továbbítani a világ körül pár másodpercen belül anélkül, hogy észrevennénk, hogy a szöveg olyan nyelven van, amelyet nem értünk. 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 egyéb üzenetek írásához a weben [2]. Pár évvel ezelőtt még az angol volt a *lingua franca* a weben – az interneten megtalálható tartalom nagy része angolul volt –, a helyzet azonban mostanra jelentősen megváltozott. A nem angol nyelvű (különösen az arab és egyéb ázsiai nyelvű) online tartalom mennyisége robbanásszerűen megnőtt.

Eddig meglepően kevés figyelmet kapott a nyilvános vitákban a nyelvi határok miatti digitális megosztottság, amely mindenhol jelentkezik; manapság azonban felvetődik az az égető kérdés, hogy mely európai nyelvek fognak boldogulni és kitartani a tudásalapú információs társadalomban.

2.2 VESZÉLYBEN A NYELVEINK

A nyomtatott sajtó megjelenése páratlan mértékű információcserét indított el Európában, ez azonban sok európai nyelv pusztulását is előidézte. A regionális és kisebbségi nyelvek alig kerültek nyomtatásba. Ennek eredményeként sok nyelv, mint például a dalmát vagy a kelta, csak beszélt formában élt tovább, és ez korlátozta további fejlődésüket és használatukat. Vajon az internetnek is hasonló hatása lesz a nyelveinkre?

Európa legfontosabb és leggazdagabb kulturális értékei közé tartozik a térségben használt csaknem 80 nyelv. Európa nyelvi sokszínűsége is hozzájárul társadalmi sikeréhez [3]. Míg a népszerű nyelvek, mint az angol vagy a spanyol biztosan megmaradnak a feltörekvő digitális társadalomban és a piacon, sok más európai nyelv el fog tűnni a digitális kommunikációból és az internetes társadalom látóköréből. Ez biztosan nem járható út. Egyrészt elveszne egy stratégiai lehetőség, és ez Európa globális helyzetét gyengítené. Másrészt az ehhez hasonló változások szemben állnak azzal az elképzeléssel, hogy az európai polgárok azonos mértékben vehessenek részt az őket érintő ügyekben, nyelvtől függetlenül.

Európa legfontosabb és leggazdagabb kulturális értékei közé tartozik nyelvi sokszínűsége.

Egy többnyelvűségről szóló UNESCO beszámoló szerint a nyelvek az alapvető jogok, mint például a politikai

önkifejezés, az oktatás vagy a társadalomban való részvétel fontos közvetítői [4].

2.3 NYELVTECHNOLÓGIA: EGY KULCSFONTOSSÁGÚ TECHNOLÓGIA

A múltban a nyelvvédő beruházások főleg a nyelvoktatásra és fordításra fókuszáltak. Példaként: becslések szerint Európának 2008-ban 8,4 milliárd eurós fordító, tolmács, szoftverlokalizációs és honlapglobalizációs piaca volt, és mindehhez még évi 10%-os növekedést vártak [5]. Azonban ez a kapacitás még mindig nem elég ahhoz, hogy kielégítse a jelenlegi és a jövőbeli igényeket. A minden területet lefedő nyelvhasználatot biztosító legizgalmasabb megoldás a holnap Európájában a megfelelő technológia használata, hiszen például a szállításhoz, az energiaiparban vagy a fogyatékkal élők életének megkönnyítéséhez szintén fejlett technológiát használunk.

A nyelvtechnológia (az írott szöveg és a beszéd minden formáját lefedve) lehetővé teszi az együttműködést, a tanulást, az üzletkötést, a tudásmegosztást és a társadalmi és politikai vitákban való részvételt, számítástechnikai tudástól és nyelvi határoktól függetlenül. Gyakran bonyolult rendszerekbe beépítve dolgozik a háttérben, segítve minket, amikor például:

- információt keresünk internetes keresővel;
- helyesírást és nyelvtant ellenőrzünk szövegszerkesztőben;
- termékajánlásokat olvasunk online vásárláskor;
- egy navigációs rendszer szóbeli utasításait hallgatjuk;
- online szolgáltatással fordítunk weboldalakat.

A nyelvtechnológiai fejlesztések tipikusan nagyobb alkalmazásokban jelennek meg. A META-NET fehér könyvek célja, hogy minden európai nyelvre vonatkozóan bemutassák, milyen készültségi állapotban vannak az azokra kidolgozott alapvető technológiák.

> A közeljövőben minden európai nyelvre elérhető, robusztus és olcsó nyelvtechnológiára van szükségünk.

Ahhoz, hogy fenntartsuk pozíciónkat a globális innováció élvonalában, a közeljövőben minden európai nyelvre elérhető, robusztus, olcsó és nagyobb szoftverkörnyezetbe integrálható nyelvtechnológiára van szükségünk. Az interaktív, multimédiás és többnyelvű internethasználat nyelvtechnológia nélkül elképzelhetetlen.

2.4 A NYELVTECHNOLÓGIA LEHETŐSÉGEI

A nyomtatás világában a szövegről készült képek gyors sokszorosítása jelentette a technológiai áttörést. Emberek végezték az információkeresés és -feldolgozás, a fordítás és az összefoglalás kemény munkáját. A beszéd rögzítésére Edison találmányáig kellett várnunk, ami megint csak analóg másolatok készítésére volt jó.

A digitális nyelvtechnológia segítségével elérhetővé válik az automatikus fordítás és tartalom-előállítás, az információfeldolgozás és a tudásmenedzsment minden európai nyelven. Emellett elősegíti az intuitív, természetesnyelv-alapú 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. 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 önellenőrző 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 tartalmú 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.

A nyelvtechnológia segíthet a nyelvi gátak ledöntésében, támogatva a szabad és nyilvános nyelvhasználatot. Emellett az innovatív, többnyelvű nyelvtechnológia segít a nemzetközi partnerekkel és a többnyelvű szervezetekkel való kommunikációban is. A nyelvtechnológiára egyfajta támogató technológiaként tekinthetünk, amely segít a nyelvi diverzitásból adódó hátrány legyőzésében és a nyelvi közösségek egymáshoz közelebb hozásában.

A kutatás aktív része a nyelvtechnológiának a katasztrófa sújtotta helyeken, mentési munkálatoknál való felhasználása is. Az ilyen, magas rizikófaktorú környezetben a fordítás pontossága élet-halál kérdése lehet, és az intelligens robotok nyelvi képességeikkel életeket menthetnek.

2.5 A NYELVTECHNOLÓGIA KIHÍVÁSAI

Bár a nyelvtechnológia nagy fejlődésen ment keresztül az utóbbi években, a termékinnováció és -fejlesztés még mindig meglehetősen lassan halad előre. A széles körben használt nyelvtechnológiai alkalmazások, mint például a szövegszerkesztők helyesírás-ellenőrzői, tipikusan egynyelvűek, és mindössze néhány nyelvre elérhetőek.

A technológiai fejlesztés még mindig meglehetősen lassan halad előre.

Az online gépi fordító szolgáltatások kitűnően használhatók arra, hogy nyersfordítást adjanak a dokumentum tartalmáról, de nem alkalmasak pontos fordításra. Az emberi nyelv komplexitásának köszönhetően nyelveink számítógépes modellezése és a való világban való tesztelése idő- és pénzigényes vállalkozás, ami hosszútávú támogatást igényel. Európának fenn kell tartania úttörő szerepét a többnyelvű közösségek igényeinek megfelelő technológiák előállításában – új módszerek kifejlesztésével, melyek Európa-szerte erősítik a fejlődést. Ezek közé tartoznak a számítógépes újítások és például a távmunka lehetősége is.

2.6 EMBERI ÉS GÉPI NYELVELSAJÁTÍTÁS

Ahhoz, hogy bemutassuk, hogyan birkóznak meg a számítógépek a nyelvvel, és miért olyan nehéz a nyelvel-

sajátítás, először egy kis kitekintést adunk arra, hogyan sajátítja el az ember az anyanyelvét, valamint idegen nyelveket, majd felvázoljuk, hogy a nyelvtechnológiai rendszerek hogyan működnek.

Két különböző módon sajátíthatunk el egy nyelvet. A gyermek először 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, amennyiben ez nem anyanyelvi közegben zajlik. Iskolás korban az idegen nyelv elsajátítása a nyelv nyelvtani szerkezetének, szókincsének és helyesírásának könyvekből és oktató anyagokból való megtanulásával zajlik, amelyek a nyelvet szabályokon, táblázatokon és példaszövegeken keresztül mutatják be. Egy idegen nyelv megtanulása sok erőfeszítést és időt igényel, és mindez az évek múlásával egyre nehezebbé válik.

A nyelvtechnológiai rendszereknek is két fő típusát különítjük el, hasonlóan az emberi nyelvelsajátításhoz. A statisztikai (vagy adatvezérelt) megközelítést követő rendszerek a nyelvtudást nagy mennyiségű szövegből nyerik. Míg az olyan alkalmazások tanításához, mint például a helyesírás-ellenőrzők, elegendő egynyelvű szövegeket gyűjteni, egy gépi fordító rendszer tanításához két- vagy többnyelvű párhuzamos szövegekre van szükség. Ezután a gépi tanuló algoritmusok olyan mintákat tanulnak meg a szövegből, amelyek azt mutatják meg, hogy a szavakat, rövid kifejezéseket és mondatokat hogyan fordítjuk le.

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 megközelítésen alapulnak. A statisztikai megközelítés nagy előnye, hogy a gépek gyorsan tanulnak, habár ennek a minősége meglehetősen változó.

Két különböző módon sajátíthatunk el egy nyelvet: példákon vagy szabályokon keresztül.

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 roppant idő- és munkaigényes feladat, amely magasan kvalifikált szakembereket igényel. A vezető szabályalapú gépi fordító rendszerek némelyike több mint húsz éve fejlesztés alatt áll. 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.

Mivel a statisztikai és a szabályalapú rendszerek előnyei és hátrányai kiegészítik egymást, a jelenlegi kutatások inkább a hibrid megközelítésre fókuszálnak, amely kombinálja a két megközelítést. Ezek a módszerek azonban az ipari környezetben kevésbé sikeresek, mint a kutatólaboratóriumban.

Ahogy ebben a fejezetben láthattuk, sok olyan alkalmazást használunk a mai információs társadalomban,

amely erősen épít a nyelvtechnológiára. Többnyelvű közösségének köszönhetően ez különösen igaz az európai gazdasági és információs térségre. Bár a nyelvtechnológia erőteljesen fejlődött az elmúlt pár évben, még mindig nagy potenciál rejlik a nyelvtechnológiai rendszerek minőségének javításában. A következőkben bemutatjuk a magyar nyelv szerepét az európai információs társadalomban, és felmérjük a magyar nyelvtechnológia jelenlegi helyzetét.

3

A MAGYAR NYELV AZ EURÓPAI INFORMÁCIÓS TÁRSADALOMBAN

3.1 ÁLTALÁNOS TÉNYEK

A magyar nyelv a legtöbb beszélővel rendelkező nem indoeurópai nyelv Európában. A Magyar Köztársaság államnyelve, ahol a 10 milliós lakosságnak kb. 97%-a magyar anyanyelvű. A szomszédos hét országban is találunk magyar nyelvű közösségeket, amelyek közül a legnagyobb a romániai, megközelítőleg másfél millió nyelvhasználóval. Becslések szerint a magyar nyelvet összesen 13 millióan beszélik, ezzel a 12. helyen áll a legtöbb beszélővel rendelkező nyelvek listáján Európában [6]. 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 – 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.

3.2 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 hegységből 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 és néhány más, Oroszországban élő népek által beszélt nyelv.

A magyar nyelv a legtöbb beszélővel rendelkező nem indoeurópai nyelv Európában.

Az uráli nyelvek néhány közös, ősi jellemzője:

- Nincsenek nemek: ugyanaz a szó (*ő*) fejezi ki a „he” és a „she” fogalmát.

- Csak két igeidő van: jelen és múlt; ezek árnyalatait, valamint a jövő időt körülírással lehet kifejezni.
- Az irányhármasság: a helyet kifejező ragokból 3x3 van, mint az 1. táblázat mutatja a *doboz* szó példáján (a névelő változatlan, és nincs egyeztetve a főnévvel).

A magyart latin betűkkel írják, de a magyar szöveg mégsem hasonlít egyik európai nyelvre sem. 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*
> *Jókedvvel, bőséggel.*
> „God, bless the Hungarian
> 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 kérdőjellel jelzett szó nem létezik a legtöbb nyelvben: a neve igekötő. 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* elöljárószó áll, ott a magyarban végződések vannak. A magyarban nincsenek elöljárószók, példánkban a *-vel*, *-gel* ragok fejezik ki azt, amit az angol *with*.

Még egy fontos magyar sajátságot említünk: a birtokviszonyt 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 (=József 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, és 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. magánhangzóharmónia (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*, míg az *ösztön* magas, ezért a többi magánhangzó is magas: *ö - ö + ö - e - ü* [6].

3.3 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,

	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

1: Az irányhármasság a *doboz* szó példáján bemutatva

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.

3.4 NYELVMŰVELÉS MAGYARORSZÁGON

Magyarországon két intézmény van, amely aktív szerepet játszik a magyar nyelv ápolásában és terjesztésében. Az egyik a Balassi Bálint Intézet, amelyet az Oktatási Minisztérium alapított. A másik a Magyar Tudományos Akadémia Nyelvtudományi Intézete.

A Balassi Intézet a magyarországi nyelvművelés egyik központja, 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 [7].

Magyarországon két intézmény van, amely aktív szerepet játszik a magyar nyelv ápolásában és terjesztésében.

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 [8].

A magyar helyesírási kérdések akadémiai szabályozás alá tartoznak: a magyar helyesírást a Nyelvtudományi Intézet Helyesírási Bizottsága szabályozza 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, amely 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.

3.5 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 országokkal került egy csoportba, mint Franciaország, Németország vagy az Egyesült Királyság [9].

3.6 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 [10].

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

3.7 A MAGYAR NYELV AZ INTERNETEN

2009-ben a magyarországi lakosság 61,6%-a volt internethasználó [11]. A fiatal generáció körében, 14-17 éves korban, ez az arány magasabb. Az internetpenetráció az európai átlag alatt van, de folyamatosan emelkedik. 2011 januárjában a .hu közdomainek alatt delegált domainek száma közel 600.000 volt [12], és határozottan növekszik. Körülbelül 70.000 regisztrált domain létezik az országban a .hu rendszeren kívül (nagy részük .com) [13].

A magyar Wikipédia a 19. legnagyobb, megelőzve más, több beszélővel rendelkező európai és világnyelveket.

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. 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.

Magyarországon nincs hivatalos törvény, amely a fogyatékkal élők esélyegyenlőségét biztosítaná, de a Fogyatékos Személyek Esélyegyenlőségéért Közalapítvány kidolgozott egy ajánlást a komplex akadálymentesítésre. Ez magában foglalja azt is, hogy a közintézményeknek a fogyatékos személyek számára is elérhetővé és használhatóvá kell tenniük a weboldalukat és internetes szolgáltatásaikat. A felhasználóbarát nyelvtechnológiai eszközök kulcsszerepet játszhatnak ezeknek a követelményeknek a teljesítésében: például a beszédszintézis a vakok számára is elérhetővé teszi a weboldalak tartalmát.

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.

4

NYELVTECHNOLÓGIA MAGYARUL

A nyelvtechnológiai rendszerek olyan szoftverek, amelyek kifejezetten a természetes emberi nyelv feldolgozására lettek specializálva. Ezért ezeket a technológiákat összefoglaló névvel természetesnyelv-feldolgozásnak is szokták nevezni. Az emberi nyelv előfordul beszélt és írott változatban is. Míg a beszéd a legősibb és legtermészetesebb módja az emberi kommunikációnak, a komplex információ, így az emberi tudás nagy része általában írott formában létezik. A beszéd- és a nyelvtechnológia az emberi kommunikációnak ezt a két különböző formáját dolgozza fel, illetve állítja elő, és mindkettőhöz használ szótárakat, nyelvtani szabályokat és szemantikát. Vagyis a nyelvtechnológia a tudásreprezentáció különféle formáit használja, amelyek függetlenek lehetnek a nyelvet közvetítő médiumtól (beszéd vagy szöveg). A 2. 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 következőkben a nyelvtechnológia fő alkalmazási területeit fogjuk tárgyalni, melyek a következők: nyelvi ellenőrzés, webes keresés, beszédtechnológia és gépi fordítás. Ezek olyan alkalmazásokat és technológiákat foglalnak magukban, mint például

- helyesírás-ellenőrzés,
- szerzői támogatási rendszerek,
- gép által támogatott nyelvtanulás,
- információ-visszakeresés,
- információkinyerés,
- szövegtömörítés,
- kérdésmegválaszoló rendszerek,
- beszédfelismerés és
- beszédszintézis.

A nyelvtechnológia kiterjedt szakirodalommal rendelkezik, melyek közül az érdeklődő olvasót a következő olvasnivalókhoz irányítjuk: [14, 15, 16, 17].
Mielőtt a fenti alkalmazási területeket tárgyalnánk, röviden bemutatjuk egy tipikus nyelvtechnológiai rendszer felépítését.

4.1 A NYELVTECHNOLÓGIAI ALKALMAZÁSOK FELÉPÍTÉSE

A tipikus nyelvtechnológiai alkalmazások több komponensből állnak össze, amelyek a nyelv egyes szintjeit tükrözik. A 3. ábra egy szövegfeldolgozó rendszer egyszerűsített felépítését mutatja. Az első három modul a bemenő szöveg szerkezetét és jelentését dolgozza fel:

1. Előfeldolgozás: adattisztítás, a formázás eltávolítása, a bemenő szöveg nyelvének megállapítása, a speciális karakterek kezelése (pl. a magyar ékezetes betűk esetében) stb.

2: Természetesnyelv-feldolgozás

2. Nyelvtani elemzés: az ige és argumentumainak megkeresése, a mondat szerkezetének feltárása.

3. Szemantikai elemzés: egyértelműsítés (adott szónak az adott kontextusban mi a jelentése), az anaforák feloldása (a névmások kire/mire vonatkoznak), 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. Mindez az alkalmazások felépítésének egyszerűsített és idealizált leírása, amely a nyelvtechnológiai alkalmazások komplexitását illusztrálja.

A legfontosabb alkalmazási területek bemutatása után rövid kitekintésben beszámolunk a nyelvtechnológiai kutatási és oktatási helyzetről, különös tekintettel a már lezárult és a folyó kutatási programokra. A fejezet végén szakértői értékelést adunk a legfontosabb nyelvtechnológiai eszközökről és erőforrásokról olyan dimenziók mentén, mint az elérhetőség, a fejlettség és a minőség. A 29. oldalon található 9. táblázat jó áttekintést ad a magyar nyelvtechnológia helyzetéről.

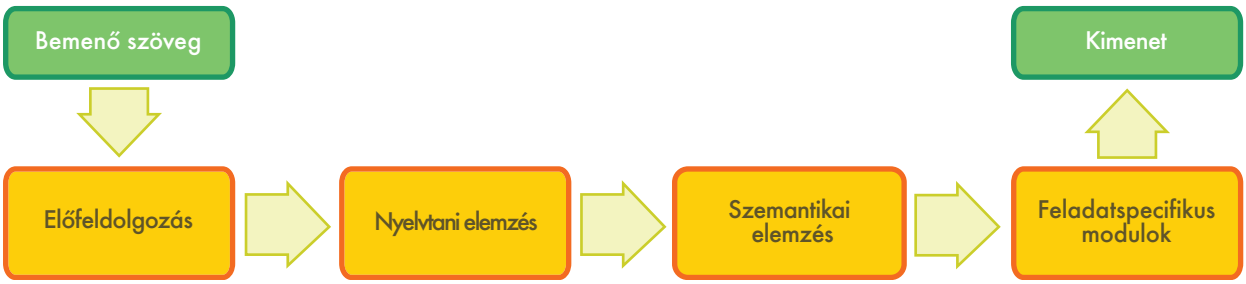
4.2 A FŐ ALKALMAZÁSI TERÜLETEK

Ebben a fejezetben a legfontosabb nyelvtechnológiai eszközökre és erőforrásokra fókuszálunk, és áttekintést adunk a magyarországi nyelvtechnológiai tevékenységről.

4.2.1 Nyelvi ellenőrzés

Mindenki, aki használt már a Microsoft Wordhöz hasonló szövegszerkesztőt, találkozott helyesírás-ellenőrző programmal, amely jelzi a helyesírási hibákat, és javítási javaslatokat tesz. Az első helyesírás-ellenőrző programok szimplán összehasonlították az ellenőrizendő szavakat a helyesen írt szavak listájával. A mai eszközök ennél sokkal kifinomultabbak. A szövegelemzéshez nyelvfüggő algoritmusokat használnak, amelyek a morfológiát (pl. a többes számú alakokat) is tudják kezelni, valamint a mondatszintű hibákat is detektálják, 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 nyelvi ellenőrző nem találna hibát a következő szövegben [18]:

> I have a spelling checker,
> It came with my PC.
> It plane lee marks four my revue
> Miss steaks aye can knot sea.



3: Egy tipikus szövegfeldolgozó alkalmazás felépítése

Az ilyen típusú hibák kezeléséhez az esetek nagy részében a kontextus elemzését is el kell végezni. A magyarban vannak olyan ragozott szavak, amelyek különböző jelentésekkel bírhatnak: például a *várunk* lehet a *vár* ige többes szám első személyű alakja, illetve a *vár* főnév birtokos személyraggal ellátott alakja.

A jelenség kezeléséhez nyelvspecifikus nyelvtani szabályok előállítására, vagyis magas szintű szakértői munkára, vagy pedig statisztikai alapú nyelvmodellekre van szükség, amelyek 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, ellenőrzött adatot tartalmazó szöveghalmazokból, más néven korpuszokból. Ez a megközelítés elsősorban angol nyelvű adatokra lett kifejlesztve, de a magyarra is alkalmazható. Azt azonban 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 is.

A nyelvi ellenőrzők használata nem csak a szövegszerkesztőkre korlátozódik, alkalmazzák még az ún. szerzői támogatási rendszerekben is.

A nyelvi ellenőrzők használata nem csak a szövegszerkesztőkre korlátozódik, alkalmazzák még az ún. szerzői támogatási rendszerekben is, olyan szoftverkörnyezetekben, amelyekben használati utasításokat és egyéb dokumentációkat írnak speciális sztenderdek alapján az információtechnológiai, az egészségügyi, a műszaki és egyéb termékek területén. 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 (fordítás, lokalizálás). A természetesnyelv-feldolgozás eredményei a szerzői támogatási rendszerekben is fejlődést hoztak: a technikai dokumentáció szerzőit szótárak, terminológiai adatbázisok és mondattani szabályok segítik, melyek követik az adott terület előírásait.

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. [19] fejlesztette ki a nyolcvanas 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 összefüg-

4: Nyelvi ellenőrzés (lent: szabályalapú, fent: statisztikai)

géseiben, 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 [20] a MySpellen alapul, és integrálva lett az OpenOffice-ba, a Mozilla Firefoxba és Thunderbirdbe, valamint a Google Chrome-ba is.

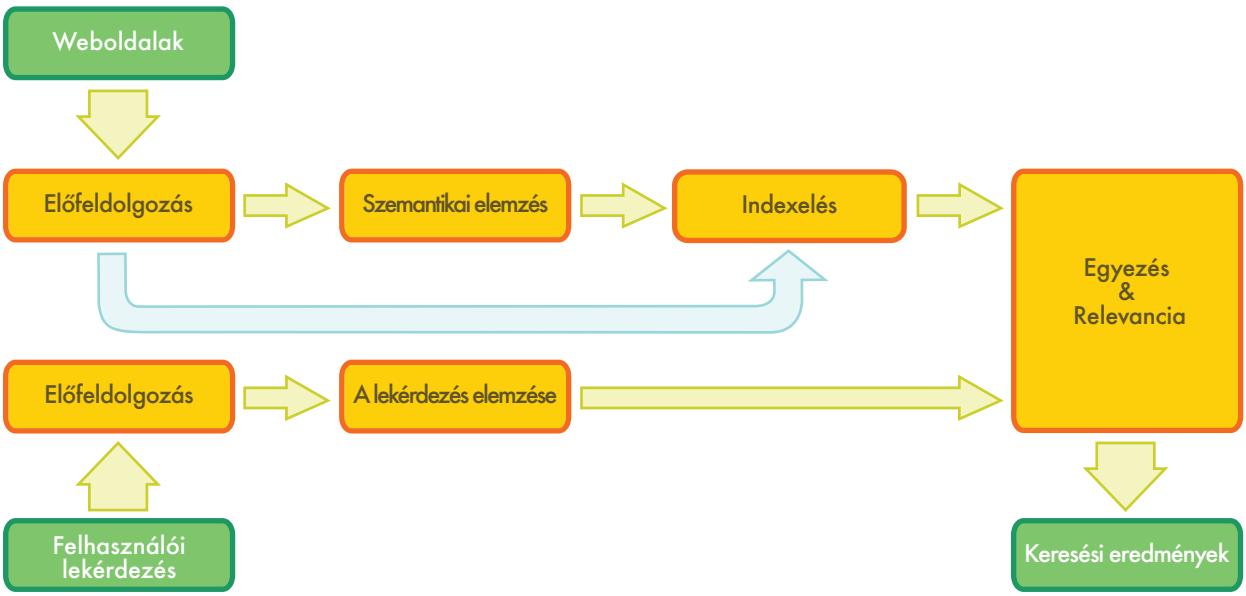
A helyesírás-ellenőrzés és a szerzői támogatás mellett a nyelvi ellenőrzés a gép által támogatott nyelvtanulás terén is fontos szerepet tölt be, továbbá a webes keresőkben is alkalmazzák a lekérdezések automatikus javítására, például a Google keresési javaslatai esetében.

4.2.2 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 [21]. 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 jelentősen az első verzió óta. A jelenlegi változatban van viszont ellenőrző program, amely az elgépeléseket javítja, továbbá nemrég alapszintű szemantikai kereső alkalmazást építettek be, amely növeli a találati pontosságot azzal, hogy kontextusban vizsgálja a keresőkifejezést [22]. A Google sikersztorija azt mutatja, hogy nagy mennyiségű adattal és hatékony indexelési technológiával a statisztikai alapú megközelítés kielégítő eredményt tud hozni.

Azonban ha bonyolultabb információhoz akarunk jutni, mélyebb nyelvi tudásra van szükségünk a szövegértelmezéshez. Az olyan lexikai erőforrások, mint a gép által olvasható tezauruszok és a WordNethez hasonló ontológiák, 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 sokkal kifinomultabb nyelvtechnológiát kell alkalmaznia, különösen az olyan esetekben, amikor a keresés kérdést vagy más típusú mondatot tartalmaz, nem csak szavak listáját. Például képzeljünk el egy olyan lekérdezést, hogy *Sorold fel azokat a cégeket, amelyeket az elmúlt öt évben vásároltak fel!* A releváns válasz megtalálásához szükség van a mondat szintaktikai és szemantikai szintű elemzésére, valamint a releváns dokumentumok gyors elérését lehetővé tevő indexelésre is. A kielégítő válaszadáshoz 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,



5: A webes keresés architektúrája

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.

A keresőmotorok új generációjának sokkal kifinomultabb nyelvtechnológiát kell alkalmaznia.

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 hívta életre 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.

Mivel a magyar nem olyan kötött szórendű, mint például az angol, a magyar mondatelemzők fejlesztése során nem tudunk pusztán 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 az MTA Nyelvtudományi Intézetének 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 nevő szabályalapú gépi fordítóba lett beépítve.

A nyelvtechnológiával foglalkozó cégek és kutatóműhelyek fő kutatási irányai között szerepel olyan trend- és szövegelemző eszközök fejlesztése, amelyek termé-

szetesnyelv-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 [23], amely lehetőséget nyújt tetszőleges számú, interneten keresztül elérhető adatbázis, forrás egyidejű lekérdezésére. A találati eredményekből közös lista készül, amelyben az elemek fontossági sorrend szerint vannak rendezve. A metakereső természetesnyelv-feldolgozási és információ-visszakeresési 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.

4.2.3 Beszédtechnológia

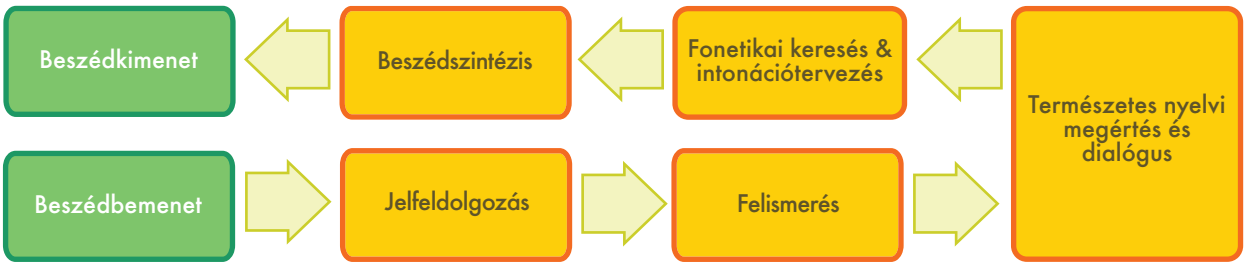
A beszédtechnológia szolgáltatja az alapot olyan interfészek előállításához, amelyek lehetővé teszik, hogy a felhasználók a gépekkel természetes emberi nyelven, és ne csak grafikus felület, billentyűzet vagy egér segítségével kommunikáljanak. Napjainkban ilyen beszédinterfészeket alkalmaznak bizonyos 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:

1. Az automatikus beszédfelismerés határozza meg, hogy milyen szavakat mondott ki a felhasználó.
2. A szintaktikai elemzés és a szemantikai interpretáció segítségével elemezhető a felhasználó közlésének szintaktikai szerkezete, valamint leképezhető annak szemantikai interpretációja az adott rendszer céljainak megfelelően.
3. A dialógusvezérlés az input nyelvi jellemzői, az adott felhasználó és feladat egyéni beállításai alapján valósítja meg a rendszer megfelelő lépését, az adatbázis-lekérdezést.
4. A beszédszintézis technológiáját alkalmazzák arra, hogy a gép előállítsa a megfelelő beszédkimenetet.

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.

Az egyik legnagyobb kihívást az automatikus beszédfelismerés jelenti, vagyis hogy a rendszer minél pontosabban felismerje a felhasználó által kiejtett szavakat. Ez kétféleképpen történhet: vagy 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. A gépi tanulási technikákat használva nyelvmodelleket állíthatunk elő automatikusan szövegkorpuszokból, vagyis audiófájlok és szöveges átirataik nagyméretű gyűjteményéből. Míg a kulcsszavas módszer merev és nehezen használható beszédinterfészt, valamint kevésbé elfogadható kimenetet eredményez, addig a nyelvmodellek előállítása és finomhangolása a költségeket emeli meg erőteljesen. Azonban



6: Egyszerű beszédalapú dialógus felépítése

a nyelvmodelleket alkalmazó beszédinterfész nagyobb elfogadottsággal rendelkezik a felhasználók körében, és előnyösebb, mint a kevésbé rugalmas rendszervezérelt megközelítés.

Ami a beszédinterfész kimeneti oldalát illeti, a vállalatok egyre inkább előre felvett 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 lejátszani kívánt tartalom, annál rosszabb lesz az elemekből összeállított mondat prozódiája az audiófájlok összevágása miatt – még akkor is, ha a mai beszédszintetizáló rendszerek egyre jobban teljesítenek, köszönhetően az egyre természetesebbé váló prozódiának.

A beszédtechnológia piacán az elmúlt évtizedekben fontos szabványosítási lépések történtek a különböző technológiai komponensek közötti interfészek, valamint az egyes alkalmazásokra épülő termékek esetében is. Intenzív 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 (USA) és a Loquendo (Olaszország), csak hogy a legprominensebbeket említsük az európai piacról. 2011-ben a Nuance bejelentette, hogy felvásárolja a Loquendót, ami egy újabb lépés a piac konszolidációja felé.

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 [24] dominálják. A legszélesebb körben használt magyar beszédszintetizátor a Profivox, amely 2002 óta elérhető, és amelyet több alkalmazásba is beépítettek: SMS- és e-mailfelolvasó szoftverbe, autós és mobiltelefonos GPS rendszerbe, valamint e-book- és képernyőolvasó szolgáltatásba, amelyek segíthetik a látássérült emberek integrációját az információs társadalomba. Egy magas szinten vezérelhető interaktív fejlesztői környezet is rendelkezésre áll speciális kutatási célok támogatására (pl. pszichológiai és prozódiai vizsgálatokra). A szoftver szövegfelolvasó (TTS) technológián alapul. Segítségével meghatározott akusztikai és prozódiai tartalommal bíró kísérleti jel hozható létre kontrollálható körülmények között. Egy 1,5 millió szóalakot tartalmazó magyar kiejtési szótár is elérhető. Ennek alapján kialakítható egy magyar szövegeket (fonetikai) szimbólumokká alakító eszköz.

Beszédfelismeréssel Magyarországon a fentebb említett és más egyetemi kutatóműhelyek (pl. a Szegedi Tudományegyetem Informatikai Tanszékcsoportja) mellett

kisebb vállalkozások is foglalkoznak, mint az Alkalmazott Logikai Laboratórium, az Aitia vagy a Digital Natives Kft. 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 folyamatosbeszéd-felismerő motort és fejlesztői környezetet, továbbá egy kötött hangsúlyozáson alapuló szóhatár-detektáló alkalmazást magyar és finn nyelvekre, melyet egy nyelvközi vizsgálat előzött meg. A már említett kutatóműhelyek közös munkájának eredményeképpen különféle orvosi leletező beszédfelismerők is készültek, melyek az orvosi vizsgálatokat közvetlen beszéd-szöveg átalakítással segítik. Továbbá létezik egy olyan alkalmazás, amely a beszédhibás gyerekek beszédtanulását segíti, és a magyar mellett több más európai nyelven is használható. A beszédfelismerő rendszerek számos további gyakorlati alkalmazást segíthetnek. Ilyen például a telefonos hívások kezelése vagy a telefonközpont-irányítás.

A jövőben jelentős változásokat fog hozni az ügyfélkapcsolatok kezelésében a hagyományos telefon, az internet és az e-mail mellett az okostelefonok terjedése, ami hatással lesz a beszédtechnológiára is. Hosszútávon kevesebb telefonalapú felhasználói interfész lesz, és a beszélt nyelv mint felhasználóbarát bemenet sokkal inkább központi szerepet fog játszani. Ez nagyrészt annak köszönhető, hogy a beszélőfüggetlen beszédfelismerés pontosságának javításában előrelépések történtek azáltal, hogy a diktáló rendszerek már most beépített szolgáltatások az okostelefon-használók számára.

4.2.4 Gépi fordítás

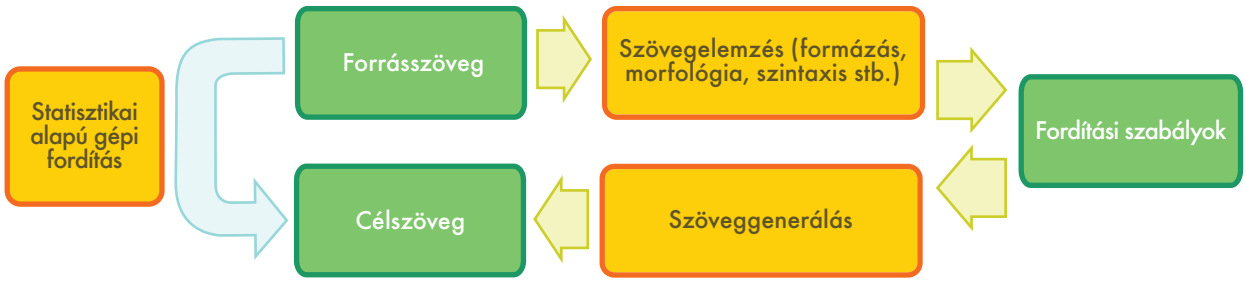
A digitális számítógépek alkalmazásának ötlete természetes nyelvek lefordítására 1946-ban merült fel először. Az ötletet az ötvenes években anyagi támogatás is követte, ami azonban csak a nyolcvanas években folytatódott. Mindemellett a gépi fordítás még mindig nem váltotta be a kezdeti nagy reményeket.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

A legalacsonyabb szinten a gépi fordítás egyszerű behelyettesítés: az egyik természetes nyelvű szót lecseréljük egy másik nyelvű szóra.

A kevésbé kötött szövegek jó fordításához nagyobb szövegegységeket (frázisokat, mondatokat vagy teljes bekezdéseket) kell illeszteni a másik nyelvű megfelelőikhez. A legfőbb problémát az okozza, hogy az emberi nyelv sokszor többértelmű, ami minden szinten kihívások elé állítja a nyelvfeldolgozókat. A lexikai szinten a szójelentés-egyértelműsítés (pl. a *nyúl* lehet cselekvés 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.
- A rendőr látta az embert a revolverrel.

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, és ez utóbbiból generálják 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, amelyek előállítása hosszú és költséges folyamat.



7: Gépi fordítás (balra: statisztikai, jobbra: szabályalapú)

A nyolcvanas évek végétől 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 párhuzamos korpuszokból lehet kiszámítani, mint amilyen az Europarl párhuzamos korpusz, amely az Európai Parlament jegyzőkönyveit tartalmazza 21 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 nyelvtanilag helytelen kimenetet produkálnak. Másrészről viszont az adatvezérelt rendszereknek több előnyük is van: amellett, hogy kevesebb emberi munkát igényelnek, a nyelv olyan különlegességeit is tudják kezelni (például 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 a magyar nyelvre különösen nehéz.

A gépi fordítás a magyar nyelvre különösen nehéz. A szabad szórend és az elváló igekötők problémát okoznak az elemzés során, a gazdag ragozási rendszer pedig kihívásokat jelent a megfelelő esetraggal rendelkező szóalakok előállításánál.

A nehézségek ellenére 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 MorphoWord angol és magyar nyelv között fordít, 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 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 mondatok illesztése 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 elérhető és lekérdezhető [25], így nyersfordítóként vagy kétnyelvű szótárként is használható.

2010 márciusában indult az iTranslate4.eu [26] projekt, melynek célja 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. 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 az MTA Nyelvtudományi Intézete, míg a szolgáltatások közös interfészét a MorphoLogic Kft. nyújtja.

A gépi fordítás nagy mértékben tudja javítani a hatékonyságot, főként ha intelligensen igazítható a felhasználóspecifikus terminológiához, illetve integrálható a megfelelő munkakörnyezetbe. A magyar nyelvre is léteznek ilyen interaktív fordításstámogató rendszerek, például a Kilgray Kft. által fejlesztett memoQ programcsomag [27].

Még mindig nagy potenciál rejlik a gépi fordító rendszerek minőségének javításában, például a nyelvi erőforrások egy adott felhasználói területre való alkalmazásával, vagy a terminológiai adatbázisok és a fordítómemóriák esetében alkalmazott munkakörnyezetek integrálásával. Problémát jelent, hogy a jelenlegi rendszerek nagy része angolközpontú, és a magyarról és magyarra való fordítás is csak angolra, illetve angolról működik. Ez fennakadásokat okoz a fordítói munkafolyamatban, és arra kényszeríti a gépi fordítást használókat, hogy különböző rendszerek különböző lexikai eszközeinek használatát is elsajátítsák.

A gépi fordító rendszerek, a különböző módszerek és a különböző nyelvpárokra működő rendszerek összehasonlítását kiértékelő kampányok segítik. A 8. táblázat, amely az Európai Bizottság Euromatrix+ projektje keretében készült, a nyelvpárokra lebontott teljesítményt mutatja a 23 hivatalos európai nyelv közül 22-re (az ír nyelv nem szerepel az összehasonlításban). Az eredményeket a BLEU-érték [28] alapján rangsorolták, amely szerint a jobb fordítás magasabb pontszámot kap. Az emberi fordítás kb. 80 pontot kapna.
A legjobb eredmények (zölddel és kékkel jelölve) azon nyelvek esetében születtek, amelyek koordinált programokban vettek részt, és amelyekre kellő mennyiségű párhuzamos korpusz áll rendelkezésre (pl. angol, francia, holland, spanyol és német). A gyengébb eredményt elért nyelvek pirossal kiemelve láthatók. Ezek vagy nélkülözni voltak kénytelenek a kutatási ráfordításokat, vagy strukturálisan különböznek a többi nyelvtől (pl. magyar, máltai és finn).

4.3 TOVÁBBI ALKALMAZÁSI TERÜLETEK

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

> A nyelvtechnológiai alkalmazások gyakran nem jelennek meg a felhasználó szintjén, hanem nagyobb rendszerekbe beépítve, a háttérben működnek.

Az egyik ilyen, aktívan kutatott terület a kérdésmegválaszolás, amelyhez annotált (nyelvi információval ellátott) korpuszokat építenek, és tudományos versenyeket rendeznek. Ennek 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

	Célnyelv – Target language																					
	EN	BG	DE	CS	DA	EL	ES	ET	FI	FR	HU	IT	LT	LV	MT	NL	PL	PT	RO	SK	SL	SV
EN	–	40.5	46.8	52.6	50.0	41.0	55.2	34.8	38.6	50.1	37.2	50.4	39.6	43.4	39.8	52.3	49.2	55.0	49.0	44.7	50.7	52.0
BG	61.3	–	38.7	39.4	39.6	34.5	46.9	25.5	26.7	42.4	22.0	43.5	29.3	29.1	25.9	44.9	35.1	45.9	36.8	34.1	34.1	39.9
DE	53.6	26.3	–	35.4	43.1	32.8	47.1	26.7	29.5	39.4	27.6	42.7	27.6	30.3	19.8	50.2	30.2	44.1	30.7	29.4	31.4	41.2
CS	58.4	32.0	42.6	–	43.6	34.6	48.9	30.7	30.5	41.6	27.4	44.3	34.5	35.8	26.3	46.5	39.2	45.7	36.5	43.6	41.3	42.9
DA	57.6	28.7	44.1	35.7	–	34.3	47.5	27.8	31.6	41.3	24.2	43.8	29.7	32.9	21.1	48.5	34.3	45.4	33.9	33.0	36.2	47.2
EL	59.5	32.4	43.1	37.7	44.5	–	54.0	26.5	29.0	48.3	23.7	49.6	29.0	32.6	23.8	48.9	34.2	52.5	37.2	33.1	36.3	43.3
ES	60.0	31.1	42.7	37.5	44.4	39.4	–	25.4	28.5	51.3	24.0	51.7	26.8	30.5	24.6	48.8	33.9	57.3	38.1	31.7	33.9	43.7
ET	52.0	24.6	37.3	35.2	37.8	28.2	40.4	–	37.7	33.4	30.9	37.0	35.0	36.9	20.5	41.3	32.0	37.8	28.0	30.6	32.9	37.3
FI	49.3	23.2	36.0	32.0	37.9	27.2	39.7	34.9	–	29.5	27.2	36.6	30.5	32.5	19.4	40.6	28.8	37.5	26.5	27.3	28.2	37.6
FR	64.0	34.5	45.1	39.5	47.4	42.8	60.9	26.7	30.0	–	25.5	56.1	28.3	31.9	25.3	51.6	35.7	61.0	43.8	33.1	35.6	45.8
HU	48.0	24.7	34.3	30.0	33.0	25.5	34.1	29.6	29.4	30.7	–	33.5	29.6	31.9	18.1	36.1	29.8	34.2	25.7	25.6	28.2	30.5
IT	61.0	32.1	44.3	38.9	45.8	40.6	26.9	25.0	29.7	52.7	24.2	–	29.4	32.6	24.6	50.5	35.2	56.5	39.3	32.5	34.7	44.3
LT	51.8	27.6	33.9	37.0	36.8	26.5	21.1	34.2	32.0	34.4	28.5	36.8	–	40.1	22.2	38.1	31.6	31.6	29.3	31.8	35.3	35.3
LV	54.0	29.1	35.0	37.8	38.5	29.7	8.0	34.2	32.4	35.6	29.3	38.9	38.4	–	23.3	41.5	34.4	39.6	31.0	33.3	37.1	38.0
MT	72.1	32.2	37.2	37.9	38.9	33.7	48.7	26.9	25.8	42.4	22.4	43.7	30.2	33.2	–	44.0	37.1	45.9	38.9	35.8	40.0	41.6
NL	56.9	29.3	46.9	37.0	45.4	35.3	49.7	27.5	29.8	43.4	25.3	44.5	28.6	31.7	22.0	–	32.0	47.7	33.0	30.1	34.6	43.6
PL	60.8	31.5	40.2	44.2	42.1	34.2	46.2	29.2	29.0	40.0	24.5	43.2	33.2	35.6	27.9	44.8	–	44.1	38.2	38.2	39.8	42.1
PT	60.7	31.4	42.9	38.4	42.8	40.2	60.7	26.4	29.2	53.2	23.8	52.8	28.0	31.5	24.8	49.3	34.5	–	39.4	32.1	34.4	43.9
RO	60.8	33.1	38.5	37.8	40.3	35.6	50.4	24.6	26.2	46.5	25.0	44.8	28.4	29.9	28.7	43.0	35.8	48.5	–	31.5	35.1	39.4
SK	60.8	32.6	39.4	48.1	41.0	33.3	46.2	29.8	28.4	39.4	27.4	41.8	33.8	36.7	28.5	44.4	39.0	43.3	35.3	–	42.6	41.8
SL	61.0	33.1	37.9	43.5	42.6	34.0	47.0	31.1	28.8	38.2	25.7	42.3	34.6	37.3	30.0	45.9	38.2	44.1	35.8	38.9	–	42.7
SV	58.5	26.9	41.0	35.6	46.6	33.3	46.6	27.4	30.9	38.9	22.7	42.0	28.2	31.0	23.7	45.6	32.2	44.2	32.7	31.3	33.5	–

8: Gépi fordítás 22 hivatalos európai nyelvre – Machine translation between 22 EU-languages [29]

a felhasználó egy konkrét kérdést tehet fel, amelyre egy konkrét választ kap. Például:

> *Kérdés: Hány éves volt Neil Armstrong, amikor a Holdra lépett?*
>
> *Válasz: 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 történik, 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, például 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 stb. A területspecifikus információkinyerés szintén kiváló példája a háttérben működő nyelvtechnológiának: 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 megtalálható például a Microsoft

Wordben is. Nagyrészt statisztikai alapon működik: a rendszer először a fontos szavakat azonosítja a szövegben (jellemzően azok számítanak fontos szavaknak, amelyek az adott 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 tulajdonképpen mondatok kiszűrésével egyenlő, ami által a szöveg mondatainak részhalmazára csökken. Minden kereskedelmi forgalomban kapható tömörítő program ezen az ötleten alapul. Egy másik módszer viszont 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. A szövegtömörítő- és generáló alkalmazások az esetek túlnyomó részében nagyobb szoftverkörnyezetbe építve jelennek meg, például klinikai információs rendszerekben, amelyben betegek adatait gyűjtik össze, tárolják és dolgozzák fel, és amelynek a jelentésgenerálás csak egy a funkciói közül.

A kérdésmegválaszolás és szöveggenerálás a magyar nyelvre sokkal kevésbé fejlett, mint az angol nyelv esetében.

A kérdésmegválaszolás és szöveggenerálás a magyar nyelvre 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 az angolra koncentráltak. Néhány versenyen többnyelvű feladatok is voltak, de a magyar ezekben soha nem szerepelt. Ennek eredményeképpen kevés olyan magyar nyelvű korpusz és egyéb erőforrás létezik, amilyen ezekhez a feladatokhoz kell. 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. Mindezek ellenére több tulajdonnévfelismerő, biológiai célú információkinyerő, trendelemző és véleménykinyerő alkalmazás is létezik a magyar nyelvre.

4.4 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 egyelőre egyetlen egyetemen sem működik számítógépes nyelvészeti tanszék. Nyelvtechnológiai oktatás ennek ellenére folyik néhány kapcsolódó tanszéken. Pár egyetemen az alap- vagy a mesterképzés szintjén tartanak kurzusokat a témában, máshol nyelvtechnológiai 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.

4.5 HAZAI PROJEKTEK

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. az MTA Nyelvtudományi Intézetének 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 szabványosítottak, 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 szabványosítás és egységesítés nemzetközi hulláma. Több, az integrációt és interoperabilitást célul tűző projekt is indult, például egy egységes magyar ontológia építését, vagy a morfológiai elemzők különböző kódolási rendszereinek harmonizálását célzó projektek.

2008-ban élenjáró magyarországi kutató-fejlesztő közösségek létrehozták a Nyelv- és Beszédtechnológiai Platformot [30] 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.

További kisebb projektekkel együtt a felsorolt projektek Magyarországon a nyelvtechnológia területén és az alapvető technológiai infrastruktúra kiépítésében egyaránt fejlődést hoztak. A nyelvtechnológiai projektek támogatása Magyarországon és Európában azonban még mindig alacsony ahhoz képest, amennyit az USA költ fordításra és többnyelvű információ-hozzáférésre [31].

Ahogy láttuk, az eddigi programok a magyar nyelvű nyelvtechnológiai eszközök és erőforrások számában növekedést hoztak. A következő fejezetben a magyar nyelvű nyelvtechnológia jelenlegi állapotát összegezzük.

4.6 AZ ESZKÖZÖK ÉS ERŐFORRÁSOK ELÉRHETŐSÉGE

A 9. táblázat összegzi a magyar nyelvtechnológia támogatottságának jelenlegi helyzetét. Az egyes technológiák és erőforrások értékelése vezető szakértők becslése alapján készült. A 7 kritériumra vonatkozó pontszámok 0-tól (nagyon alacsony) 6-ig (nagyon magas) terjedő skálán mozognak.

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. Van azonban egy 1,2 millió tokent tartalmazó, manuálisan szintaktikailag annotált korpusz, amely ingyenesen hozzáférhető, és amely számos alkalmazás alapjául szolgált már.
- Az erőforrások nagy része nem szabványosított; létrehozásukkor a fenntarthatóság nem szerepelt a tervek között. Szervezett programok keretében, a megfelelő előírásokat követve szabványosítani kellene a meglévő adatbázisokat.
- A sztenderd előfeldolgozó lépések (tokenizálás, morfológiai elemzés, felszíni szintaktikai elemzés stb.) már megoldottnak tekinthetők a magyarra, de a bonyolultabb szemantikai feldolgozás még további kutatásokat igényel.
- Minél több szemantikát alkalmaz egy eszköz, annál nagyobb hiányok mutatkoznak a technológiában (lásd a szövegelemzés és -értelmezés különbségét). A mélyebb nyelvi elemzés több erőfeszítést igényel.
- A kutatás sikeres volt egyes kiváló minőségű szoftverek létrehozásában, de az erőforrások nagy része nem szabványos és nem hatékonyan fenntartható. Az adatok szabványosításához és a közös adatcserélő formátumok kialakításához szervezett programok szükségesek.
- Szintaktikai, szemantikai és diskurzusszerkezeti annotációval ellátott korpuszokból hiány mutatkozik. Minél több szemantikát alkalmaz egy eszköz, annál nehezebb a fejlesztéséhez megfelelő adatokat előállítani.
- A világról való tudásunkat leképező tudásbázisokhoz szükséges szemantikai sztenderdek (RDF, OWL stb.) léteznek ugyan, de nehezen alkalmazhatók a természetesnyelv-feldolgozási feladatokra.
- 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.

Összegzésképpen elmondható, hogy a magyar nyelvű kutatás több specifikus területén rendelkezünk limitált funkcionalitású szoftverekkel. Nyilvánvaló, hogy további kutatások kellenek ahhoz, hogy pótolni tudjuk a jelenlegi hiányosságokat a mélyebb szemantikai szintű szövegfeldolgozásban és a beszédfelismeréshez szükséges beszédkorpuszok létrehozásában.

4.7 NYELVEK KÖZÖTTI ÖSSZEHASONLÍTÁS

A nyelvtechnológia helyzete jelentősen különbözik az egyes nyelvi közösségek esetében. Annak érdekében, hogy a nyelvek helyzetét össze tudjuk hasonlítani, ebben a fejezetben értékelést adunk két példa alkalmazási területről (gépi fordítás és beszédtechnológia), valamint egy alaptechnológiáról (szövegelemzés) és a nyelvtechnológiai alkalmazások építéséhez szükséges erőforrásokról.

A nyelvek az alábbi ötelemű skála alapján lettek csoportosítva:

	Mennyiség	Elérhetőség	Minőség	Lefedettség	Fejlettség	Fenntarthatóság	Alkalmazhatóság
Nyelvtechnológia: Eszközök, Technológiák és Alkalmazások							
Beszédfelismerés	3	0	4	2	4	3	3
Beszédszintézis	4	3	4	4	5	3	3
Nyelvtani elemzés	4,5	2	4	4,5	4	3	4,5
Szemantikai elemzés	0,6	2	2,5	0,5	0	0	2
Szöveggenerálás	0	0	0	0	0	0	0
Gépi fordítás	6	1	4	3	6	5	6
Nyelvi erőforrások: Erőforrások, Adatok és Tudásbázisok							
Szövegkorpuszok	3,5	6	5,5	5,5	6	6	4
Beszédkorpuszok	2	2	4	2	4	4	0
Párhuzamos korpuszok	6	4	4,5	2,5	6	6	6
Lexikai erőforrások	3	1	3,5	3,5	3,5	3,5	4,5
Nyelvtanok	3	3	5	5	6	4	3

9: A magyar nyelvtechnológia helyzete

- 1. klaszter: kiváló nyelvtechnológiai támogatás
- 2. klaszter: jó támogatás
- 3. klaszter: közepes támogatás
- 4. klaszter: töredékes támogatás
- 5. klaszter: gyenge vagy semmi támogatás

A nyelvtechnológiai támogatás csoportosítása az alábbi kritériumok szerint történt:

- Beszédtechnológia: A meglévő beszédfelismerő és -szintetizáló technológiák minősége, a területek lefedettsége, a meglévő beszédkorpuszok száma és mérete, az elérhető beszédalapú alkalmazások száma és változatossága.
- Gépi fordítás: A meglévő gépi fordító technológiák minősége, a lefedett nyelvpárok száma, a nyelvi jelenségek és területek lefedettsége, a létező párhuzamos korpuszok minősége és mérete, az elérhető gépi fordító alkalmazások száma és változatossága.
- Szövegelemzés: A meglévő szövegelemző technológiák (morfológia, szintaxis, szemantika) minősége és lefedettsége, a nyelvi jelenségek és területek lefedettsége, az elérhető alkalmazások száma és változatossága, a létező (annotált) szövegkorpuszok minősége és mérete, a létező lexikai erőforrások (pl. WordNet) és nyelvtanok minősége és lefedettsége.
- Erőforrások: A létező szöveg-, beszéd- és párhuzamos korpuszok minősége és mérete, a létező lexikai erőforrások és nyelvtanok minősége és lefedettsége.

A 10., 11., 12. és 13. táblázatok azt mutatják, hogy – köszönhetően az elmúlt évtizedek nyelvtechnológiai támogatásainak – a magyar nyelv viszonylag kedvező helyzetben van. Azonban a magyar nyelvű erőforrások és eszközök a minőség és a lefedettség tekintetében nem érik el a megfelelő angol nyelvű erőforrások és eszközök szintjét, amelyek majdnem minden nyelvtechnológiai területen vezetnek. És persze az angol nyelvi erőforrások, különösen a magas minőségű alkalmazások tekintetében is vannak hiányosságok.

Ami a beszédtechnológiát illeti, a jelenlegi technológiák elég jól teljesítenek ahhoz, hogy sikeresen integrálják őket olyan ipari alkalmazásokba, mint például a dialógus- és diktáló rendszerek. A ma elérhető szövegelemző komponensek és nyelvi erőforrások a magyar nyelvben megfigyelhető jelenségek egy részét lefedik, és több sekély nyelvi elemzést nyújtó alkalmazás, mint például a helyesírás-ellenőrzés és néhány információkinyerési feladat részét képezik.

Bonyolultabb alkalmazások, például gépi fordító építéséhez azonban olyan erőforrásokra és technológiákra van szükség, amelyek a nyelvi jelenségek szélesebb körét fedik le, és a szöveg mély szemantikai elemzését is lehetővé teszik. Az alapvető erőforrások és technológiák minőségének és lefedettségének javításával új lehetőségeket nyithatunk további alkalmazási területek, így a jó minőségű gépi fordítás előtt.

4.8 ÖSSZEGZÉS

A fehér könyvek sorozatával megtörtént az első fontos lépés afelé, hogy átfogó felmérést készítsünk 30 európai nyelv nyelvtechnológiájáról, és összehasonlítását is adjuk ezen nyelveknek. A hiányosságok és igények feltárásával az európai nyelvtechnológiai közösségnek és a kapcsolódó területek vezetőinek mostmár lehetősége van egy olyan kutatási-fejlesztési program összeállítására, amelynek célja egy valóban többnyelvű Európa létrehozása.

Láttuk, hogy Európa nyelvei között nagy különbségek vannak. Miközben néhány nyelvre és alkalmazási területre jó minőségű szoftverek és erőforrások léteznek, mások (általában a „kisebb" nyelvek) esetében alapvető hiányosságok vannak. Több nyelvre hiányoznak a szövegelemzéshez szükséges alapvető technológiák és az ezek kifejlesztéséhez elengedhetetlen erőforrások. Másoknak megvannak ezek az eszközei és erőforrásai, de a szemantikai feldolgozás itt is nehézségeket okoz. Nagymértékű erőfeszítés szükséges annak az ambiciózus célnak az eléréséhez, hogy jó minőségű gépi fordítást tudjunk nyújtani minden európai nyelvre.

A magyarországi nyelvtechnológiai helyzet óvatos optimizmusra ad okot. Nagyrészt állami támogatással, de létezik nyelvtechnológiai kutatás Magyarországon. Számos technológia és erőforrás áll rendelkezésre a magyar nyelvre, bár közel sem annyi, mint az angolra, és ezek nem elégségesek egy valódi többnyelvű tudásalapú társadalom igényeinek kielégítésére.

Az angol nyelvre kifejlesztett és arra optimalizált technológiák nehezen vihetők át a magyarra. A magyar nyelv speciális jellege miatt a szintaktikai elemzéshez kifejlesztett angolalapú rendszerek jellemzően rosszul teljesítenek a magyar szövegeken.

A magyar nyelvtechnológiai piac relatíve kicsi. A magyar természetesnyelv-feldolgozás piacát elsősorban egyetemi kutatócsoportok és akadémiai intézetek uralják, de mellettük kisebb cégek is léteznek a piacon.

A fentiekből világossá válik, hogy a magyarországi kutatás-fejlesztéshez, 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 kutatás-fejlesztési támogatások folytonossága nem megfelelő. Rövid távú programok váltják egymást alacsony támogatású időszakokkal, és az EU-s országok és

az Európai Bizottság programjainak koordinációjában is általános hiányosságok mutatkoznak.

Összegzésképpen elmondhatjuk, hogy nagy szükség van egy, a különböző európai nyelvek nyelvtechnológiai felkészültségében mutatkozó különbségek meghaladására fókuszáló, jól koordinált programra, amely az európai nyelveket egy egységként kezeli.

A META-NET hosszútávú célja jó minőségű nyelvtechnológia bevezetése minden nyelvre, a politikai és gazdasági egység elérésének érdekében. A technológia segíthet a meglévő határok ledöntésében és hidak építésében Európa nyelvei között. Ennek érdekében a jövőben minden döntéshozónak, a politika, a kutatás, a gazdaság és a társadalom terén egyaránt, egyesítenie kell erőit.

Kiváló támogatás	Jó támogatás	Közepes támogatás	Töredékes támogatás	Gyenge/semmi támogatás
	angol	cseh finn francia holland német olasz portugál spanyol	baszk bolgár dán észt galíciai görög ír katalán lengyel **magyar** norvég svéd szerb szlovák szlovén	horvát izlandi lett litván máltai román

10: Nyelvi klaszterek a beszédtechnológiában

Kiváló támogatás	Jó támogatás	Közepes támogatás	Töredékes támogatás	Gyenge/semmi támogatás
	angol	francia spanyol	holland katalán lengyel **magyar** német olasz román	baszk bolgár cseh dán észt finn galíciai görög horvát ír izlandi lett litván máltai norvég portugál svéd szerb szlovák szlovén

11: Nyelvi klaszterek a gépi fordításban

Kiváló támogatás	Jó támogatás	Közepes támogatás	Töredékes támogatás	Gyenge/semmi támogatás
	angol	francia holland német olasz spanyol	baszk bolgár cseh dán finn galíciai görög katalán lengyel **magyar** norvég portugál román svéd szlovák szlovén	észt horvát ír izlandi lett litván máltai szerb

12: Nyelvi klaszterek a szövegelemzésben

Kiváló támogatás	Jó támogatás	Közepes támogatás	Töredékes támogatás	Gyenge/semmi támogatás
	angol	cseh francia holland lengyel **magyar** német olasz spanyol svéd	baszk bolgár dán észt finn galíciai görög horvát katalán norvég portugál román szerb szlovák szlovén	ír izlandi lett litván máltai

13: Nyelvi klaszterek az erőforrások esetében

5

A META-NET-RŐL

A META-NET az Európai Bizottság által alapított kiválósági hálózat, amelynek jelenleg 54 tagja van 33 országból[32]. A META-NET támogatja a META-t (Multilingual Europe Technology Alliance), amely az európai nyelvtechnológiával foglalkozó szakértők és intézmények egyre növekvő közössége. 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 hálózat egy olyan Európát támogat, amely egységes digitális piacként és információs térként működik. A META-NET ösztönzi a többnyelvű technológiák létrehozását minden európai nyelvre. Ezek a technológiák az alkalmazások széles körében elérhetővé teszik az automatikus fordítást, az információfeldolgozást és a tudásmenedzsmentet, továbbá intuitív nyelvalapú interfészeket biztosítanak a háztartási elektronikától kezdve a gépészeten és szállítmányozáson keresztül a robotikáig minden területen.

A META-NET 2010. február 1-én alakult azzal a céllal, hogy továbbvigye a már megkezdett tevékenységet három fő irányvonalon. Ezek: a META-VISION, a META-SHARE és a META-RESEARCH.

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 fragmentált és elszigetelt csoportjainak találkozását. Ezt az összefogást példázza az is, hogy jelen fehér könyv a 29 másik nyelvű kiadvánnyal közösen lett előkészítve, a technológiai víziót pedig három különböző ágazati csoport állította össze. A META technológiai tanácsa annak érdekében alakult, hogy megvitassák és előkészítsék a közös vízión alapuló stratégiai kutatási tervet, szoros együttműködésben a nyelvtechnológiai közösséggel.

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 szabványosított 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-RESEARCH** hidakat épít a kapcsolódó technológiai területek között. Ez az irányvonal más területek innovatív fejlesztési eredményeit próbálja meg átemelni a nyelvtechnológiába, és kamatoztatni azokat. Ennek az irányvonalnak a tevékenysége elsősorban a kiváló minőségű gépi fordítás fejlesztésére, adatgyűjtésre és adatelőkészítésre, valamint az eszközök kiértékeléséhez szükséges nyelvi erőforrások előállítására fókuszálódik. További céljai között szerepel a rendelkezésre álló eszközök és módszerek leltárjának elkészítése, valamint workshopok és tréningek szervezése a közösség tagjainak.

office@meta-net.eu – http://www.meta-net.eu

1

EXECUTIVE SUMMARY

Information technology changes our everyday lives. We typically use computers for writing, editing, calculating, and information searching, and increasingly for reading, listening to music, viewing photos and watching movies. We carry small computers in our pockets and use them to make phone calls, write emails, get information and entertain ourselves, wherever we are. How does this massive digitization of information, knowledge and everyday communication affect our language? Will our language change or even disappear?

All our computers are linked together into an increasingly dense and powerful global network. The girl in Ipanema, the officer in Budapest and the engineer in Delhi can all chat with their friends on Facebook, but they are unlikely ever to meet one another in online communities and forums. If they are worried about how to treat earache, they will all check Wikipedia to find out all about it, but even then they won't read the same article. When Europe's netizens discuss the effects of the Fukushima nuclear accident on European energy policy in forums and chat rooms, they do so in cleanly-separated language communities. What the Internet connects is still divided by the languages of its users. Will it always be like this?

Many of the world's 6,000 languages will not survive in a globalised digital information society. It is estimated that at least 2,000 languages are doomed to extinction in the decades ahead. Others will continue to play a role in families and neighbourhoods, but not in the wider business and academic world. What are the Hungarian language's chances of survival?

With its approx. 13 million speakers, Hungarian is 12th on the list of the most populous European languages. It is the official language of the Republic of Hungary, where ca. 97% of the population of 10 million claims Hungarian as their native language. It is also spoken by Hungarian communities in the seven neighbour countries, the largest one being an approx. 1.5 million community in Romania. Additionally, emigrant communities use it worldwide, primarily in the United States, Canada and Israel.

The Hungarian language is an island in Europe – most European languages belong to the Indo-European family of languages, but not Hungarian. It is a Finno-Ugric language, related to Finnish, Estonian and a number of minority languages spoken in the Baltic states and in Russia. It is the most widely spoken non-Indo-European language in Europe, but contrary to world languages such as English and Chinese, or to more commonly used European languages such as German and French, Hungarian does not play a prominent role on the international scene.

There are plenty of complaints in Hungary about the ever-increasing use of Anglicisms, and some even fear that the Hungarian language will become riddled with English words and expressions. But our study suggests that this is misguided. The Hungarian language has already survived the impact of new words and terms from the Old Turkish on the steppes, then later from the Slavs in the Carpathian basin. Moreover, Hungary was a part of the Ottoman Empire for 150 years in the 16th-17th century, then a part of the Habsburg Empire till the first

half of the 20th century. In those times the Latin and German influence was the strongest. One good antidote to losing our lovely little Hungarian words and phrases is to actually use them – frequently and consciously; linguistic polemics about foreign influences and government regulations do not usually help. Our main concern should be not the gradual Anglicisation of our language, but rather its complete disappearance from major areas of our personal lives. Not science, aviation and the global financial markets, we mean the many areas of life in which it is far more important to be close to a country's citizens than to international partners – domestic policies, for example, administrative procedures, the law, culture and shopping.

The status of a language depends not only on the number of speakers, but also on the presence of the language in the digital information space and software applications. The existence of a quite active Hungarian-speaking web community is well demonstrated by the fact that the Hungarian Wikipedia is the 19th largest, ranking higher than commonly used European languages such as Turkish, Romanian or Danish, and world languages such as Arabic or Korean. A few important international software product is available in Hungarian versions, however, due to the special characteristics of Hungarian, the adaptation of English-based applications is quite difficult. Another reason that hinders the development of expensive technologies for Hungarian is the fact that the Hungarian market is quite small.

In the field of language technology, we can be cautiously optimistic about the current state of Hungarian language technology support. There is a viable LT research community in Hungary, which has been supported in the past by national and recently, increasingly, European funding. Currently both of the two EU-funded projects that are coordinated by Hungary in the competitive ICT field come from the language technology domain. A number of large-scale resources and state-of-the-art technologies have been produced and distributed for Hungarian. However, the scope of the resources and the range of tools are still very limited when compared to the resources and tools for the English language, and they are simply not sufficient in quality and quantity to develop the kind of technologies required to support a truly multilingual knowledge society.

Information and communication technology are now preparing for the next revolution. After personal computers, networks, miniaturisation, multimedia, mobile devices and cloud-computing, the next generation of technology will feature software that understands not just spoken or written letters and sounds but entire words and sentences, and supports users far better because it speaks, knows and understands their language. Forerunners of such developments are the free online service Google Translate that translates between 57 languages, or its European counterpart itranslate4.eu (the product of a Hungarian led consortium), IBM's supercomputer Watson that was able to defeat the US-champion in the game of "Jeopardy", and Apple's mobile assistant Siri for the iPhone that can react to voice commands and answer questions in English, German, French and Japanese.

The next generation of information technology will master human language to such an extent that human users will be able to communicate using the technology in their own language. Devices will be able to automatically find the most important news and information from the world's digital knowledge store in reaction to easy-to-use voice commands. Language-enabled technology will be able to translate automatically or assist interpreters; summarise conversations and documents; and support users in learning scenarios.

The next generation of information and communication technologies will enable industrial and service robots (currently under development in research laboratories) to faithfully understand what their users want

them to do and then proudly report on their achievements.

This level of performance means going way beyond simple character sets and lexicons, spell checkers and pronunciation rules. The technology must move on from simplistic approaches and start modelling language in an all-encompassing way, taking syntax as well as semantics into account to understand the drift of questions and generate rich and relevant answers.

However, there is a yawning technological gap between English and Hungarian, and it is currently getting wider. There is a lack of continuity in research and development funding. Short-term coordinated programmes tend to alternate with periods of sparse or zero funding. In addition, there is an overall lack of coordination with programmes in other EU countries and at the European Commission level. As a result, Hungary (and Europe in general) lost several very promising high-tech innovations to the US, where there is greater continuity in their strategic research planning and more financial backing for bringing new technologies to the market. In the race for technology innovation, an early start with a visionary concept will only ensure a competitive advantage if you can actually make it over the finish line. Otherwise all you get is an honorary mention in Wikipedia.

Nevertheless, there is still a high research potential in Hungary and the EU. Apart from internationally renowned research centres and universities, there are a number of innovative small- and medium-sized language technology companies that manage to survive through sheer creativity and immense efforts, despite the lack of venture capital or sustained public funding.

Although Hungary has supported important developments in corpus building and language resource generation, language technology resources and tools for Hungarian clearly do not yet reach the quality and coverage of comparable resources and tools for the English language, which is in the lead in almost all language technology areas. Every international technology competition tends to show that results for the automatic analysis of English are far better than those for Hungarian. This holds true for extracting information from texts, grammar checking, machine translation and a whole range of other applications.

Many researchers reckon that these setbacks are due to the fact that, for fifty years now, the methods and algorithms of computational linguistics and language technology application research have first and foremost focused on English. However, other researchers believe that English is inherently better suited to computer processing. And languages such as Spanish and French are also a lot easier to process than Hungarian using current methods. This means that we need a dedicated, consistent, and sustainable research effort if we want to be use the next generation of information and communication technology in those areas of our private and work life where we live, speak and write Hungarian.

Summing up, despite the prophets of doom the Hungarian language is not in danger, even from the prowess of English language computing. However, the whole situation could change dramatically when a new generation of technologies really starts to master human languages effectively. Through improvements in machine translation, language technology will help in overcoming language barriers, it will only be able to operate between those languages that have managed to survive in the digital world. If there is adequate language technology available, then it will be able to ensure the survival of languages with very small populations of speakers. If not, even 'larger' languages will come under severe pressure.

The dentist jokingly warns: "Only brush the teeth you want to keep". The same principle also holds true for research support policies: You can study every language under the sun all you want, but if you really intend to keep them alive, you also need to develop technologies to support them.

2

LANGUAGES AT RISK: A CHALLENGE FOR LANGUAGE TECHNOLOGY

We are witnesses to a digital revolution that is dramatically impacting communication and society. Recent developments in information and communication technology are sometimes compared to Gutenberg's invention of the printing press. What can this analogy tell us about the future of the European information society and our languages in particular?

The digital revolution is comparable to Gutenberg's invention of the printing press.

After Gutenberg's invention, real breakthroughs in communication were accomplished by efforts such as Luther's translation of the Bible into vernacular language. In subsequent centuries, cultural techniques have been developed to better handle language processing and knowledge exchange:

- the orthographic and grammatical standardisation of major languages enabled the rapid dissemination of new scientific and intellectual ideas;
- the development of official languages made it possible for citizens to communicate within certain (often political) boundaries;
- the teaching and translation of languages enabled exchanges across languages;
- the creation of editorial and bibliographic guidelines assured the quality of printed material;
- the creation of different media like newspapers, radio, television, books, and other formats satisfied different communication needs.

In the past twenty years, information technology has helped to automate and facilitate many processes:

- desktop publishing software has replaced typewriting and typesetting;
- Microsoft PowerPoint has replaced overhead projector transparencies;
- e-mail allows documents to be sent and received more quickly than using a fax machine;
- Skype offers cheap Internet phone calls and hosts virtual meetings;
- audio and video encoding formats make it easy to exchange multimedia content;
- web search engines provide keyword-based access;
- online services like Google Translate produce quick, approximate translations;
- social media platforms such as Facebook, Twitter and Google+ facilitate communication, collaboration, and information sharing.

Although these tools and applications are helpful, they are not yet capable of supporting a fully-sustainable, multilingual European society in which information and goods can flow freely.

2.1 LANGUAGE BORDERS HOLD BACK THE EUROPEAN INFORMATION SOCIETY

We cannot predict exactly what the future information society will look like. However, there is a strong likelihood that the revolution in communication technology is bringing together people who speak different languages in new ways. This is putting pressure both on individuals to learn new languages and especially on developers to create new technology applications to ensure mutual understanding and access to shareable knowledge. In the global economic and information space, there is increasing interaction between different languages, speakers and content thanks to new types of media. The current popularity of social media (Wikipedia, Facebook, Twitter, YouTube, and, recently, Google+) is only the tip of the iceberg.

A global economy and information space confronts us with different languages, speakers and content.

Today, we can transmit gigabytes of text around the world in a few seconds before we recognise that it is in a language that we do not understand. According to a recent report from the European Commission, 57% of Internet users in Europe purchase goods and services in non-native languages; English is the most common foreign language followed by French, German and Spanish. 55% of users read content in a foreign language while 35% use another language to write e-mails or post comments on the Web [2]. A few years ago, English might have been the lingua franca of the Web – the vast majority of content on the Web was in English – but the situation has now drastically changed. The amount of online content in other European (as well as Asian and Middle Eastern) languages has exploded.

Surprisingly, this ubiquitous digital linguistic divide has not gained much public attention; yet, it raises a very pressing question: Which European languages will thrive in the networked information and knowledge society, and which are doomed to disappear?

2.2 OUR LANGUAGES AT RISK

While the printing press helped step up the exchange of information in Europe, it also led to the extinction of many European languages. Regional and minority languages were rarely printed and languages such as Cornish and Dalmatian were limited to oral forms of transmission, which in turn restricted their scope of use. Will the Internet have the same impact on our modern languages?

The wide variety of languages in Europe is one of its richest and most important cultural assets.

Europe's approximately 80 languages are one of our richest and most important cultural assets, and a vital part of this unique social model [3]. While languages such as English and Spanish are likely to survive in the emerging digital marketplace, many European languages could become irrelevant in a networked society. This would weaken Europe's global standing, and run counter to the strategic goal of ensuring equal participation for every European citizen regardless of language. According to a UNESCO report on multilingualism, languages are an essential medium for the enjoyment of fundamental rights, such as political expression, education and participation in society [4].

2.3 LANGUAGE TECHNOLOGY IS A KEY ENABLING TECHNOLOGY

In the past, investments in language preservation focussed primarily on language education and translation. According to one estimate, the European market for translation, interpretation, software localisation and website globalisation was €8.4 billion in 2008 and is expected to grow by 10% per annum [5]. Yet this figure covers just a small proportion of current and future needs in communicating between languages. The most compelling solution for ensuring the breadth and depth of language usage in Europe tomorrow is to use appropriate technology, just as we use technology to solve our transport and energy needs among others.

Language technology targeting all forms of written text and spoken discourse can help people to collaborate, conduct business, share knowledge and participate in social and political debate regardless of language barriers and computer skills. It often operates invisibly inside complex software systems to help us already today to:

- find information with a search engine;
- check spelling and grammar in a word processor;
- view product recommendations in an online shop;
- follow the spoken directions of a navigation system;
- translate web pages via an online service.

Language technology consists of a number of core applications that enable processes within a larger application framework. The purpose of the META-NET language white papers is to focus on how ready these core enabling technologies are for each European language.

Europe needs robust and affordable language technology for all European languages.

To maintain our position in the frontline of global innovation, Europe will need language technology, tailored to all European languages, that is robust and affordable and can be tightly integrated within key software environments. Without language technology, we will not be able to achieve a really effective interactive, multimedia and multilingual user experience in the near future.

2.4 OPPORTUNITIES FOR LANGUAGE TECHNOLOGY

In the world of print, the technology breakthrough was the rapid duplication of an image of a text using a suitably powered printing press. Human beings had to do the hard work of looking up, assessing, translating, and summarising knowledge. We had to wait until Edison to record spoken language – and again his technology simply made analogue copies.

Language technology can now simplify and automate the processes of translation, content production, and knowledge management for all European languages. It can also empower intuitive speech-based interfaces for household electronics, machinery, vehicles, computers and robots. Real-world commercial and industrial applications are still in the early stages of development, yet R&D achievements are creating a genuine window of opportunity. For example, machine translation is already reasonably accurate in specific domains, and experimental applications provide multilingual information and knowledge management, as well as content production, in many European languages.

As with most technologies, the first language applications such as voice-based user interfaces and dialogue systems were developed for specialised domains, and often exhibit limited performance. However, there are huge market opportunities in the education and entertainment industries for integrating language technologies into games, edutainment packages, libraries, simu-

lation environments and training programmes. Mobile information services, computer-assisted language learning software, eLearning environments, self-assessment tools and plagiarism detection software are just some of the application areas in which language technology can play an important role. The popularity of social media applications like Twitter and Facebook suggest a need for sophisticated language technologies that can monitor posts, summarise discussions, suggest opinion trends, detect emotional responses, identify copyright infringements or track misuse.

Language technology helps overcome the "disability" of linguistic diversity.

Language technology represents a tremendous opportunity for the European Union. It can help to address the complex issue of multilingualism in Europe – the fact that different languages coexist naturally in European businesses, organisations and schools. However, citizens need to communicate across the language borders of the European Common Market, and language technology can help overcome this final barrier, while supporting the free and open use of individual languages. Looking even further ahead, innovative European multilingual language technology will provide a benchmark for our global partners when they begin to support their own multilingual communities. Language technology can be seen as a form of "assistive" technology that helps overcome the "disability" of linguistic diversity and makes language communities more accessible to each other. Finally, one active field of research is the use of language technology for rescue operations in disaster areas, where performance can be a matter of life and death: Future intelligent robots with cross-lingual language capabilities have the potential to save lives.

2.5 CHALLENGES FACING LANGUAGE TECHNOLOGY

Although language technology has made considerable progress in the last few years, the current pace of technological progress and product innovation is too slow. Widely-used technologies such as the spelling and grammar correctors in word processors are typically monolingual, and are only available for a handful of languages. Online machine translation services, although useful for quickly generating a reasonable approximation of a document's contents, are fraught with difficulties when highly accurate and complete translations are required. Due to the complexity of human language, modelling our tongues in software and testing them in the real world is a long, costly business that requires sustained funding commitments. Europe must therefore maintain its pioneering role in facing the technological challenges of a multiple-language community by inventing new methods to accelerate development right across the map. These could include both computational advances and techniques such as crowdsourcing.

Technological progress needs to be accelerated.

2.6 LANGUAGE ACQUISITION IN HUMANS AND MACHINES

To illustrate how computers handle language and why it is difficult to program them to process different tongues, let's look briefly at the way humans acquire first and second languages, and then see how language technology systems work.

Humans acquire language skills in two different ways. Babies acquire a language by listening to the real interactions between their parents, siblings and other family members. From the age of about two, children produce

their first words and short phrases. This is only possible because humans have a genetic disposition to imitate and then rationalise what they hear.

Learning a second language at an older age requires more cognitive effort, largely because the child is not immersed in a language community of native speakers. At school, foreign languages are usually acquired by learning grammatical structure, vocabulary and spelling using drills that describe linguistic knowledge in terms of abstract rules, tables and examples.

Humans acquire language skills in two different ways: learning from examples and learning the underlying language rules.

Moving now to language technology, the two main types of systems 'acquire' language capabilities in a similar manner. Statistical (or 'data-driven') approaches obtain linguistic knowledge from vast collections of concrete example texts. While it is sufficient to use text in a single language for training, e. g., a spell checker, parallel texts in two (or more) languages have to be available for training a machine translation system. The machine learning algorithm then "learns" patterns of how words, short phrases and complete sentences are translated.

This statistical approach usually requires millions of sentences to boost performance quality. This is one reason why search engine providers are eager to collect as much written material as possible. Spelling correction in word processors, and services such as Google Search and Google Translate, all rely on statistical approaches. The great advantage of statistics is that the machine learns quickly in a continuous series of training cycles, even though quality can vary randomly.

The second approach to language technology, and to machine translation in particular, is to build rule-based systems. Experts in the fields of linguistics, computational linguistics and computer science first have to encode grammatical analyses (translation rules) and compile vocabulary lists (lexicons). This is very time consuming and labour intensive. Some of the leading rule-based machine translation systems have been under constant development for more than 20 years. The great advantage of rule-based systems is that the experts have more detailed control over the language processing. This makes it possible to systematically correct mistakes in the software and give detailed feedback to the user, especially when rule-based systems are used for language learning. However, due to the high cost of this work, rule-based language technology has so far only been developed for a few major languages.

As the strengths and weaknesses of statistical and rule-based systems tend to be complementary, current research focusses on hybrid approaches that combine the two methodologies. However, these approaches have so far been less successful in industrial applications than in the research lab.

As we have seen in this chapter, many applications widely used in today's information society rely heavily on language technology, particularly in Europe's economic and information space. Although this technology has made considerable progress in the last few years, there is still huge potential to improve the quality of language technology systems. In the next section, we describe the role of Hungarian in European information society and assess the current state of language technology for the Hungarian language.

3

THE HUNGARIAN LANGUAGE IN THE EUROPEAN INFORMATION SOCIETY

3.1 GENERAL FACTS

Hungarian is the most widely spoken non-Indo-European language in Europe. It is the official language of the Republic of Hungary, where ca. 97% of the population of 10 million claims Hungarian as their native language. It is also spoken by Hungarian communities in the seven neighbour countries, the largest one being an approximately 1.5 million community in Romania. Additionally, emigrant communities use it worldwide, primarily in the United States, Canada and Israel. With its 13 million speakers, Hungarian is 12th on the list of the most populous European languages [6]. Abroad, Hungarian is an official language in Vojvodina, as well as in three municipalities in Slovenia. Hungarian is officially recognised as a minority or regional language in Austria, Croatia, Romania, Ukraine, and Slovakia.

It is interesting that Hungarian barely has any major variety: its dialects differ very little from each other and from the standard, and spelling is particularly uniform. This may be the result of a long-term co-existence, which – by continuously clashing with other languages – may have launched speakers on a road to standardisation. According to the traditional categorisation, there are seven dialects identified in the present area of Hungary. These dialects are, for the most part, mutually intelligible. Two additional Hungarian dialects exist in Romania: Székely and Csángó.

There is scant difference between the Hungarian used in the Republic of Hungary and that used in neighbouring countries. Of course, minor but characteristic differences are present. While the variety in Hungary developed under fundamentally German influence, Romanian Hungarian has been mostly influenced by Romanian. The Csángó minority group has been largely isolated from other Hungarian people, and they therefore preserved a dialect closely resembling medieval Hungarian.

> Hungarian is the most widely spoken non-Indo-European language in Europe.

3.2 PARTICULARITIES OF THE HUNGARIAN LANGUAGE

Most European languages belong to the Indo-European family of languages, but not Hungarian! It is a Uralic language, part of the Ugric group, related to Finnish, Estonian and a number of minority languages spoken in the Baltic states and in Russia.

Uralic languages share a few ancient characteristics, such as:

- There is no gender in Hungarian: the same word (*ő*) expresses the concepts of both 'he' and 'she'.
- There are only two verb tenses: present and past. Their variations and the future tense may be circumscribed.

- The so-called 'direction triad': there are 3x3 of each set of location cases, as shown by table 1 using the word *doboz* ('box') (the determiner *a* ('the') not being subject to declension).

Hungarian is written in Roman letters; nonetheless, Hungarian texts do not resemble any other European language. Below are two lines from a classic poem (from Ferenc Kölcsey's 1823 poem *Hymnus*, forming the lyrics of the Hungarian national anthem), in simple literal translation:

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

Not a single word in the text is recognisable on the basis of the average European vocabulary; not only do Hungarians refer to God as *Isten*, they do not even call themselves 'Hungarian'; they call themselves *magyar*. But there is more to this than differences in individual words:

Isten	áldd	meg	a	magyart
God	bless	?	the	Hungarian

The word denoted with the question mark does not exist in most languages: its name is *igekötő* ('verbal prefix'). It plays a multitude of roles, here expressing the perfect tense, i. e., it indicates a completed action. One of the beauties (and difficulties) of the Hungarian language lies precisely within the usage of verbal prefixes. Now let us examine the second line:

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

Where English uses the preposition *with*, Hungarian uses suffixes. Hungarian does not feature any prepositions. In this example the suffixes *-vel* and *-gel* express what English expresses by means of *with*.

Another important feature of Hungarian is the possessive structure, the reverse of its counterparts in Indo-European languages. For example, in *Paul's radio*, Hungarian does not attach a suffix to the possessor (Paul), but rather to his possession, the radio: *Pál rádió-ja*, literally: 'Paul radio-his'.

It is more of a cultural-historical rather than a linguistic curiosity that in Hungarian the family name comes first, with the 'utónév' ('given name, Christian name') behind, thus the regular order is Liszt Ferenc (=Franz Liszt), Bem József (=József Bem), Bartók Béla, Márai Sándor, etc.

Hungarian is called a synthetic language: for the most part, it expresses grammatical elements in a single word form using affixes, as opposed to isolating languages, which tend to employ separate words, e. g., prepositions, pronouns, auxiliaries, for expressing grammatical phenomena. For example, the Hungarian equivalent of the English auxiliary *can* is the suffix *-hat/-het*.

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

Suffixes, often multiple ones, must be attached to the word stem in strict order, thus words may grow to stunning lengths. This type of synthetic word formation is called agglutination (meaning 'gluing of words'). For example: *bolondozhattunk* "we could fool [around]" (='fool-verb-can-past-we'); *ösztönözhettünk* "we could stimulate" (='stimulus-verb-can-past-we'). The structure of the two words is identical – the apparent difference is caused by the difference of vowels, which is due to the so-called vowel harmony (also known as assimilation). The vowels are relegated into one of two classes: "deep": *a o u*, and "high": *e i ö ü*. In the suffixes, the vowel appears to fit the stem: *bolond* is deep, thus the vowels in the suffixes are also deep: *o - o + o - a - u*, while *ösztön* is high, therefore the other vowels are high as well: *ö - ö + ö - e - ü* [6].

	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

1: The so-called 'direction triad' using the word *doboz* ('box')

3.3 RECENT DEVELOPMENTS

In a way, Hungarian has always been a minority language that continuously adopted words into its vocabulary from other peoples present in the Carpathian basin: Slavs (primarily Slovaks, Serbs, Croats), and later German, Romanian, Jewish and Roma populations. Latin was used as the official language as late as the beginning of the 19th century, being the language of public administration and science. The Hungarian Parliament introduced legislative sessions in Hungarian only from 1844 onward.

Hungarian has always been more of an importer than an exporter. The current vocabulary contains numerous words borrowed from Slavic, Latin, Romanian, and Italian. The German influence was the strongest, since Hungary was a part of the Habsburg Empire for 400 years. There is a vast number of words of German origin, including *tánc* 'dance' and *hering* 'herring'. Lexical borrowing continues to this day: from French *fritőz* 'friteuse', *bagett* 'baguette'; from Italian *maffiózó* 'Mafioso', *paparazzi*; from English *fitnesz* 'fitness', *szerver* 'server', etc. Nowadays loan words are usually Anglicisms, due to the strong influence of American films, popular music, and technology (including the Internet).

3.4 OFFICIAL LANGUAGE PROTECTION IN HUNGARY

Hungary has two main institutional bodies that play an active role in the promotion of the Hungarian language: the Balassi Bálint Institute, which was founded by the Ministry of Education, and the Research Institute for Linguistics of the Hungarian Academy of Sciences.

Hungary has two main institutional bodies that play an active role in the promotion of the Hungarian language.

The Balassi Bálint Institute, founded in 2002, was launched to promote Hungarian language culture, analogously to the well-known British Council and Goethe Institute. It contributes to the teaching of Hungarian language and Hungarian studies for foreigners living in Hungary. In co-operation with the international network of institutions for Hungarian studies, it promotes the education and research of the Hungarian language and culture abroad. The Balassi Institute foresees the cultivation of the Hungarian language and education of Hungarians living beyond the borders of Hungary, participates in the linguistic and terminological follow-up training of teachers of Hungarian and other

experts abroad, as well as organises courses of Hungarian studies and minority rights [7].

The Research Institute for Linguistics is among the leading institutions in the field of research on the Hungarian language. It was founded in 1949, and placed under the direction of the Hungarian Academy of Sciences in 1951. Its primary tasks include research in Hungarian linguistics, general, theoretical and applied linguistics, Uralic linguistics, and phonetics. The Institute's tasks include the preparation of a comprehensive dictionary of the Hungarian language, and the maintenance of its archival materials. Its research projects investigate various aspects of Hungarian as well as minority languages in and outside Hungary, and deal with issues of language policy within the framework of the European integration. Further activities include the compilation of linguistic corpora and databases, and the laying of the linguistic groundwork for language technology applications. Besides, the Institute operates a public counselling service on language and linguistics, and runs the Theoretical Linguistics undergraduate and PhD programmes, jointly with Eötvös Loránd University [8].

Hungarian orthography is under strict academic control: the rules of Spelling Committee of the Hungarian Academy of Sciences are intended to use. The regulations are not obligatory, but misspellings can certainly cause loss of prestige.

These days, many enthusiastic traditionalists argue that the neologisms originating from the English language threaten the Hungarian rather than enrich it. As a result of their "language protecting" activities, the so-called "language law" was ratified in 2002, which demands that English advertisements and slogans must be replaced by Hungarian equivalents. Additional measures for protecting the status of the Hungarian language have also been taken: for example, a television and radio quota regulating the percentage of music sung in Hungarian was introduced at the beginning of 2011.

3.5 LANGUAGE IN EDUCATION

From 1844, when Hungarian became the official language of public administration, science and education, elementary school children have the possibility to have lessons in Hungarian. After the Education Reform Act of 1868, Hungarian became the language of higher education. Diplomas in Hungarian may be earned at numerous institutions of higher education beyond the border at Hungarian universities and colleges: from Nyitra (Nitra, Slovakia) all the way to Újvidék (Novi Sad, Serbia) or Kolozsvár (Cluj-Napoca, Romania).

From the 19th century onward, Hungarian language and literature have played an important role in education. The study of Hungarian is compulsory from age 6 to 18. In elementary school, from age 6 to 10, the teaching requirements are divided into key areas of reading, writing and composition. After age 10, grammar and literature are taught separately.

According to the PISA 2009 study that aims to measure literary reading skills of teenagers, Hungary became the member of countries whose results are not statistically significantly different from the OECD average. The overall reading score in Hungary is comparable with those of Germany, France and the UK [9].

3.6 INTERNATIONAL ASPECTS

Hungary has a great number of world famous physicists (Ede Teller, Jenő Wigner and Leó Szilárd, who contributed to the Manhattan Project), mathematicians (Alfréd Rényi, Paul Erdős, the latter being the author of the Erdős number), and musicians (Franz Liszt, Béla Bartók). Hungarian scientists have won several Nobel prizes in physics, chemistry, and medicine.

As everywhere in the scientific world, Hungarian scholars face a great deal of pressure to publish in international, English-language journals, which leads to a self-perpetuating cycle that stresses the importance of Eng-

lish. The situation is similar in the business world: in many large and internationally active companies, English has become the *lingua franca*, both in written and oral communication.

However, according to a survey in 2005 [10], the number of foreign language speakers in Hungary is below the European average: the percentage of Hungarian people who speak at least one foreign language is 35%. Language technology can address this challenge from a different perspective by offering services such as machine translation or cross-lingual information retrieval, and thereby help diminish personal and economic disadvantages naturally faced by non-native speakers of English.

3.7 HUNGARIAN ON THE INTERNET

In 2009, 61.6% of the people in Hungary were Internet users [11]. Among young people aged 14-17, the proportion was even higher. The Internet penetration was below the European average, but it has been increasing steadily. In January 2011, the number of domains under the .hu public domains was near 600,000, and is rising [12]. About 70,000 domains exist in the country outside of the .hu system (most of them .com) [13].

> The Hungarian Wikipedia is the 19th largest, before more commonly used European and world languages.

According to a European study in 2010, the usage of community pages such as Facebook is above the European average – which may be due to the pre-existence of a quite popular Hungarian community site called iWiW. The existence of a quite active Hungarian-speaking web community is also reflected by the fact that the Hungarian Wikipedia is the 19th largest, before more commonly used European languages such as Turkish, Romanian or Danish, and world languages such as Arabic or Korean.

The growing importance of the Internet is critical for language technology. The vast amount of digital language data is a key resource for analysing the usage of natural language, in particular, for collecting statistical information about patterns. And the Internet offers a wide range of application areas for language technology. The most commonly used web application is search, which involves the automatic processing of language on multiple levels (see Chapter 4). Web search involves sophisticated language technology that differs for each language. For Hungarian this comprises taking into account the different inflectional endings of nouns, adjectives and verbs, and different stem forms like *ló* ('horse, single') and *lovak* ('horses, plural').

In Hungary there is no officially ratified law on equal opportunities for the disabled, but a design guide for the implementation of complex accessibility has been developed by the Public Foundation for Equal Opportunities of Persons with Disabilities. It recommends public agencies to ensure that the disabled can use their websites and Internet services without any restrictions. User-friendly language technology tools are a key solution by offering for example speech synthesis to enunciate the content of web pages for the blind.

Internet users and providers of web content can also use language technology in less obvious ways, for example, by automatically translating web page contents from one language into another. Despite the high cost of manually translating this content, comparatively little language technology has been developed and applied to the issue of website translation in light of the supposed need. This may be due to the complexity of the Hungarian language and to the range of different technologies involved in typical applications.

The next chapter gives an introduction to language technology and its core application areas, together with an evaluation of current language technology support for Hungarian.

4

LANGUAGE TECHNOLOGY SUPPORT FOR HUNGARIAN

Language technology is used to develop software systems designed to handle human language and are therefore often called "human language technology". Human language comes in spoken and written forms. While speech is the oldest and in terms of human evolution the most natural form of language communication, complex information and most human knowledge is stored and transmitted through the written word. Speech and text technologies process or produce these different forms of language, using dictionaries, rules of grammar, and semantics. This means that language technology (LT) links language to various forms of knowledge, independently of the media (speech or text) in which it is expressed. Figure 1 illustrates the LT landscape.

When we communicate, we combine language with other modes of communication and information media – for example speaking can involve gestures and facial expressions. Digital texts link to pictures and sounds. Movies may contain language in spoken and written form. In other words, speech and text technologies overlap and interact with other multimodal communication and multimedia technologies.

In this section, we will discuss the main application areas of language technology, i. e., language checking, web search, speech interaction, and machine translation. These applications and basic technologies include

- spelling correction,
- authoring support,
- computer-assisted language learning,
- information retrieval,
- information extraction,
- text summarisation,
- question answering,
- speech recognition and
- speech synthesis.

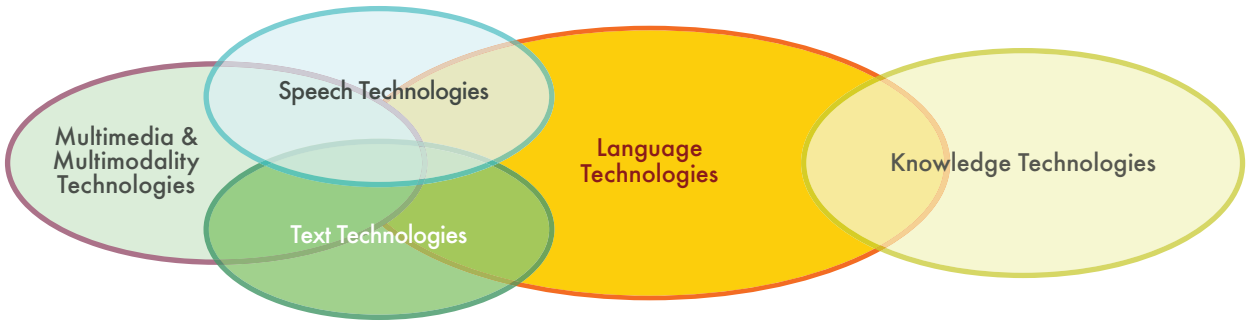
Language technology is an established area of research with an extensive set of introductory literature. The interested reader is referred to the following references: [14, 15, 16, 17].

Before discussing the above application areas, we will briefly describe the architecture of a typical LT system.

4.1 APPLICATION ARCHITECTURES

Software applications for language processing typically consist of several components that mirror different aspects of language. While such applications tend to be very complex, figure 2 shows a highly simplified architecture of a typical text processing system. The first three modules handle the structure and meaning of the text input:

1. Pre-processing: cleans the data, analyses or removes formatting, detects the input language, handles the specific accented letters (*á, é, í, ó, ö, ő, ú, ü, ű*) in Hungarian, and so on.



2: Language technologies

2. Grammatical analysis: finds the verb, its objects, modifiers and other sentence elements; detects the sentence structure.

3. Semantic analysis: performs disambiguation (i. e., computes the appropriate meaning of words in a given context); resolves anaphora (i. e., which pronouns refer to which nouns in the sentence); represents the meaning of the sentence in a machine-readable way.

After analysing the text, task-specific modules can perform other operations, such as automatic summarisation and database look-ups.

In the remainder of this section, we firstly introduce the core application areas for language technology, and follow this with a brief overview of the state of LT research and education today, and a description of past and present research programmes. Finally, we present an expert estimate of core LT tools and resources for Hungarian in terms of various dimensions such as availability, maturity and quality. The general situation of LT for the Hungarian language is summarised in figure 7 (p. 61) at the end of this chapter. LT support for Hungarian is also compared to other languages that are part of this series.

4.2 CORE APPLICATION AREAS

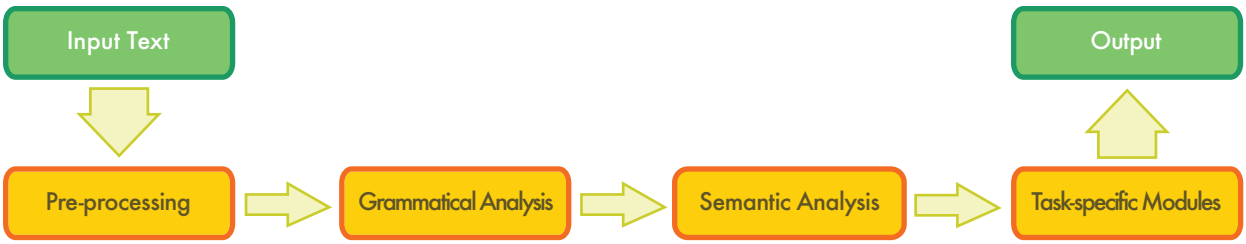
In this section, we focus on the most important LT tools and resources, and provide an overview of LT activities in Hungary.

4.2.1 Language Checking

Anyone who has used a word processor such as Microsoft Word knows that it has a spell checker that highlights spelling mistakes and proposes corrections. The first spelling correction programs compared a list of extracted words against a dictionary of correctly spelled words. Today these programs are far more sophisticated. Using language-dependent algorithms for grammatical analysis, they detect errors related to morphology (e. g., plural formation) as well as syntax–related errors, such as a missing verb or a conflict of verb-subject agreement (e. g., *she *write a letter*). However, most spell checkers will not find any errors in the following text [18]:

> I have a spelling checker,
> It came with my PC.
> It plane lee marks four my revue
> Miss steaks aye can knot sea.

Handling these kinds of errors usually requires an analysis of the context. For example: there are inflected word forms in Hungarian that can hold several meanings, e. g.,



3: A typical text processing architecture

the word *várunk* can be an inflected form of the verb *vár*, or the noun *vár* with possessive inflection.

This type of analysis either needs to draw on language-specific grammars laboriously coded into the software by experts, or on a statistical language model. In this case, a model calculates the probability of a particular word as it occurs in a specific position (e. g., between the words that precede and follow it). For example: *várunk* is probably not a verb if the sentence contains an other finite verb. A statistical language model can be automatically created by using a large amount of (correct) language data (called a text corpus). Most of these two approaches have been developed around data from English. Neither approach can transfer easily to Hungarian because it has a flexible word order and a richer inflection system.

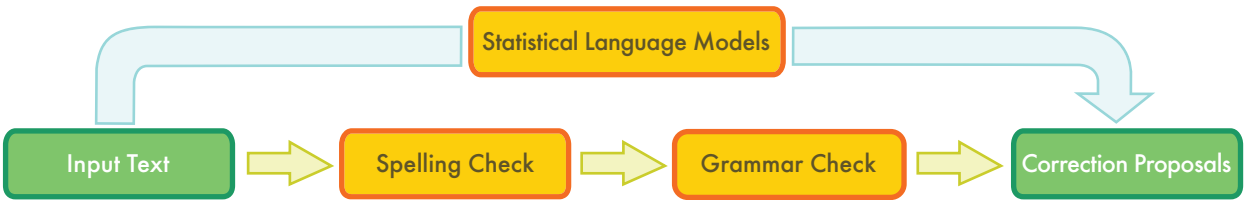
Language checking is not limited to word processors; it is also used in authoring support systems, i. e., software environments in which manuals and other documentation are written to special standards for complex IT, healthcare, engineering and other products. Fearing customer complaints about incorrect use and damage claims resulting from poorly understood instructions, companies are increasingly focusing on the quality of technical documentation while targeting the international market (via translation or localisation) at the same time. Advances in natural language processing have led to the development of authoring support software, which helps the writer of technical documentation use vocabulary and sentence structures that are consistent with industry rules and (corporate) terminology restrictions.

The use of language checking is not limited to word processors. It also applies to authoring support systems.

As Hungarian is a highly agglutinative language, a Hungarian spell checker must contain a morphological analyzer that handles the great number of affixes and complex words. The first spell checker for Hungarian has been developed by combining a spell checking system and a morphological model by a Hungarian SME called MorphoLogic [19], in the late 80s. Their program (*Helyes-e?*) is available for MS Office, QuarkXPress, Adobe InDesign and other desktop publisher packages. MorphoLogic developed grammar and style checkers that recognise spelling errors based on the context. The program indicates possible mistakes and leaves it to the user to decide whether it is a real mistake.

An open source spell checker for Hungarian exists as well. Hunspell [20] is based on MySpell, and it has been integrated into OpenOffice, Mozilla Firefox, Mozilla Thunderbird and Google Chrome.

Besides spell checkers and authoring support, language checking is also important in the field of computer-assisted language learning. And language checking applications also automatically correct search engine



4: Language checking (top: statistical; bottom: rule-based)

queries, as found in Google's *Did you mean...* suggestions.

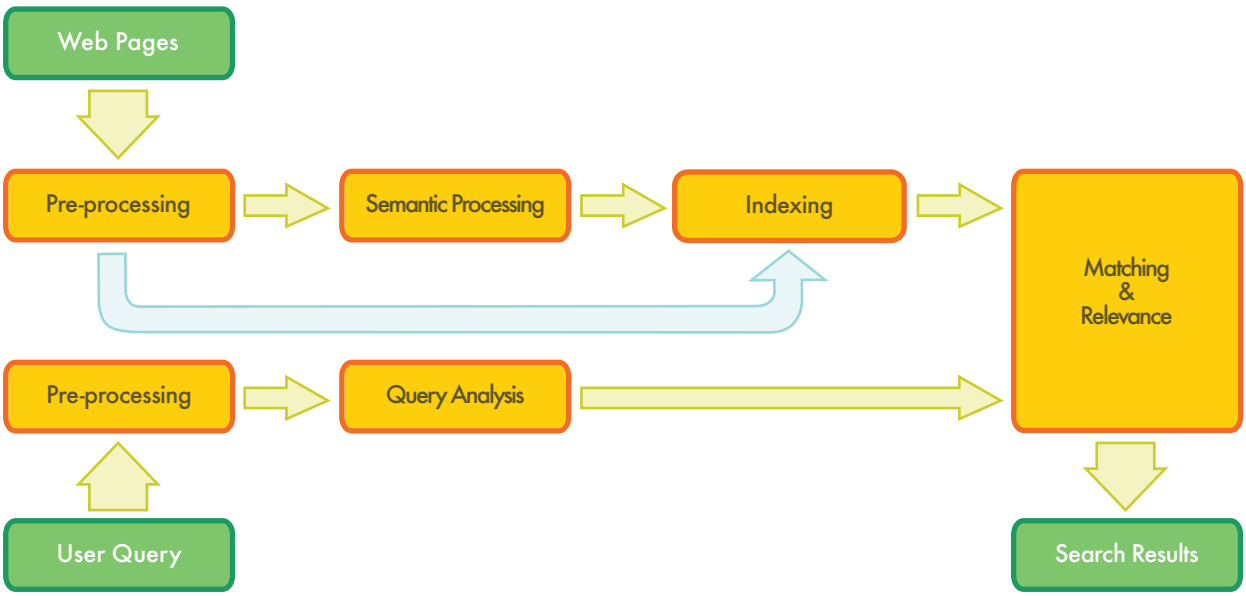
4.2.2 Web Search

Searching the Web, intranets or digital libraries is probably the most widely used yet largely underdeveloped language technology application today. The Google search engine, which started in 1998, now handles about 80% of all search queries [21]. The verb *guglizni* is commonly used in Hungarian, even though it has not made its way into printed dictionaries yet. The Google search interface and results page display has not significantly changed since the first version. Yet in the current version, Google offers spelling correction for misspelled words and has now incorporated basic semantic search capabilities that can improve search accuracy by analysing the meaning of terms in a search query context [22]. The Google success story shows that a large volume of available data and efficient indexing techniques can deliver satisfactory results for a statistically-based approach.

The next generation of search engines will have to include much more sophisticated language technology.

For more sophisticated information requests, it is essential to integrate deeper linguistic knowledge for text interpretation. Experiments using lexical resources such as machine-readable thesauri or ontological language resources (e. g, WordNet for English or Hungarian WordNet for Hungarian) have demonstrated improvements in finding pages using synonyms of the original search terms, such as *atomenergia* [atomic energy], *magenergia* [atomic power] and *nukleáris energia* [nuclear energy], or even more loosely related terms.

The next generation of search engines will have to include much more sophisticated language technology, in particular in order to deal with search queries consisting of a question or other sentence type rather than a list of keywords. For the query, *Give me a list of all companies that were taken over by other companies in the last five years*, the LT system needs to analyse the sentence syntactically and semantically as well as provide an index to quickly retrieve relevant documents. A satisfactory answer will require syntactic parsing to analyse the grammatical structure of the sentence and determine that the user wants companies that have been acquired, not companies that acquired other companies. For the expression *last five years*, the system needs to determine the relevant years. And, the query needs to be matched against a huge amount of unstructured data to find the piece or pieces of relevant information the user wants. This is called information retrieval, and involves searching and ranking relevant documents. To generate a list of companies, the system also needs to recognise a particular string of words in a document as a company name, a process called named entity recognition.



5: Web search architecture

A more demanding challenge is matching a query in one language with documents in another language. Cross-lingual information retrieval involves automatically translating the query into all possible source languages and then translating the results back into the target language.

Now that data is increasingly found in non-textual formats, there is a need for services that deliver multimedia information retrieval by searching images, audio files and video data. In the case of audio and video files, a speech recognition module must convert the speech content into text (or into a phonetic representation) that can then be matched against a user query.

For inflectional languages like Hungarian, it is important to be able to search for all the inflected forms of a word simultaneously, instead of having to enter each different form separately. For this purpose, several morphological parsers exist for Hungarian. NP chunkers for identifying noun phrases provide higher level parsing: a statistical and a rule-based application have been developed for Hungarian.

Due to the variable word order characteristic of Hungarian, we cannot rely on exploiting particular linear configurations alone when syntactic parsers are developed. On the other hand, Hungarian is an agglutinative language with rich case marking, and morphological case markers and postpositions lend themselves to being used as cues for parsing. A database of Hungarian verbs and case markers of their arguments was developed at the Research Institute for Linguistics, which has been built in higher level parsing applications, e. g., for automatic acquisition of verb argument frames, or rule-based syntactic parsing. More syntactic parsers for Hungarian exist – one of them was built in the Hungarian treebank (Szeged Treebank) and in a rule-based Hungarian-English machine translation program (MetaMorpho).

Focus on development for HLT companies and research institutes lies on providing trend- and text-analysis tools which integrate natural language processing tools to find the relevant information in unstructured text. For this purpose part-of-speech taggers, dependency parsers

and named entity recognisers have been developed for Hungarian, which are mostly based on statistical learning algorithms.

A meta-search and clustering engine is PolyMeta [23]. It enables organisations and individuals to simultaneously search diverse information resources on the Web with a common interface. It employs natural language processing and information retrieval algorithms in its query analysis and refinement, search strategy, relevancy ranking, focused drill-down and exploration of multi-dimensional information spaces.

Certainly, not only SMEs try to extract information by natural language processing tools. Several projects have been running in the academia with the aim of developing a model-based semantic search system, creating the framework of a unified Hungarian ontology, or creating a semantically structured, general purpose Hungarian concept set on the basis of the results and formalism of EuroWordNet language ontology (Hungarian WordNet).

4.2.3 Speech interaction

Speech interaction is one of many application areas that depend on speech technology, i. e., technologies for processing spoken language. Speech interaction technology is used to create interfaces that enable users to interact in spoken language instead of using a graphical display, keyboard and mouse.

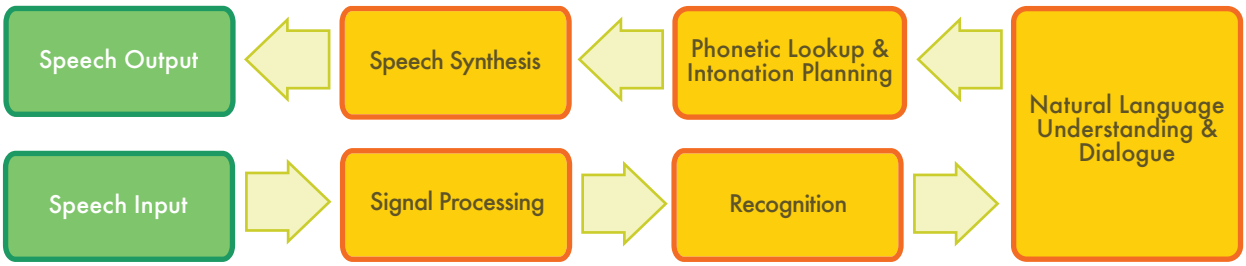
Speech interaction is the basis for creating interfaces that allow a user to interact with spoken language instead of a graphical display, keyboard and mouse.

Today, these voice user interfaces (VUI) are used for partially or fully automated telephone services provided by companies to customers, employees or partners. Business domains that rely heavily on VUIs include banking, supply chain, public transportation, and telecommunications. Other uses of speech interaction technology include interfaces to car navigation systems and the use of spoken language as an alternative to the graphical or touchscreen interfaces in smartphones.

Speech interaction technology comprises four technologies:

1. Automatic speech recognition (ASR) determines which words are actually spoken in a given sequence of sounds uttered by a user.
2. Natural language understanding analyses the syntactic structure of a user's utterance and interprets it according to the system in question.
3. Dialogue management determines which action to take given the user input and system functionality.
4. Speech synthesis (text-to-speech or TTS) transforms the system's reply into sounds for the user.

One of the major challenges of ASR systems is to accurately recognise the words a user utters. This means restricting the range of possible user utterances to a limited set of keywords, or manually creating language models that cover a large range of natural language utterances. Using machine learning techniques, language models can also be generated automatically from speech corpora, i. e., large collections of speech audio files and text transcriptions. Restricting utterances usually forces people to use the voice user interface in a rigid way and can damage user acceptance; but the creation, tuning and maintenance of rich language models will significantly increase costs. VUIs that employ language models and initially allow a user to express their intent more flexibly – prompted by a *How may I help you?* greeting – tend to be automated and are better accepted by users. Companies tend to use utterances pre-recorded by professional speakers for generating the output of the voice user interface. For static utterances where the wording



6: Speech-based dialogue system

does not depend on particular contexts of use or personal user data, this can deliver a rich user experience. But more dynamic content in an utterance may suffer from unnatural intonation because different parts of audio files have simply been strung together. Today's TTS systems are getting better (though they can still be optimised) at producing natural-sounding dynamic utterances.

Interfaces in speech interaction have been considerably standardised during the last decade in terms of their various technological components. There has also been strong market consolidation in speech recognition and speech synthesis. The national markets in the G20 countries (economically resilient countries with high populations) have been dominated by just five global players, with Nuance (USA) and Loquendo (Italy) being the most prominent players in Europe. In 2011, Nuance announced the acquisition of Loquendo, which represents a further step in market consolidation.

Due to the specific characteristics of Hungarian, the widely used methods in speech interaction technology are difficult or impossible to adapt for Hungarian. However, the methods developed for Hungarian can be applied for similar languages, e. g., Finnish, Turkish, Arabic, in the field of TTS and ASR.

The Hungarian TTS market is dominated by research groups at Budapest University of Technology and Economics [24]. The most widely used TTS system is Profivox, available since 2002, which has been built into SMS- and email-reader softwares, into in-car and mobilephone GPS systems, and into e-book- and screen-reader applications which can help the integration of blind people into information society. A high level interactive development tool is also available for supporting special research (psychology and prosody research). The software supports a supervised generation of speech stimuli with predefined acoustic and prosodic content, based on TTS technology. Hungarian pronunciation electronic dictionary also exists for 1.5 million word forms. This may be the basis of the development of a Hungarian TTS symbol conversion tool.

On the Hungarian ASR market there are additional smaller companies, such as Applied Logic Laboratory, Aitia, Digital Natives, as well as academic research groups, e. g., at the University of Szeged. In spite of the linguistic difficulties mentioned above, several speech recogniser applications for Hungarian have been developed over the last few years. One of them is a prosodic recogniser that was prepared by a cross-lingual study for agglutinative, fixed stressed languages, such as Hungarian and Finnish, about the segmentation of continuous speech on word level by examination of supra-segmental parameters. Another system helps the work of doctors: during examining the patient they dictate the diagnosis which will be automatically transcribed. Yet another one is a Hungarian computer assisted speech pronun-

ciation learning and training system for speech handicapped and for language learning, which is adapted for some European languages, as well. Further application areas are call centres, dialogue systems, or indexing and searching media databases.

Looking ahead, there will be significant changes, due to the spread of smartphones as a new platform for managing customer relationships, in addition to fixed telephones, the Internet and e-mail. This will also affect how speech interaction technology is used. In the long term, there will be fewer telephone-based VUIs, and spoken language apps will play a far more central role as a user-friendly input for smartphones. This will be largely driven by stepwise improvements in the accuracy of speaker-independent speech recognition via the speech dictation services already offered as centralised services to smartphone users.

4.2.4 Machine Translation

The idea of using digital computers to translate natural languages can be traced back to 1946 and was followed by substantial funding for research during the 1950s and again in the 1980s. Yet machine translation (MT) still cannot deliver on its initial promise of providing across-the-board automated translation.

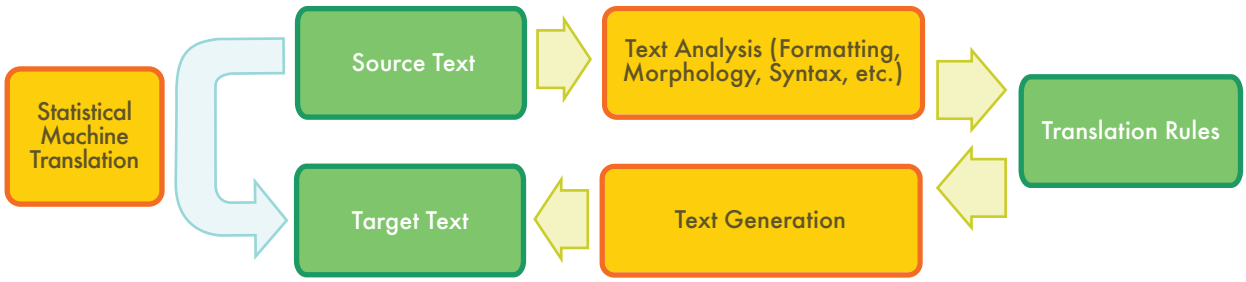
At its basic level, machine translation simply substitutes words in one natural language with words in another language.

The most basic approach to machine translation is the automatic replacement of the words in a text written in one natural language with the equivalent words of another language. This can be useful in subject domains that have a very restricted, formulaic language such as weather reports. However, in order to produce a good translation of less restricted texts, larger text units (phrases, sentences, or even whole passages) need to be matched to their closest counterparts in the target language. The major difficulty is that human language is ambiguous. Ambiguity creates challenges on multiple levels, such as word sense disambiguation at the lexical level (a *jaguar* is a brand of car or an animal) or the assignment of case on the syntactic level, for example:

- *A rendőr látta az embert a távcsővel.*
  'The policeman saw the man with the telescope.'
- *A rendőr látta az embert a revolverrel.*
  'The policeman saw the man with the revolver.'

One way to build an MT system is to use linguistic rules. For translations between closely related languages, a translation using direct substitution may be feasible in cases such as the above example. However, rule-based (or linguistic knowledge-driven) systems often analyse the input text and create an intermediary symbolic representation from which the target language text can be generated. The success of these methods is highly dependent on the availability of extensive lexicons with morphological, syntactic, and semantic information, and large sets of grammar rules carefully designed by skilled linguists. This is a very long and therefore costly process.

In the late 1980s when computational power increased and became cheaper, interest in statistical models for machine translation began to grow. Statistical models are derived from analysing bilingual text corpora, parallel corpora, such as the Europarl parallel corpus, which contains the proceedings of the European Parliament in 21 European languages. Given enough data, statistical MT works well enough to derive an approximate meaning of a foreign language text by processing parallel versions and finding plausible patterns of words. Unlike knowledge-driven systems, however, statistical (or data-driven) MT systems often generate ungrammatical output. Data-driven MT is advantageous because less human effort is required, and it can also cover special par-



7: Machine translation (left: statistical; right: rule-based)

ticularities of the language (e. g., idiomatic expressions) that are often ignored in knowledge-driven systems.
The strengths and weaknesses of knowledge-driven and data-driven machine translation tend to be complementary, so that nowadays researchers focus on hybrid approaches that combine both methodologies. One such approach uses both knowledge-driven and data-driven systems, together with a selection module that decides on the best output for each sentence. However, results for sentences longer than, say, 12 words, will often be far from perfect. A more effective solution is to combine the best parts of each sentence from multiple outputs; this can be fairly complex, as corresponding parts of multiple alternatives are not always obvious and need to be aligned.

Machine Translation is particularly challenging for the Hungarian language.

Machine translation is particularly challenging for the Hungarian language. The free word order and split verb constructions pose problems for analysis; and extensive inflection is a challenge for generating words with proper case markings.
There are knowledge- and data-driven solutions on the Hungarian MT market, as well. MorphoLogic, a private R&D company offers both desktop machine translation programs and online services. MorphoWord translates between English and Hungarian. Their MT system integrates rule-based and statistical methods, its main component being a parser that creates an intermediary representation, from which it produces the text in the target language.

Availability of large amounts of bilingual texts is actually the key in statistical MT. The Hunglish Corpus is a free sentence-aligned Hungarian-English parallel corpus of about 54.2 million words in 2.07 million sentences. At present this is the largest Hungarian-English parallel corpus. Sentence alignment was performed with hunalign, which is one of the most widely used sentence level aligners, developed by Hungarian researchers at Budapest University of Technology and Economics. The corpus may be searched through an online sentence search service [25], which can be used as a raw translator or a smart bilingual lexicon.

The iTranslate4.eu [26] project started off in March 2010, which intends to provide online translation solution for all European languages. It does not only offer full coverage of EU languages, but also provides for each language pair the best quality available at the time, and mediates easy transfer to professional translators. The project is carried out by a consortium of European MT companies that have developed the best translation system for at least one language pair. The project has two Hungarian participants: the consortium leader is the Research Institute for Linguistics of Hungarian

Academy of Sciences, and MorphoLogic provides the common API for the services.

The use of machine translation can significantly increase productivity provided the system is intelligently adapted to user-specific terminology and integrated into a workflow. Special systems for interactive translation support were developed, for example, at Kilgray [27]. They provide translation tools, web-based terminology management systems and translation memories.

There is still a huge potential for improving the quality of MT systems. The challenges involve adapting language resources to a given subject domain or user area, and integrating the technology into workflows that already have term bases and translation memories. Another problem is that most of the current systems are English-centred and only support English from and into Hungarian. This leads to friction in the translation workflow and forces MT users to learn different lexicon coding tools for different systems.

Evaluation campaigns help to compare the quality of MT systems, the different approaches and the status of the systems for different language pairs. Figure 8 (p. 25), which was prepared during the EC Euromatrix+ project, shows the pair-wise performances obtained for 22 of the 23 official EU languages (Irish was not compared). The results are ranked according to a BLEU score, which indicates higher scores for better translations [28]. A human translator would normally achieve a score of around 80 points.

The best results (in green and blue) were achieved by languages that benefit from a considerable research effort in coordinated programmes and the existence of many parallel corpora (e. g., English, French, Dutch, Spanish and German). The languages with poorer results are shown in red. These languages either lack such development efforts or are structurally very different from other languages (e. g., Hungarian, Maltese and Finnish).

4.3 OTHER APPLICATION AREAS

Building language technology applications involves a range of subtasks that do not always surface at the level of interaction with the user, but they provide significant service functionalities "behind the scenes" of the system in question. They all form important research issues that have now evolved into individual sub-disciplines of computational linguistics. Question answering, for example, is an active area of research for which annotated corpora have been built and scientific competitions have been initiated. The concept of question answering goes beyond keyword-based searches (in which the search engine responds by delivering a collection of potentially relevant documents) and enables users to ask a concrete question to which the system provides a single answer. For example:

> *Question: How old was Neil Armstrong when he stepped on the moon?*
>
> *Answer: 38.*

While question answering is obviously related to the core area of web search, it is nowadays an umbrella term for such research issues as which different types of questions exist, and how they should be handled; how a set of documents that potentially contain the answer can be analysed and compared (do they provide conflicting answers?); and how specific information (the answer) can be reliably extracted from a document without ignoring the context.

Language technology applications often provide significant service functionalities "behind the scenes" of larger software systems.

Question answering is in turn related to information extraction (IE), an area that was extremely popular and influential when computational linguistics took a statistical turn in the early 1990s. IE aims to identify spe-

cific pieces of information in specific classes of documents, such as the key players in company takeovers as reported in newspaper stories. Another common scenario that has been studied is reports on terrorist incidents. The task here consists of mapping appropriate parts of the text to a template that specifies the perpetrator, target, time, location and results of the incident. Domain-specific template-filling is the central characteristic of IE, which makes it another example of a "behind the scenes" technology that forms a well-demarcated research area, which in practice needs to be embedded into a suitable application environment.

Text summarisation and text generation are two borderline areas that can act either as standalone applications or play a supporting role. Summarisation attempts to give the essentials of a long text in a short form, and is one of the features available in Microsoft Word. It mostly uses a statistical approach to identify the "important" words in a text (i. e., words that occur very frequently in the text in question but less frequently in general language use) and determine which sentences contain the most of these "important" words. These sentences are then extracted and put together to create the summary. In this very common commercial scenario, summarisation is simply a form of sentence extraction, and the text is reduced to a subset of its sentences. An alternative approach, for which some research has been carried out, is to generate brand new sentences that do not exist in the source text.

For the Hungarian language, research in most text technologies is much less developed than for the English language.

This requires a deeper understanding of the text, which means that so far this approach is far less robust. On the whole, a text generator is rarely used as a stand-alone application but is embedded into a larger software environment, such as a clinical information system that collects, stores and processes patient data. Creating reports is just one of many applications for text summarisation.

For the Hungarian language, research in these text technologies is much less developed than for the English language. Question answering, information extraction, and summarisation have been the focus of numerous open competitions in the USA since the 1990s, primarily organised by the government-sponsored organisations DARPA and NIST. These competitions have significantly improved the start-of-the-art, but their focus has mostly been on the English language. As a result, there are hardly any annotated corpora or other special resources needed to perform these tasks in Hungarian. When summarisation systems use purely statistical methods, they are largely language-independent and a number of research prototypes are available. For text generation, reusable components have traditionally been limited to surface realisation modules (generation grammars) and most of the available software is for the English language. There are, however, several named entity recognizer, biomedical information extracting, trend analysing and opinion mining systems for the Hungarian language.

4.4 EDUCATIONAL PROGRAMMES

Language technology is a very interdisciplinary field that involves the combined expertise of linguists, computer scientists, mathematicians, philosophers, psycholinguists, and neuroscientists among others. As a result, it has not acquired a clear, independent existence in the Hungarian faculty system yet, so in Hungary there is no university with an established department of Computational Linguistics. However, there are relevant programmes offered by related departments, such as the faculty of computer science or the faculty of linguistics. Some universities offer Master or Bachelor courses

only, or modules in Language Technology to students of other courses of study. Many of these programs and courses have only been introduced recently. Currently, six Hungarian universities offer at least courses in the field of Language Technology.

In spite of the efforts in recent years to find the way of regular teaching of CL into the Hungarian faculty system, the education of next generation computational linguists does not achieve the required level. The aim of the Hungarian CL community is to develop a high quality curriculum of BSc-MSc-PhD sequence, which fits into the European standards. The relatively low salaries and scholarships of young researchers pose further problems, which could partly be solved by strengthening the relationships between research and industry.

4.5 NATIONAL PROJECTS AND INITIATIVES

The beginnings of natural language processing in Hungary are connected with Machine Translation. The first attempts were made in the 1960s – in those years from Russian to Hungarian. In the 1970s-1980s the lexicographers' work gave the impetus that led to the development of the first computational morphosyntactic systems for Hungarian. In those years there were no regular nationally financed projects, moreover, Hungary was separated from European support.

After the political change, in the 1990s new sections were formed at universities (e. g., the Natural Language Processing Group at Szeged University) and in research institutes (e. g., the Department for Corpus Linguistics at the Research Institute for Linguistics). Since 2000, there has been a significant increase in the number of projects supported by European funds and nationally financed projects, supported mainly by the Fund of the Ministry of Education, or the Agency for Research Fund Management and Research Exploitation.

As a consequence, over the past decade a number of important electronic language resources (dictionaries, corpora, lexical databases) as well as processing resources (spell checking, morphological analyser etc.) have been developed. Activities however have not been synchronized, and not uncommonly similar resources have been developed in parallel at different locations (e. g., there are at least three morphological analysers for Hungarian). A range of different formalisms or standards have been used in these, which in the majority of cases are either incompatible or difficult to convert from; there is also a lack of documentation and in many cases copyright issues are unclear. Nevertheless, in recent years the international trends of standardisation and uniformisation of existing resources have reached Hungary as well. Several projects started off with the objective of integration and interoperability, e. g., creating a unified Hungarian ontology, or harmonising the different coding systems of separately developed morphological analysers.

In 2008, prominent Hungarian academic institutions and R&D companies formed the Hungarian Platform for Speech and Language Technology [30], which aims to help sharing and integration of high quality knowledge accumulated in centres that worked in isolation beforehand; to work out detailed strategic and implementation plans and to help their subsequent implementation; to disseminate its analyses and proposals among the members of the IT sector; to represent the Hungarian interests at the international level; and to disseminate the achievements of the Platform among the potential users of the technology. Hungarian institutions have also been involved in the CLARIN project.

Along with many smaller scale projects that have now been completed, the above projects have led to the development of wide-ranging competence in the field of language technology as well as the creation of a basic technological infrastructure for Hungarian language

tools and resources. Public funding for LT projects in Hungary and in Europe is still relatively low, however, when compared to the amount of money the USA spends on language translation and multilingual information access [31].

As we have seen, previous programmes have led to the development of a number of LT tools and resources for the Hungarian language. The following section summarises the current state of LT support for Hungarian.

4.6 AVAILABILITY OF TOOLS AND RESOURCES

Figure 7 provides a rating for language technology support for the Hungarian language. This rating of existing tools and resources was generated by leading experts in the field who provided estimates based on a scale from 0 (very low) to 6 (very high) using seven criteria.

The key results for Hungarian language technology can be summed up as follows:

- While several specific corpora of high quality exist, a very large syntactically annotated corpus is not available. However, there is a syntactically highly elaborately annotated corpus for Hungarian (1.2 million tokens). The corpus is available for free, which results in a wide range applications developed based on the corpus.
- Many of the resources lack standardisation, i. e., even if they exist, they are not addressing sustainability; concerted programs and initiatives are needed to standardise data and interchange formats.
- The standard preprocessing steps (tokenisation, morphology, shallow parsing, etc.) are completed for Hungarian, but treating the more difficult semantics requires further research.
- The more linguistic and semantic knowledge a tool draws on, the more gaps there are in the technology (text analysis versus text interpretation). There is a need for far more effort to support deep linguistic processing.
- Research has been successful in designing particularly high quality software, but many of the resources are not standardised and cannot be sustained effectively. A concerted program is required to standardise data and interchange formats.
- There is a lack of annotated corpora with syntactic, semantic and discourse structure mark-up. Again, the situation gets worse as the need for more deep linguistic and semantic information grows.
- Standards do exist for semantics in the sense of world knowledge (RDF, OWL, etc.); they are, however, not easily applicable to NLP tasks.
- Speech Recognition and Machine Translation of Hungarian is studied at several universities and workplaces, but free tools and data are currently not available. It is a typical phenomenon at the Hungarian NLP market that the number of free databases and open source programs is quite low.

To conclude, in a number of specific areas of Hungarian language research, we have softwares with limited functionality available today. Obviously, further research efforts are required to meet the current deficit in processing texts on a deeper semantic level and to address the lack of resources such as speech corpora for speech recognition.

4.7 CROSS-LANGUAGE COMPARISON

The current state of LT support varies considerably from one language community to another. In order to compare the situation between languages, this section will present an evaluation based on two sample application areas (machine translation and speech processing)

	Quantity	Availability	Quality	Coverage	Maturity	Sustainability	Adaptability
Language Technology: Tools, Technologies and Applications							
Speech Recognition	3	0	4	2	4	3	3
Speech Synthesis	4	3	4	4	5	3	3
Grammatical analysis	4.5	2	4	4.5	4	3	4.5
Semantic analysis	0.6	2	2.5	0.5	0	0	2
Text generation	0	0	0	0	0	0	0
Machine translation	6	1	4	3	6	5	6
Language Resources: Resources, Data and Knowledge Bases							
Text corpora	3.5	6	5.5	5.5	6	6	4
Speech corpora	2	2	4	2	4	4	0
Parallel corpora	6	4	4.5	2.5	6	6	6
Lexical resources	3	1	3.5	3.5	3.5	3.5	4.5
Grammars	3	3	5	5	6	4	3

8: State of language technology support for Hungarian

and one underlying technology (text analysis), as well as basic resources needed for building LT applications. The languages were categorised using the following five-point scale:

1. Excellent support
2. Good support
3. Moderate support
4. Fragmentary support
5. Weak or no support

LT support was measured according to the following criteria:

Speech Processing: Quality of existing speech recognition technologies, quality of existing speech synthesis technologies, coverage of domains, number and size of existing speech corpora, amount and variety of available speech-based applications.

Machine Translation: Quality of existing MT technologies, number of language pairs covered, coverage of linguistic phenomena and domains, quality and size of existing parallel corpora, amount and variety of available MT applications.

Text Analysis: Quality and coverage of existing text analysis technologies (morphology, syntax, semantics), coverage of linguistic phenomena and domains, amount and variety of available applications, quality and size of existing text corpora, quality and coverage of existing lexical resources (e. g., WordNet) and grammars.

Resources: Quality and size of existing text corpora, speech corpora and parallel corpora, quality and coverage of existing lexical resources and grammars.

Figures 8 to 11 show that, thanks to large-scale LT funding in recent decades, the Hungarian language is quite well equipped compared to other. But LT resources and

tools for Hungarian clearly do not yet reach the quality and coverage of comparable resources and tools for the English language, which is in the lead in almost all LT areas. And there are still plenty of gaps in English language resources with regard to high quality applications. For speech processing, current technologies perform well enough to be successfully integrated into a number of industrial applications such as spoken dialogue and dictation systems. Today's text analysis components and language resources already cover the linguistic phenomena of Hungarian to a certain extent and form part of many applications involving mostly shallow natural language processing, e. g., spelling correction and some information extraction tasks.

However, for building more sophisticated applications, such as machine translation, there is a clear need for resources and technologies that cover a wider range of linguistic aspects and allow a deep semantic analysis of the input text. By improving the quality and coverage of these basic resources and technologies, we shall be able to open up new opportunities for tackling a vast range of advanced application areas, including high-quality machine translation.

4.8 CONCLUSIONS

In this series of white papers, we have made an important effort by assessing the language technology support for 30 European languages, and by providing a high-level comparison across these languages. By identifying the gaps, needs and deficits, the European language technology community and its related stakeholders are now in a position to design a large scale research and development programme aimed at building a truly multilingual, technology-enabled communication across Europe.

The results of this white paper series show that there is a dramatic difference in language technology support between the various European languages. While there are good quality software and resources available for some languages and application areas, others, usually smaller languages, have substantial gaps. Many languages lack basic technologies for text analysis and the essential resources. Others have basic tools and resources but the implementation of for example semantic methods is still far away. Therefore a large-scale effort is needed to attain the ambitious goal of providing high-quality language technology support for all European languages, for example through high quality machine translation.

In the case of the Hungarian language, we can be cautiously optimistic about the current state of language technology support. There is a viable LT research community in Hungary, which has been supported in the past mostly by national funds. And a number of large-scale resources and state-of-the-art technologies have been produced and distributed for Hungarian. However, the scope of the resources and the range of tools are still very limited when compared to the resources and tools for the English language, and they are simply not sufficient in quality and quantity to develop the kind of technologies required to support a truly multilingual knowledge society.

Nor can we simply transfer technologies already developed and optimised for the English language to handle Hungarian. English-based systems for parsing (syntactic and grammatical analysis of sentence structure) typically perform far less well on Hungarian texts, due to the specific characteristics of the Hungarian language.

There is a relatively small language technology industry at work on the Hungarian language. Thus the Hungarian NLP market is dominated by research groups at universities and academic institutes, however there are additional smaller companies on the market.

Our findings lead to the conclusion that the only way forward is to make a substantial effort to create language technology resources for Hungarian, as a means to drive forward research, innovation and development. The need for large amounts of data and the extreme com-

plexity of language technology systems makes it vital to develop an infrastructure and a coherent research organisation to spur greater sharing and cooperation.

Finally there is a lack of continuity in research and development funding. Short-term coordinated programmes tend to alternate with periods of sparse or zero funding. In addition, there is an overall lack of coordination with programmes in other EU countries and at the European Commission level.

The long term goal of META-NET is to enable the creation of high-quality language technology for all languages. This requires all stakeholders – in politics, research, business, and society – to unite their efforts. The resulting technology will help tear down existing barriers and build bridges between Europe's languages, paving the way for political and economic unity through cultural diversity.

Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
	English	Czech Dutch Finnish French German Italian Portuguese Spanish	Basque Bulgarian Catalan Danish Estonian Galician Greek **Hungarian** Irish Norwegian Polish Serbian Slovak Slovene Swedish	Croatian Icelandic Latvian Lithuanian Maltese Romanian

9: Speech processing: state of language technology support for 30 European languages

Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
	English	French Spanish	Catalan Dutch German **Hungarian** Italian Polish Romanian	Basque Bulgarian Croatian Czech Danish Estonian Finnish Galician Greek Icelandic Irish Latvian Lithuanian Maltese Norwegian Portuguese Serbian Slovak Slovene Swedish

10: Machine translation: state of language technology support for 30 European languages

Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
	English	Dutch French German Italian Spanish	Basque Bulgarian Catalan Czech Danish Finnish Galician Greek **Hungarian** Norwegian Polish Portuguese Romanian Slovak Slovene Swedish	Croatian Estonian Icelandic Irish Latvian Lithuanian Maltese Serbian

11: Text analysis: state of language technology support for 30 European languages

Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
	English	Czech Dutch French German **Hungarian** Italian Polish Spanish Swedish	Basque Bulgarian Catalan Croatian Danish Estonian Finnish Galician Greek Norwegian Portuguese Romanian Serbian Slovak Slovene	Icelandic Irish Latvian Lithuanian Maltese

12: Speech and text resources: State of support for 30 European languages

ABOUT META-NET

META-NET is a Network of Excellence funded by the European Commission. The network currently consists of 54 members from 33 European countries [32]. META-NET fosters the Multilingual Europe Technology Alliance (META), a growing community of language technology professionals and organisations in Europe. META-NET fosters the technological foundations for a truly multilingual European information society that:

- makes communication and cooperation possible across languages;
- provides equal access to information and knowledge in any language;
- offers advanced and affordable networked information technology to European citizens.

The network supports a Europe that unites as a single digital market and information space. It stimulates and promotes multilingual technologies for all European languages. These technologies support automatic translation, content production, information processing and knowledge management for a wide variety of applications and subject domains. They also enable intuitive language-based interfaces to technology ranging from household electronics, machinery and vehicles to computers and robots. Launched on 1 February 2010, META-NET has already conducted various activities in its three lines of action META-VISION, META-SHARE and META-RESEARCH.

META-VISION fosters a dynamic and influential stakeholder community that unites around a shared vision and a common strategic research agenda (SRA). The main focus of this activity is to build a coherent and cohesive LT community in Europe by bringing together representatives from highly fragmented and diverse groups of stakeholders. The present White Paper was prepared together with volumes for 29 other languages. The shared technology vision was developed in three sectorial Vision Groups. The META Technology Council was established in order to discuss and to prepare the SRA based on the vision in close interaction with the entire LT community.

META-SHARE creates an open, distributed facility for exchanging and sharing resources. The peer-to-peer network of repositories will contain language data, tools and web services that are documented with high-quality metadata and organised in standardised categories. The resources can be readily accessed and uniformly searched. The available resources include free, open source materials as well as restricted, commercially available, fee-based items.

META-RESEARCH builds bridges to related technology fields. This activity seeks to leverage advances in other fields and to capitalise on innovative research that can benefit language technology. In particular, the action line focuses on conducting leading-edge research in machine translation, collecting data, preparing data sets and organising language resources for evaluation purposes; compiling inventories of tools and methods; and organising workshops and training events for members of the community.

office@meta-net.eu – http://www.meta-net.eu

HIVATKOZÁSOK REFERENCES

[1] Aljoscha Burchard, Markus Egg, Kathrin Eichler, Brigitte Krenn, Jörn Kreutel, Annette Leßmöllmann, Georg Rehm, Manfred Stede, Hans Uszkoreit, and Martin Volk. *Die Deutsche Sprache im Digitalen Zeitalter – The German Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer, 2012.

[2] Directorate-General Information Society & Media of the European Commission. User Language Preferences Online, 2011. http://ec.europa.eu/public_opinion/flash/fl_313_en.pdf.

[3] European Commission. Multilingualism: an Asset for Europe and a Shared Commitment, 2008. http://ec.europa.eu/languages/pdf/comm2008_en.pdf.

[4] Directorate-General of the UNESCO. Intersectoral Mid-term Strategy on Languages and Multilingualism, 2007. http://unesdoc.unesco.org/images/0015/001503/150335e.pdf.

[5] Directorate-General for Translation of the European Commission. Size of the Language Industry in the EU, 2009. http://ec.europa.eu/dgs/translation/publications/studies.

[6] Ádám Nádasdy. Did you know? Educational publication about the Hungarian language.

[7] http://www.bbi.hu/index.php?id=99&fid=110.

[8] http://www.nytud.hu/eng/index.html.

[9] PISA 2009 Results: What Students Know and Can Do: Student Performance in Reading, Mathematics and Science (Volume I). http://www.oecd.org/document/61/0,3343,en_2649_35845621_46567613_1_1_1_1,00.html.

[10] http://www.tarki.hu/tarkitekinto/20050412.html.

[11] http://www.google.com/publicdata?ds=wb-wdi&met_y=it_net_user_p2&idim=country:HUN&dl=hu&hl=hu&q=internethaszn%C3%A1lat.

[12] http://www.nic.hu/English/statisztika/domain-teljes.html.

[13] http://www.webhosting.info/registries/country_stats/HU.

[14] Daniel Jurafsky and James H. Martin. *Speech and Language Processing (2nd Edition)*. Prentice Hall, 2009.

[15] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[16] Language Technology World (LT World). http://www.lt-world.org/.

[17] Ronald Cole, Joseph Mariani, Hans Uszkoreit, Giovanni Battista Varile, Annie Zaenen, and Antonio Zampolli, editors. *Survey of the State of the Art in Human Language Technology (Studies in Natural Language Processing)*. Cambridge University Press, 1998.

[18] Jerrold H. Zar. Candidate for a Pullet Surprise. *Journal of Irreproducible Results*, page 13, 1994.

[19] http://www.morphologic.hu/.

[20] http://hunspell.sourceforge.net/.

[21] Spiegel Online. Google zieht weiter davon (Google is still leaving everybody behind), 2009. http://www.spiegel.de/netzwelt/web/0,1518,619398,00.html.

[22] Juan Carlos Perez. Google Rolls out Semantic Search Capabilities, 2009. http://www.pcworld.com/businesscenter/article/161869/google_rolls_out_semantic_search_capabilities.html.

[23] http://www.weblib.com/.

[24] http://www.tmit.bme.hu/home.

[25] http://szotar.mokk.bme.hu/hunglish/search/corpus.

[26] http://itranslate4.eu/.

[27] http://kilgray.com/.

[28] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation). In *Proceedings of the 40th Annual Meeting of ACL*, Philadelphia, PA, 2002.

[29] Philipp Koehn, Alexandra Birch, and Ralf Steinberger. 462 Machine Translation Systems for Europe. In *Proceedings of MT Summit XII*, 2009.

[30] http://hlt-platform.hu/.

[31] Gianni Lazzari. Sprachtechnologien für Europa (Language Technology for Europe), 2006. http://tcstar.org/pubblicazioni/D17_HLT_DE.pdf.

[32] Georg Rehm and Hans Uszkoreit. Multilingual Europe: A challenge for language tech. *MultiLingual*, 22(3):51–52, April/May 2011.

B

META-NET TAGOK META-NET MEMBERS

Ausztria	Austria	Zentrum für Translationswissenschaft, Universität Wien: Gerhard Budin
Belgium	Belgium	Computational Linguistics and Psycholinguistics Research Centre, Univ. of Antwerp: Walter Daelemans
		Centre for Proc. Speech and Images, Univ. of Leuven: Dirk van Compernolle
Bulgária	Bulgaria	Inst. for Bulgarian Lang., Bulgarian Academy of Sciences: Svetla Koeva
Ciprus	Cyprus	Lang. Centre, School of Humanities: Jack Burston
Csehország	Czech Republic	Inst. of Formal and Applied Linguistics, Charles Univ. in Prague: Jan Hajic
Dánia	Denmark	Centre for Lang. Technology, Univ. of Copenhagen: Bolette Sandford Pedersen, Bente Maegaard
Egyesült Királyság	UK	Inst. for Lang., Cognition and Computation, Center for Speech Technology Research, Univ. of Edinburgh: Steve Renals
		Research Inst. of Informatics and Lang. Proc., Univ. of Wolverhampton: Ruslan Mitkov
		School of Computer Science, Univ. of Manchester: Sophia Ananiandou
Észtország	Estonia	Inst. of Computer Science, Univ. of Tartu: Tiit Roosmaa
Finnország	Finland	Computational Cognitive Systems Research Group, Aalto Univ.: Timo Honkela
		Dept. of General Linguistics, Univ. of Helsinki: Kimmo Koskenniemi, Krister Linden
Franciaország	France	Centre National de la Recherche Scientifique, Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur: Joseph Mariani
		Evaluations and Lang. Resources Distribution Agency: Khalid Choukri
Görögország	Greece	Inst. for Lang. and Speech Proc., R. C. "Athena": Stelios Piperidis
Hollandia	Netherlands	Utrecht Inst. of Linguistics, Utrecht Univ.: Jan Odijk
		Computational Linguistics, Univ. of Groningen: Gertjan van Noord
Horvátország	Croatia	Inst. of Linguistics, Faculty of Humanities and Social Science, Univ. of Zagreb: Marko Tadić
Írország	Ireland	School of Computing, Dublin City Univ.: Josef van Genabith
Izland	Iceland	School of Humanities, Univ. of Iceland: Eirikur Rögnvaldsson
Lengyelország	Poland	Inst. of Computer Science, Polish Academy of Sciences: Adam Przepiórkowski, Maciej Ogrodniczuk

		Univ. of Łódź: Barbara Lewandowska-Tomaszczyk, Piotr Pęzik
		Dept. of Computer Linguistics and Artificial Intelligence, Adam Mickiewicz Univ.: Zygmunt Vetulani
Lettország	Latvia	Tilde: Andrejs Vasiljevs
		Inst. of Mathematics and Computer Science, Univ. of Latvia: Inguna Skadina
Litvánia	Lithuania	Inst. of the Lithuanian Lang.: Jolanta Zabarskaitė
Luxemburg	Luxembourg	Arax Ltd.: Vartkes Goetcherian
Magyarország	Hungary	Research Inst. for Linguistics, Hungarian Academy of Sciences: Tamás Váradi
		Dept. of Telecommunications and Media Informatics, Budapest Univ. of Technology and Economics: Géza Németh, Gábor Olaszy
Málta	Malta	Dept. Intelligent Computer Systems, Univ. of Malta: Mike Rosner
Németország	Germany	DFKI (German Research Centre for Artificial Intelligence): Hans Uszkoreit, Georg Rehm
		Human Lang. Technology and Pattern Recognition, RWTH Aachen Univ.: Hermann Ney
		Dept. of Computational Linguistics, Saarland Univ.: Manfred Pinkal
Norvégia	Norway	Dept. of Linguistic, Literary and Aesthetic Studies, Univ. of Bergen: Koenraad De Smedt
		Dept. of Informatics, Lang. Technology Group, Univ. of Oslo: Stephan Oepen
Olaszország	Italy	Consiglio Nazionale Ricerche, Istituto di Linguistica Computazionale "Antonio Zampolli": Nicoletta Calzolari
		Human Lang. Technology, Fondazione Bruno Kessler: Bernardo Magnini
Portugália	Portugal	Dept. of Informatics, Univ. of Lisbon: Antonio Branco
		Spoken Lang. Systems Lab., Inst. for Systems Engineering and Computers: Isabel Trancoso
Románia	Romania	Research Inst. for Artificial Intelligence, Romanian Academy of Sciences: Dan Tufis
		Faculty of Computer Science, Univ. Alexandru Ioan Cuza: Dan Cristea
Spanyolország	Spain	Barcelona Media: Toni Badia
		Institut Universitari de Lingüistica Aplicada, Univ. Pompeu Fabra: Núria Bel
		Aholab Signal Proc. Lab., Univ. of the Basque Country: Inma Hernaez Rioja
		Center for Lang. and Speech Technologies and Applications, Technical Univ. of Catalonia: Asunción Moreno
		Dept. of Signal Proc. and Communications, Univ. of Vigo: Carmen García Mateo

Svájc	Switzerland	Idiap Research Inst.: Hervé Bourlard
Svédország	Sweden	Dept. of Swedish Lang., Univ. of Gothenburg: Lars Borin
Szerbia	Serbia	Faculty of Mathematics, Belgrade Univ.: Dusko Vitas, Cvetana Krstev, Ivan Obradovic
		Pupin Inst.: Sanja Vranes
Szlovákia	Slovakia	Ludovit Stur Inst. of Linguistics, Slovak Academy of Sciences: Radovan Garabik
Szlovénia	Slovenia	Jozef Stefan Inst.: Marko Grobelnik

Több mint 100 nyelvtechnológus szakértő – a META-NET-ben részt vevő országok és nyelvek képviselői – vitatta meg és véglegesítette a fehér könyvek sorozat főbb kérdéseit egy META-NET találkozón Berlinben, 2011. október 21-22-én. — About 100 language technology experts – representatives of the countries and languages represented in META-NET – discussed and finalised the key results and messages of the White Paper Series at a META-NET meeting in Berlin, Germany, on October 21/22, 2011.

A META-NET FEHÉR KÖNYVEK SOROZAT

THE META-NET WHITE PAPER SERIES

angol	English	English
baszk	Basque	euskara
bolgár	Bulgarian	български
cseh	Czech	čeština
dán	Danish	dansk
észt	Estonian	eesti
finn	Finnish	suomi
francia	French	français
galíciai	Galician	galego
görög	Greek	ελληνικά
holland	Dutch	Nederlands
horvát	Croatian	hrvatski
ír	Irish	Gaeilge
izlandi	Icelandic	íslenska
katalán	Catalan	català
lengyel	Polish	polski
lett	Latvian	latviešu valoda
litván	Lithuanian	lietuvių kalba
magyar	Hungarian	magyar
máltai	Maltese	Malti
német	German	Deutsch
norvég bokmål	Norwegian Bokmål	bokmål
norvég nynorsk	Norwegian Nynorsk	nynorsk
olasz	Italian	italiano
portugál	Portuguese	português
román	Romanian	română
spanyol	Spanish	español
svéd	Swedish	svenska
szerb	Serbian	српски
szlovák	Slovak	slovenčina
szlovén	Slovene	slovenščina

In everyday communication, Europe's citizens, business partners and politicians are inevitably confronted with language barriers. Language technology has the potential to overcome these barriers and to provide innovative interfaces to technologies and knowledge. This white paper presents the state of language technology support for the hungarian language. It is part of a series that analyses the available language resources and technologies for 31 European languages. The analysis was carried out by META-NET, a Network of Excellence funded by the European Commission. META-NET consists of 54 research centres in 33 countries, who cooperate with stakeholders from economy, government agencies, research organisations, non-governmental organisations, language communities and European universities. META-NET's vision is high-quality language technology for all European languages.

A mindennapi kommunikáció Európa polgárai, mind az üzleti, mind a politikai szférában elkerülhetetlenül nyelvi akadályokba ütközik. A nyelvtechnológia hozzá tud járulni ezen akadályok legyőzéséhez, továbbá kapcsolódási pontokat nyújt az innovatív technológiák és tudás felé. Ez a fehér könyv a magyar nyelvtechnológia helyzetét mutatja be, egyben egy sorozat részét képezi, amely az elérhető nyelvi erőforrásokról és technológiákról ad elemzést 31 európai nyelvre. A felmérést a META-NET, az Európai Bizottság által alapított hálózat végezte. A META-NET 33 ország 54 kutatóközpontjából áll, akik gazdasági döntéshozókkal, kormányzati szervekkel, kutatószervezetekkel, nyelvi közösségekkel és európai egyetemekkel dolgoznak együtt. A META-NET jövőképe: kiváló minőségű nyelvtechnológia minden európai nyelvre.

"META-NET is making a significant contribution to innovation, research and development in Europe and to an effective implementation of the European idea."
– Valéria Csépe (Deputy General Secretary of Hungarian Academy of Sciences)

"A META-NET jelentős mértékben hozzájárul az innovációhoz és a kutatás-fejlesztéshez, valamint az európai eszme hatékony megvalósításához."
– Csépe Valéria (főtitkárhelyettes, MTA)