NYELV- ÉS
BESZÉDTECHNOLÓGIAI
PLATFORM

GÉPEKKEL - EMBERI NYELVEN

www.hlt-platform.hu

Megvalósítási Terv

Budapest, 2010. szeptember 22.

Vezetői összefoglaló

A Nyelv- és Beszédtechnológiai Platform az NKTH által 2007-ben kiírt, magyar technológiai platformok alakítására felhívó első pályázati körben nyert támogatást. A Platform élenjáró magyarországi kutató-fejlesztő műhelyek szövetsége, mely azzal a céllal jött létre, hogy összehangolt munkával erősítse és elősegítse az innovációt a nyelv- és beszédtechnológia területén, így hozzájáruljon a magyar technológia fejlődéséhez, a nemzetgazdaság versenyképességének növeléséhez.

Vállalásaink között szerepelt a nyelv- és beszédtechnológia legfontosabb fejlesztési és kutatási irányainak a feltérképezése, valamint részletes stratégiai és arra épülő megvalósítási tervek kidolgozása. A Platform eddigi tevékenysége során elkészítette a magyarországi nyelv- és beszédtechnológia jelenlegi helyzetének áttekintését, felvázolta a Platform jövőképét, majd ezekre támaszkodva létrehozta a Stratégiai Kutatási Tervet.

Jelen Megvalósítási Terv a Stratégiai Kutatási Tervben bővebben kifejtett stratégiai célokat emeli ki prioritásokként, azokon belül pedig a főbb kutatási irányokat ismerteti. Minden egyes kutatási irány esetében bemutatunk néhány olyan projekttervet, amelyek jól példázzák az adott terület főbb problémaköreit. A Megvalósítási Terv tartalmazza ezen projektek rövid leírását, várható eredményeit, ütemezését, erőforrásigényét. Tartalmazza az erőforrások és infrastruktúra biztosítására, a megfelelő szabályozási környezet kialakítására, a résztvevő és bevonandó szervezetekre, a kutatási eredmények hasznosításának módjára vonatkozó terveket, elképzeléseket, ajánlásokat.

Kiemelt prioritásként azonosítottuk a nyelv- és beszédtechnológiai kutatási infrastruktúra kiépítését, a nyelvi információ feldolgozására vonatkozó fejlesztéseket, a természetes ember-gép kommunikációt, hangsúlyoztuk a magyar nyelv- és beszédtechnológia értékőrző és értékmentő, az esélyegyenlőség és életminőség javításában és a nyelvi korlátok leküzdésében betöltött szerepét. Megkülönböztetett figyelemmel kezeltük a kutatásszervezés és -finanszírozás, valamint az együttműködés témáit, mivel ezek olyan területek, amelyek a nyelv- és beszédtechnológia minden egyes részfeladatára, céljára „ráterülnek”, egyfajta horizontális prioritást képezve azok felett.

A hét fő stratégiai célon belül a Megvalósítási Terv 18 kutatási irányt és 34 konkrét projekttervet fogalmaz meg a 2011-től 2020-ig tartó időtartamra – összesen 3105 millió forint értékben.

Tartalomjegyzék

1 Bevezetés

Ha olyan cégeket, termékeket és szolgáltatásokat keresünk, amelyek emblematikusan jelenítik meg a 21. századi általános társadalmi és életmódváltozásokat, akkor valószínűleg mindenkinek elsősorban a web fellegvárai, a Google, a Facebook, a Twitter, a Wikipédia és társaik jutnak eszébe, és olyan termékek, mint az iPhone vagy a Wii. A jövőt illetően a web megalkotója, Tim Berners-Lee tette népszerűvé az értő (szemantikus) webet mint olyan célt, ami felé az összes multimodális irány (szöveg, hang, álló- és mozgókép, zene) konvergálni fog. A megértés itt elsősorban olyan virtuális ágensek létrehozását jelenti, amelyek (sőt: akik) ugyanúgy terjesztik ki az emberek képességeit a virtuális világban, ahogy a fizikai világban az autó kiterjesztette messzejárási, a daru súlyemelési, az írni-olvasni tudás pedig emlékezési képességünket. A szemantikus web tehát nem passzív weblapokból, hanem aktív, a kívánságainkat értő ágensekből fog állni.

A nyelv- és beszédtechnológia e jövőkép központi eleme. A beszélő és az embert megértő számítógép lassan, fanfár nélkül átvonul a tudományos fantasztikumból a hétköznapi életbe: ma már nehéz úgy információt szerezni, szolgáltatást vagy terméket venni, hogy első körben ne számítógéppel kerülnénk kapcsolatba, akár a weben, akár telefonon. Azokba a széles sodrású világfolyamatokba, amelyek ezt a fejlődést táplálják, nehéz, talán lehetetlen is beavatkoznunk. Annak az elérése viszont, hogy ne csak angolul vagy kínaiul, hanem magyarul is elérhető legyen mindez, egyáltalán nem lehetetlen, bár kétségtelenül aktív kutatás- és fejlesztéstámogatási politikát igényel, melynek kialakítása akkor is állami feladat, ha a megvalósítás egyes elemeit az ipari szféra résztvevőire bízzuk.

A nyelv- és beszédtechnológia egyik legfontosabb feladata, hogy a nyelvi információt hordozó digitális adatfolyamokat (nyomtatott oldalak beszkennelt képét, írott szöveget, beszédet tartalmazó hangfelvételt, videót) automatikus módszerekkel feldolgozva olyan – az eredeti anyagban explicit formában nem szereplő – további információval lássa el, amely lehetővé teszi a szövegben kódolt információ, illetve tudás minél többféle szempont szerinti megtalálását (az intelligens keresést), strukturált adatbázisokba szervezését és a felhasználó számára optimális prezentációját. Az optimális prezentálás magában foglalja többek között a legrelevánsabb információ kiemelését és annak a felhasználónak leginkább megfelelő modalitásban (írás, beszéd, esetleg animált szájmozgással kísérve, jelelés stb.) és nyelven (pl. automatikus fordítással) való visszaadását. A nyelvi tartalmak hatékony szűrése, a lényeges információ megtalálása és kiemelése az adattengerből, a nehezen érthető információnak a felhasználó számára jobban értelmezhető formában való visszaadása alapvető fontosságú feladatok, melyek megoldásával a nyelv- és beszédtechnológia nélkülözhetetlen hát-

térinfrastruktúrát ad a többi tudományterületnek. Az intézmények, a vállalkozások számára a (szöveges-, hangzó- és videóanyagokból automatikusan kinyert) tudásbázisok komoly versenyelőnyt jelentenek információs társadalmunkban, és az állampolgárok számára is az élet megjobbításának alapvető eszközeivé válnak. Kiemelkedő szerepet játszhat különböző hátrányos helyzetű csoportok (siketek, gyengénlátók, baleset következtében beszédkészségüket elvesztők, idegen nyelveket nem tudók) életminőségének javításában.

A Nyelv- és Beszédtechnológiai Platformot élenjáró magyarországi kutató-fejlesztő közösségek hozták létre azzal a céllal, hogy összehangolt munkával erősítsék és elősegítsék az innovációt a nyelv- és beszédtechnológia területén, így hozzájáruljanak a magyar technológia fejlődéséhez, a nemzetgazdaság versenyképességének növeléséhez. Vállalásaink között szerepelt a nyelv- és beszédtechnológia legfontosabb fejlesztési és kutatási irányainak a feltérképezése, valamint részletes stratégiai és arra épülő megvalósítási tervek kidolgozása.

A Platform eddigi tevékenysége során elkészítette a magyarországi nyelv- és beszédtechnológia jelenlegi helyzetének áttekintését, felvázolta a Platform jövőképét, majd ezekre támaszkodva létrehozta a Stratégiai Kutatási Tervet.

1.1. A Stratégiai Kutatási Terv összefoglalása

A Nyelv- és Beszédtechnológiai Platform Stratégiai Kutatási Tervének célja az volt, hogy megfogalmazza a hazai nyelv- és beszédtechnológia fejlődésének irányait, e technológiák nyelvfüggő elemeinek „kötelező” hazai feladatait, rámutasson a nemzetközi kitörési lehetőségekre, és meghatározza az ezek realizálásához szükséges lépéseket. A Terv iránymutatásként szolgál a szakpolitika, a gazdasági, kormányzati döntéshozók és az ágazati szereplők számára.

Ma alapvető fontosságúak azok a módszerek, melyek segítségével tájékozódni tudunk a nagy tömegű természetes nyelven megfogalmazott információ között, könnyebben és gyorsabban tudjuk *elérni a számunkra fontos információt*, és csak azt. A nyelv- és beszédtechnológia éppen ebben tud segíteni. E tudományterület célja, hogy olyan új technológiákat és alkalmazásokat állítson elő, melyek az emberi kommunikációt természetesen és hatékonyan szolgálják ki, a természetes nyelven történő információáramlást gépi eszközökkel hatékonyan támogatják. A jövő tudásalapú gazdaságának és társadalmának *nélkülözhetetlen* alkotóelemei ezek a technológiák. A szektor mai gazdasági, társadalmi környezete a hajtóerők, a motivációk tekintetében tehát nagyon ígéretes.

A Stratégiai Kutatási Terv . . .

- meghatározza a kiemelt *stratégiai célokat* és *kutatás-fejlesztési területeket*, ahová a közvetlen hasznosulás és a versenyképesség növelése érdekében a ráfordításokat irányítani érdemes;
- vázolja azokat az *oktatási* és *kutatásfinanszírozási* kereteket, melyek biztosítják a tartós eredményességet;

- összefoglalja az eredmények hatékonyabb közzététele, alkalmazása, gyakorlati *hasznosítása* érdekében végzendő teendőket.

A stratégiai célok a következők:

1. **Nemzeti kutatási infrastruktúra kialakítása és szolgáltatása a nyelv- és beszédtechnológia területén**. A megfelelő erőforrások, írott és beszélt *nyelvi adatbázisok*, szótárak, alapvető sztenderdizált *feldolgozó eszközök* a nyelv- és beszédtechnológia elengedhetetlen szükségletei a fejlesztésben és az elért eredmények kiértékelésében egyaránt. Fontos, hogy a létrejövő technológiák, erőforrások megfeleljenek a meglévő nemzetközi *szabványoknak*.
2. **Kutatásszervezés**. Szükséges az *oktatás* színvonalának emelése, a képzési erőforrások koncentrációja és egységesítése. A fiatal kutatók számára ösztöndíjakat kell létesíteni, lehetővé kell tenni képzésük egy részének kihelyezését ipari szereplőkhöz. Szükséges ösztönözni az ipari szereplőket saját kutatás-fejlesztési ráfordításaik növelésében. Fontos a *kutatásfinanszírozási* lehetőségek bővítése, a keretek hosszú távú meghatározása, a tehetséges fiatal kutatókat megtartani képes új kutatóhelyek létrehozása. Célzott pályázati kiírásokra van szükség, erős szakmai kontroll alkalmazásával előnyben kell részesíteni a valódi innovációt tartalmazó pályaműveket. Létre kell hozni a terület *technológiatranszfer-központját*, amely a kialakítandó nemzeti kutatási infrastruktúrát egységes keretben szolgáltatja, és hozzáférhetővé teszi mind a kutatási, mind az ipari szereplők, illetve akár a nagyközönség számára is. A széles körű hasznosíthatóság érdekében kívánatos, hogy az általános célú nyelv- és beszédtechnológiai szoftvereszközök *nyílt forráskódúak* legyenek, és ennek megfelelően államilag támogatott projektek keretében valósuljanak meg.
3. **A nyelvi információ kezelése, tárolása és feldolgozása** stratégiai cél két alfejezetre bomlik.
   - 3.1. **Nyelvalapú tudásmenedzsment**. Megbízható nyelvfeldolgozó eszközök szükségesek ahhoz, hogy hozzáférjünk a *nagy mennyiségű* hangzó vagy szöveges nyelvi adatban rejlő információhoz. A szemantikai keresés megvalósítása pedig különféle szemantikai elemző modulok kifejlesztését igényli.
   - 3.2. **A nyelvi kulturális örökség digitális korba való átmentése**. Az automatikus szövegfeldolgozás technológiái segítséget nyújtanak abban, hogy az ország írott kulturális örökségét a digitális korba átmentsük. A magyarországi és határon túli magyar nyelvváltozatokat feltérképező kutatásokban jelentős szerepet játszik a nyelv- és beszédtechnológia a beszélt és írott nyelvváltozatok digitális rögzítése és automatikus feldolgozása terén. A beszédfelismerési technológiák a nagy nemzeti hang/film/multimédia archívumok szövegtartalom szerinti kereshetőségét biztosítják.
4. **A természetes nyelven történő kommunikáció számítógépes támogatása** stratégiai cél három alfejezetet tartalmaz.
   - 4.1. **Természetes ember-gép kommunikáció**. Célunk, hogy *élőszó* és/vagy gesztusok segítségével is lehetővé váljon az internet böngészése és ál-

talában az emberi inputot igénylő számítógépes programok irányítása. A gépek részben vagy egészben át is vehetnek bizonyos funkciókat, például az információfeldolgozás és -megjelenítés területén.

4.2. **Fogyatékkal élők és hátrányos helyzetűek információs társadalmi integrációjának elősegítése**. A beszédszintézisre és -felismerésre alapuló technológiák, amelyek *más médiumokra* „fordítanak" és tesznek elérhetővé információt, mind a siketek és nagyothallók, mind a vakok és gyengénlátók számára ezt az integrációs lépést könnyítik meg.

4.3. **Többnyelvűség az Európai Unióban, a nyelvi korlátok leküzdése**. A kommunikáció célja, hogy a befogadó az információt *meg tudja érteni*. A nyelv- és beszédtechnológiai kutatások egyik stratégiai célja pedig az (akár teljesen ismeretlen) *idegen* nyelven megfogalmazott információ megértésének számítógépes támogatása az automatikus gépi fordítás, fordítástámogató eszközök, automatikus tolmácsolás révén.

Nemzetközi kitörési pontokat jelenthetnek például az alábbi *kutatási területek*:

- a robusztus beszédfelismerési technikák fejlesztése, a nagyszótáras, folyamatos többnyelvű gépi beszédfelismerés hatásfokának javítása;
- az idegen nyelvű szövegek megértését támogató gépi fordítás fejlesztése;
- a szövegek tartalmi elemzését végző szemantikus technológiákra irányuló fejlesztés;
- az emberi beszédértés, a kogníció nemzetközi szinten előrehaladott kutatásaiba történő bekapcsolódás;
- az interdiszciplináris és integratív kutatások előtérbe helyezése.

Az Európai Unió kiemelt figyelmet fordít a nyelv- és beszédtechnológiai fejlesztésekre. A kérdés prioritását jelzi, hogy e törekvések az európai információs társadalom előmozdítására irányuló i2010 kezdeményezés részévé váltak. Az i2010 által megjelölt három kiemelt fontosságú területen a nyelv- és beszédtechnológiának kulcsszerep jut:

- *Kutatási ráfordítás és innováció.* A nyelv- és beszédtechnológiai kutatás hatékonyan használja fel a forrásokat, előreviszi az innovációt; az alkalmazások versenyelőnyhöz juttatják a többnyelvű környezetben működő cégeket, serkentik az európai gazdaságot.
- *Társadalmi integráció.* Hozzájárulnak a hátrányos helyzetűek integrációjához, az esélyegyenlőség biztosításához, a különböző eletronikus tartalmak és közszolgáltatások mindenki számára hozzáférhetővé tételéhez, a nyelvi korlátok leküzdéséhez.
- *Információs tér.* Hozzájárulnak a minőségi elektronikus tartalom és szolgáltatások széles körének kialakításához.

A szektor társadalmi szerepének, nemzetgazdasági jelentőségének fontosságát mutatja, hogy az európai szintű stratégiákban a nyelv- és beszédtechnológia kiemelt helyen szerepel. A kitűzött stratégiai célok ennek a kiemelt szerepnek igyekeznek megfelelni. Hangsúlyoztuk: a magyar nyelv- és beszédtechnológia támogatása hosszú távú, kormányzati szintű elkötelezettséget kíván. Ehhez kívánt szakmai támogatást nyújtani a Nyelv- és Beszédtechnológiai Platform Stratégiai Kutatási Terve.

1.2. A Megvalósítási Terv áttekintése

Jelen Megvalósítási Terv a Stratégiai Kutatási Tervben bővebben kifejtett stratégiai célokat emeli ki, azokon belül pedig a főbb kutatási irányokat ismerteti. Minden egyes kutatási irány esetében bemutatunk néhány olyan projekttervet, amelyek jól példázzák az adott terület főbb problémaköreit. A Megvalósítási Terv tartalmazza ezen projektek rövid leírását, várható eredményeit, ütemezését. Meghatározza továbbá a feladatok elvégzésének pénzügyi erőforrásigényét, infrastruktúra-szükségletét. Tartalmazza az erőforrások és infrastruktúra biztosítására, a megfelelő szabályozási környezet kialakítására, a feladatok megoldásában résztvevő, illetve bevonandó szervezetekre, a kutatási eredmények hasznosításának módjára vonatkozó terveket, elképzeléseket, ajánlásokat.

A fő prioritások meghatározásánál néhány ponton eltértünk a Stratégiai Kutatási Tervtől. Kisebb változtatás, hogy a főbb stratégiai célok alá rendelt alfejezetek egy szinttel feljebb kerültek. A Kutatásszervezés című stratégiai célt megkülönböztetett figyelemmel kezeltük, mivel a technológiatranszfer, a kommunikáció, a szabványosítás, az oktatói, kutatói utánpótlás, a kutatásfinanszírozás és az együttműködés olyan területek, amelyek a nyelv- és beszédtechnológia minden egyes részfeladatára, céljára „ráterülnek”, egyfajta horizontális prioritást képezve azok felett. Ezért ez a fejezet a tanulmány végén kapott helyet.

A Megvalósítási Terv stratégiai céljai tehát a következők:

1. Nemzeti kutatási infrastruktúra kialakítása és szolgáltatása a nyelv- és beszédtechnológia területén
2. A nyelvi információ kezelése, tárolása és feldolgozása
3. A nyelvi kulturális örökség digitális korba való átmentése
4. Természetes ember-gép kommunikáció
5. Környezeti intelligenciával segített élet
6. Többnyelvűség, a nyelvi korlátok leküzdése
7. Kutatásszervezés

Célunk az, hogy a következő évek finanszírozási terveihez adjunk szakmai anyagot. Rövidtávról 2 éves, középtávról 5 éves, hosszútávról ennél hosszabb időhorizont esetén fogunk beszélni. Terveinket a 2011-2020-as időintervallumra készítettük.

Az anyag összeállításában nem kizárólag a nyelv- és beszédtechnológia határain belül maradtunk, hanem kerestük az integratív kutatási irányokat is, amelyekben különféle célok érdekében hatékonyan és sikeresen tudnak együtt dolgozni az egyes szakterületek képviselői. Ennek megfelelően a projekttervek kidolgozásánál igénybe vettük nyelvtörténészek, neuro- és pszicholingvisták, kognitív tudósok, pszichológusok, informatikusok segítségét is. Továbbá igyekeztünk feltárni az ipari szférával és a társplatformokkal való együttműködés lehetőségeit is.

2 A stratégiai célok

2.1. Nemzeti kutatási infrastruktúra kialakítása és szolgáltatása a nyelv- és beszédtechnológia területén

A nyelv- és beszédtechnológia területén sikerrel alkalmazható módszerek és eljárások jellegéből következik, hogy korszerű kutatási eredmények és alkalmazások nem jöhetnek létre a megfelelő erőforrások, írott és beszélt nyelvi adatbázisok, alapvető sztenderdizált feldolgozó eszközök nélkül; ezek a nyelv- és beszédtechnológia elengedhetetlen szükségletei a fejlesztésben és az elért eredmények kiértékelésében is. A magyarországi nyelv- és beszédtechnológia fennállása óta a különböző műhelyekben szép számú adatbázist és szövegfeldolgozó eszközt fejlesztettek ki. Manapság viszont egyre erőteljesebben jelenik meg az egységesítés igénye: nem újabb és újabb, elszigetelten működő, egymással össze nem egyeztethető formátumú fejlesztésekre kéne költeni a pénzt, hanem a meglévőeket egységesíteni, és mindenki számára könnyen használhatóvá tenni.

Helyzetkép. A 90-es évek során Magyarországon megalakult nyelv- és beszédtechnológiai műhelyek, melyek sokáig elszigetelten működtek, az utóbbi években felismerték az együttműködés fontosságát, és annak szükségét, hogy az Európai Unió irányelveivel megegyező, egységes, mindenki számára elérhető és könnyen kiterjeszthető kutatási infrastruktúrákat hozzanak létre. Maga a Nyelv- és Beszédtechnológiai Platform létrejötte is ennek a törekvésnek az egyik eredménye. A szakmai fejlődést támogató kutatási infrastruktúra kialakításának több sarokköve is van. Ilyenek

- a megfelelő nyelvi erőforrások (korpuszok, ontológiák) folyamatos korszerűsítése és fenntartása, illetve
- az ezen erőforrások és a feldolgozásukhoz szükséges feldolgozóeszközök sztenderdizálása, valamint
- a létrehozott erőforrások terjesztése, illetve, amennyiben lehet, szabadon elérhetővé tétele.

A magyarországi nyelv- és beszédtechnológia elmúlt években, évtizedekben létrehozott erőforrásainak folyamatos korszerűsítése azonban anyagi nehézségekbe ütközik, mivel a kutatás-fejlesztés finanszírozási kereteiben az új erőforrások létrehozása nagyobb prioritásként szerepelt, mint a meglévő erőforrások továbbfejlesztése. Ennek ellenére a kutatás-fejlesztéshez szükséges korpuszok, általános és specifikus ontológiák léteznek, a feladat ma már valójában inkább ezek továbbfejlesztése és egységesítése. Az együttműködést és az erőforrások elérhetővé tételét célzó infrastruktúra

kialakítása az utóbbi években elsőrendű prioritássá vált, melyet egyértelműen jelez Magyarország részvétele olyan európai projektekben, mint a CLARIN (Common Language Resources and Technology Infrastructure), a FLARENET (Fostering Language Resources Network), a DARIAH (Digital Research Infrastructure for the Arts and Humanities) és az ESFRI (European Strategy Forum on Research Infrastructure).

Elérendő célok. A Nyelv- és Beszédtechnológiai Platform stratégiai céljainak eléréséhez az első és legfontosabb lépés a még hiányzó erőforrások megteremtése, valamint ezzel párhuzamosan a már meglévők egységesítése, szabványosítása és szabadon hozzáférhetővé tétele. Az egységesítés fontos momentuma olyan adatbázisok létrehozása, melyek minden magyarországi – egyetemi vagy ipari szférában dolgozó – fejlesztőnek kiindulási és összehasonlítási alapul szolgálhatnak.

Kutatási irányok.

1. Nagyméretű, különböző nyelvi információval ellátott spontánbeszéd-adatbázisok létrehozása
2. Magyar nyelvtechnológiai alapeszközkészlet kifejlesztése és közzététele
3. Kiértékeléshez szükséges adatbázisok és fórumok létrehozása

2.1.1. kutatási irány: Nagyméretű, különböző nyelvi információval ellátott spontánbeszéd-adatbázisok létrehozása

A könnyebben előállítható tervezett (olvasott) beszédadatbázisok mellett fontos a spontán vagy ahhoz közeli beszédet tartalmazó adatbázisok létrehozása, hiszen az ilyen jellegű beszéd szöveggé alakítása a tipikus élet- és alkalmazásközeli feladat. A spontán beszéd esetében a hangkapcsolat-eloszlást nem lehet előre tervezni, ezért csak a nagy (tipikusan több mint 100 órás) adatbázisméret tesz lehetővé reprezentatív mintavételt. Lényeges, hogy a beszélők száma, kora, neme stb. is jól kövesse a megcélzott réteget.

Mivel az alapvető emberi kommunikáció spontán beszéddel történik, ezért kiemelt jelentőségű, hogy kellően nagyméretű, szöveges leirattal rendelkező, reprezentatív tartalmú spontánbeszéd-adatbázisok készüljenek magyar nyelven. A legjobb beszédfelismerési hatásfok a legspecifikusabb feladatok terén érhető el. Ehhez pedig illeszkedő témaspecifikus adatbázisok szükségesek. A gépi beszédfelismerés mellett beszélőazonosításra, dialógusmodellezésre és általános fonetikai, morfológiai, korpusznyelvészeti kutatásokra is jól használhatók az ilyen nyelvi erőforrások.

Helyzetkép. A Platform tagjainak közreműködésével már számos beszédadatbázis készült, de ezek egyrészt csak tervezett beszédet tartalmaznak, másrészt a nemzetközi szinten elfogadott adatbázisméretektől egy-két nagyságrenddel le vannak maradva. A gépi beszédfelismerésnél, ahol kizárólag statisztikai modelleket használnak, különösen fontos az adatbázis mérete az akusztikai modellek jobb becsülhetősége és így a felismerési pontosság növelése érdekében.

Elérendő célok. Célunk tehát nemzetközi szinten is elfogadott méretű spontánbeszéd-adatbázis létrehozása magyar nyelvre, mind általános célú, mind témaspecifikus szövegek tekintetében.

Kapcsolódó projektek.

- A **Beszélt nyelvi Adatbázis** (BEA) fonetikailag megalapozott többfunkciós spontánbeszéd-adatbázis fejlesztésének célja a mai budapesti beszélők beszédének rögzítése, továbbá anyag biztosítása különféle kutatásokhoz és gyakorlati alkalmazásokhoz. A fejlesztés 2007 óta folyik az MTA Nyelvtudományi Intézetének Fonetikai Osztályán.
- A **Magyar Referencia Beszédadatbázist** (MRBA) a BME TMIT Beszédakusztikai Laboratóriuma és az SZTE Informatikai Tanszékcsoportja hozta létre 2004-ben. A cél egy olyan, olvasott folyamatos szöveget tartalmazó beszédadatbázis létrehozása volt, amely alkalmas PC-s beszédfelismerők betanítására, tesztelésére.

Projekttervek.

1. Reprezentatív, nagyméretű, szövegesen lejegyzett spontánbeszéd-adatbázis létrehozása, folyamatos fejlesztése
2. Témaspecifikus, nagyméretű, szövegesen lejegyzett tervezett- és spontánbeszéd-adatbázisok létrehozása, folyamatos fejlesztése

2.1.1.1. projektterv: Reprezentatív, nagyméretű, szövegesen lejegyzett spontánbeszéd-adatbázis létrehozása, folyamatos fejlesztése

Az emberi kommunikáció alapvető eszközének, a spontán beszédnek az általános vizsgálatához kiemelt jelentőségűek a kellően nagyméretű, szöveges leirattal rendelkező, különféle beszédstílusokat és témaköröket magukban foglaló beszédadatbázisok, magyar nyelven is. Az adatokat jó akusztikai viszonyok között célszerű rögzíteni, mert a tiszta felvétel általánosabban használható, mint a témaspecifikus zajokat tartalmazó. Fontos, hogy az adatbázist folyamatosan kell bővíteni, frissíteni tartalom és feldolgozottság szempontjából, hiszen a nyelvünk folyamatosan változik.

Várható eredmény, hatás, hasznosulás.

- Mindenfajta spontánbeszéd-felismerési feladatban használható a nyelvi és akusztikai modellek tanítására, javítására.
- Gyorsabb alkalmazásfejlesztés a beszédtechnológiában.
- Magasabb beszédfelismerési hatásfok.
- Fonetikai alap- és alkalmazott kutatások alapanyaga, megsokszorozódó eredmények.

2.1.1.2. projektterv: Témaspecifikus, nagyméretű, szövegesen lejegyzett tervezett- és spontánbeszéd-adatbázisok létrehozása, folyamatos fejlesztése

Az alkalmazásközeli eredményekhez nélkülözhetetlen a megfelelő témaspecifikus adatbázisok folyamatos létrehozása, ugyanis kiemelkedően jó beszédfelismerési hatásfok a legspecifikusabb feladatok terén érhető csak el. Ennek állami támogatása azért célszerű, hogy a technológiai kutatások valódi adatokon alapulva folyhassanak, és ne hátráltassa ezeket az ipari szféra bizalmatlansága (csak arra a technológiára áldoznak, ami már bizonyított, de a bizonyításhoz drága adatbázisokat kellene a kutató-fejlesztőknek saját forrásból előállítani). Továbbá a versengő technológiák korrekt összehasonlításához is szükség van publikus, a kutató-fejlesztők számára alacsony költséggel elérhető témaspecifikus, és így demonstratív, alkalmazásközeli adatbázisokra.

Várható eredmény, hatás, hasznosulás.

- Lényegesen pontosabb felismerési modellek, ezáltal jobb eredmények.
- Gyorsabb alkalmazásfejlesztés a beszédtechnológiában.
- Ipari alkalmazások gyors, akár exponenciális felfutású terjedése.

2.1.2. kutatási irány: Magyar nyelvtechnológiai alapeszközkészlet kifejlesztése és közzététele

A magyarországi nyelvtechnológia elmúlt éveiben az egyes kutatóműhelyekben kellő mennyiségű és minőségű szövegfeldolgozó eszköz került kifejlesztésre, ám ezek jellemzően egymástól elszigetelten jöttek létre. Így egyrészt sok a felesleges átfedés (pl. három különböző magyar nyelvű morfológiai elemző létezik), másrészt az eszközök nem kompatibilisek egymással, ami megakadályozza az összehasonlíthatóságot és az egészséges versenyszellem kialakulását. A már meglévő eszközök az egyes műhelyekben sok esetben elzárva „porosodnak”, mivel az utómunkálatokra, karbantartásra nem, csak új termékek kifejlesztésére szóltak a pályázati pénzek. Célunk ezen erőforrások és eszközök „leporolása”, publikálásra való felkészítése és mindenki számára egyszerűen és szabadon használhatóvá tétele.

Helyzetkép. A magyar nyelv erősen ragozó jellege miatt speciális helyzetben vannak a magyar fejlesztők, ugyanis a nagyobb európai nyelvekre, elsősorban az angolra kidolgozott technológiák magyarra való adaptálása nem feltétlenül célravezető. Ezért a magyar nyelvtechnológusok az elmúlt években saját szövegfeldolgozó alapeszközöket fejlesztettek ki: létezik magyar tokenizáló, mondatra bontó, morfológiai elemző és egyértelműsítő, főnévicsoport-azonosító, tulajdonnév-felismerő, mondatelemző. Mivel a kommunikációáramlás nem mindig megfelelő az egyes műhelyek között, sokszor előfordul, hogy ugyanarra a célra szolgáló eszközt több helyen is fejlesztenek.

A már kifejlesztett eszközök közül több a kutatóhelyeken „porosodik”, sok közülük nincs megfelelően publikálva, így nem használhatók. Az eddigi pályázati kiírások mind új termékek létrehozását célozták, a már meglévők karbantartására, utógondozására,

publikálására nem fektettek elég hangsúlyt. A nyelvtechnológia elég dinamikusan fejlődő szakterület ahhoz, hogy a pár évvel ezelőtti fejlesztések mára már elavultnak számítsanak, vagyis az idő előrehaladtával egyre nagyobb erőfeszítést igényel az erőforrások felfrissítése. Természetesen vannak már most is szabadon elérhető erőforrások, ám azok használatához az esetek nagy részében komoly informatikai szaktudás szükséges.

Elérendő célok. A kutatási irány legfőbb célkitűzése, hogy a már meglévő eszközöket és erőforrásokat egy mindenki által könnyen használható és szabadon elérhető adatbázisba szervezzük. Célunk illeszkedik a nemzetközi trendekhez: világszintű (BLARK (Basic Language Resource Kit)), európai (ESFRI, CLARIN) és hazai (NEKIFUT) kezdeményezések mutatják ennek stratégiai fontosságát.

Az eddig jellemző egymástól elszigetelt, sokszor átfedő fejlesztések helyett egy nagy közös adatbázis létrehozásával a kutatóhelyek közötti kommunikáció erősítése mellett emberi és pénzügyi erőforrást is megtakaríthatunk.

A közös nyelvtechnológiai alapeszközkészlet kifejlesztése kiváló lehetőséget nyújt arra, hogy a már elkészült erőforrásokat felfrissítsük, mindenki számára könnyen használhatóvá és szabadon elérhetővé tegyük.

Kapcsolódó projektek.

- A **CLARIN** projekt nagyszabású (26 országból 32 partnert tömörítő) kutatási infrastruktúra projekt, melynek célja a tudományos kutatás támogatása a nyelvtechnológia, a nyelvi erőforrások könnyen elérhetővé tételével. Ezt a jelenlegi, elszigetelten működő központok egységes hálózatba szervezésével kívánja elérni. Az eszközök és a szolgáltatások a weben mindenki számára egyszerűen, bárhonnan elérhetőek lesznek. Az infrastruktúra kiterjeszthető, vagyis új erőforrások és szolgáltatások könnyen hozzáadhatók.
- A **HunCLARIN** a vezető hazai nyelvtechnológiai kutatás-fejlesztést végző tudásközpontok stratégiai jelentőségű kutatási infrastruktúrája. A HunCLARIN hálózat a nemzetközi CLARIN projekt szerves része; a nyelvi erőforrásokat és eszközöket azzal összhangban, szabványos módon egyetlen, ellenőrzött belépési ponton át egy közös hálózatba szervezve teszi elérhetővé a kutatók számára.

Projektterv.

1. Magyar BLARK nyelvtechnológiai alapeszközkészlet kifejlesztése és közzététele

2.1.2.1. projektterv: Magyar BLARK nyelvtechnológiai alapeszközkészlet kifejlesztése és közzététele

A BLARK (Basic Language Resource Kit) iniciatívát 1998-ban hozták létre azzal a céllal, hogy összeállítsanak egy mátrixot a létező nyelvi erőforrásokról – annyi nyelvre, amennyire lehetséges. Ezzel egyrészt egy helyről elérhetővé teszik a meglevő erőforrásokat, másrészt jól láthatóvá válnak a pótolnivaló hiányosságok, ami így utat mutathat a későbbi fejlesztéseknek.

A magyar BLARK alapeszközkészlet kifejlesztésének első nagy lépése a már meglévő eszközök feltérképezése, felfrissítése, majd publikálása; ezután következik a hiányzó eszközök létrehozása. A magyar nyelvre már kifejlesztett eszközök mennyiségét és minőségét tekintve megállapítható, hogy a munka döntő része nem az új eszközök fejlesztése, hanem a meglévők felfrissítése lesz. Fontos a már publikált eszközkészlet folyamatos karbantartása és frissítése – ennek hiányában az erőforrások fejlesztése idővel ugyanolyan esetlegessé és széttöredezetté válna, mint most.

Várható eredmény, hatás, hasznosulás.

- A felesleges átfedések kiküszöbölésével időt, munkát és pénzt takarítunk meg.
- Erősítjük a kutatóhelyek közötti kommunikációt.
- Bárki számára elérhetővé és könnyen használhatóvá tesszük a magyar nyelvi erőforrásokat.

2.1.3. kutatási irány: Kiértékeléshez szükséges adatbázisok és fórumok létrehozása

A nyelvtechnológia nemzetközi színterén szinte már közhelyszámba megy, hogy a nyelvfeldolgozás hatékony fejlődéséhez elengedhetetlen egy olyan nyelvtechnológiai infrastruktúra, amelynek két fő komponense a *fejlesztés*hez és a teszteléshez szükséges nyelvi erőforrások létrehozása, valamint a *kiértékelés*, amely a kifejlesztett eszközök minőségének mérésén keresztül lehetővé teszi azok továbbfejlesztését és összehasonlítását. A nyelvfeldolgozó algoritmusok és eszközök kiértékelése elméleti szempontból a mögöttes hipotézisek igazolásához és a kutatási irányok közötti választáshoz, mérnöki szempontból pedig a legmegfelelőbb technológia kiválasztásához szükséges.

Ennek megfelelően az elmúlt évtizedben egyre több nemzetközi kezdeményezés tűzte ki céljául az egymással versengő módszerek és eszközök összehasonlíthatóvá tételét. Ehhez első lépésben létre kell hozni azokat a célnak megfelelő annotált korpuszokat, amelyeken a különféle nyelvtechnológiai eljárások kiértékelhetők, második lépésben pedig időről időre szükség van olyan célzott kiértékelési fórumok szervezésére, amelyek ösztönzik a tudományterület résztvevőit a többiekkel való megmérettetésre.

Helyzetkép. Mint azt a 2.1.2. kutatási irányban is kifejtettük, az elmúlt években számos nyelvtechnológiai alkalmazás került kifejlesztésre speciálisan a magyar nyelvre, ám ezek sok esetben ugyanazt a feladatot hajtják végre. Ebből következik, hogy – a nemzetközi trenddel összhangban – a magyar nyelvre is szükséges olyan korpuszok létrehozása, amelyek lehetővé teszik ezen eszközök összehasonlítását. A magyar nyelvtechnológiai kutatások fejlődéséhez nagyban hozzájárulhat a kiértékeléshez szükséges adatbázisok, illetve fórumok létrehozása, elsősorban az alacsonyabb nyelvfeldolgozási szintekre (pl. morfológiai elemzés és egyértelműsítés, részleges szintaktikai elemzés, tulajdonnév-felismerés).

Mindazonáltal fontos látni, hogy egy megfelelő minőségű kiértékelő korpusz nyelvi információval való ellátása (annotálása) csakis kézzel történhet. Figyelembe véve egy megfelelő kiértékelő korpusz méretét, a szükséges nyelvi annotáció gazdagságát, valamint azt a tényt, hogy az annotálást csak szakképzett nyelvészek végezhetik el, ezt a munkát egyetlen intézmény sem képes önerőből finanszírozni. További nehézséget jelent, hogy általában a kiértékelendő rendszerek más és más annotációs sémát használnak egy adott nyelvi jelenség leírására, amelyek mögött esetenként eltérő elméleti előfeltevések húzódnak meg. Így a kiértékelő korpusszal szemben elvárás, hogy legalább a már meglévő eszközök annotációs sémáját kezelje.

Célzott kiértékelési versenyek a nemzetközi nyelvtechnológia különböző területein igen régóta zajlanak: CLEF (információ-visszakeresés), TREC (információkinyerés), MT-NIST (gépi fordítás), EASY (szintaktikai elemzés). Ezek a rendszeresen megrendezésre kerülő versenyek hatalmas lökést adnak a fejlesztéseknek: egyrészt a kiértékelő korpuszok létrehozásával és publikálásával, másrészt egységes annotációs sémák kialakításával.

Elérendő célok. A fentiek értelmében célunk egy olyan nagyméretű referenciakorpusz létrehozása, amely alapján összemérhetőek a már kifejlesztésre került eszközök. Szándékaink szerint a kialakításra kerülő korpusz lehetővé teszi a különböző morfológiai elemzők, tulajdonnév-felismerők, részleges szintaktikai elemzők, szemantikai egyértelműsítők és koreferenciafeloldók kiértékelését is. Továbbá olyan ösztönzőket kívánunk bevezetni, amelyek motiválják a magyar nyelvtechnológia szereplőit arra, hogy az általuk kifejlesztett eszközök hatékonyságát összemérjék. Ilyen ösztönző lehet egy rendszeres időközönként szervezett kiértékelési verseny vagy egy olyan webhely, amely nemcsak az egy-egy nyelvfeldolgozási feladat elvégzésére alkalmas eszközöket gyűjti össze, hanem az eszközök minőségéről is leírást nyújt, lehetőleg azok tipikus hibáival (ld. 2.1.2.).

Kapcsolódó projektek.

- A **Szeged Korpusz és Treebank** építésének munkálatai 1999-ben kezdődtek, azóta két verziót publikált az SZTE Informatikai Tanszékcsoportja. A legújabb, 2.0 verzió egy morfoszintaktikailag elemzett és kézzel egyértelműsített szöveges adatbázis, amely a publikálása (2005) óta referencia-adatbázisként szolgált számos magyar természetesnyelv-feldolgozással foglalkozó kutatáshoz.
- A **Szegedi NER korpusz** a Szeged Treebank gazdasági rövidhíreket tartalmazó alkorpuszának tulajdonnév-annotált része, amely az eddigi összes magyar nyelvű tulajdonnév-felismerő rendszer építéséhez és kiértékeléséhez referenciakorpuszként szolgált.

Projektterv.

1. Kiértékeléshez szükséges referenciakorpusz létrehozása

2.1.3.1. projektterv: Kiértékeléshez szükséges referenciakorpusz létrehozása

A létrehozni kívánt kiértékelő korpusz javasolt mérete egymillió szövegszó. Fontos követelmény, hogy a referenciakorpusz reprezentatív legyen, azaz általános célú szövegeket éppúgy tartalmazzon, mint speciális szaknyelvi szövegeket, különös tekintettel azokra a területekre, ahol a jövőben intenzív nyelvtechnológiai fejlesztések várhatóak (orvosi, jogi, biológiai, műszaki szövegek).

A tervezett korpusz olyan többszintű nyelvi annotációt tartalmaz, amely lehetővé teszi a morfológiai egyértelműsítők, részleges szintaktikai elemzők, tulajdonnév-felismerők, referenciafeloldók és jelentésegyértelműsítők kiértékelését. A már meglévő eszközök felhasználásával az annotáció automatizálható, de a megfelelő minőség biztosítása érdekében szükséges a kézi ellenőrzés. A minőségi annotáció másik biztosítéka az annotátorok közötti magas egyetértés, amelyet úgy kívánunk garantálni, hogy minden annotálási feladatot 3 különböző személlyel is elvégeztetünk.

Mivel a kiértékelendő rendszerek általában más és más annotációs sémát követnek, egy olyan annotációs sémát kell kifejleszteni, amelyre minden már létező nyelvfeldolgozó szoftver annotációja leképezhető. A korpusz formátuma a nemzetközi szabványokhoz illeszkedő TEI-kompatibilis XML.

A kiértékelésben fontos a motiváció megteremtése is, hogy a cégek, kutatóintézetek a konkrét fejlesztéseken túl áldozzanak még annyi további energiát a munkafolyamatra, amely a tényleges kiértékelés elvégzéséhez szükséges. A motiváció megteremtéséhez szükségesnek látjuk rendszeres időközönként (kb. 3 évente) kiértékelő kampányok szervezését, hogy az elkészült alkalmazások hatékonyságát összevethessük.

Várható eredmény, hatás, hasznosulás.

- A kutatási hipotézis igazolása (kutatás).
- Az előrehaladás értékelése (kutatás).
- A kutatási alternatívák közötti választás (kutatás).
- A legígéretesebb technológia kiválasztása (piac).
- A finanszírozó intézmények objektív tájékoztatása (állami intézmények, Európai Bizottság).
- Szoftverek minősítése.

2.2. A nyelvi információ kezelése, tárolása és feldolgozása

A nyelvi információ kezelése és gépi feldolgozása hosszú távon valószínűleg az egyik legnagyobb érdeklődésre számot tartó és széles felhasználói rétegek által egyik leginkább keresett kutatási terület a nyelv- és beszédtechnológiában. Mutatja ezt például a Google és a Yahoo! keresőmotoroknak mint az információ-visszakeresés nemzetközi

Projektterv		Típus	Résztvevők	Költségforrás	Időzítés	Költség
1.1.1.	Reprezentatív, nagyméretű, szövegesen lejegyzett spontánbeszéd-adatbázis létrehozása, folyamatos fejlesztése	alkalmazott	kutatóintézet, egyetem	100% pályázat	2011-től folyamatosan	120M Ft
1.1.2.	Témaspecifikus, nagyméretű, szövegesen lejegyzett tervezett- és spontánbeszéd-adatbázisok létrehozása, folyamatos fejlesztése	alkalmazott	kutatóintézet, egyetem, kkv, multi	70% pályázat, 30% ipari támogatás	2011-től folyamatosan	120M Ft
1.2.1.	Magyar BLARK nyelvtechnológiai alapeszközkészlet kifejlesztése és közzététele	fejlesztő kutatás	kutatóintézet, egyetem, kkv	100% pályázat	2011-től folyamatosan	70M Ft
1.3.1.	Kiértékeléshez szükséges referenciakorpusz létrehozása	fejlesztő kutatás	kutatóintézet, egyetem, kkv	100% pályázat	2011–2013	90M Ft

2.1. táblázat. A 1. stratégiai cél projektterveinek összegzése.

élenjáróinak felfelé ívelő története. A böngészők következő generációjának már a szemantikai keresés és a lekérdezett információ strukturált megjelenítése lesz a feladata. Minden ilyen irányú lépéssel eggyel közelebb hozzuk az emberekhez a szemantikus webet, vagyis az értő számítógép világát.

Helyzetkép. A feladat egyik fő nehézségét az adja, hogy az ember az értelmezés során számos nehezen formalizálható információt is figyelembe vesz, amelyeket egy gép számára csak korlátozott módon és nehezen lehet elérhetővé tenni. Ilyenek többek között a megnyilatkozás körülményei (hol, mikor, kikkel), valamint azok többletjelentése (pl. ígéret, fenyegetés), amely szintén hatással van arra, hogy hogyan értelmezünk egy üzenetet. A nyelv- és beszédtechnológia feladata azonban egyelőre nem az ilyen jellegű többletinformáció figyelembevétele, hanem csakis a szövegfolyamban detektálható releváns információ adott célnak megfelelő feldolgozása. Mint minden, a nyelvben tárolt információ – bizonyos fokú – megértését magában foglaló, tehát szemantikai célkitűzéssel bíró feladat, ez is rengeteg, önmagában is kihívást jelentő részfeladatot tartalmaz.

Jelenleg ezeknek a sokrétű előfeldolgozási feladatoknak a kezelése zajlik a magyar nyelv- és beszédtechnológiában: a szöveg alkotóelemeinek azonosítása (tokenizálás, mondatra bontás, morfológiai elemzés és egyértelműsítés) már megoldottnak tekinthető, de folyamatban van a mondatok közötti összefüggések felismerése (referenciafeloldás), a tulajdonnév-felismerés és a jelentésegyértelműsítés is. Folyik továbbá az ezeket a feladatokat integráló kutatási irányok, mint a webbányászat és az információkinyerés, illetve -visszakeresés alapszintű kezelése.

Elérendő célok. Azokba a széles sodrású világfolyamatokba, amelyek a szemantikus web felé vezető fejlődést táplálják, talán lehetetlen beavatkoznunk. Annak az elérése viszont, hogy ne csak angolul vagy kínaiul, hanem magyarul is elérhető legyen mindez, egyáltalán nem lehetetlen. A továbblépéshez az eddigi fejlesztéseken túl szükséges újabb tudástárak építése, valamint a hangzó anyagok feldolgozásához a gépi beszédfelismerés és az információkinyerési technikák nyelvi elemeinek összehangolása.

Az intézmények, a vállalkozások számára a szöveges, hangzó és videóanyagokból automatikusan kinyert tudásbázisok komoly versenyelőnyt jelentenek információs társadalmunkban. Ezért azt gondoljuk, hogy ez az a stratégiai cél, ahol a legtöbb lehetőség adódik a kutatói és az ipari szféra együttműködésére.

Kutatási irányok.

1. Magyar nyelvű szöveges anyagokból történő információkinyerés és -visszakeresés
2. Magyar nyelvű hangzó anyagokból történő információkinyerés és -visszakeresés
3. Webbányászat

2.2.1. kutatási irány: Magyar nyelvű szöveges anyagokból történő információkinyerés és -visszakeresés

A nyelvtechnológia hosszú távú célja, hogy képesek legyünk szimulálni azt a tudást, amelyet az ember mozgósít, amikor szövegeket értelmez. A nagy mennyiségű információ kezelhetővé tétele alatt egyelőre az ilyen mélységű és finomságú megértésnél alapvetőbb célt ért a nyelvtechnológia: adott adattömegből megtalálni a kereső számára releváns információt, és csak azt, az eredményt pedig strukturálva megjeleníteni. Ez a feladat éppoly fontos az átlag felhasználó számára, mint specifikus (pl. vállalati) környezetben a munkafolyamatok megkönnyítésére. A dokumentumok rendszerezése, az információ áramlásának követhetővé tétele jelenleg rengeteg emberi munkát igényel, ám ezeket hosszú távon jelentősen meg lehet támogatni gépi módszerekkel.

Az ontológiák, tudástárak építése, amelyekkel a világismereti és nyelvi tudásunkat szimulálhatjuk, éppoly fontos, mint azoknak a speciális technológiáknak a fejlesztése, amelyekkel bizonyos típusú szövegekben adott információt be lehet azonosítani. Az elektronikus formában rendelkezésre álló feldolgozatlan állományok sok hasznos információt rejtenek, amelyekből szövegbányászati technikák segítségével statisztikák gyűjthetők vagy következtetések vonhatók le. A közeljövő technológiájára támaszkodva megvalósíthatóvá válnak a jelenleginél jobb eredményt nyújtó trendelemző szoftverek, amelyek segítségével, különböző aggregált statisztikákra támaszkodva az érdeklődők reális képet kaphatnak egy-egy iparágra vagy szakterületre jellemző helyzetről.

Helyzetkép. A karaktersorozatok felismerésén túlmutató keresés az egyik legösszetettebb és legnehezebben definiálható nyelvtechnológiai feladat. A releváns információ kinyerése, valamint annak strukturálása megfelelő relációk mentén olyan „mozgó

célpont", melynek elérése a kutatás előrehaladtával egyre távolabb kerül. A már megvalósított célok után mindig lesz hova fejlődni ezen a kutatási irányon belül: az ideális minőségű eredménynek csak a képzelet szab határt.

Magyarországon egyelőre a részfeladatok kidolgozásán van a hangsúly, illetve megtörtént már az első olyan komplex információkinyerő rendszerek prototípusának megalkotása is, amelyek ígéretes eredményt mutatnak. Az információkinyeréshez és -visszakereséshez elengedhetetlen részfeladatok közül a mondatfeldolgozás különböző alaplépései már tulajdonképpen megoldottnak tekinthetők a magyar nyelvre – annál nagyobb kihívást jelent a mondaton belüli és a mondatok közötti összefüggések automatikus felismerése. Az ehhez szükséges legtöbb részfeladat területén már történtek előrelépések, de a minőségi áttöréshez hosszú távú fejlesztésre van szükség. Jelenleg még nem megoldott a mondatok közötti időbeli és ok-okozati összefüggések feltárása, szereplőinek azonosítása. Ehhez szükséges mind a tudástárak típusainak bővítése, mind a gépi tanulási módszerek fejlesztése.

Elérendő célok. Célunk a magas színvonalú információkinyerés és -visszakeresés elengedhetetlen feltételét képező háttérmunka elvégzése, amely magában foglalja mind tudástárak fejlesztését, illetve létrehozását, mind felügyelt és félig felügyelt gépi tanulási módszerek fejlesztését, valamint ezen módszereknek az alkalmazását olyan területeken, ahol értékelhető eredmények várhatók középtávon. A megfelelő háttéradatbázisokkal, illetve gépi tanulási módszerekkel 10 év múlva éppúgy nem lesz probléma egy természetes nyelven feltett kérdésre megfelelő választ adó következtetőrendszert elkészíteni, mint automatizálni az álláspiaci munkafolyamatok jelentős részét.

Kapcsolódó projektek.

- A **Magyar WordNet** ontológia építése és alkalmazása információkinyerő rendszerekben című projekt 2005 és 2007 között zajlott három konzorciumi partner közreműködésével (SZTE Informatikai Tanszékcsoport, MTA Nyelvtudományi Intézet, MorphoLogic Kft.) egy GVOP pályázat keretén belül.
- Az **Automatikus információszerzés rövid (politikai, üzleti, piaci) hírekből** elnevezésű projekt legfontosabb kutatás-fejlesztési célja olyan tartalomelemzési és információkinyerési technológia kifejlesztése volt, amelynek segítségével szöveges dokumentumokból strukturált formában kivonható a releváns információ. (NKFP, 2001)
- Az SZTE Mesterséges Intelligencia Kutatócsoportja kidolgozott egy információkinyerő rendszert, amely emberi fehérjék interakcióira fókuszál. A **MEDLINE** adatbázis szövegeiből kinyert fehérje-interakciókat egy gráf formájában jelenítik meg, amely rendszerezi a kinyert tudást, a biológus szakértők számára könnyen értelmezhetővé téve azt. (GVOP, 2004)

Projekttervek.

1. Magyar FrameNet
2. Információkinyerés és trendelemzés az álláspiachoz kapcsolódóan

3. Klinikai információkinyerés

2.2.1.1. projektterv: Magyar FrameNet

A Berkeley egyetemen készülő FrameNet adatbázis egy olyan erőforrás, amely egyes igék argumentumainak szemantikai annotációján túlmutató általánosításokat ragad meg. Az ún. keretszemantika (frame semantics) elméletén belül kidolgozott módszertan szerint nem egy-egy ige pontos leírása történik meg, hanem több igét (de akár főnevet vagy melléknevet) is összefogó struktúrákat, kereteket írnak le, és meghatározzák az adott keretre jellemző szemantikai szerepeket. A keretek és keretelemek közötti kapcsolatokon túl a FrameNet az egyes keretek között öröklődéses hierarchikus viszonyokat is tartalmazhat, ezzel jelenítve meg bizonyos általánosításokat. Az az enciklopédikus tudás, amit minden beszélő mozgósít, amikor szöveget ért meg, ilyen formában lenne tárolható, és a számítógép számára feldolgozható.

A FrameNetet már több európai nyelvre elkezdték adaptálni különböző módokon – vagy egy meglévő erőforrás és az angol nyelvű FrameNet összekapcsolásával, vagy egy saját nyelvi FrameNet elkészítésével. A FrameNet magyar nyelvre történő adaptálásával lehetővé válnának mélyebb szemantikai elemzések. Az elkészülő rendszernek a meglévő Magyar WordNettel, illetve a tervezett Magyar Lexikográfiai Referencia-adatbázissal (2.6.1.1.) való összekötése felbecsülhetetlen értékű lexikális erőforrást eredményezne.

Várható eredmény, hatás, hasznosulás.

- Hatékonyabb módszerek alkalmazása az információkinyerés és -visszakeresés területén.
- Több innovációs eredmény.
- Az eredmények piaci felhasználhatósága.

2.2.1.2. projektterv: Információkinyerés és trendelemzés az álláspiachoz kapcsolódóan

A jelenlegi álláspiaci szoftverek nem alkalmasak arra, hogy a felhasználók folyton változó igényeit nyomon kövessék, így az adatok mögött rejlő naprakész információkat sem képesek kiaknázni. A munkáltatók nem a megfelelő embereket választják ki elsőre, vagy a megfelelők nem kerülnek a látószögükbe. Egy betöltendő új álláshoz nemcsak a leendő munkatárs eddigi munkahelyeit, végzettségét, hobbijait vagy vezetői engedélyének betűjelét kellene ismerni. Sokkal fontosabb és lényegesebb információ, hogy a leendő munkatárs eddig pontosan milyen munkákon dolgozott, hány emberrel és milyen módon tartott kapcsolatot, milyen beosztásban végezte a tevékenységét, és milyen problémákat kellett megoldania a munkája során.

A munkapiaci folyamatokat jól reprezentáló különböző aggregált statisztikákra támaszkodva az érdeklődők reális képet kaphatnak a munkapiac aktuális helyzetéről. Így könnyen elérhetők lehetnek olyan, egyes iparágakra, szakterületekre jellemző adatok, amelyek nagyban megkönnyíthetik az álláskereső dolgát. Többek között választ

kaphatnak az elvárható fizetések nagyságáról, vagy arról, hogy melyik régióban keresett egy adott szakterület. Lehetőség nyílhat a különböző végzettségek összehasonlítására, így nem utolsósorban az egymással versengő felsőoktatási intézmények rangsorolására is kiváló eszköz lehet. Másrészről a humánerőforrás-szakemberek munkáját jelentősen segítheti egy, az önéletrajzokból automatikusan felépített adatbázis, amely igen nagy mértékben megkönnyíti a megfelelő jelöltek kiválasztását.

Az álláshirdetések és önéletrajzok számos értékes információt rejtenek, de mivel ezek rendezetlen formában vannak, gépi feldolgozásuk csak nyelvtechnológiai eredmények felhasználásával lehetséges. A projekt keretében álláshirdetések és önéletrajzok elemzésére alkalmas rendszert dolgozunk ki. Ezen dokumentumok egy része az internet publikus részén elérhető, ezeket ún. webrobotok gyűjtenék. Egy robot folyamatosan begyűjti és elemzi azokat a site-okat, ahol álláshirdetés vagy önéletrajz megjelenhet. Egy másik robot monitorozza a potenciálisan hasznos információt tartalmazó oldalakat, és ha ott új hirdetés jelenik meg, azt leképezi az adatbázisba. A folyamatos monitorozás biztosítja, hogy az adatbázis naprakész maradjon. Az elemzés fő lépésében szükséges az adott hirdetésben, illetve önéletrajzban található lényeges információk automatikus felkutatása. Az azonos információtartalmú, de eltérő megfogalmazású szövegrészekből ugyanazokat a jellemzőket szeretnénk kinyerni: ehhez különféle ontológiák, szinonimaszótárak stb. beépítése szükséges a rendszerbe.

Várható eredmény, hatás, hasznosulás.

- Folyamatosan frissülő, naprakész munkapiaci adatbázis álláskeresőknek és hirdetőknek.
- Automatikusan előállítható statisztikák a munkaerőpiacról.
- A nyelvtechnológia eredményeinek gyakorlati alkalmazása, szinergia az ipari szférával.

2.2.1.3. projektterv: Klinikai információkinyerés

A kórházakban felhalmozódó vizsgálati dokumentumok, zárójelentések stb. elektronikus formában rendelkezésre álló, de feldolgozatlan állományok. Ezek sok hasznos tudást rejtenek, gyakran a folyó szöveges részekben. Szövegbányászati technikák segítségével statisztikák gyűjthetők, következtetések vonhatók le ezekből a dokumentumokból.

Az automatikus kódolás feladata, hogy a szövegekben előforduló tüneteket, diagnózisokat, kezeléseket azonosítsa, majd egy kódrendszerben elhelyezze azokat. A klinikai dokumentumok emberi pontossághoz közelítő eredményes feldolgozása nem lehetetlen célkitűzés a napjainkban rendelkezésre álló eszközökkel. Ebben a projektben a következő típusú dokumentumok automatikus kódolását tervezzük megvalósítani: műtéti leírás, anamnézis és nővéri tevékenységek ápolási dokumentációi.

A nemzetközi sztenderd kódrendszerekben szereplő fogalmakon túl a kórházi dokumentumok rengeteg hasznos információt tartalmaznak. Erre remek példát szolgáltatnak a beteg káros szenvedélyeiről – amennyiben azokra a vizsgálatok során fény

derül, illetve a beteg panaszaival összefüggésben lehet – szóló szövegrészek. Ezen információk összegyűjtése után olyan lekérdezések azonnali megválaszolására lesz lehetőség, mint például: „A 2008-ban tüdőrákban elhunyt betegek hány százaléka volt aktív dohányos?”

Várható eredmény, hatás, hasznosulás.

- Az információkinyerési technikák továbbfejlesztése, illetve alkalmazása egészségügyi célokra.
- A nyelvtechnológia alkalmazása egészségügyi célokra a hazai technológiát organikus módon támogatná.

2.2.2. kutatási irány: Magyar nyelvű hangzó anyagokból történő információkinyerés és -visszakeresés

Folyamatosan növekszik az interneten közzétett hangzó anyagok (videók, hanganyagok) mennyisége. Fontos terület az ezekben rejlő információ kinyerése és adatbázisba rendezése. A hangzó anyagokból történő információkinyerés és -visszakeresés messze leghatékonyabb módja az, ha először egy folyamatos nagyszótáras gépi beszédfelismerővel előállítjuk a közelítő pontosságú szövegátiratot, majd azon az írott szövegekre kidolgozott információkinyerési és -visszakeresési technikákhoz hasonlóakat alkalmazunk.

Helyzetkép. Megfelelő méretű spontánbeszéd-adatbázisok hiányában jelenleg nincs lehetőség az ilyen irányú fejlesztések megvalósítására. A 2.1. részben leírt adatbázisok (legalább részbeni) kialakítása révén nyílhatna meg a gyors fejlődés útja ezen a területen.

Elérendő célok. Hosszú távú célunk, hogy hangzó anyagokban ugyanolyan könnyen lehessen információkat keresni és megtalálni, ahogy szövegben. Mint azt a 2.2.1. részben bemutattuk, a szövegből történő információkinyerés terén már történtek előrelépések. A hangzó anyagokban történő keresés legalább kétszer annyira nehéz, hiszen a hangzó anyagokból először szöveges átiratot kell készíteni, majd azon lehet alkalmazni a szöveges információkinyerési technikákat.

Kapcsolódó projektek.

- Az AITIA és a BME TMIT kidolgozott egy természetes nyelvet feldolgozó, beszédfelismerésen alapuló szókereső eljárást (**Voxearch**), amelynek segítségével lehetőség nyílik az archívumok automatikus indexelésére. (GVOP, 2004)

Projektterv.

1. Magyar nyelvű hangzó anyagokból történő információkinyerés és -visszakeresés morf alapon

2.2.2.1. projektterv: Magyar nyelvű hangzó anyagokból történő információkinyerés és -visszakeresés morf alapon

Magyar nyelvre — annak erősen ragozó jellege miatt — a szó helyett morfémaszerű nyelvi egységekre (szótő, toldalék) alapozott gépi beszédfelismerés lényegesen pontosabb lehet, mint a szóalapú, ezenfelül az információkinyerési és -visszakeresési technikák is tipikusan szónál kisebb nyelvi egységekkel dolgoznak. Reményeink szerint a gépi beszédfelismerés és az utána következő feldolgozó technikák által használt nyelvi alapegységek összehangolása javít a magyar nyelvű hangzó anyagokból történő információkinyerés és -visszakeresés hatékonyságán. A projektben statisztikai és szabályalapú, illetve hibrid módszerrel meghatározott morfémaszerű egységeket vinnénk végig a magyar nyelvű hangzó anyagokból történő információkinyerés teljes folyamatán.

Várható eredmény, hatás, hasznosulás.

- Pontosabb, kisebb erőforrásigényű információkinyerés és -visszakeresés.
- Ipari alkalmazások gyorsabb terjedése.

2.2.3. kutatási irány: Webbányászat

Az interneten szinte korlátlan mennyiségben található információ, amelynek több mint 80%-a strukturálatlan szöveg formájában áll rendelkezésre. A webbányászat célja az információ automatikus kinyerése ezekből a szövegekből. A nagy mennyiségű szöveg kezelése mindenképpen gépi feldolgozást követel meg. Így fontos feladat olyan algoritmusok kifejlesztése és implementálása, amely az információ hatékony keresését teszi lehetővé. A webbányászat gazdaságilag és kormányzati szempontból egyaránt fontos területe az automatikus véleménykinyerés, amely jól használható minden olyan esetben, amikor fontos az internethasználók véleményének gyors felmérése.

Helyzetkép. Nemzetközi szinten a webbányászatnak számos alkalmazásával találkozunk. Ezek közé tartozik például a személyre szabott marketing: a cégek jobban megérthetik a vásárlók igényeit, és gyorsabban reagálhatnak a szükségleteikre. Kormányzati szempontból a webbányászat használható a bűnelkövetés vagy a terrorizmus megfékezésére. Magyarországon is történtek már előrelépések a webbányászat területén, de a terület gazdasági hasznosságát figyelembe véve mindenképpen szükségesek további fejlesztések.

Elérendő célok. Céljaink közé tartozik olyan algoritmusok kifejlesztése és implementálása, amelyek költséghatékonyan teszik elérhetővé a keresett információt. A webbányászat különösen fontos területének tartjuk az automatikus véleménykinyerést, amely mind gazdasági, mind kormányzati szereplők számára közvetlenül hasznosítható.

Kapcsolódó projektek.

- **TEXTREND**: Gazdasági és kormányzati döntéshozást támogató keretrendszer létrehozása trendelemző és szövegfeldolgozó eszközökkel (NKTH, Jedlik Ányos Program). A pályázat célja az integrált TEXTREND Platform létrehozása és demonstrációs alkalmazásainak kidolgozása a gazdasági és szakpolitikai döntéshozatal területein. A platform célja a világhálón található dinamikus és hatalmas mennyiségű információ mély elemzése.
- Szintén a TEXTREND projekt keretében az SZTE Informatikai Tanszékcsoportja kifejlesztett egy olyan rendszert, amellyel egy népszavazással kapcsolatos fórumon a hozzászólók automatikusan osztályozhatók aszerint, hogy az illető el fog-e menni szavazni, és ha elmegy, mit fog voksolni.
- Ipari megrendelésre és ipari támogatással az SZTE Informatikai Tanszékcsoportja 2009-ben megvalósította az [**origo**] híreinek automatikus címkézését, amely lehetővé teszi a cikkek szemantikai alapú kategorizálását.

Projekttervek.

1. Véleménykinyerés közösségi tartalmakból
2. Interaktív gépi tanulási technikák a webbányászatban

2.2.3.1. projektterv: Véleménykinyerés közösségi tartalmakból

Ahogy az internet egyre szélesebb társadalmi csoportokhoz jut el, úgy növekszik a nem csupán tényszerű adatokat közlő, hanem az emberek személyes tapasztalatain alapuló vélemények közlésére szolgáló oldalak (pl. fórumok, blogok) száma és ezzel együtt a személyes véleményekkel, érzelmi töltettel telített adatok mennyisége is. Az elmúlt években kellő mennyiségű szöveges dokumentum gyűlt össze – megnyitva az utat az emberi vélemények detektálására specializálódott rendszerek fejlesztése előtt.

A véleménydetektáló rendszerek számos alkalmazása képzelhető el: a termékekkel kapcsolatos fogyasztói elégedettség automatizált felmérésére irányuló alkalmazásoktól kezdve a politikai pártok, közszereplők támogatottságának automatikus felderítésére irányuló rendszerekig. A véleménydetekció a rendelkezésre álló szöveges információ mennyisége miatt gépi erőforrás igénybevétele nélkül elképzelhetetlen lenne.

Ebben a projektben magyar nyelvű közösségi tartalmak automatikus véleményelemzésére szolgáló módszerek kidolgozására és empirikus kiértékelésére törekszünk. A megcélzott rendszer képes egy adott egyedhez (pl. termék, szervezet) kapcsolódó vélemények kiemelésére a teljes magyar blog- és fórumszférából, illetve a vélemények osztályozásával időbeli tetszésiindex-változásokra is rá tud világítani.

Várható eredmény, hatás, hasznosulás.

- Fogyasztói visszajelzések megismerése.
- Keresőmotorok informáltságának növelése.
- A rendszer használható piackutatásra és hírcsoportok, fórumok monitorozására is.

2.2.3.2. projektterv: Interaktív gépi tanulási technikák a webbányászatban

A gépi tanulás a mesterséges intelligencia azon ága, ahol egy emberi szakértő által adott példákban (ebben az esetben szövegbeli jelölésekben) keresünk mintázatokat, rejtett összefüggéseket statisztikai módszerekkel. Ha rendelkezésre áll egy megfelelő méretű kézzel annotált tanító adatbázis, akkor gépi tanulási módszerek felhasználásával igen nagy pontosságú automatikus elemzőrendszereket lehet építeni. Azonban minden egyes új feladat és tématerület új, szakértők által annotált adatbázis építését követeli meg, ami igen költséges és időigényes feladat.

A projekt célja olyan algoritmusok kutatása és implementálása, amelyek a felhasználó és a gép interakcióján alapulnak. Ebben a megközelítésben a szakértő néhány példája alapján a gép tanul egy modellt, majd megkérdi a felhasználót, hogy bizonyos esetekben jó döntést hozott-e. A folyamat iteratív, vagyis a felhasználó válaszai alapján tovább tanul a gép. Összességében a szakértőnek egy nagyságrenddel kevesebb példát kell mutatnia a gépnek, mint a klasszikus tanítóadatbázis-építési megközelítésben. Így egy nyelvtechnológiai alkalmazás kifejlesztésének, illetve testreszabásának költsége drasztikusan csökkenthető.

Az algoritmusokat egy bárki által tanítható webbányász alkalmazásban teszteljük és validáljuk. A kidolgozandó algoritmusok segítségével bármely (nem szakértő) felhasználó taníthat saját maga igényei szerint webbányász rendszert úgy, hogy néhány példát mutat az általa kinyerni kívánt információtípusra (böngészőbe épülő pluginnel), majd a gép felkéri bizonyos esetek ellenőrzésére. Ezt az interaktív protokollt folytatva költséghatékonyan előáll egy megfelelő pontosságú webbányász rendszer.

Várható eredmény, hatás, hasznosulás.

- A módszer lehetővé teszi, hogy az új témakörökhoz tartozó információkhoz költséghatékonyan jussunk hozzá.
- A nyelvtechnológiai alkalmazások kifejlesztésének költsége drasztikusan csökkenthető.
- Könnyen testreszabható, egyedi webbányászati alkalmazások.
- A kifejlesztett gépi tanulási technikák más alkalmazási területeken is felhasználhatók.

2.3. A nyelvi kulturális örökség digitális korba való átmentése

Az Európai Unió egyik legfőbb célkitűzése Európa gazdag kulturális örökségének ápolása, megőrzése és minél szélesebb közönséggel való megismertetése. Az EU kulturális dimenzióval egészítette ki számos szakpolitikáját, így például az oktatást, a tudományos kutatást, az informatikai és kommunikációs technológiák támogatását, valamint a társadalmi és regionális fejlesztést. Ebbe az irányvonalba illeszkedik ez a stratégiai célunk, hiszen a nyelv- és beszédtechnológia fejlesztései szinte kínálják magukat a nyelvi kulturális örökség megőrzésének céljaira.

Projektterv		Típus	Résztvevők	Költségforrás	Időzítés	Költség
2.1.1.	Magyar FrameNet	alap	kutatóintézet, egyetem	80% pályázat, 20% önerő	2011–2013	80M Ft
2.1.2.	Információkinyerés és trendelemzés az álláspiachoz kapcsolódóan	alkalmazott	kutatóintézet, egyetem, kkv, multi	100% ipari támogatás	2012–2014	60M Ft
2.1.3.	Klinikai információkinyerés	alkalmazott	kutatóintézet, egyetem, multi, kórház	80% pályázat, 20% ipari támogatás	2011–2016	200M Ft
2.2.1.	Magyar nyelvű hangzó anyagokból történő információkinyerés és -visszakeresés morf alapon	alkalmazott	kutatóintézet, egyetem, kkv	70% pályázat, 30% önerő	2015–2016	40M Ft
2.3.1.	Véleménykinyerés közösségi tartalmakból	alkalmazott	egyetem, kkv, multi	100% ipari támogatás	2011–2013	150M Ft
2.3.2.	Interaktív gépi tanulási technikák a webbányászatban	alap, alkalmazott	kutatóintézet, egyetem, kkv	100% pályázat	2013–2015	60M Ft

2.2. táblázat. A 2. stratégiai cél projektterveinek összegzése.

Helyzetkép. Nyelvi kulturális örökségünk digitális korba való átmentése három fő szempont szerint zajlik párhuzamosan: egyrészt bármely magyar nyelvű nyelvi kulturális tartalom digitalizálása értékmentő, értékőrző feladat, valamint a kereshetővé tételen át szolgálja ezen kultúrkincsek könnyebb tudományos feldolgozását is. Az ilyen nyelvemlékek digitalizálása évek óta folyamatban van. Az Országos Széchényi Könyvtár, a NAVA és más hasonló nagy könyvtár/archívum elsődleges felhasználási helye lehet a nyelv- és beszédtechnológiai fejlesztéseknek. Másrészt elsősorban a magyar nyelvtechnológiától várható a kevés beszélővel rendelkező rokon nyelvek nyelvi rendszerének dokumentálása, írott és hangzó megnyilatkozásainak digitalizálása, valamint egy- és többnyelvű szótárainak, korpuszainak és egyéb erőforrásainak fejlesztése. E tekintetben is már számos előrelépés történt, ám az értékmentő munka korántsem tekinthető befejezettnek. Harmadrészt a jelenlegi, határontúli magyar nyelvváltozatok feltérképezése, dokumentálása és megőrzése is társadalmi prioritás kell, hogy legyen. Az írott alaperőforrások tekintetében jól áll a nyelvtechnológia, ám a beszélt nyelvi adatok rögzítése és feldolgozhatóvá tétele hossszú távú kihívás.

Elérendő célok. A tervezett értékmentő munka egyik célkitűzése a nyelvi kultúrkincs digitalizása (nyelvemlékek, irodalmi szövegek, nyelvváltozatok rögzítése), valamint kereshetővé tétele, amely nagyban segíti a magyar kulturális örökséget feldolgozó kutatói munkát. A forrásfelhasználás gazdaságosabbá tétele érdekében szükséges az eddig elszigetelten működő műhelyek munkáját szinkronizálni.

Kutatási irányok.

1. A régi magyar nyelvemlékek digitális korba való átmentése
2. A magyarországi és határon túli magyar nyelvváltozatok feltérképezése és adatbázisba szervezése
3. Nagy nemzeti hang/film/multimédia archívumok szövegtartalom szerinti kereshetővé tétele
4. Rokon nyelvek nyelvi erőforrásainak fejlesztése

2.3.1. kutatási irány: A régi magyar nyelvemlékek digitális korba való átmentése

A magyar nyelv- és beszédtechnológia szövegfeldolgozó és korpuszépítő eredményei jelentős segítséget nyújthatnak abban is, hogy az ország írott kultúrkincsét a digitális korba átmentsük. Ennek egy fontos részfeladata a régi magyar nyelvemlékek feldolgozása és digitalizálása. Mivel a feladat több tudományterület képviselőit is igényli (nyelvtörténészek, elméleti nyelvészek, informatikusok, könyvtárosok stb.), csak nemzeti összefogással hozható létre.

Helyzetkép. Az ország több intézményében folynak a régi magyar irodalmi és nyelvemlékek feldolgozását célzó projektek, mégsem mondhatjuk el, hogy akár már az ómagyar anyag teljeskörűen feldolgozásra került volna. A mai napig létezik még olyan ómagyar kódex, amely egyáltalán nem lett kiadva sem nyomtatásban, sem elektronikusan. Az idegen nyelvű szövegekbe beágyazottan előforduló első magyar szavak, az ún. szórványemlékek csak hosszas kutakodás után találhatók meg, ha egyáltalán. A kutatások egymással szinte alig kommunikáló műhelyekben zajlanak, így több párhuzamos projekt fut, amelyek külön-külön több pályázati forrást emésztenek fel, mintha egy nagy közös cél érdekében fognának össze. Ebből a helyzetből következik az is, hogy azok a szövegek, amelyek már megvannak elektronikusan, elszórtan lelhetők fel kiadóknál, egyetemeken, kutatóintézetekben, könyvtárakban.

Elérendő célok. Célunk a régi magyar nyelvemlékek nagy közös nemzeti adatbázisának létrehozása, amely első körben az ómagyar anyagot tartalmazná teljes egészében. A későbbiekben ez szolgálhatna mintaprojektként a középmagyar és akár a modern magyar irodalom digitalizálásához is. Az eddig elszigetelten működő műhelyek munkáját szinkronizálva a kommunikáció is fejleszthető, a forrásfelhasználás gazdaságosabbá tehető.

Az egyes szövegek többszintű feldolgozása a bennük található felbecsülhetetlen értékű nyelvi információt könnyen hozzáférhetővé teszi a nyelvtörténészek, irodalmárok, elméleti nyelvészek és magyartanárok számára, ami nemcsak a kutatást, de az oktatást is fellendítheti.

Kapcsolódó projektek.

- A **Magyar Generatív Történeti Szintaxis** OTKA-projekt az MTA Nyelvtudományi Intézetében zajlik (2009–2013). A projekt egyik célja egy elektronikus nyelvtörténeti adatbázis felállítása, mely tartalmazni fogja a már kiadott ómagyar kori nyelvi anyag kereshető elektronikus változatát, továbbá arányos válogatást a középmagyar kor nyelvemlékeiből.
- Az ELTE BTK Régi Magyar Irodalomtörténeti Tanszékének **Sermones Compilati** kutatócsoportja OTKA-támogatással készíti internetes szövegközléseit, melyek középpontjában a késő középkori és kora újkori magyarországi prédikációirodalom forrásszövegei állnak.
- A **Magyar Nyelvemlékek** című szolgáltatás 2009 őszén indult útnak az Országos Széchényi Könyvtár kezdeményezésére. A kezdeményezés célja, hogy – előmozdítva a dokumentumok digitalizálását – hozzáférhetővé tegye a legjelentősebb magyar nyelvemlékeket, hogy gyűjtőoldala legyen a velük kapcsolatos tudományos szövegközléseknek (bibliográfiák, szövegkiadások), és hogy egyúttal korszerű segédanyagot biztosítson diákok és tanárok számára egyaránt.

Projektterv.

1. Nemzeti Ómagyar Adatbázis

2.3.1.1. projektterv: Nemzeti Ómagyar Adatbázis

A projekt célja egy nagy közös, mindenki számára szabadon hozzáférhető és könnyen használható adatbázis, amelyben egy helyen megtalálható és kereshető minden ómagyar korból származó nyelvi adat. Az adatbázis tartalmazná az összes típusú nyelvemléket a szórványemlékektől a teljes kódexekig.

A projekt első lépése egy szakértői grémium létrehozása, amelynek feladata az együttműködés feltételeinek és annak az egységes formátumnak a kialakítása, amelybe minden feldolgozandó szöveg illeszkedik. Következő lépés a már elektronikus formában meglévő anyagok összegyűjtése és egységesítése. Ehhez szükség van a kiadók, egyetemek, kutatóintézetek, könyvtárak együttműködésére.

Minden ómagyar szövegnek többszintű változata is elkészülne: az eredeti kézirat és a nyomtatott kiadás(ok) beszkennelve, a kereshető betűhű szöveg, a mai magyar olvasó számára könnyebben érthetővé tett változat, a morfológiailag elemzett és egyértelműsített változat. A teljes szöveg elemzésével és kereshetővé tételével elérhetővé válik a mai magyar kifejezések és nyelvi jelenségek diakrón vizsgálata akár egy gombnyomással. Minél részletesebb nyelvi információval van ellátva a szöveg, annál több – ortográfiai, fonológiai, morfológiai, szintaktikai – vizsgálatot tesz lehetővé. Ebben tudnak hathatós segítséget nyújtani a nyelvtechnológiai eszközök.

Az adatbázis egy helyről, mindenki számára szabadon hozzáférhetővé tételével nagyot lendíthetne a nyelvtörténeti kutatásokon, és akár iskolai tananyagok bemenete is lehet.

Várható eredmény, hatás, hasznosulás.

- Az eddig egymástól elszigetelten működő kutatóműhelyek együttműködésétől a kommunikáció javulása várható.
- Az egy helyről, szabadon hozzáférhető adatbázis ismertebbé teheti a kutatók, a fiatalok és az érdeklődők számára is a magyar nyelv történetét.
- A párhuzamosan futó projektek helyett egy közös projekt idő és erőforrás szempontjából is gazdaságosabb.
- A nemzetközi szabványok alkalmazásával csatlakoznánk a nemzetközi nyelvtörténeti adatbázisok sorába, lehetővé téve a magyar nyelv kutatását külföldiek számára is.

2.3.2. kutatási irány: A magyarországi és határon túli magyar nyelvváltozatok feltérképezése és adatbázisba szervezése

A magyar nyelv különböző, még élő változatainak digitális rögzítése sürgető feladat, hiszen a ritkább nyelvváltozatok beszélői kiöregednek, és a vidéki lakosság városokba áramlásából fakadóan egyre inkább megfigyelhető a dialektusok eltűnése, illetve a határon túli nyelvváltozatok esetében a nyelvvesztés folyamata. A nyelvváltozatok feltérképezése és adatbázisba szervezése nagyszabású, központi koordinációt, valamint szociolingvisták és nyelvtechnológusok összefogását és nemzetközi együttműködést igénylő feladat.

Helyzetkép. A magyar nyelvváltozatok kutatása az elmúlt húsz évben túlnyomórészt nyelvtechnológiai támogatás nélkül zajlott, az elmúlt években azonban a szociolingvisztikai kutatások egyéb területein megkezdődött az együttműködés bizonyos kutatócsoportok között. A nyelvtechnológiai eszközökkel segített értékőrző tevékenység nagy része egyelőre az írott szövegek digitalizálására terjed ki, a beszélt nyelvi korpuszok elkészítésének technikai komplexitása és munkaigényes volta miatt a beszélt nyelv hangzó formában való rögzítése és kereshetővé tétele az elkövetkezendő évek súlypontja lehet. Amellett, hogy a hangzó formában való rögzítés értékmentő szereppel is bír, a beszélt nyelvi anyagok hangzó korpusza a szöveges átiratokhoz képest mindig magában hordoz további kutatási lehetőségeket (pl. fonetikai vizsgálatok).

Elérendő célok. A magyar nyelv különböző nyelvváltozatainak rögzítésével digitális lenyomatot készíthetünk a már eltűnőben lévő dialektusokról, megnyitva az utat ezzel a szociolingvisztikai vizsgálatok széles spektruma előtt. A felvett anyagok hangzó és szövegesen lejegyzett változata egyaránt fontos, hiszen más és más vizsgálatok alapjául szolgálhatnak.

Kapcsolódó projektek.

- A **Budapesti Szociolingvisztikai Interjú** (BUSZI) egy nagyszabású vizsgálat, amely a budapesti lakosok reprezentatív mintáján végzett magnetofonos felmérés alapján megbízható adatokat gyűjt a magyar nyelv Budapesten beszélt válto-

zatáról. A munkálatok 1987 óta többszörös OTKA-támogatással folynak az MTA Nyelvtudományi Intézetében.

- A **Magyar Nemzeti Szövegtár** munkálatai 2002-től a Kárpát-medencei Magyar Nyelvi Korpusz projekt keretében kiegészültek a teljes Kárpát-medence magyar nyelvhasználatára kiterjedő gyűjtéssel. 2005 novemberében készült el a szlovákiai, kárpátaljai, erdélyi és vajdasági nyelvváltozatokkal kiegészült, valóban nemzetivé váló Magyar Nemzeti Szövegtár. A Nyelvi Irodák és az MTA Nyelvtudományi Intézete együttműködésének köszönhetően az első olyan magyar nyelvi korpusz jött létre, amely a magyarországiak mellett a határon túli magyar nyelvváltozatokat is felöleli. A szövegtár összeállítását az OTKA, az internetes megjelenést az IHM támogatta, a Kárpát-medencei Magyar Nyelvi Korpusz munkálatai pedig egy NKFP pályázat keretében folytak.

Projektterv.

1. Magyar beszélt nyelvi dialektális adatbázis építése

2.3.2.1. projektterv: Magyar beszélt nyelvi dialektális adatbázis építése

A projekt célja a magyar nyelv dialektusaiból egy reprezentatív mintaanyag beszélt nyelvi korpusz formájában történő megőrzése. A mintaanyag magában forglalná a magyarországi és a határon túli nyelvváltozatok egyes elkülöníthető dialektusainak egy-egy reprezentatív mintáját, és kereshető formában lenne szabadon hozzáférhetővé téve a kutatóközösség számára. A korpuszépítést a szociolingvisztikai alapkutatásokra támaszkodó nyelvész szakértők munkája előzi meg, akik a rögzítendő dialektusok, beszélők kiválasztását, az elkészülő korpusz annotációját tervezik meg. A korpusz annotációja, illetve kereshetővé tétele a projekt nyelvtechnológus résztvevőinek feladata. A korpusz értékőrző szerepe nyilvánvaló, a rögzített anyagok mind a kutatásban, mind az oktatásban hasznosulhatnak.

Várható eredmény, hatás, hasznosulás.

- Az eltűnőben lévő dialektusokat az utolsó pillanatban rögzíthetjük.
- Az adatbázis további szociolingvisztikai és fonetikai kutatások bemenete lehet.

2.3.3. kutatási irány: Nagy nemzeti hang/film/multimédia archívumok szövegtartalom szerinti kereshetővé tétele

A beszédfelismerési technológia a nagy nemzeti hang/film/multimédia archívumok szövegtartalom szerinti kereshetőségéhez is hozzájárulhat. Az alaptechnológia már ma is elérhető magyar nyelven, azonban a speciális tartalmakhoz (pl. régi filmhíradók vagy parlamenti beszédek hanganyagához) való adaptáció jelentősen növelheti a használhatóságot.

Helyzetkép. Ma hangarchívumok szövegtartalom szerinti keresésére több elérhető megoldás is van a piac kínálatában. Ezek a nagyszótáras folyamatos beszédfelismerésre épülő rendszerek lényegesen magasabb szófelismerési pontosságra képesek, mint a korábbi általános fonémasorozat-hasonlóságra épülő megközelítések, melyek lényegében nem alkalmasak éles helyzetben való használatra. Azonban a modern technológiának az a jellegzetessége, hogy a tanítóadatbázis akusztikus és nyelvi viszonyaihoz jól illeszkedő felvételeket képesek csak nagy pontossággal felismerni és tartalom szerint indexelni. Minél jobban eltér a tesztfelvétel szóhasználata, háttérzaja, hangrögzítésének módja a betanításétól, annál kisebb lesz a felismerési pontosság. Mivel a ma elérhető tanítóadatbázisok csak nagyon szűk szeletét képviselik a nemzeti hang/film/multimédia archívumoknak, így az alkalmazásuk is csak hasonlóan korlátozott lehet, azaz az archívumok csak egy viszonylag kis részét képesek nagy pontossággal indexálni.

Elérendő célok. Célunk az, hogy a jelentős értéket képviselő nemzeti hangarchívumok hanganyagának nagy többségét magas (70% feletti szó-, 90% feletti karakterpontossággal) tudjuk felismerni és indexálni. Ezáltal a szövegtartalom szerinti keresés használhatóságának mértéke megszokszorozódhat.

Kapcsolódó projektek.

- A **mindroom** egy beszédfelismerővel támogatott videóarchívum. A rendszer alapfunkciója, hogy a hangfájlokból egy beszédfelismerő motor segítségével szöveges leiratot készít, majd abból egy kereshető adatbázist épít, így a látogatók ugyanúgy tudnak keresni a videó- és hanganyagokban, mint egy cikk szövegében. A mindroomot egy akadémiai hátterű projektcég hozta létre egy uniós támogatású KMOP projekt keretében.
- A **SEARCHPERT-ALL** az ALL által kifejlesztett rendszer, amely audiovizuális archívumokban hangzó (beszéd) formában megőrzött, szöveges átirattal nem rendelkező anyagok tartalmának visszakeresését teszi lehetővé.

Projekttervek.

1. Egy nemzeti hang/film/multimédia archívum szövegtartalom szerinti kereshetővé tétele
2. Parlamenti beszédek tartalmi kereshetősége, folyó beszédek élő feliratozása

2.3.3.1. projektterv: Egy nemzeti hang/film/multimédia archívum szövegtartalom szerinti kereshetővé tétele

Egy archívummal szoros együttműködésben, a már elérhető szöveges lejegyzett tartalmainak felhasználásával részben öntanuló algoritmusokkal segítve lényegesen pontosabb beszédfelismerést tennénk lehetővé. Reprezentatív mintát véve az archívum tartalmából, azt szövegesen lejegyeznénk, és így segítenénk a tanuló algoritmusokat, tennénk kiértékelhetővé a felismerési, indexálási eredményeket.

Az előálló rendszer közhasznú lenne, ezért indokolt az állami támogatás, de az eddigi tapasztalatok alapján a kutatói szféra mellett a kis- és középvállalkozásokat is be lehet vonni ilyen típusú projektbe.

Várható eredmény, hatás, hasznosulás.

- A nemzeti hang/film/multimédia archívum „életre keltése” azáltal, hogy bárminemű szöveges tartalom kereshetővé válik.
- A nemzeti film-, tévé- és hangzó történelem jobban részévé válhat a kultúrának.
- Más beszédfelismerési, szövegtartalom szerinti keresési feladatokra továbbvihetők a tudományos-technológiai eredmények.

2.3.3.2. projektterv: Parlamenti beszédek tartalmi kereshetősége, folyó beszédek élő feliratozása

A Magyar Parlament beszédeinek szöveges kereshetősége olyan nyilvánvaló közérdek, hogy kezdettől megtörténik ezek szöveges lejegyzése. A videófelvételek is rögzítésre kerülnek, azonban a szöveges és a multimédia tartalmak összerendelése már a felhasználó feladata, ami néha elég körülményes. Azonban a mai gépi beszédfelismerési technológiákkal megoldható, hogy a már lejegyzett videófelvételek szövegtartalom szerint kereshetők, egy kattintással lejátszhatók, sorrendezhetők legyenek, méghozzá közel 100%-os pontossággal. Továbbá a rendelkezésre álló nagy mennyiségű lejegyzett hanganyag viszonylag alacsony költséggel teszi lehetővé, hogy a még nem lejegyzett élő felszólalások nagy pontosságú szöveges átalakítása online megtörténhessen, hiszen jó minőségű modellek építhetők a már meglévő anyagokból. A parlamenti beszédeket folyamatosan a képernyőre is lehet írni, de nagyon gyors információkivonatolás is elvégezhető, illetve hosszabb távú projekt keretében akár másik nyelvre is lefordítható.

Várható eredmény, hatás, hasznosulás.

- A társadalom jobban részt vehet a törvényalkotási folyamatokban, erősebb társadalmi kontroll valósítható meg.

2.3.4. kutatási irány: Rokon nyelvek nyelvi erőforrásainak fejlesztése

A ma beszélt finnugor nyelvek legtöbbje erősen veszélyeztetett, azonban sokuk nyelvi rendszere még nincs leírva modern keretben. Nemzetközi szinten folyik az a munka, amely a kihalásra ítélt vagy tipológiai jellegzetességeik miatt figyelmet igénylő nyelvi rendszerek lejegyzését célozza meg. A napjainkban lejegyzett anyagok feldolgozása nyelvtechnológiai hátteret igényel – legyen szó korpuszépítésről vagy később akár ezek alapján szótárak kiadásáról. A nyelvrokonság kérdése azonban nemcsak napjainkban fontos téma – sok lejegyzett anyag kéziratos formában a 19. századból maradt ránk, illetve a 19-20. században jelent meg könyv formában. Az ilyen magyarországi anyagok feldolgozása, digitalizálása és elérhetővé tétele kizárólag itthon történhet meg, ám nemzetközileg is fontos kutatási feladatoknak az előfeltétele.

Helyzetkép. Napjainkban nemzetközi együttműködésben zajlik több kis rokon nyelvre szótárak kiadása (elektronikus is), ám további munka szükséges a nyelvek teljes rendszerének feltárásához és dokumentálásához. Fontos e szempontból, hogy a régebbi gyűjtések anyagai hozzáférhetőek legyenek. Jelenleg Magyarország több kutatóhelyén találhatók olyan kéziratok, amelyeknek forrásértéke és nemzetközi jelentősége indolkolttá teszi kiemelt kezelésüket. A 19. századból és a 20. század elejéről származó könyvek, kéziratok nagy része az MTA könyvállományának részét képezi, de megtalálhatók egyes példányok az ország több kutatóhelyén és könyvtáraiban is. Ezek a könyvek ma nehezen elérhetőek, illetve használhatatlan állapotban vannak. Feldolgozásuk a digitalizálás hiánya miatt sokszor még várat magára.

Elérendő célok. Célunk az erősen veszélyeztetett vagy kihalt rokon nyelvek nyelvi adatait modern eszközökkel rögzíteni, illetve a meglévő anyagokat digitalizálni, elérhetővé és kereshetővé tenni. Ennek a kulturális és tudományos szempontból is értékmentő tevékenységnek az első lépése lehet az ország különböző kéziratait, illetve fizikailag rossz állapotban levő könyveit a digitalizálás által megmenteni a szó szerinti elporladástól.

Kapcsolódó projektek.

- **A nganaszan nyelv számítógépes morfológiai elemzése** (2006-2009) című OTKA-projekt célja az volt, hogy elemzőprogram fejlesztésével, a hiányos és ellentmondásos adatok, paradigmák tisztázásával, újabb szövegek, hangzó anyagok terepen történő gyűjtésével a nganaszan nyelv teljes fonológiai és morfológiai leírását adja.
- Az **Obi-ugor morfológiai elemzők és korpuszok** (2008-2010) című OTKA-projekt munkatársai két obi-ugor nyelv három nyelvjárásának morfológiailag elemzett korpuszát hozzák létre.

Projektterv.

1. Digitalizált Reguly-archívum.

2.3.4.1. projektterv: Digitalizált Reguly-archívum

Reguly Antal a 19. század derekán hatalmas mennyiségű anyagot gyűjtött össze nyelvrokonainktól. Reguly gyűjtése a nemzetközi tudományosság egyik legnagyobb teljesítménye, ami akkor nyerhetne igazi jelentőséget, ha a világ bármely pontján elérhető lenne számos tudományág kutatói számára.

Ennek ellenére a jelenleg Magyarországon elérhető Reguly-művek közül (melyek több kötetet tesznek ki) egy sem digitalizált. A projekt célja a Reguly által gyűjtött művek digitalizálása és elérhetővé tétele. A szövegek digitális formában való közreadása mind nyelvészeti, mind néprajzi szempontból nagy érdeklődésre tarthat számot a nemzetközi tudományosságban, illetve az érintett kis népek körében is. A projekt egyik fontos lépése az átírások nemzetközi szabványokat követő formátumra hozása.

Várható eredmény, hatás, hasznosulás.

- A szabadon hozzáférhető adatbázis hatalmas lendületet adhat a nyelvészeti, néprajzi kutatásoknak.
- A másfél évszázada porladó gyűjtemény végre bárki számára szabadon elérhető lenne.

Projektterv		Típus	Résztvevők	Költségforrás	Időzítés	Költség
3.1.1.	Nemzeti Ómagyar Adatbázis	alap	kutatóintézet, egyetem, kkv, állami intézmények	100% pályázat	2011–2020	360M Ft
3.2.1.	Magyar beszélt nyelvi dialektális adatbázis építése	alap	kutatóintézet, egyetem, kkv	70% pályázat, 30% önerő	2011–2016	170M Ft
3.3.1.	Egy nemzeti hang/film/multimédia archívum szövegtartalom szerinti kereshetővé tétele	alkalmazott	kutatóintézet, egyetem, kkv	100% pályázat	2012–2014	130M Ft
3.3.2.	Parlamenti beszédek tartalmi kereshetősége, folyó beszédek élő feliratozása	alkalmazott	kutatóintézet, egyetem, kkv	100% pályázat	2011–2013	80M Ft
3.4.1.	Digitalizált Reguly-archívum	alap	kutatóintézet, egyetem, kkv	100% pályázat	2011-2012	30M Ft

2.3. táblázat. A 3. stratégiai cél projektterveinek összegzése.

2.4. Természetes ember-gép kommunikáció

A természetes ember-gép kommunikáció rendkívül sokrétű, szerteágazó témakör. Tágabb értelemben véve az ember-gép együttélésen azt értjük, hogy az ember többletképességeket kaphat a gépektől. A gépek segítenek bizonyos funkciókat, például az értékelés, a diagnosztika (2.5.1.) vagy a döntés-előkészítés (2.2.1.) területén. Ebben a stratégiai célban a szűkebb értelemben vett ember-gép kommunikációt, vagyis elsősorban a beszédalapú ember-gép kapcsolati technológiákat, főként a gépi beszédfelismerést és -előállítást ismertetjük.

Helyzetkép. A magyarországi beszédtechnológia egyes részterületein világszínvonalú, azonban általánosságban véve inkább követő jellegű. Sok mindenre már használható a beszédtechnológia, de a teljesen általános, széles körű elterjedés a mai technológiai szint mellett még nem lehetséges.

Elérendő célok. Rövidtávú célunk a jól teljesítő élvonalbeli technológiák továbbfejlesztése és az általános technológiai színvonal felzárkóztatása az élbolyhoz. Középtávú cél, hogy a más nyelveken már sikeres alkalmazásokat magyarra is átültessük, illetve a magyar nyelven már sikeres technológiákat más nyelvekre is alkalmazzuk. Hosszú távú cél az emberi kommunikációt valamilyen értelemben elérő, illetve meghaladó ember-gép kapcsolati kommunikációs technológia, amihez más tudományos-technológiai területeknek (pl. agykutatás, számítógépek kapacitása) is fejlődnie kell, továbbá az emberi nyelvnek, gondolkodásnak is alkalmazkodnia kell a gépekhez (pl. internetes keresési kultúra).

Kutatási irányok.

1. Robusztus beszédfelismerési technikák
2. Skálázható kiejtésátíró szoftver és kiejtési szótárak fejlesztése

2.4.1. kutatási irány: Robusztus beszédfelismerési technikák

A beszédfelismerési technológia szélesebb körű elterjedésének egyik kulcsa a zajrezisztencia javítása lehet. Itt található az egyik legkritikusabb szűk keresztmetszet („bottleneck”), ami gátolja a beszédfelismerésre épülő számos egyéb technológia és alkalmazás (dialógusrendszerek, hangzó anyagokból történő információkinyerés és -visszakeresés (2.2.2.), élő fordítás stb.) közhasznú megvalósulását. Fontos látni, hogy itt nem pusztán a jelfeldolgozó technikák kutatás-fejlesztéséről van szó, egyes kutatási irányok már a szemantikai szintet is bekapcsolják a magasabb zajérzéketlenség eléréséért folyó munkálatokba. Noha a feladat nehézsége nyilvánvaló, minden kisebb, de stabil javulás alkalmazások tízezreit teheti lényegesen jobban használhatóvá nyelvtől gyakorlatilag függetlenül. A magyarországi kutatóhelyek eddig is aktívan részt vettek az ezirányú kutatásokban, azonban a megfelelő méretű beszédadatbázisok elkészültével (ld. 2.2.1. és 2.2.2.), a megfelelő eszközpark kialakításával és versenyképes kutatói jövedelmek hosszabb távú biztosításával eltűnhet a mostani jelentős versenyhátrányunk a nyugati szereplőkhöz képest.

Helyzetkép. A klasszikus, közel 30 éves beszédfelismerési lényegkiemelési technika (MFCC: Mel Frequency Cepstral Coefficients) hegemóniája mára kezd megszűnni, a kifinomultabb (PLP: Perceptual Linear Prediction), tanulást is alkalmazó lényegkiemelési technikák pedig kezdenek előtérbe kerülni. Ugyanakkor a javulás még mindig nem túl nagy, messze elmarad a hatásfok az emberi hallás zajrezisztenciájától.

Elérendő célok. Középtávú célunk a jelenlegi state-of-the-art megjavítása. Bármily kicsi, de következetes javulás világszerte fokozott érdeklődés tárgya, gyakran rövid időn belül alkalmazásba vihető. A hosszú távú cél új paradigmák vizsgálatával lényeges javulás elérése, az emberi képességeket – legalább bizonyos körülmények között – elérő vagy meghaladó zajrobusztusság megvalósítása.

Kapcsolódó projektek.

- Az **ALL** beszédfelismerője nagyszótárral működő rendszer, amely egy adott nyelven elhangzó, lexikálisan nyílt (nem korlátozott), folyamatos beszéd automatikus felismerését (beszédről szövegre való leképezését) teszi lehetővé. A rendszer használata automatizálni tudja a munka nagy részét, amikor digitálisan rögzített beszélt anyag utólagos leírására van szükség.
- A **BME TMIT** műhelye kifejlesztett egy statisztikai alapú folyamatosbeszéd-felismerő motort és fejlesztői környezetet. Az eszközzel középszótáras, valós időben működő beszédfelismerők készíthetők (IKTA, OTKA, 2004).

Projekttervek.

1. Robusztus lényegkiemelő technikák vizsgálata gépi beszédfelismeréshez
2. Új paradigmák vizsgálata a zaj- és beszélőrobusztus gépi beszédfelismerés érdekében

2.4.1.1. projektterv: Robusztus lényegkiemelő technikák vizsgálata gépi beszédfelismeréshez

A közelbeszélő mikrofon melletti beszédfelismerés pontossága igen magas lehet – ameddig a háttérzaj szintje lényegesen alacsonyabb, mint a felismerendő beszédé. Amint ez megváltozik, a szófelismerési pontosság rohamosan csökkenni kezd. Ennek egyik alapvető oka az, hogy az alkalmazott jelfeldolgozás meg sem közelíti az emberi hallás lényegkiemelési képességeit. Ez a gépi beszédfelismerés szűk keresztmetszete, ezért ennek a területnek a kutatása kiemelten fontos. Az sem elhanyagolható tény, hogy az emberi hallás fizikája, fiziológiája sincsenek kellő mértékben feltérképezve. További nehézséget jelent, hogy az összetett pszichofizikai-matematikai modellek olyan nagy számításigényűek, hogy néhány évvel ezelőttig nem is volt reális esély a kivitelezésükre.

A projektben elsősorban az akusztikai előfeldolgozást vizsgálnánk. Ez az a modul, ahol véleményünk szerint biztosan szükség van előrelépésre, és ahol nagy valószínűséggel elérhető bizonyos javulás, ami, még ha nem is ugrásszerű, de ha általános érvényű, az világszintű érdeklődésre tarthat számot.

Várható eredmény, hatás, hasznosulás.

- A beszédfelismerési alkalmazások pontosságának, használhatóságának növekedése.
- A beszédfelismerésen alapuló technikák szélesebb körű elterjedése.

2.4.1.2. projektterv: Új paradigmák vizsgálata a zaj- és beszélőrobusztus gépi beszédfelismerés érdekében

Szinte bizonyosnak látszik, hogy a jelenlegi state-of-the-arttól különböző technikai megoldásokat kell alkalmazni ahhoz, hogy ugrásszerű javulást érjünk el a robusz-

tusság tekintetében. Valószínű, hogy a jelenleg viszonylag jól elkülönülő beszédfelismerési főmodulok – a lényegkiemelő és a mintaillesztő modul – szoros, esetleg teljes összefonódása fog megtörténni a jövőben. Ennek kutatása tehát reményteli feladat. Azonban ennél is jóval perspektivikusabban kell gondolkozni ahhoz, hogy megvalósulhasson a távlati cél, az emberi percepciós képességeket elérő vagy azt meghaladó gépi beszédfelismerés. Érdemes megfontolni az agykutatókkal történő együttműködést, hogy az emberi percepciós működés elvét jobban megértve tudjuk a gépi beszédészlelést fejleszteni. Felmerülhet a szemantika hibajavító képességének már alacsony felismerési szinten történő hasznosítása, melyre reménykeltő kezdeti próbálkozások történtek már a világban. Az új paradigmák vizsgálata a gépi beszédfelismerésben nehéz, nagy szakértelmet igénylő hosszabb távú feladat, azonban az átütő javulás kis esélyű bekövetkezte is akkora értéket képvisel, ami miatt indokolt ezen kutatások állami finanszírozása.

Várható eredmény, hatás, hasznosulás.

- A beszédfelismerési alkalmazások pontosságának, használhatóságának növekedése.
- A beszédfelismerésen alapuló technikák ugrásszerű elterjedése.
- A hazai kutatás fellendülése nemcsak a beszédtechnológia, de akár az akusztika, a fizika, a fiziológia és az agykutatás területén is.

2.4.2. kutatási irány: Skálázható kiejtésátíró szoftver és kiejtési szótárak fejlesztése

A gépi beszéd-előállítást sokan már megoldott problémának tekintik, ám az emberével teljesen összetéveszthető gépi beszéd szintetizálása még mindig távoli cél. Egyes szűkebb témakörökben és sok kézi munka árán élethű beszéd állítható elő, de a hibátlan témafüggetlen automatikus szöveg-beszéd átalakítás még utópia.

Helyzetkép. Kevés az olyan terület, ahol le lehet mondani az automatikus felolvasók folyamatos emberi tanításáról, támogatásáról. Tipikus probléma, hogy a bemenő szövegben mindig lehetnek olyan részek, amelyeknek a kiejtését még nem rögzítették elektronikusan: ezek a kivételes írásmódú és ejtésű szavak.

Ennek ellenére már ma is több magyar nyelvű beszédszintetizátor áll rendelkezésünkre, melyek nem külföldi rendszerek adaptációi, hiszen a magyar nyelv erősen ragozó jellege miatt ezek alkalmazása nem célravezető. Ám ezek a rendszerek mind valamilyen egy-egy szűkebb területre specifikáltak (pl. időjárásjelentés-, SMS- vagy menetrend-felolvasó).

Elérendő célok. A fentiekből következik, hogy a hibamentes automatikus szövegfelolvasás eléréséhez némi emberi támogatásra sokáig szükség lesz, ennek csökkentésére átfogó kutatásra van szükség. A korszerű megoldásokhoz itt is nagyméretű és több szinten pontosan annotált beszédadatbázisokat kell felépíteni.

A gépi szövegfelolvasás megítélésének három fő kritériuma van: helyes-e a kiejtés, helyes-e a hangsúlyozás, a beszéddallam és a ritmus, valamint hogy emberi hangszínezete van-e a szintetizátornak. Ezen kritériumrendszer első elemét érinti a korrekt hangátírás. Magyar nyelvre jelenleg még nem létezik olyan szoftver, amely tesztelt és minősített kiejtési átírást valósít meg, esetleg hangolható, skálázható.

A skálázható kiejtésátíró szoftver elsősorban a gépi beszéd-előállítás, illetve egyes fonológiai kutatások számára lehet hasznos. A nyomtatott formában már elérhető idegen vagy idegen eredetű szavak magyar kiejtésének elektronikus formában történő elérhetősége azonban az általános gépi beszédfelismerés szempontjából is nagy segítséget jelentene.

Kapcsolódó projektek.

- A **Profivox** beszédszintetizátor tekinthető az első olyan magyar nyelvű beszéd-előállító programnak, amely teljesíti a korszerű beszédszintetizátoroktól elvárható alapkövetelményeket. A program további alkalmazások hátteréül szolgál (pl. e-mail-, könyv-, SMS-felolvasó).
- Beszédadatbázis a magyar **beszédhang-kapcsolódások** szerkezeti bemutatására: minden hangkapcsolat egy mintaszón keresztül kerül bemutatásra. A mintaszavak szöveges és hangátírásos formáját, valamint hangidőtartamait is tartalmazza az adatbázis.

Projektterv.

1. Központi kiejtési adatbank létrehozása

2.4.2.1. projektterv: Központi kiejtési adatbank létrehozása

A skálázható kiejtésátíró szoftver létrehozásához az egyes szakmákat érintő szakszavak kiejtési szótárait kell elektronikus, egységes, szabványosított formában elkészíteni. Ezzel a munkával csak csökkenteni lehet a jövőbeni emberi támogatás nagyságát, azt teljesen kiküszöbölni nem lehet, mert mindig lesznek olyan szavak, kifejezések, amelyeknek a kiejtését legalább egy alkalommal meg kell határozni. Javasoljuk egy központi kiejtési adatbank létrehozását, ahonnan a jövő nyelv- és beszédtechnológiai rendszerei lekérdezhetik a szükséges adatokat. Az adatbank közösségi karbantartású lenne, mivel több kutatóműhely számára jelent fontos információt az idegen szavak és nevek kiejtési módja. Például a szövegekben található ragozott külföldi tulajdonnevek felismerésénél és elemzésénél is nagy segítséget nyújtana egy ilyen adatbank. Az adatbank folyamatos, napi szintű frissítésével a hírműsorok automatikus feliratozása is tovább pontosítható lenne.

Várható eredmény, hatás, hasznosulás.

- Javuló eredmények a gépi beszédszintézis terén.
- Pontosabb tulajdonnév-felismerés szövegekben és hangzó anyagokban egyaránt.

Projektterv		Típus	Résztvevők	Költségforrás	Időzítés	Költség
4.1.1.	Robusztus lényegkiemelő technikák vizsgálata gépi beszédfelismeréshez	alap	egyetem, kutatóintézet	100% pályázat	2011–2020	170M Ft
4.1.2.	Új paradigmák vizsgálata a zaj- és beszélőrobusztus gépi beszédfelismerés érdekében	alap, alkalmazott	egyetem, kutatóintézet	100% pályázat	2011–2020	210M Ft
4.2.1.	Központi kiejtési adatbank létrehozása	alkalmazott	egyetem, kutatóintézet	100% pályázat	2011–2012	30M Ft

2.4. táblázat. A 4. stratégiai cél projektterveinek összegzése.

2.5. Környezeti intelligenciával segített élet

A nyelv- és beszédtechnológia fejlesztéseit alkalmazó eszközök az esélyegyenlőség és az életminőség javításának is komoly elősegítői lehetnek. A fogyatékkal élők társadalmi integrációjának szempontjából az ember-gép kommunikáció megkönnyítése kulcsfontosságú — ezek egyrészt a mindennapi életüket könnyítik meg, másrészt olyan információhoz is hozzájuthatnak, amelyeket számukra primér módon nem hozzáférhető médiumokon keresztül közvetítenek.

A technológia eredményei a beszéd-, hallás- és nyelvkészség zavarainak diagnózisában és terápiájában is fontos szerepet játszanak, elsősorban a beszédalapú diagnosztika, a hallássérültek beszédterápiájának szoftveres támogatása és a rehabilitáció terén.

Külön kiemeljük a környezeti intelligenciával segített élet jelentős nyelvtechnológiai vonzatát, hiszen erre a lakosság elöregedésével egyre inkább szükség lesz.

Helyzetkép. Jelenleg a magyar nyelv- és beszédtechnológia azokat az alapkutatásokat végzi, amelyekre alapozva megszülethetnek majd azok a főleg beszédszintézisre és -felismerésre alapuló alkalmazások, amelyek a fogyatékkal élők számára is hozzáférhető médiumokra „fordítanak” információt. Számos beszédadatbázis áll már rendelkezésünkre, valamint többfajta feladatspecifikus beszédszintetizátor, ill. -felismerő, amelyek egyelőre nagyrészt témaspecifikus modulokkal rendelkeznek (időjárásjelentés-, menetrend-, név- és címfelolvasó). Az, hogy a számítógép bármilyen jellegű szöveget képes legyen az emberi szintet megközelítő minőségben felolvasni, illetve „megérteni”, egyelőre még távlati célnak tekintendő.

A hallássérültek beszédterápiájának szoftveres támogatásához a Platform már jelenleg is ad technológiát, de ennek további fejlesztéséhez a személyi állomány bővítésére lenne szükség.

Elérendő célok. Céljaink között szerepel olyan specifikus beszédtechnológiai alkalmazások fejlesztése, melyekkel akár személyre szabott támogatást nyújthatunk a fo-

gyatékkal élőknek, de legalábbis megkönnyíthetjük az életüket intelligens eszközökkel.

Ez a stratégiai cél több terület célkitűzéseivel is összehangolható: neuro- és pszicholingvisták, kognitív és alkalmazott pszichológusok, orvosok, fonetikusok, nyelv- és beszédtechnológusok együttműködésével az egészségügy olyan alapcéljait lehet megvalósítani, mint egyes betegségek korai diagnózisa és későbbi terápiája.

Kutatási irányok.

1. Nyelvalapú diagnosztika
2. Beszédterápiai kutatások
3. Fogyatékkal élők életének nyelvtechnológiai alkalmazásokkal való segítése

2.5.1. kutatási irány: Nyelvalapú diagnosztika

Az egészségügy alapcéljai olyan célok, melyekhez relatíve kicsi nyelvtechnológiai befektetéssel lényegesen közelebb lehet jutni, és ennek meglenne az az előnye is, hogy ezek a hazai technológiát organikus módon támogatnák. Magától értetődő, hogy a beszéd-, hallás- és nyelvkészség zavarainak diagnózisában és terápiájában a technológia eredményei fontos szerepet játszanak. Az egészségügyi alkalmazások egyik fő területe a beszédalapú diagnosztika, mellyel nemcsak a hangképzési rendellenességek vizsgálata és automatikus diagnosztizálása lehetséges, hanem a beszéd részletes vizsgálatával sok egyéb betegség is előrejelezhető (pl. az Alzheimer-kór). A kutatások célja a kóros eseteket tükröző ún. nyelvi markerek keresése beszédben vagy szövegben. A szöveges anyagok tartalomelemzésével kóros pszichológiai eltérések és nyelvi zavarok is tetten érhetők.

Helyzetkép. A beszédalapú diagnosztika célja alapvetően a hangképzési rendellenességek vizsgálata és automatikus diagnosztizálása. A beszéd rendszeres ellenőrzése főleg kisgyermekek esetében fontos, hogy minél előbb kiderüljenek azok a rejtett problémák, amelyek a normál, elvárható beszédhallás fejlődését bármilyen formában megakadályozzák. Az életkorspecifikus beszédészlelés és -megértés az iskolai tanulás alapja, ezek a működések pedig a megfelelő beszédhalláson alapszanak.

Az egészségügyi célú diagnosztikai rendszerek általános célú beszédfelismerő és -szintetizáló adaptálásával állíthatók elő. A beszéd részletes vizsgálatával a hangképzési rendellenességeken kívül sok egyéb betegség (pl. az Alzheimer-kór) is előrejelezhető.

A nyelv és a beszéd, valamint a lelki élet közötti összefüggések tárhatók fel az elbeszélő szövegek tartalmi elemzése során. A pszichológusok a nyelvtechnológia eredményeire támaszkodva az elmúlt években az emberi társas alkalmazkodás pszichológiai folyamataira vonatkozó, empirikusan ellenőrizhető tudományos ismeretek birtokába juthattak.

A nyelvtechnológia a pszicholingvisztikai kutatásokban is támogató szerephez jut. A tipikus fejlődésű gyerekek spontán megnyilatkozásainak anyaga egyrészt az általános

emberi kognitív folyamatokról ad információt, másrészt az atipikus nyelvi fejlődésű csoportok diagnózisában segít.

Elérendő célok. A beszédalapú diagnosztika elsőrendű célja a hangképzési rendellenességek automatikus feltárása, melynek segítségével elsősorban a gyermekek beszédészlelési problémáit lehet felismerni és ezáltal megakadályozni, hogy a fejlődésben lemaradjanak. A beszéd egyes paraméterei, az ún. nyelvi markerek alapján további betegségek is diagnosztizálhatóak. Célunk ezeknek a markereknek a feltérképezése, valamint az általános célú beszédfelismerő és -szintetizáló rendszerek adaptálása diagnosztikai célokra.

További célunk a pszichológusok, kognitív tudósok és nyelvtechnológusok több éves közös munkájának folytatása, és újabb kutatási eredmények felmutatása a lelki és nyelvi zavarok diagnosztizálása terén.

Kapcsolódó projektek.

- A **MONDOM-2000** beszédhallást vizsgáló szűrőkészülékkel a beszédhallás szűréses ellenőrzését lehet elvégezni. A szűrést speciális szerkezetű szintetizált szavak alkalmazása teszi lehetővé. A készüléket a Nikol Kkt. fejlesztette ki.
- A BME Kognitív Tudományi Tanszékén folyt egy kutatás egy OTKA-projekt keretében, amelynek a célja az **analogikus általánosítási folyamatok** feltérképezése volt a gyereknyelvben. Mivel a korpuszalapú vizsgálathoz nem volt elegendő magyar nyelvű gyereknyelvi adat, további hangkazetták kerültek begépelésre, de az eredmények így is azt mutatták, hogy az adatritkasági problémával kell szembenéznie annak, aki korpuszalapú vizsgálatokat akar folytatni magyar gyereknyelvi szövegeken.
- Az MTA Pszichológiai Kutatóintézete és a Moszkvai Orvosbiológiai Intézet jelenleg a magyar szakemberek által korábban kifejlesztett tartalomelemzési módszert alkalmazza orosz és angol nyelvre adaptálva a **Mars500** elnevezésű nemzetközi űranalóg szimulációs kísérletben.

Projekttervek.

1. Az Alzheimer-kór korai diagnosztizálása beszédtechnológiai fejlesztésekkel
2. Pszichodiagnosztikai vizsgálatok nyelvtechnológiai támogatása
3. Magyar gyereknyelvi korpusz építése
4. Beszédképző szervek zavarainak diagnosztizálása beszédtechnológiai fejlesztésekkel

2.5.1.1. projektterv: Az Alzheimer-kór korai diagnosztizálása beszédtechnológiai fejlesztésekkel

A lakosság életkorának a kitolódásával egyre nő a demenciák (kóros leépülési folyamatok), azon belül pedig az Alzheimer-kór előfordulási aránya, így egyre nagyobb

szükség lenne az ilyenfajta betegségek automatikus nyelvi szűrésére, mely az orvostechnikai eszközökhöz képest (CT, PET, MRI, fMRI) lényegesen költségkímélőbb a társadalomnak. Általánosan alkalmazott módszer a beszéd egyes paramétereinek (pl. a szünetek gyakoriságának és hosszának, az agrammatikus kifejezések számának) mérése, melyek a rövidtávú munkamemória teljesítményéről árulkodnak. Az Alzheimer-kór felismerésének nyelvi tünetei közé tartoznak a megnövekedett idejű hezitálások és a szótalálási problémák. A spontán beszéd fonetikai jellemzői közül a beszédtempó lassulása és a hezitációk megszaporodása markerként működhet a kór preklinikai szakaszában és korai diagnosztizálásában.

Egy általános célú beszédfelismerő rendszer adaptálásával, neurolingvisták, kognitív pszichológusok és beszédtechnológusok együttműködésével kifejleszthető lenne egy olyan rendszer, amellyel előrejelezhető az Alzheimer-kór és egyéb demenciák. A megfelelő nyelvi markerek feltárásával és alkalmazásával továbbá egyes afáziatípusok, a Parkinson-kór és a Huntington-kór is diagnosztizálható.

Várható eredmény, hatás, hasznosulás.

- Az Alzheimer-kór és egyéb demenciák korai diagnózisának EU-szinten is mérhető társadalmi és gazdasági haszna van.
- Az integrált kutatás különböző tudományterületek képviselőit hozza össze: neurolingvisták, kognitív pszichológusok, fonetikusok, beszédtechnológusok, informatikusok, orvosok.
- Viszonylag kis beszédtechnológiai ráfordítással lényegesen közelebb lehet jutni az egészségügy egyes alapcéljaihoz.

2.5.1.2. projektterv: Pszichodiagnosztikai vizsgálatok nyelvtechnológiai támogatása

A pszichodiagnosztika és a nyelvtechnológia integráns összefüggésének logikai alapja az, hogy az egyének és a csoportok pszichológiai állapotai és folyamatai (érzelmek, gondolkodásmód, szándékok stb.) nem csupán a fizikai, hanem a verbális viselkedésben is kódolódnak. E kódok nyelvi markerek formáját öltik. Ekképp az elektronikusan rögzített kommunikáció tartalomelemzése révén az egyének és a csoportok pszichológiai folyamatai diagnosztizálhatók, ezek időbeli változásai statisztikailag mérhetők és feltérképezhetők, tovább az egyes egyének és csoportok egymással kvantitatíven összehasonlíthatók.

A nyelvi markerek és a pszichológiai állapotok és folyamatok összefüggéseit mintegy hatvan éve kutatják. Klasszikus példa, hogy az egyes szám első személyű igék és névmások és a tagadószavak együttes túlzott használata a depresszió nyelvi tünete lehet, illetve a veszélyes küldetést teljesítő legénységek csoporton belüli konfliktusa, illetve a távoli irányító személyzettel való szembefordulása a kommunikációból előrejelezhető.

A projekt alapkutatási szakaszában a vizsgálni kívánt szó- és kifejezéskategóriák összeállítása, a keresőalgoritmusok létrehozása a cél. A projekt alkalmazott kutatási szakaszában következik a vizsgálni kívánt minta meghatározása, a szövegkorpusz

kialakítása és a tartalomelemzési tevékenység elvégzése. Mindkét szakaszhoz intézményi együttműködés szükséges – az elsőhöz akadémiai intézetek között, a másodikhoz oktatási, pszichiátriai, egészségügyi intézményekkel, illetve olyan hatóságokkal, amelyek célvezérelt kiscsoportok pszichológiai folyamatainak feltérképezését, előrejelzését igénylik.

Várható eredmény, hatás, hasznosulás.

- Az alap- és alkalmazott kutatások innovációs eredményei nemzetközi együttműködéseket is lehetővé tesznek.
- A magyar nyelvű mintákon végzett alapkutatások nemzetközi eredményekkel összemérhetőekké válnak.
- Az automatikus tartalomelemző módszerek részben kiválthatják a személyiség- és klinikai pszichológiában jelenleg használatos teszteket.
- Lehetővé válik a célvezérelt kiscsoportok távoli, automatikus pszichodinamikai monitorozása.

2.5.1.3. projektterv: Magyar gyereknyelvi korpusz építése

A magyarban az anyanyelv elsajátítási lépéseinek és a későbbi nyelvi fejlődésnek a részletes és számszerű adatokkal is alátámasztott leírása egyelőre hiányzik. Egy ilyen leíró munka, amely gyerekek spontán beszélgetéseinek az elemzésén és azok különböző életkorokon keresztüli követésén alapul, számos gyakorlati szempontból elengedhetetlen. A tipikus fejlődésű gyerekek spontán beszédmintáinak vizsgálata, a részletes nyelvi mutatók számolása és azok fejlődési változásainak felrajzolása alapvető fontosságú az atipikus nyelvi fejlődésű csoportok – a nyelvfejlődési zavarral, autizmussal, az értelmi sérülés különböző formáival (Down-szindróma, Williams-szindróma, magzati alkohol szindróma stb.) élő gyerekek – nyelvi diagnózisában és a megfelelő nyelvi fejlesztés kidolgozásában.

A nyelvelsajátítási vizsgálatokhoz elengedhetetlenül szükséges egy kellően nagy méretű, nemzetközi szabványokat követő gyereknyelvi korpusz megépítése. Az MTA Nyelvtudományi Intézete őrzi Réger Zita világhírű magyar pszicholingvista gyűjteményét, amely gyerekek spontán beszélgetéseinek hangfelvételeiből áll. A gyűjtemény megközelítőleg 1300 darab hang- és videókazettából, illetve jegyzetekből áll. A teljes gyűjtemény digitalizálásával és átírásával egy olyan gyereknyelvi korpusz építhető, amely kiindulási alapja lehet minden magyar gyereknyelvi kutatásnak. Az átírások a CHILDES nemzetközi gyereknyelvi korpusz formátumát követnék, ezáltal a magyar adatok a külföldi kutatók számára is elérhetőek lennének.

Várható eredmény, hatás, hasznosulás.

- A Réger Zita-gyűjtemény végre méltó módon lenne kezelve és feldolgozva.
- A korpusz megépítése nagyban előremozdítaná a magyar nyelvelsajátítási kutatásokat.

2.5.1.4. projektterv: Beszédképző szervek zavarainak diagnosztizálása beszédtechnológiai fejlesztésekkel

A beszédfeldolgozás új eredményei lehetőséget adnak új módszerek kidolgozására a hang kóros elváltozásainak mennyiségi becsléséhez és általában a hangminőség felméréséhez. Másfelől a klinikai orvosok által gyűjtött adatok lehetővé teszik a beszédjelek statisztikai akusztikai feldolgozását és új módszerek kidolgozását az automatikus hangdiagnosztizálás területén.

Hazánkban és külföldön is egyre nagyobb az igény egy, a beszédképző szervek zavarainak automatikus diagnosztizálására alkalmas rendszer iránt, amely segít a hang kóros elváltozásainak felderítésében. Az automatikus hangdiagnosztikai rendszer kifejlesztéséhez szükséges egy kórosan elváltozott folyamatos beszédet tartalmazó beszédadatbázis, amely gégészeti elváltozásokat reprezentál. Az adatbázist szakorvosokkal együttműködésben kívánjuk elkészíteni.

Várható eredmény, hatás, hasznosulás.

- Az elkészülő rendszer segíteni fogja a klinikai orvosok munkáját és a rák korai stádiumban való kimutatását.
- Különböző patológiai esetekkel foglalkozó akusztikai felderítéseink segíthetnek a beszéd fizikai és patopszichológiai szempontjainak megértésében.
- A beszédadatbázis nyilvános lesz, ezzel alapvető hanganyagot szolgáltatva e terület jövőbeni kutatásaihoz és fejlesztéseihez.

2.5.2. kutatási irány: Beszédterápiai kutatások

Az emberi kommunikáció legfontosabb csatornája, a társadalmi beilleszkedés fontos eszköze a helyes, érthető beszéd. Éppen ezért minden társadalomnak nagy súlyt kell fektetnie beszédsérült tagjainak beszédoktatására, a beszédterápiára. A beszédterápiai kutatások közé tartozik az artikulációs hibák korrekciója, a megkésett beszédfejlődés terápiája, a cochleáris implantátummal rendelkezők rehabilitációja, a fonológiai problémák javítása vagy az idegen akcentus csökkentése.

Helyzetkép. Hatékony beszédoktatás a beszédsérültek esetében ma már a beszédtechnológiai eszközök használata nélkül elképzelhetetlen. A számítógéppel segített jelenlegi beszédterápiás eszközök elsősorban az egyes beszédhangok, sziszegők, nazálisok, magánhangzók helyes kiejtésére koncentrálnak. Ezen eszközök korlátja az, hogy a beszéd szegmentális tulajdonságaira koncentrál, és nem vagy alig veszi figyelembe a beszédet mint egy nagyobb hosszantartó szupraszegmentális tulajdonságokkal rendelkező egységet. Holott a folyamatos, szépen tagolt érthető beszéd az, amit a beszédoktatás során ki kell alakítani.

Elérendő célok. Már rövid távon is kifejleszthetők olyan rendszerek, eszközök és terápiai eljárások, amelyek a mai elvárásoknak megfelelő hatékonysággal képesek a szép magyar tagolt, könnyen érthető folyamatos beszédet kialakítani.

Kapcsolódó projektek.

- A **VARÁZSDOBOZ** egy multiszenzoros beszédoktató rendszer, amely segítséget nyújt beszédhibás, valamint nagyothalló gyermekek és felnőttek helyes beszédképzésének kialakításában.
- Egy IKTA projekt keretében az SZTE kifejlesztett egy számítógéppel segített beszédjavítás-terápiára és olvasásfejlesztésre alkalmas eszközt, a **Beszédmestert**.

Projektterv.

1. A beszéd prozódiai jegyeit oktató és gyakorló audiovizuális rendszer kifejlesztése beszédhibás gyermekek részére

2.5.2.1. projektterv: A beszéd prozódiai jegyeit oktató és gyakorló audiovizuális rendszer kifejlesztése beszédhibás gyermekek részére

A prozódia adja a beszéd kifejező erejét, melyben a kiejtés sebessége, a ritmus, a hangerő, a hangszínezet és a hangsúlyozás a döntő tényezők. A különböző prozódiai jellemzők megfelelő használata nagyon fontos a beszédben, segítségükkel válik érthetővé a kommunikáció két fél között; ezáltal a helyes kialakításukra és tanításukra kiemelkedő figyelmet kell fordítani.

A kialakítandó oktatórendszer audiovizuális visszacsatolás alapján fog működni, ahol a beszéd dallama, a hangerő dinamikája, színképe, tempója stb. a képernyőn jelenik meg. Ezáltal a hallássérült vagy beszédhibás tanuló vizuális visszacsatolást kap beszédproduktumának milyenségéről a gátolt auditív visszacsatolás helyett.

Az önálló tanulást fogja elősegíteni a már kifejlesztett prozódiai felismerő beillesztése a rendszerbe, amely előzetes betanítás segítségével különbséget tud tenni a prozódiai szempontból helyes és helytelen kiejtésű mondatok között. Továbbá automatikus úton utasítást tud adni arról, hogy mit csináljon a használó ahhoz, hogy kiejtése jobb legyen.

Várható eredmény, hatás, hasznosulás.

- A rendszer önálló tanulásra, gyakorlásra ad lehetőséget.
- Egy már kifejlesztett beszédterápiás rendszert egészít ki új modullal.

2.5.3. kutatási irány: Fogyatékkal élők életének nyelvtechnológiai alkalmazásokkal való segítése

A fogyatékkal élő polgártársaink életminőségének meghatározó eleme, hogy mennyire tudnak fogyaték nélküli társaikhoz hasonló életet élni. A nyelv- és beszédtechnológia eszközrendszere a modern infokommunikációs technológiákkal kiegészítve számos esetben nyitja meg az aktív tanulás és munka terét számukra. Például a képernyőolvasó technológia általános elérhetősége a vakok és gyengénlátók számára megteremtette a számítógép használatának lehetőségét.

Attól függően, hogy a fogyaték mikor keletkezett, egészen eltérő megközelítés szükséges. Ezért nagyon fontos, hogy a különböző fogyatékcsoportokra megfelelő technológiai megoldásokat dolgozzunk ki.

Helyzetkép. A beszédszintézisre és -felismerésre alapuló technológiák mind a siketek és nagyothallók, mind a vakok és gyengénlátók számára megkönnyítik az információs társadalmi integrációt. A beszédtechnológusok és több civil kezdeményezésű szervezet közös munkája eredményeképpen az elmúlt években több ilyen eszköz is elkészült. Létezik SMS-, e-mail-, menetrend-, időjárásjelentés-felolvasó vakok számára; a siketek és hallássérültek számára pedig beszédterápiai szoftver (ld. 2.5.2.). Ezek nagy része a már meglévő technológia adaptációjával viszonylag egyszerűen előállítható, ám a fogyatékkal élők életét nagyban segíti.

Elérendő célok. Általánosságban elmondható, hogy a meglévő technológiából egy sor további eszköz lenne fejleszthető, amennyiben azok létrehozását cégek vagy állami intézmények támogatnák. Az Informatika a Látássérültekért Alapítvánnyal és a Nemzeti Építésügyi Technológiai Platformmal egyeztetve a következő projektötleteket tartjuk kivitelezhetőnek:

- helymeghatározási és útvonal-tanácsadási szolgáltatást nyújtó mobiltelefonos alkalmazás, akár részletes térinformatikai adatbázissal és integrált menetrendi információkkal helyi és távolsági közlekedés esetére is;
- intelligens információs terminál beszédfelismeréssel és -szintézissel;
- tananyag-felolvasó szoftver;
- akadálymentes multimédiás oktatószoftver;
- meglévő e-learning keretrendszerek akadálymentessé tétele;
- nemzeti ajánlás kidolgozása a vakok közlekedésének megkönnyítése érdekében.

Az alábbiakban részletesebben az Olvasó Telefon ötletét fejtjük ki.

Kapcsolódó projektek.

- Automatikus név-, cégnév- és postacím-felolvasás Magyarország egész területére. Jelenlegi egyik alkalmazása: **Automatikus szám szerinti tudakozó**.
- 2003-ban indult el a **VilágHalló** elnevezésű, magyar nyelvű hangos online elektronikus könyvtár. Már most több mint ezer magyar és világirodalmi művet hallgathatnak meg mesterségesen generált beszéd segítségével azok, akik letöltik a felolvasóprogramot. Nagy könnyebbség a vak és látáskorlátozott vagy idős, fáradó szemű embereknek az, hogy a VilágHalló Elektronikus Könyvtárat vizuális információ nélkül, a látókkal közel egyenértékűen használhatják.
- **Gyógyszervonal**: információs rendszer a gyógyszerekhez kiadott gyógyszertájékoztatók elérésének biztosítására tértől és időponttól függetlenül, a nap bármely szakában. Elérhető telefonon, weben és wapon. A hívást gép fogadja, a

beszédet gép ismeri fel, az információt gép olvassa fel. Működő szolgáltatás. (GVOP, 2004)

Projektterv.

1. Olvasó Telefon

2.5.3.1. projektterv: Olvasó Telefon

A látássérültek körében gyermek, felnőtt és időskori felhasználásra egyaránt alkalmazható mobilfelolvasó szoftver. A modern infokommunikációs technológiát használva, nem helyhezkötött, hanem mobiltelefonon futó optikai karakterfelismerő (OCR) alkalmazás, amely a lefotózott kép alapján szöveges állományt készít, majd a szövegfelolvasó segítségével felolvassa azt. Két változata lehetséges:

1. Helyben, a készüléken futó OCR. Előnye, hogy nincs szükség netkapcsolatra.
2. A lefotózott képet az interneten továbbítja egy szervernek. Ott a képet feldolgozza az OCR, majd a szöveget visszaküldi a mobiltelefonra, ahol a szövegfelolvasó segítségével a készülék felolvassa azt. Előnye a nagyobb felismerési pontosság.

Várható eredmény, hatás, hasznosulás.

- Viszonylag kis befektetéssel sokat lehet tenni a látássérültek akadálymentesítése terén.
- A már meglévő technológiák széles körű felhasználása.

2.6. Többnyelvűség, a nyelvi korlátok leküzdése

Az Európai Unió fontos elve az esélyegyenlőség biztosítása az Unión belül. Ennek fontos feltétele a nyelvek sokféleségének tisztelete és a nyelvi alapon történő megkülönböztetés tilalma. Ennek megfelelően az EU szándéka, hogy ösztönözze a nyelvtanulást, elősegítse a többnyelvű gazdaság fejlődését, és lehetővé tegye, hogy körülményeitől függetlenül valamennyi európai polgár élvezhesse az információs társadalom előnyeit, és saját nyelvén jusson hozzá az uniós információkhoz.

Helyzetkép. A Platform számos tagja foglalkozik gépi fordítással, és van érdeklődés a közvetlen beszédfordítás iránt is. Amint azt több példa is mutatja, a feladat nehéz, és irreális lenne arra számítani, hogy egy húszfős cég önerőből megoldja. Ezért itt is, mint a nyelvtechnológia számos területén, elengedhetetlenül szükséges a hagyományos kutatásfinanszírozási keretek átlépése. Hangsúlyozzuk, hogy az ország jövője, gazdasági versenyképessége szempontjából kardinális kérdésről van szó, olyanról, aminek megoldását nem várhatjuk a Google-tól, hiszen a magyar nyelv csak nekünk igazán fontos.

A legfontosabb, a gépi fordításhoz nélkülözhetetlen és az emberi fordítást is támogató eszközök a kétnyelvű szótárak. Bár a hagyományos szótárak digitalizálása terén

Projektterv		Típus	Résztvevők	Költségforrás	Időzítés	Költség
5.1.1.	Az Alzheimer-kór korai diagnosztizálása beszédtechnológiai fejlesztésekkel	alkalmazott	kutatóintézet, egyetem	100% pályázat	2011–2012	30M Ft
5.1.2.	Pszichodiagnosztikai vizsgálatok nyelvtechnológiai támogatása	alap, alkalmazott	kutatóintézet, egyetem, külföldi akadémiai intézet	100% pályázat	2011–2013	40M Ft
5.1.3.	Magyar gyereknyelvi korpusz építése	alap	kutatóintézet, egyetem	100% pályázat	2011–2013	40M Ft
5.1.4.	Beszédképző szervek zavarainak diagnosztizálása beszédtechnológiai fejlesztésekkel	alkalmazott	egyetem, kutatóintézet, kórház	80% pályázat, 20% önerő	2011–2012	20M Ft
5.2.1.	A beszéd prozódiai jegyeit oktató és gyakorló audiovizuális rendszer kifejlesztése beszédhibás gyermekek részére	alkalmazott	egyetem, kutatóintézet, kkv, iskola, kórház	80% pályázat, 20% önerő	2011–2013	30M Ft
5.3.1.	Olvasó Telefon	kísérleti fejlesztés	egyetem, multi, kkv	100% ipari támogatás	2011–2012	20M Ft

2.5. táblázat. Az 5. stratégiai cél projektterveinek összegzése.

már komoly a fejlődés, tudomásul kell venni, hogy ez a folyamat ebben a formában nem vezet, és a szerzői jogi korlátok miatt nem is vezethet a gépi fordítás elterjedéséhez. Szükség van olyan szabad felhasználású, nyílt forráskódú tartalmak és kereső-eszközök kifejlesztésére, amelyek az eddigieknél jóval erősebben formalizáltak, és támogatni kell az ilyenek automatikus építését.

Elérendő célok. Cél tehát, hogy bármely nyelven nyilvánosságra hozott közlemény az EU bármely polgára számára egyenlő eséllyel hozzáférhető és megérthető legyen, illetve a releváns tartalom belőle egyszerűen kinyerhető legyen. A nyelvtechnológiai kutatások egyik stratégiai célja éppen ez: a természetes nyelven megfogalmazott információ megértésének számítógépes támogatása. Ez a nyelvtechnológia számos területén megvalósulhat. Ide tartozik többek között a számítógéppel segített szótárkészítés, a többnyelvű információkinyerés és -visszakeresés, valamint az automatikus gépi fordítás.

Kutatási irányok.

1. Számítógépes lexikográfia
2. Többnyelvű információkinyerés és -visszakeresés
3. Automatikus gépi fordítás

2.6.1. kutatási irány: Számítógépes lexikográfia

Mindig szükség lesz a nyelvek változó szókincsét követő és bemutató újabb és újabb szótárakra. A számítógépes lexikográfia fő célja, hogy a modern információ- és nyelvtechnológiai eszközök és eljárások felhasználásával támogassa egy- és többnyelvű szótárak előállítását.

A magyarul beszélők száma egyértelműen meghatározza hazánk szótári piacát: egyrészt csak a nagy beszélőközösség által beszélt nyelvekre megtérülő befektetés szótárakat létrehozni, másrészt ezen szótárakat nem éri meg aktualizálni. Ez a gyakorlat azonban azt eredményezi, hogy – megfelelő szótár hiányában – egy magyar beszélő a kevéssé használt nyelvek nagy részét csak egy nagyobb közvetítő nyelven keresztül tudja megtanulni. Magyarország földrajzi és nyelvi adottságait figyelembe véve időszerű lenne naprakész és reprezentatív kétnyelvű szótárakat létrehozni legalább a régió nyelveire, hiszen ezekre feltétlenül szükség van, hogy a két külön nyelvvel rendelkező beszédközösség kapcsolata erősödjön, így segítve elő a kereskedelmet, a turizmust és a külföldi munkavállalást. A fenti igényekkel összhangban a számítógépes lexikográfia lehetővé teszi a szótárak költséghatékonyabb és gyorsabb előállítását, valamint folyamatos fejlesztését és aktualizálását.

Helyzetkép. A lexikográfiában a számítógép szerepe kezdetben arra korlátozódott, hogy a szótárakat elektronikus formában is kiadták, ami a papírváltozat hasonmásának volt tekinthető. Aztán az elektronikus forma elszakadt a papír által diktált rendszertől, a számítógép a szótári anyag strukturálásában kapott szerepet, kialakultak az ún.

„intelligens szótárak”. Ezekben például nincs szükség utaló szócikkekre; az automatikus morfológiai elemzés segítségével a szavak toldalékolt formájukban is kereshetők; a többtagú kifejezések bármely tagjuknál megtalálhatók; és a szótárak „kifordításával” mindinkább eltűnik a forrás- és a célnyelv közötti különbség is. Manapság pedig a kutatás főárama arra irányul, hogy hogyan tudjuk a szótárkészítés bizonyos lépéseit automatizálni különféle nyelvtechnológiai erőforrások és eszközök segítségével, azaz hogyan tudjuk magát a szótár anyagát automatikusan vagy félig automatikusan előállítani.

Az a tény, hogy hazánk nem „szótárkészítő nagyhatalom”, nemcsak a piacra kerülő szótárak számát, hanem a minőségét is veszélyeztetheti, hiszen egyfelől állami ráfordítás nélkül csak a világnyelveket lefedő szótárakba éri meg invesztálni, másfelől – metalexikográfiai kutatások és megfelelő lexikográfiai infrastruktúra híján – az elkészült szótárak minősége sok esetben nem éri el a világszínvonalat. A szótárak minősége nagyban függ a szótárkészítés módszertanától, a kialakítandó szótár szerkezetétől és kereshetőségétől. A fejlett lexikográfiai hagyományokkal rendelkező országokban ma már általánosan bevett gyakorlat, hogy a lexikográfiai munka során egy referenciakorpuszból indulnak ki, és a szótárban rögzítendő jelenségeket nagy mennyiségű nyelvi adatból nyerik ki nyelvtechnológiai eszközökkel. Ezen lexikai adatbázisok jellemzője, hogy bennük nem a szó az alapegység, hanem a jelentéses nyelvi elem, mely egyes esetekben több szóból áll. A szótárkészítés egyik mai trendje a többszavas kifejezések központi szerepének felismerése és kezelése.

További probléma, hogy Magyarországon a nemzetközi összehasonlításban is meglehetősen kevés lexikográfiai munka teljesen decentralizáltan zajlik. Ennek két káros hatása is van. Egyfelől a szakmai tapasztalatok áramlása nem megoldott, így fordulhat elő, hogy szerényebb lexikográfiai tartalmú, de professzionális technológiákra épülő elektronikus eszközök állnak szemben gyengébb szoftvermegoldásokon alapuló, ám professzionális szótártartalmakkal. Másfelől, az egymástól elszigetelt szótárkészítési munkálatokban ugyanazon feladatok újra és újra elvégzésre kerülnek. A már elvégzett munkák újrahasznosíthatóságát elvileg is lehetetlenné teszi, hogy a szótárkészítés módszertana is változik projektről projektre. A fentiekből következik, hogy a szükséges ráfordítás már középtávon is jelentősen csökkenthető egy szabadon hozzáférhető és újrahasznosítható lexikográfiai infrastruktúra kialakításával.

Elérendő célok. Általános szándék, hogy a nemzetközi trendekhez kapcsolódjunk az automatizálás és a többszavas kifejezések középpontba állítása terén is, és modern eszközökkel jó minőségű szótárakat hozzunk létre.

Célunk egy olyan lexikográfiai infrastruktúra és szótárkészítési metodológia kialakítása, amely magas színvonalú kétnyelvű szótárak költséghatékony előállítását teszi lehetővé. A modern szótárak egy reprezentatív nyelvi adatbázisból kiindulva, emberi, lexikográfusi munkát alkalmazva (korpuszalapú módszertan), vagy a lexikográfiailag releváns információ szövegekből való kinyerésére képes nyelvtechnológiai algoritmusok használatával (korpuszvezérelt módszertan) készülnek.

Ezen metódusok elterjesztéséhez mindenekelőtt egy Magyar Lexikográfiai Referen-

cia-adatbázis létrehozása szükséges. Ez az infrastruktúra központi eleme, amely lexikográfiailag releváns nyelvi adatoknak lexikográfiai, nyelvészeti és informatikai szempontból is egységes, nyelvtechnológiai eszközökkel jól kezelhető adatbázisa. Rövidtávú cél továbbá olyan nyelvfüggetlen, korpuszalapú, automatikus szótárépítő eljárások kidolgozása, melyek segítségével dinamikusan készíthetünk szótárt, legyen az speciális szaknyelvi vagy többnyelvű szótár. Szintén rövidtávú cél, hogy a hivatásos fordítók igényét kielégítve létrehozzuk a Magyar Kollokációs Szótárat, amely a többszavas kifejezéseket térképezi fel, a magyar szavak szokásos kapcsolódási viszonyait mutatja be.

Kapcsolódó projektek.

- A Nemzeti Nyelvi Intézetek Európai Szövetsége (EFNIL) támogatásával folyó projekt (**EFNILEX**) célja, hogy korpuszvezérelt módszerek adaptációjával a szótárkészítési folyamatot idő- és költséghatékonyabbá tegye. Egy ilyen megközelítésnek különösen az olyan nyelvpárok esetében van kiemelt jelentősége, ahol a hagyományos nagy emberimunka-ráfordítást igénylő szótárkészítési eljárás üzletileg nem kifizetődő. Tipikusan ez a helyzet a kevésbé használt uniós nyelveknél, így ez a projekt összhangban van az európai nyelveket egyenrangúnak tekintő EU-s alapelvvel.
- Az MTA Nyelvtudományi Intézetében készül az **Igei szerkezetek összehasonlító gyakorisági szótára**. Ennek az egynyelvű szótárnak az anyaga korpuszvezérelt módon áll elő: egy speciális algoritmus a korpusz alapján közvetlenül a nyers szócikkeket állítja elő, melyek utána emberi munkával szerkeszthetők.

Projekttervek.

1. Magyar Lexikográfiai Referencia-adatbázis létrehozása
2. Magyar Kollokációs Szótár létrehozása
3. Általános szótárkészítő eszközkészlet kialakítása

2.6.1.1. projektterv: Magyar Lexikográfiai Referencia-adatbázis létrehozása

A Magyar Lexikográfiai Referencia-adatbázis (MLRA) egy olyan lexikográfiai infrastruktúra része, amely lehetővé teszi, hogy magas színvonalú szótárakat költséghatékonyan állítsunk elő. Mivel egy ilyen erőforrás létrehozása költséges, kialakításánál fontos szempont, hogy más jellegű nyelvtechnológiai célokra is felhasználható legyen (pl. információkinyerés, jelentésegyértelműsítés). Az MLRA céljaink szerint különböző nyelvű és méretű szótárak kiindulópontjául szolgálhatna, ezáltal csökkentve egy lexikográfiai projekt költségeit. Hogy ezt a szerepét betölthesse, az adatbázisnak formailag és tartalmilag is konzisztensnek kell lennie. Az adatbázis formai konzisztenciáját a nemzetközi szabványoknak való megfelelés, tartalmi konzisztenciáját pedig a munka hátteréül szolgáló nyelvtechnológiai fejlesztések kiaknázása biztosítja.

A tervezett adatbázis a magyar nyelv 45 ezer jelentéses egységét tartalmazná, amely egy középszintű szótár méretének felel meg. A jelentéses egységek a Magyar

Nemzeti Szövegtárban található nagy mennyiségű nyelvi adat lexikográfiai elemzése eredményeképpen kerülnek az adatbázisba, így garantálva, hogy az MLRA a magyar nyelvnek egy reprezentatív szelete legyen. Fontos szempont, hogy az egyes jelentések felvételét az egyes nyelvi elemek meghatározott módszertan szerinti kimerítő nyelvészeti elemzése előzi meg, amely minimalizálja az emberi intuíció szerepét a jelentéses egységek felvétele során. Az alkalmazni kívánt módszer lényege, hogy a szavak, illetve többszavas kifejezések egyes jelentéseit azon kontextusok alapján különböztetjük meg, amelyekben előfordulhatnak. Az összes releváns kontextus figyelembevételét egy olyan szoftverrel kívánjuk biztosítani, amelyben a megfelelő szlotok már előre definiálva vannak, a lexikográfusok feladata ezen előre definiált helyek kitöltése nyelvi adatokkal. Az MLRA előnye a hagyományos egynyelvű szótárakkal szemben, hogy korpuszalapú, ami biztosítja a mai szókincs megfelelő lefedettségét, másfelől a nyelvi elemzés garantálja, hogy a megfelelő jelentések szerepeljenek az adatbázisban.

Az adatbázis fenti két tulajdonsága lehetővé teszi, hogy az MLRA bármilyen célnyelvhez forrásnyelvi kiindulópontként szolgáljon, hiszen célnyelvtől függetlenül lett kidolgozva, másrészt hogy a fontos célnyelvi megfelelőket is tartalmazza, abban az esetben, ha a magyar a célnyelv. Az adatbázis további előnye, hogy a szerepeltetendő gyakorisági adatok alapján kisszótárak kiindulásául is szolgálhat. Például egy 15 ezer bejegyzést tartalmazó szótár esetében elegendő a leggyakoribb 15 ezer jelentéses egységet figyelembe venni.

Egy egységes korpuszalapú szótárírói módszertan kidolgozása fontos eleme az újrahasznosíthatóságnak. Az egységes módszertan kidolgozásába beletartozik az adatbázis mikro- és makrostruktúrájának meghatározása, a jelentéses egységek elkülönítéséhez szükséges nyelvi információk körének meghatározása, a számítógépes háttér biztosítása, a többszavas kifejezések automatikus kezelésének megoldása.

Az MLRA kialakításánál már meglévő nyelvtechnológiai fejlesztésekre kívánunk támaszkodni. A fentebb leírt korpuszalapú és korpuszvezérelt módszerek bármelyikét is használjuk, a rendkívül időigényes szótárírói munka automatizálásával az emberi munkaerőt váltjuk ki, amennyire lehetséges.

Várható eredmény, hatás, hasznosulás.

- Nagy potenciál a jövőbeni szótárak elkészítésében, minőségi javulás, új nyelvpárokra gyorsabb megtérülés.
- Nagyobb ismertség. A más országokban is használt sztenderdek használata lehetővé teszi, hogy kooperáljunk más országok lexikográfusaival.
- Újrafelhasználhatóság más projektekben, több innovációs eredmény.

2.6.1.2. projektterv: Magyar Kollokációs Szótár létrehozása

A kollokációs szótár a szavak szokásos kapcsolódási viszonyait mutatja be, ezáltal képet ad a nyelv többszavas kifejezéseiről, melyek fontos, jellegzetes és gyakori elemei a

nyelvnek. Fordítás során csak a többszavas kifejezések megfelelő használatával biztosítható az adott nyelvre jellemző kifejezésmód, mely nemcsak nyelvtanilag helyes, hanem megszokott és gördülékeny is.

Jelenleg nem létezik modern magyar kollokációs szótár. Egy ilyen szótár nagyon hasznos segédeszköz lenne általános és szakfordítók, valamint a magyart mint idegen nyelvet oktatók számára is: olyan esetekben segíthetne megtalálni a legmegfelelőbb megfogalmazást, ahol az anyanyelvi beszélő nyelvérzéke is ingadozik.

A Magyar Kollokációs Szótár modern korpuszalapú és/vagy korpuszvezérelt félautomatikus módszertannal készülne. Anyaga nem csak konkrét szavak kombinációira terjedne ki, hanem magában foglalná az igei vonzatkereteket és a szavak és a különböző nyelvtani szerkezetek (pl. tagadás, igeidő) között meglévő kapcsolatokat is.

A kiadás elsődleges médiuma az elektronikus forma lenne. A online szótári szolgáltatás folyamatos fenntartásához a projekt lezárulta után szükséges kis mértékű finanszírozás, ezzel biztosítható lenne az általános, ingyenes hozzáférés. Mivel az alkalmazott technológia nagyrészt nyelvfüggetlen, lehetőséget adna arra, hogy a jövőben egyéb nyelvekre is hasonló szótárak készüljenek magyar szakmai műhelyekben.

Várható eredmény, hatás, hasznosulás.

- Egy hasznos fordításstámogató termék.
- Jobb minőségű fordítások.
- Potenciál hasonló jellegű, idegen nyelvű szótárak létrehozására.

2.6.1.3. projektterv: Általános szótárkészítő eszközkészlet kialakítása

A nyelvtechnológiai eszközök használata jelentősen leegyszerűsíti a szótárkészítési folyamatot, aminek elsősorban az olyan kevéssé használt nyelvek esetében van jelentősége, mint amilyen a magyar is. Ám teljesen automatizált módszer nem létezik, egy megfelelő lefedettségű és pontosságú szótár előállítása mindenképpen igényel emberi utószerkesztési munkálatokat. Jelen projektterv célja a lexikográfusok számára olyan erőforrások teljesen automatikus előállítása, amelyek a lehető legjobban csökkentik a szótárak elkészítéséhez szükséges munkát.

A módszer alapját szövegek forrás- és célnyelvi fordításainak mondatszintű megfeleltetéseit tartalmazó ún. párhuzamos korpuszokon végzett automatikus szóillesztés képezi. Bár az automatikus szóillesztést széles körben használják elsősorban a gépi fordítás területén, ezt a megközelítést eddig csak hazánkban használták lexikográfiai projektekben emberi felhasználásra szánt szótárak készítésének támogatására. A javasolt módszer egyik előnye, hogy biztosítja, hogy a legfontosabb fordítások benne legyenek a szótárban. Ezenfelül az automatikusan generált fordítási valószínűségek lehetővé teszik, hogy a fordítási jelölteket rangsoroljuk a szótárban, és így a leggyakrabban előforduló fordítások szerepeljenek elöl egy szótári bejegyzésen belül. További előny, hogy a releváns példamondatok is rendelkezésre állnak, így mind a lexikográfusok, mind a szótárhasználók valós nyelvi adatok alapján alkothatnak képet arról, hogy milyen szövegkörnyezetben használható adekvátan egy-egy fordítás. Mindezen

tulajdonságok miatt a javasolt módszer különösen alkalmas aktív (az idegen nyelvű szövegalkotást segítő) szótárak előállításának magas szintű támogatására.

Fontos előfeltétel, hogy a szóban forgó nyelvpárokra nagyméretű párhuzamos korpuszok álljanak rendelkezésre. Ezen korpuszok tartalmának függvényében a javasolt módszer alkalmas lehet terminológiai protoszótárak előállítására is, melyekből lexikográfusi utószerkesztéssel jogi, informatikai vagy építészeti szakszótárak is előállíthatók.

A legnagyobb nehézséget a javasolt módszer számára a párhuzamos korpuszok csekély mennyisége jelenti a kevéssé használt nyelvekre. Ezért a projekt egyik célkitűzése, hogy releváns mennyiségű párhuzamos korpuszt építsen a kijelölt nyelvpárokra. Az elégtelen mennyiségű bemenő adat egy másik megoldása lehet, hogy párhuzamos korpuszok helyett összehasonlítható korpuszokkal dolgozunk, amelyek esetében a forrás- és célnyelvi szövegek nem pontos fordításai egymásnak, de a szövegek témája ugyanaz.

Várható eredmény, hatás, hasznosulás.

- Nagy potenciál a jövőbeni szótárak elkészítésében, minőségi javulás, új nyelvpárokra gyorsabb megtérülés.
- Új szakterületekre is könnyen adaptálható a módszer, így elősegíti a terminológiai szótárak készítését is.
- Bekapcsolódás a nemzetközi lexikográfiai gyakorlatba és kutatási trendekbe.

2.6.2. kutatási irány: Többnyelvű információkinyerés és -visszakeresés

Bár a magyar web más nyelvekkel összehasonlítva is dinamikusan fejlődik, a gyakorlatban nagyon sokszor okoz problémát, hogy a felhasználó számára releváns dokumentum csak valamely más nyelven áll rendelkezésre. A gépi fordítási technológiák fejlődésével ezek az információk akkor is elérhetővé válnak a felhasználó számára, ha nem beszéli megfelelő szinten a kérdéses célnyelvet. Tehát amellett az evidens használati eset mellett, amikor a felhasználó megérti egy másik célnyelv dokumentumait, és a csak egyetlen nyelven megfogalmazott lekérdezésére egy többnyelvű találati listát szeretne, lehetőség van a felhasználó nyelvi képzettségén túlmutató alkalmazásokra is az idegen nyelvű találatok lefordításával.

Helyzetkép. Az utóbbi években a nemzetközi nyelvtechnológia egyre inkább megoldhatónak tekint olyan komplex kihívásokat is, amelyek több nyelvtechnológiai alkalmazás – jelen esetben az információkinyerés és -visszakeresés, valamint a gépi fordítás – ötvözését igénylik. Ennek ellenére mind Magyarországon, mind nemzetközi szinten a jelenlegi nyelvtechnológiai alkalmazások többsége az egy nyelvre fejlesztett eredményeket aknázza ki.

Elérendő célok. Célunk az információkinyerés és -visszakeresés területén a lehető legszélesebb felhasználási területeken alkalmazható rendszerek fejlesztése, amelyek

nemcsak a magyar dokumentumok tartalmában végeznek keresést, de az idegen nyelven elérhető találatokat a felhasználó számára érthető módon visszadják. Ezek a komplex rendszerek a hazai nyelvtechnológia legaktuálisabb fejlesztéseinek eredményeit fogják ötvözni, megkönnyítve a felhasználó számára a kívánt információ elérését, esetleges nyelvi akadályok ellenére.

Kapcsolódó projektek.

- **CACAO** (Cross-Language Access to Catalogues and On-line Libraries) (2007-2009): az EU által támogatott projekt célja egy olyan rendszer létrehozása volt, mely lehetőséget ad arra, hogy a felhasználók a saját nyelvükön megadott lekérdezéseikre tetszőleges nyelven rendelkezésre álló releváns dokumentumokat kapjanak eredményül különböző európai könyvtári adatbázisokból.

Projektterv.

1. Nyelvközi információ-visszakeresés

2.6.2.1. projektterv: Nyelvközi információ-visszakeresés

Ebben a projektben magyar nyelvre megvalósított nyelvközi információ-visszakereső megoldások kutatását és fejlesztését tervezzük, egy többnyelvű kereső prototípusának fejlesztésével. Ha a felhasználó nem beszél más nyelvet a saját anyanyelvén kívül, de a számára fontos információ csak más nyelven érhető el, szüksége lehet olyan alkalmazásra, amely a találati listában szereplő dokumentumokat lefordítja. Erre a jelenleg rendelkezésre álló gépi fordítási megoldások már megfelelő megoldást biztosítanak (szövegmegértéshez elégséges minőségű fordítást adnak). De lehetőség van az idegen nyelvű találatok előfeldolgozására és strukturált megjelenítésére is: ebben az esetben fordításra sincs szükség. Például ha a felhasználó egy nemzetközi termékkel kapcsolatos felhasználói véleményekre kíváncsi, joggal feltételezhető, hogy a weben fellelhető releváns tartalmak túlnyomó része nem magyar nyelven érhető el. Ezen információk hasznosítása korábban elképzelhetetlen távlatot nyit a webes információ-visszakeresés terén, és ez esetben, mivel a felhasználó csak a különböző forrásokból begyűjtött, aggregált eredményeket kapja meg (összesített vélemény vagy válasz a felhasználói kérdésre), nincs szükség a találatok visszafordítására.

Ezen technológiák alapvető fontosságú lépése a felhasználói lekérdezés idegen nyelvre történő leképezése, mely (megfelelő kontextus hiányában) sokszor nehezebb a kulcsszavak idegen nyelvre történő lefordításánál. A nemzetközi eredmények azt mutatják, hogy a nyelvközi információ-visszakeresés hatékonysága a legtöbb esetben eléri vagy meghaladja az egynyelvű keresés hatékonyságának 75%-át. Ez, valamint az idegen nyelven elérhető információ mennyisége széles körű alkalmazási lehetőségeket kínál.

Várható eredmény, hatás, hasznosulás.

- Új, innovatív eredmények az információ-visszakeresésben.
- Széles körű alkalmazási lehetőségek.

2.6.3. kutatási irány: Automatikus gépi fordítás

Az automatikus gépi fordítás célja, hogy akár nagy mennyiségű szövegnek egy gombnyomásra emberi vagy azt megközelítő minőségű fordítását állítsa elő. Az információáramlás fontossága miatt nem véletlen, hogy ma ez az egyik vezető irányzat a nyelvtechnológián belül. Számos nagy IT-vállalat rendelkezik saját fordítóval (pl. Google, Microsoft), de számos olyan cég is van világszerte, amelyeknek fő tevékenységei közé tartozik az automatikus gépi fordítórendszerek fejlesztése (pl. Systran, Linguatec).

Helyzetkép. Módszertanilag a gépi fordítóknak két nagy csoportját különíthetjük el. A szabályalapú fordítók nyelvtani szabályok szerint elemzik a forrásnyelvi mondatokat és célnyelvi szabályokra képezik le ezeket, míg a statisztikai fordítók nem készítenek elemzést a mondatokhoz, hanem a szövegben megtalálható gyakori szókombinációkhoz keresnek célnyelvi megfelelőket valószínűségi alapon. A szabályalapú és a statisztikai fordítóknak egyaránt vannak erősségeik és gyengeségeik. A szabályalapú fordítók egyik előnye, hogy a ritkán előforduló nyelvi szerkezeteket, szókapcsolatokat is képesek helyesen lefordítani. Ezzel szemben egy statisztikai fordító elkészítése kevesebb emberi munkát, nyelvészeti szaktudást igényel, ezért a fejlesztése jóval költséghatékonyabb.

A magyar nyelv szempontjából nem kedvező, hogy a statisztikai fordítók jellemzően azon nyelvpárokra működnek jól, amelyekre sok bemeneti adat áll rendelkezésre, a nyelvekben előforduló szavaknak csak viszonylag kevés alakjuk van, illetve a forrás- és a célnyelv hasonlítanak egymásra (akár szerkezetileg, akár szókincsben). Ezekből az következik, hogy a magyarra a közeljövőben vélhetően nem fog jól működő statisztikai gépi fordító születni.

Elérendő célok. Elsődleges célunk a már meglévő szabályalapú magyar-angol, illetve angol-magyar gépi fordító minőségének javítása. A szabályalapú gépi fordítóknak két alapvető problémájuk lehet: vagy nem értik a szöveget, vagy túl sokféleképpen értik, és rosszul választanak. Ezekre a problémákra kínálunk megoldást az alábbi projekttervekben.

Kapcsolódó projektek.

- **Magyar-angol gépi fordító rendszer**. Konzorciumi tagok: MTA Nyelvtudományi Intézet, MorphoLogic Kft., SZTE Informatikai Tanszékcsoport. (NKFP, 2004-2007)
- **iTranslate4 – Internet Translators for all European Languages**: a projekt célja, hogy segítse az európai nyelvek közötti fordítást. Ehhez egyfelől a már létező gépi fordítók közvetítő nyelveken keresztüli összekötése szükséges, másfelől az egyes fordítórendszerek részletes kiértékelése, hogy a lehetséges fordítási alternatívák közül a legjobbat választhassuk. (ICT PSP, 2010-2012)

Projekttervek.

1. Nyelvi adatbázis gépi fordításhoz

2. Jelentésegyértelműsítés a gépi fordításban

2.6.3.1. projektterv: Nyelvi adatbázis gépi fordításhoz

A szabályalapú gépi fordítás nagyméretű nyelvi adatbázisok segítségével dolgozik. Ezek az adatbázisok nyelvpárokra készülnek, minden nyelvpárhoz kettő: az oda és a vissza irányú fordításhoz. Ilyen adatbázispár magyar nyelvre jelenleg csak az angol vonatkozásában létezik. A nyelvi adatbázisokat jellemzően a forrásnyelv határozza meg, de nem teljesen függetlenek a célnyelvtől sem. Vagyis amennyiben egyszer elkészül egy jó magyar elemző például az angol fordításhoz, akkor az átdolgozással, de hasznosítható például a német nyelvű fordításhoz is.

A jelenlegi magyar-angol gépi fordító elemzőjében működtetett adatbázis 150 ezer szabályt tartalmaz. A szabályok az egyes szófaji kategóriákra jellemző nyelvi jelenségeket írják le. A magyar-angol adatbázis fejlesztése szófajonként zajlott, így történhetett meg, hogy a magyar határozói kifejezéseket tartalmazó lexikon mind a mai napig nem készült el.

Angol-magyar irányban a szabályok és szótárak ilyen alapvető hiányosságot nem mutatnak, itt viszont probléma az egyes szabályok kidolgozatlansága. Ez abban nyilvánul meg, hogy a szabályok nem veszik figyelembe azokat a speciális helyzeteket, amikor őket mégsem, vagy mégsem úgy kellene alkalmazni, ahogyan azt ők diktálják. A konkrét munkát két tényező határozná meg. Egyrészt az elméleti oldal, amely tételesen leírja a magyar grammatikában kezelendő jelenségeket, másrészt a gyakorlat, amely szövegvizsgálati módszerek alapján kijelöli az aktuálisan legtöbb megértésbeli problémát okozó nyelvi jelenséget.

Várható eredmény, hatás, hasznosulás.

- Közvetlenül a gépi fordítás minőségének javítása.
- Bármilyen magyar, illetve angol nyelv elemzésén alapuló alkalmazás támogatása.

2.6.3.2. projektterv: Jelentésegyértelműsítés a gépi fordításban

A jó minőségű gépi fordítás előállításához elengedhetetlen annak beazonosítása, hogy a forrásnyelvi szövegben előforduló többjelentésű szavak melyik jelentésükkel szerepelnek. A fordítás szempontjából például nem mindegy, hogy a *nap* szó egy adott szövegrészletben az égitestre vagy az időegységre utal-e. Az ehhez hasonló kérdések megválaszolása a jelentésegyértelműsítés feladata. A jelentésegyértelműsítéshez szükség van egy olyan jelentéstárra, amely tartalmazza a releváns jelentéseket és a megfelelő kontextuális információkat, másfelől pedig egy olyan algoritmusra, amely lehetővé teszi, hogy a mondatban szereplő szavakhoz, kifejezésekhez a jelentéstárban található jelentések közül a megfelelőt rendeljük.

A jelentésegyértelműsítésben a legfőbb problémát a megfelelő jelentéstár kialakítása okozza. Ennek egyik oka, hogy a jelentés fogalma nem egyértelműen meghatározott a nyelvtudományban, ezért a készítők elsősorban saját intuíciójukra vannak

utalva. Ám a meglévő jelentéstárak alapján a jelentésegyértelműsítés feladata még az emberek számára is nehéz: az eddigi nemzetközi és hazai kísérletek azt mutatják, hogy a hagyományos lexikográfiai módszerekkel készülő jelentéstárak jelentéseinek bizonyos szövegkörnyezetben előforduló szavakhoz való hozzárendelése csak igen alacsony egyetértést mutat a kísérleti személyek között.

Egy konzisztens adatbázis létrehozása tehát egy megfelelő módszertan kidolgozása nélkül lehetetlen. A jelentéstár építése során két megközelítést ötvöznénk: egyrészt korpuszalapú vizsgálatokat folytatnánk, másrészt gépi tanuló algoritmusokat alkalmaznánk. Vélhetően a két módszer kombinációja fogja a megfelelő eredményt nyújtani.

Várható eredmény, hatás, hasznosulás.

- Közvetlenül a gépi fordítás minőségének javítása.
- Közvetetten az információáramlás elősegítése.
- Minden olyan alkalmazás számára is hasznos fejlesztés, amely a szövegek értelmezését célozza meg.

Projektterv		Típus	Résztvevők	Költségforrás	Időzítés	Költség
6.1.1.	Magyar Lexikográfiai Referencia-adatbázis létrehozása	alap, alkalmazott	kutatóintézet, egyetem, kkv	100% pályázat	2011–2015	190M Ft
6.1.2.	Magyar Kollokációs Szótár léterhozása	alap, alkalmazott	kutatóintézet, egyetem, kkv	100% pályázat	2012–2014	40M Ft
6.1.3.	Általános szótárkészítő eszközkészlet kialakítása	alkalmazott	kutatóintézet, egyetem, kkv	100% pályázat	2011–2013	40M Ft
6.2.1.	Nyelvközi információ-visszakeresés	alkalmazott	egyetem, kkv, multi	100% ipari támogatás	2011–2014	80M Ft
6.3.1.	Nyelvi adatbázis gépi fordításhoz	alkalmazott	kutatóintézet, kkv	100% pályázat	2012–2013	70M Ft
6.3.2.	Jelentésegyértelműsítés a gépi fordításban	alap, alkalmazott	kutatóintézet, egyetem, kkv	100% pályázat	2011–2015	60M Ft

2.6. táblázat. A 6. stratégiai cél projektterveinek összegzése.

2.7. Kutatásszervezés

A kutatás-fejlesztési eredmények gyakorlati alkalmazását, az ezen tevékenységekből származó tudás hozzáférhetőségét és felhasználhatóságát, a kutatás és az ipar közötti hatékony kommunikációt, az oktatást segítő tevékenységek valamennyi kutatási

területet érintik. A fejezet tartalmazza azokat az általános programokat, amelyek eredményei kedvezőbb feltételeket teremtenek a hazai kutatás-fejlesztéshez, a kutatási eredmények gyakorlati hasznosításához, növelik a kutatói és az üzleti szféra közötti együttműködés hatékonyságát, javítják a kutatói utánpótlás képzésének a körülményeit. Ezen horizontális prioritás kiemelt témáit a Stratégiai Kutatási Terv a következőkben határozta meg:

- Technológiatranszfer, kommunikáció
- Szabványosítás
- Oktatói, kutatói utánpótlás
- Kutatásfinanszírozás
- Együttműködés

A fentiek körében megfogalmazott javaslatok sokszor mintegy a követendő szemlélet megerősítőiként, alapelvekként, iránymutatásokként szolgálva valamennyi kutatás-fejlesztési tevékenység nélkülözheteten kiegészítő részeként jelennek meg, nem feltétlen fogalmazhatók meg önmagukban független projektekként. Bizonyos területeken azonban a horizontális célok is összefoghatók, meghatározhatók mint önálló programok – ezek kerülnek ismertetésre az alábbiakban.

Helyzetkép. A terület fejlődését két alapvető gátló tényező hátráltathatja. A **technológiai dömping** lényege bizonyos kulcstechnológiák áron alul (esetenként: ingyen) való adása annak érdekében, hogy a versenytársak ne juthassanak a saját fejlesztéseiket támogató piaci haszonhoz. Az információtechnológia világában először az IBM, később a Microsoft alkalmazott sok éven át dömpingen alapuló versenystratégiát, de ma már első számú alkalmazója nem ez az (antitröszt perek által mára már kordába fogott) két multicég, hanem a Google, mely hosszú távú piaci stratégiája révén olyan egyedülálló technológiai és adatinfrastruktúrát épített ki, amelynek révén a nyelvtechnológiával kapcsolatos szinte minden területen (is) a felhasználók széles tömege számára „ingyenes” szolgáltatásokat nyújt. (A szolgáltatás természetesen nem ingyenes, hanem annak fejében történik, hogy a felhasználók a szolgáltatások használata közben minden információt, amelyet feldolgoztatnak, végleg és ellenszolgáltatás nélkül a Google rendelkezésére bocsátanak). A bárki számára hozzáférhető és általában legalábbis elfogadható minőségű szolgáltatásokkal a kevésbé tőkeerős társaságok nem tudják felvenni a versenyt még akkor sem, ha a Google-nél jobb minőségű szolgáltatást nyújtanak, mert a Google piaci stratégiája hatására az emberek ingyenes szolgáltatásokat várnak, és megelégszenek azzal, amit „ingyen” kapnak. Fennáll ugyanakkor az a veszély, hogy a Google szolgáltatásainak bármilyen okból való hozzáférhetetlenné válása nehéz helyzetbe hozza mindazokat, akik a Google szolgáltatásait használják.

A probléma megoldásához két út vezet: az egyik a peres út (az Európai Bizottsághoz már több antitröszt panasz futott be a Google ellen), mely azonban csupán *hosszú távon* vezethet célhoz, a másik a magyar nyelv- és beszédtechnológia aktív (kormányzati) támogatása.

A másik gátló tényező a **tömegtájékoztatással** kapcsolatos. A terület hazai (el)ismertsége nem megfelelő, javarészt azért, mert a nyelv- és beszédtechnológia még mindig nem igazán illeszkedik bele a hazai tudományos intézményrendszerbe. Ma még sem a döntéshozók, sem a széles felhasználói kör számára nem világosak a fentiekben leírt dömpingstratégiának a nemzeti technológiai bázisra gyakorolt bénító hatásai, mert az ingyenes (de rosszabb minőségű) megoldások az esetek egy részében eltéríthetik akár a kormányzat figyelmét is. Ma a legnagyobb veszély a nyelv- és beszédtechnológia ellehetetlenülése, holott ezt helyettünk más nem fogja jól megcsinálni. A probléma megoldásához szükség van nemcsak új intézmények kialakítására és a meglevő intézmények jobb bevonására, hanem aktív segítőmechanizmusokra különösen ott, ahol a piac önmagában ezt nem termeli ki. Az állam maga is igen jelentős gazdasági szereplő, és mint ilyen a közbeszerzéseknél előírhatja bizonyos nyelv- és beszédtechnológiai erőforrások használatát, illetve ezek folyamatos karbantartását. A kormányzat, ha ismerné a technológiai lehetőségeket, sokat tehetne szabályozással, illetve a kormányzati portál szolgáltatásainak és folyamatainak nyelvtechnológiai irányú fejlesztésével.

Veszélyes, hogy a nyelvtechnológiai fejlesztéseket a sajtó sokszor bulvártémává teszi, ami árt a fejlesztéseknek, mert vagy késznek ítélik, ami még nem az, vagy örökre használhatatlannak, ami még nincs kész. Időről időre megjelennek tipikus (ál)hírek arról, hogy egy adott problémát már megoldotta valamelyik nagy multi (Google, Microsoft, „a japánok” stb.), anélkül, hogy a megoldás peremfeltételeiről, környezetéről szó esne. Ennek a problémának a megoldásához a Platform pontosabb, szakszerűbb tájékoztatás nyújtásával már eddig is hozzájárult, de *hosszabb távon* tartós megoldás csak az intenzív szakemberképzés, a nyelvtechnológiai oktatás átstrukturálása lehet.

Az ágazat szereplői, a kutatóhelyek és szervezetek két szempontból osztályozhatók: nonprofit vagy profitérdekeltségű, illetve alkotó vagy felhasználó műhelyek. Természetesen egyik kategória sem tökéletes, hiszen ipari cégek is folytatnak tisztán tudományos, profitot még hosszú távon sem ígérő tevékenységet, és akadémiai/egyetemi környezetből is gyakran nőnek ki az ottani kutatások eredményeire támaszkodó profitérdekeltségű cégek. Az alkotás és a felhasználás között is igen bonyolult az átmenet, különösen azért, mert ugyanaz a műhely egyszerre lehet alkotó egyes területeken, és felhasználó másokon, akár egyetlen projekt keretén belül is. Az az ellentmondás is mindennapos, hogy a pályázatokban az elvégzett tevékenységek megítélése nem tényleges tartalmuk, hanem az őket végrehajtó intézmény besorolása alapján történik. Tehát hiába alapkutatás valami, ha azt vállalkozás végzi, nem az alapkutatásnak megfelelő támogatásban részesül, hanem a vállalkozásokéban, míg ha egy költségvetési intézmény termékközeli megoldással áll elő, több jár neki, mint egy valódi vállalkozásnak. A magyar nyelvtechnológiai cégektől a hazai piac méretei miatt nem realisztikus arra számítani, hogy kutatási potenciáljukat önerőből lényegesen fel tudják futtatni. Ezért kifejezetten fontos, hogy ipari cégek is részesülhessenek alapkutatási pénzekben, méghozzá *minimális vagy zéró önrész vállalása mellett*. Ennek fényében már *rövid távon* is fontos olyan pályázatok kiírása, illetve az európai kírásokhoz illeszkedő *matching grantek* olyan rendszerének kialakítása, amelyek kifejezetten a kis- és kö-

zépvállalkozásokra szabottak, és a hazai ipar kutatóbázisának növelését tűzik ki célul.

A nyelv- és beszédtechnológiai alkalmazások sajátossága az igen nagy, tipikus esetben domináns hozzáadott érték. Nem ritka az az eset, amikor az ilyen technológiák a *sine qua non*, ami nélkül az egész rendszer nem is működne: például ha nincs beszédfelismerés, akkor nincs automatikus telefontudakozó. Ugyanakkor, éppen központi szerepük miatt, az ilyen technikák beépítése, rendszerbe integrálása maga is nagyon igényes „high tech" feladat, amely rutinszerű informatikai eszközökkel és átlagosan képzett programozókkal csak a legritkábban oldható meg. A nyelv- és beszédtechnológia alkalmazói tehát maguk is egyfajta elitet képeznek, és ennek a technológiai élcsapatnak a fenntartása, bővítése, nemcsak az alkalmazó cégeknek, hanem az országnak is érdeke. A probléma (amely egyébként ebben a formában nem csak a nyelv- és beszédtechnológiát, hanem minden csúcstechnológiai területet érint) lényege az, hogy a kis cégek egy-két sztárprogramozót tudnak legfeljebb megfizetni, a többit elszívják a nagy (általában multinacionális) cégek.

A hazai felsőoktatásban, az utóbbi évek jelentős erőfeszítéseinek ellenére (lásd pl. BME, PPKE), a jövendő nemzedékek nyelv- és beszédtechnológusainak oktatása ma még nem áll az európai élvonalat jelentő olyan programok szintjén, mint amilyenek Edinburghban (School of Informatics), Saarbrückenben (Saarland University) vagy Amsterdamban (UvA) találhatók. A javulás érdekében a kutatói utánpótlás képzését koordinálni kell, az egyes területek legkiválóbb szakembereit be kell vonni az oktatásba.

Elérendő célok.

- A kutatói és az ipari szféra közötti párbeszéd, az interdiszciplináris együttműködés feltételeinek javítása érdekében az ágazat technológiatranszfer-központjának kialakítása.
- A kutatásfinanszírozási rendszer átalakítása, hatékonyabbá, vállalkozásbarátabbá tétele, a kutatói díjazás és a számonkérés európai normához való közelítése.
- Az általános európai rendszerbe illeszkedő BA-MA-PhD szekvencia tantervének kidolgozása, a magasabb (MA/MSc, PhD) fokozatok megszerzésének ipari gyakorlathoz való kötése, ahol a befogadó cégeken keresztül a gyakornokok nem szimbolikus, hanem az uniós ipari fizetésekkel összemérhető ösztöndíjban részesülnek.

Projekttervek.

1. Technológiatranszfer-központ
2. Nyelv- és Beszédtechnológiai Inkubátor
3. Lexikográfiai Központ
4. Kutatói utánpótlás képzése

2.7.1. Technológiatranszfer-központ

A magyar nyelv- és beszédtechnológiának saját, kutatási eredményeket és ismereteket összesítő, az információátadást elősegítő, a nemzeti kutatási infrastruktúrát szolgáltató központja. Feladata az ipari szereplők kutatás-fejlesztési igényeinek felmérése, valamint az országban rendelkezésre álló tudás és a hozzáférhető eredmények, módszerek feltérképezése és az információ közvetítése a lehetséges partnerek, a kutatási, ipari szereplők, illetve akár a nagyközönség felé. A hasonló európai kezdeményezésekkel (pl. CLARIN, FLARENET) összhangban, a korszerű hálózati technológiákat kihasználva működő virtuális központ.

Várható eredmény, hatás, hasznosulás.

- Párhuzamos fejlesztések kiküszöbölése, kutatási tevékenység hatékonyságának növekedése.
- Az ágazati szereplők közötti kommunikáció és együttműködés szembetűnő javulása.
- Az interdiszciplináris kutatások rendszerének elterjedése, a határterületekkel és az alkalmazókkal való együttműködés erősödése.
- Dokumentációk és gyakorlati útmutatók kidolgozásának megszervezése, ezek hozzáférhetővé tétele.
- Médiaesemények, szakmai rendezvények koordinálása.

2.7.2. Nyelv- és Beszédtechnológiai Inkubátor

A magyar nyelvtechnológiai kutatások egy része tőkeerős nemzetközi cégekhez kapcsolódik. Ugyanakkor a magyar nyelv nem nagy piac, ezért a terület olyan óriásainak, mint az IBM vagy a Microsoft, nem érdeke az erre irányuló kutatások elaprózása. Javasoljuk tehát egy Nemzeti Nyelv- és Beszédtechnológiai Inkubátor létrehozását. *Hosszú távon*, egy 5-7 éves periódus végére, az inkubátornak már önfenntartónak kell lennie, legalábbis ami a bér- és működési költségeket illeti. A helyiségigényét (becsléseink szerint 20-30 kutató elhelyezéséről lehet szó) az akadémiai intézethálózat sokfelé megüresedett helyiségeiből lehet fedezni. Az inkubátor működési modellje azon alapul, hogy a kutatókat, legalábbis egy részüket, a támogató multinacionális cégek delegálják, tehát fizetésüket is ők fedezik. Az inkubátor nem termékeket állít elő, tevékenysége kizárólag a magyarral (és esetleg a terület egyéb kisnyelveivel) kapcsolatos alapkutatásokra korlátozódik, az előállított szoftverek és egyéb erőforrások pedig nyílt forráskódúak és szabad felhasználásúak lesznek.

A szabad felhasználhatóság és ingyenesség persze sokszor kompromisszumokkal jár: a különböző projektek keretében létrejött nyílt forráskódú szoftverek nagy része gazdátlanná válik a projekt lezárulásával, a dokumentáció sokszor hiányos és elavult, nincs, aki a felhasználóknak támogatást nyújtson, és az elvileg már csak viszonylag kis befektetést igénylő javítások, további fejlesztések, adaptációk elvégzésére már nem vállalkoznak a korábbi fejlesztők. Ezért gondoskodni kell a fontos alapinfrastruktúrát

építő projektek hosszú távú életben maradásáról, és az inkubátor formáció erre *közép- és hosszútávon* igen alkalmas. Mindez tipikus *industry-led* formájú együttműködésre ad lehetőséget, ahol az alapítók csupán a kereteket, elsősorban a helyiséget és a fundraising időszakra szükséges támogatást adják, ezt szakmai tartalommal az ipar által delegált vendégkutatók töltik ki. Ez az intézmény adhatja annak a magas hozzáadott értékű kutatásnak a szellemi bázisát, amiről a helyzetértékelésben már szó volt. Az inkubátort a hozzájáruló ipari cégek delegálta szakértőkből álló igazgatótanács irányítja.

Várható eredmény, hatás, hasznosulás.

- A kutatási hatékonyság ugrásszerű növekedése.
- Koncentrált, magas színvonalú kutatás-fejlesztési tevékenység.
- Az ipari kutatóbázis megerősítése.

2.7.3. Lexikográfiai Központ

A szótár (adatbázis) a nyelv- és beszédtechnológia kulcseleme: gépi fordító nem képzelhető el többnyelvű szótár nélkül, de a szemantikai webalkalmazásokhoz is szükség van legalább egynyelvű értelmező szótárra. Az elmúlt évtizedekben magánerőből létrejött nem egy nyelvtechnológiai és/vagy lexikográfiai indíttatású intézmény. Ilyen például a MorphoLogic Kft., mely hatékony szótárkészítési és elektronikus megjelenítési technológiát fejlesztett ki, vagy a Tinta Könyvkiadó, amely rengeteg új egynyelvű szótárat készített, illetve a korszerű lexikográfiai alapokra épülő kétnyelvű szótárakat piacra hozó Grimm Kiadó. A magyarországi szótárpiac mérete nehezen teszi lehetővé, hogy a nyelvek, nyelvpárok többségére egymással versengő szótárak készüljenek. Ez az elektronikus világban sincs másképp, sőt a két terület – a papíralapú és az elektronikus kiadványok – megkülönböztetése sem igazán időszerű már 2010-ben. Ennek oka, hogy a szótári anyagok létrehozására irányuló technológia gyakorlatilag az utolsó lépésig (nyomtatás vagy elektronikus megjelenítés) megegyezik. Ezért a számítógépes és a hagyományos szótárkiadás sem versenyezhet egymással.

Az adott történelmi és piaci helyzetben értelmetlennek tűnik fenntartani egy olyan versenyhelyzetet, melyben a magas szintű technológia és a kiváló hazai tartalom küzd egymással. Az utóbbi években ugyanis ez a furcsa helyzet állt elő a hazai szótáriparban: szerényebb lexikográfiai tartalmú, de professzionális technológiákra épülő elektronikus eszközök állnak szemben gyengébb szoftvermegoldásokon alapuló, ám professzionális szótártartalmakkal. Ennek a helyzetnek már rövid távon egy mindenki számára előnyös átalakítása volna időszerű. A megoldást egy önálló, dedikált intézmény létrehozásában látjuk, mely a mostani széttagolt szótárkészítői munkálatokat integrálni képes.

Célunk egy olyan, modern nyelvtechnológián alapuló Lexikográfiai Központ megalapítása és működtetése, mely a hazai (és a nemzetközi) szótárvilág területén kiemelkedő mennyiségű és minőségű tartalommal és a legkorszerűbb nyelvtechnológiai megoldásokkal bír. Ez az intézmény a felhasználók széles köre által hatékonyan

használható, egyben a professzionális szótárírók és -szerkesztők magas színvonalú munkavégzését is segítő szótárkészítő eszközrendszerrel (2.6.1.3.), valamint a tetszőleges formátumú forrásanyagoknak a szótári rendszerbe integrálására szolgáló szoftvereszközökkel és a világviszonylatban is kiemelkedőnek számító szótármegjelenítési technológiákkal is rendelkezni fog. Meglevő egy- és kétnyelvű szótárak bevonásával és a Magyar Lexikográfiai Referencia-adatbázis megépítésével (2.6.1.1.) ez a központ alkalmas lesz új nyelvpárok kétnyelvű szótárainak korszerű, az eddiginél lényegesen gyorsabb létrehozását garantáló technológia kifejlesztésére és használatára.

Az erőforrások újrafelhasználása által az új, illetve az eddig gazdaságilag nehezen menedzselhető nyelvpárok esetében sem feltétlenül veszteséges a szótárak létrehozására irányuló tevékenység. A befektetendő munkamennyiség a megfelelő technológiák alkalmazásával jelentősen csökkenthető. Az elektronikus kiadás nem igényel külön befektetést, az esetleges papírkiadást pedig a mindenkori aktuális állapotból, egyedileg, önköltségi áron lehet vállalni. Ehhez viszont a hagyományostól eltérő együttműködésre van szükség a mindenkori szerkesztőségek és a velük a jövőben folyamatos kapcsolatot tartó szerzők, szótárírók között. Mindezek eredményeképpen nem egyedi kiadások születnek majd, amelyek változatlan formában hosszú ideig maradnak a piacon, miközben elavulnak, hanem folyamatosan javuló, bővülő, fejlődő anyagok lehetnek a tartós piaci siker zálogai. A Lexikográfiai Központ szervezetileg alapulhat a résztvevő nyelvtechnológiai és lexikográfiai intézmények közreműködésével kialakítandó PPP-megoldásra, vagy lehet ez egy erre a célra kialakított új központi/akadémiai intézmény is, akár bizonyos ipari-gazdasági vetületekkel.

Finanszírozását hosszú távon is nonprofit alapon képzeljük el. Ennek felel meg irányítási struktúrája is: elismert kutatókból álló akadémiai grémium felügyelete alatt tervezzük működtetni. Az MTA Szótári Munkabizottság tömöríti a hazai szótártudomány legnagyobb tapasztalatú szakembereit, akiknek szaktudását a Lexikográfiai Központ működtetésében jól lehetne kamatoztatni. Így a jelenleg elsősorban formális bizottsági munka a napi gyakorlatban is hasznosulhatna. Rövid távon szükséges a Központ megalapításáról, koncepciójának, üzleti modelljének kidolgozásáról szóló pályázat kiírása; illetve rövid- és hosszútávon a folyamatos működtetéshez szükséges elsődleges erőforrások (kutatói státuszok, iroda, hardver, szoftver) garantálása. Az alapítói pályázatok elbírálásánál különösen fontos szempont kell legyen az, hogy a hazai nyelvtechnológia legfontosabb intézményei ehhez a leendő egységhez milyen tudással, milyen (esetenként copyrighttal védett) tartalmakkal és milyen piaci tapasztalatokkal tudnak hozzájárulni.

Várható eredmény, hatás, hasznosulás.

- Professzinális szótárkészítő technológiák és szótártartalmak integrálása.
- Korszerű szótárkészítő eszközrendszer és szótármegjelenítés.
- Új nyelvpárok kétnyelvű szótárainak eddiginél gyorsabb létrehozása.
- A meglévő lexikográfiai adatkincs átmentése.
- Világszínvonalú szótárkészítés Magyarországon.

2.7.4. Kutatói utánpótlás képzése

A Platform *középtávú* célja egy az általános európai rendszerbe illeszkedő BA-MA/MSc-PhD szekvencia tantervének kidolgozása. Ehhez *rövid távon* a Platform szervezésében egy konferencia megrendezése szükséges, ahol nemcsak az egyetemek, hanem a jövendőbeli munkaadók (ipari cégek) is beleszólhatnak a fejlődés irányának kijelölésébe. Az oktatás hatékonyságát növeli a képzési erőforrások koncentrációja és egységesítése: azonos ismeretek oktatásához közös tananyagmodulok kidolgozása, ezek kommunikációs hálózatokon keresztül történő szabad hozzáférhetősége. A fiatal kutatók számára versenyképes ösztöndíjakat kell létesíteni, az ipar és az oktatási intézmények közötti kapcsolat megerősítésének keretében lehetővé kell tenni képzésük egy részének kihelyezését ipari szereplőkhöz.

Várható eredmény, hatás, hasznosulás.

- Nemzetközi szinten versenyképes fiatal kutatók.

Projektterv		Típus	Résztvevők	Költségforrás	Időzítés	Költség
7.1.	Technológiatranszfer-központ	—	egyetem, kutatóintézet, kkv, multi	70% pályázat, 30% önerő	2011-2013	50M Ft
7.2.	Nyelv- és Beszédtechnológiai Inkubátor	—	egyetem, kutatóintézet, kkv, multi	25% pályázat, 75% önerő	2011-től folyamatosan	100M Ft
7.3.	Lexikográfiai Központ	—	kutatóintézet, egyetem, kkv	100% pályázat	2011-től folyamatosan	100M Ft
7.4.	Kutatói utánpótlás képzése	—	egyetem, kutatóintézet, kkv, multi	75% pályázat, 25% önerő	2011-2014	25M Ft

2.7. táblázat. A 7. stratégiai cél projektterveinek összegzése.

3 Összefoglaló

A Megvalósítási Terv egyes fejezeteiben a Stratégiai Kutatási Tervben részletesebben kifejtett stratégiai célokat emeltük ki proritásokként. Az egyes stratégiai célokon belül azokat a kutatási irányokat ismertettük, amelyeket a magyarországi nyelv- és beszédtechnológia további fejlődése és a nemzetgazdaság szempontjából is kiemelten fontosnak tartunk. Továbbá konkrét projektterveken keresztül mutattuk be, hogy melyek azok a kutatási feladatok, melyek kitörési pontokként szolgálhatnak az elkövetkező 10 év során.

Kiemelt területként azonosítottuk a kutatási infrastruktúra kiépítését, a nyelvi információ feldolgozására vonatkozó fejlesztéseket, hangsúlyoztuk a magyar nyelv- és beszédtechnológia értékőrző és értékmentő, valamint az esélyegyenlőség és életminőség javításában betöltött szerepét.

A felsorakoztatott célokból, kutatási irányokból, feladatokból megállapítható, hogy a magyarországi nyelv- és beszédtechnológia a helyzetképekben felvázolt kedvezőtlen jelenségek ellenére jelentős erősségekkel rendelkezik. A különböző kutatóműhelyekben összegyűlt szaktudás, a kutatók motivációja, az aktív hazai és nemzetközi kapcsolatok mind olyan paraméterei a területnek, amelyekre a sikeres fejlődés alapozható.

Az anyag összeállításában nem kizárólag a nyelv- és beszédtechnológia határain belül maradtunk, hanem kerestük az integratív kutatási irányokat is, amelyekben különféle célok érdekében hatékonyan és sikeresen tudnak együtt dolgozni az egyes szakterületek képviselői. Ennek megfelelően a projekttervek kidolgozásánál igénybe vettük nyelvtörténészek, neuro- és pszicholingvisták, kognitív tudósok, pszichológusok, informatikusok segítségét is.

Igyekeztünk feltárni az ipari szférával és a társplatformokkal való együttműködés lehetőségeit is. Elsősorban a világtrendekhez illeszkedő területeken – mint a beszédfelismerés, az információkinyerés és -visszakeresés, valamint a webbányászat – lehetséges az ipari szféra nagyarányú bevonása a fejlesztésekbe.