A Hunglish korpusz és szótár

Halácsy Péter[1], Kornai András[1], Németh László[1], Sass Bálint[2],
Varga Dániel[1], Váradi Tamás[2], Vonyó Attila

[1] BME – Média Oktató és Kutató Központ
{hp, kornai, nemeth, daniel}@mokk.bme.hu
[2] MTA Nyelvtudományi Intézet,
{joker, varadi}@nytud.hu

Kivonat: Cikkünkben a Budapest Műszaki Egyetem és az MTA Nyelvtudományi Intézet által épített angol–magyar Hunglish korpuszt és Hunglish szótárat mutatjuk be.

Bevezetés

A Budapesti Műszaki Egyetemen 2004 nyarán indult el a Hunglish projekt [5], melynek fő célja egy statisztikai elven működő gépi nyersfordító rendszer kifejlesztése volt. A feladat megoldásához létre kellett hoznunk egy mondatszinten illesztett, magyar–angol párhuzamos korpuszt. A mondatszintű illesztés a korpusz minden forrásnyelvi mondatához hozzárendeli annak célnyelvi fordítását, amely esetenként akár több mondatból is állhat. A mondatszinten illesztett korpusz legalapvetőbb felhasználási területe a gépi fordítás, amely a fordítás statisztikai modelljének paramétereit képes beállítani a korpusz alapján.

A Hunglish projekt eredményeként tehát — az angol-magyar nyersfordító prototípus mellett — elkészült egy automatikus mondatillesztő program, létrejött egy angol-magyar párhuzamos korpusz, illetve kialakult egy teljes párhuzamos korpusz építésére alkalmas eszközkészlet és módszertan. A projekt minden eredményét a Creative Commons “nevezd meg” licensz alatt tettük elérhetővé, vagyis minden termékünk bárki számára szabadon hozzáférhető, felhasználható, átdolgozható. A korpusz felhasználási lehetőségei változatosak: felhasználható nyelvtechnológiai, számítógépes nyelvészeti alkalmazásokban (lásd például ebben a kötetben Miháltz és Pohl 2005), fordítástámogatásban, kétnyelvű terminológiai adatbázis építésben, sőt talán még a nyelvoktatásban és fordítóképzésben is. Nyelvfüggetlen, pontosabban nyelvpárfüggetlen eszközkészletünk egyszerűvé teszi további párhuzamos korpuszok építését.

Már az adatbázis építése során is — de a gépi nyersfordító fejlesztésekor különösen — felhasználtuk Vonyó Attila magyar–angol szótárát. (Ez az anyag képezi a sokak által használt magyar–angol Sztaki-szótár[27] alapját is.) A szótárt más forrásokból gyűjtött terminológiai és egyéb kétnyelvű adatbázisokkal összefésültük, s a pár-

[27] http://dict.sztaki.hu

huzamos mondatok felhasználásával a szótár minőségét javítottuk. Ettől is azt reméljük, hogy sokan sokféle célra használni tudják majd, és a jövőben több tőlünk független kutatás vagy szolgáltatás részeként meg fog jelenni.

1. A korpusz alapanyaga

A webes források automatikus felkutatása és feldolgozása [7] óta igen gyakran alkalmazott módszer párhuzamos korpusz építésére. Az eljárás angol és másik világnyelv esetén ígéretes eredményeket adott [2,8]. Chen és Nie angol–arab nyelvpárra 2000 párhuzamos weboldalról mindösszesen 2,3 millió szövegszót, Resnik és Smith angol–francia nyelvpár esetén 2491 párhuzamos oldalt tudott felkutatni automatikus módszerekkel.

A mi célunk ennél legalább egy nagyságrenddel nagyobb korpusz építése volt. Tapasztalatunk szerint azonban a magyar weben egyszerűen nincs elegendő automatikusan felkutatható párhuzamos weboldal.

Ezért egy manuális, de sokkal hatékonyabb módszert választottunk. Ugyan fő forrásunk továbbra is a web, de nem próbáltunk automatikusan fordításpárokat találni. Bár így a dokumentumok felkutatása komoly munkabefektetést jelentett, a dokumentumok manuális letöltése és párosítása után a további lépéseket már automatizáltan végeztük. A korpusz az alábbi fő forrásokból épül fel:

Irodalmi szövegek

Az irodalmi szövegek fő forrása a *Project Gutenberg*[28] és a *Magyar Elektronikus Könyvtár*[29] , amelyek adatbázisait összevetve kikerestük azokat a műveket, amelyek egymás fordításai. Innen majdnem száz klasszikus irodalmi mű tölthető le Jane Austentól Tolsztojig. Ugyanilyen fontosak a weben szép számmal fellelhető még szerzői jog védelme alatt álló modern művek is.

A szövegek között tehát voltak olyanok (a modern irodalmi anyagon kívül a később említett filmfeliratok is), amelyeknek a változatlan formában történő újrapublikálása jogi problémákba ütközne. Itt említjük meg, hogy ezeket a szövegeket a párhuzamosítás után összefűztük, majd angol–magyar mondatpárok véletlenszerűen (pontosabban ábécérendbe) rendezett halmazává alakítva publikáltuk. Ezzel az eredeti szövegek rekonstruálását lehetetlenné tettük, megvédve így a szövegek jogtulajdonosainak szerzői jogait. Ugyanakkor a korpusz legfontosabb általunk megcélzott felhasználásai a mondatok sorrendjét nem veszik figyelembe, tehát ezen célokra a keverési művelet a korpusz értékét nem csökkenti.

Jogi adatbázis

Legnagyobb forrásunk az EU közösségi jogszabályok *CELEX adatbázisa* és az *Európai Alkotmány*. [30]

[28] http://www.gutenberg.org

[29] http://mek.oszk.hu/indexeng.phtml

[30] A forrás érdekessége, hogy az adatbázis elérhető a közösség minden hivatalos nyelvén. A soknyelvű mondatszintű párhuzamosítás elvégzését tervbe vettük.

Nyílt forráskódú szoftverek dokumentációi

A nyílt forráskódú szoftverek honosításainak eredményeit tudomásunk szerint először [9] használta párhuzamos korpusz alapanyagaként. A Hunglish korpuszba a KDE, Gnome, OpenOffice, Mozilla és GNU eszközök dokumentációit építettük be.

Filmfeliratok

Az internetről letölthető *filmfeliratoknak* csak egy részét (mintegy 400 film) vettük be az adatbázisunkba, főleg kísérleti jelleggel. Ezek, bár sok esetben elég rossz minőségű fordítások, bizonyos nyelvhasználatra és szókészletre (pl. szleng és káromkodás) olyan kiváló forrásanyagot tartalmaznak, amely a formálisabb forrásokból nem lenne kinyerhető.

Hírek, magazinok

Jó minőségű, de az internetről nem letölthető, ezért nehezen beszerezhető anyagok származhatnak *kétnyelvű magazinokból*, illetve magazinok magyar nyelvre fordított kiadásaiból. Mi a National Geographic és a Diplomacy and Trade magazin néhány magyarra fordított számát dolgoztuk fel.

Sajtófigyelés

A Magyar Telekom Rt. szabad felhasználásra a rendelkezésünkre bocsátott nagy mennyiségű távközlési témájú sajtóanyagot, amelyet fordítóik ültettek át angol nyelvre.

További, még fel nem dolgozott források

A fentieken kívül megkezdtük további források korpuszba építését is. *Tőzsdei cégek nyilvános éves jelentései* sokszor elérhetőek angol nyelven is a vállalat weboldalán. Mi három vállalat 19 éves jelentését töltöttük le. *Vallási szövegek*: a katolikus egyház által sok nyelven publikált pápai ediktumok feldolgozását már megkezdtük.

1. Táblázat: A korpusz összetétele szövegtípusok szerint

forrás	Angol tokenek (millió)	Magyar tokenek (millió)
irodalom	14,6	11,5
jog	24,1	18,3
filmfelirat	2,5	1,9
szoftver	0,8	0,7
magazinok	0,3	0,3
sajtó	2,1	1,7
összesen	**44,5**	**34,5**

2. A korpusz feldolgozása

Első lépésben pontosan párosítottuk a letöltött vagy más úton szerzett nyers dokumentumokat. Ezután kinyertük belőle a nyers szöveget, formátumuk, karakterkészletük konvertálásával. Ez az egyszerűnek hangzó lépéssor a korpuszépítés nagy mennyiségű manuális munkát és szakértelmet igénylő fázisa.

A forrásként szolgáló állományok formátuma és karakterkódolása igen változatos, így például PDF, Postscript, DOC, RTF, HTML, SXW, TEX, valamint különböző szöveges formátumok automatikus konvertálását kellett megoldanunk, amihez nyílt forráskódú programkönyvtárakat és segédprogramokat használtunk.

A táblázatokat és szigorú tördelést tartalmazó, Quark Express formátumban lévő, szigorúan a nyomtatás előtti fázisból hozzánk került magazinok feldolgozása alig automatizálható (sokszor csak az OCR programok jönnek számításba). Szerencsére a szövegek nagy része nem ilyen problematikus: az `antiword`, `catword`, `html2text` és más hasonló nyílt forráskódú programok megfelelően alkalmazhatóak.

A karakterkódolás meghatározó jelentőséggel bír a szöveges adatok tárolásában. Tapasztalataink szerint érdemes a nehezebben kezelhető, de veszteségmentes Unicode karakterkódolást választani. A Hunglish esetében mégis 8 bites karakterkódolást alkalmaztunk, mert egyes eszközeink (például a mondatra szegmentáló) csak ezt támogatják, és a veszteség a korpusz legfontosabb várható alkalmazásait (szótárépítés, gépi fordító tanítása) véleményünk szerint nem hátráltatja. A korpusz magyar oldalán ISO 8859-2 kódolást, az angol oldalán ISO 8859-1 kódolást alkalmaztunk. Emiatt egyes speciális szimbólumokat le kellett cserélnünk, de ezek az esetek a korpuszban rendkívül ritkán fordultak elő.

3. Mondatszintű párhuzamosítás

Ha a dokumentumszinten párosított nyers szövegek már elkészültek, mindössze néhány perc elkészíteni egy párhuzamos dokumentumpár mondatszintű illesztését.

Ehhez első lépés a mondathatár-azonosítás, amit a szabályalapú `huntoken` [6] programmal végeztünk, magyar és angol kivételszótárakkal.

A szóhatárolás, amelyre szintén szükség van az illesztés előtt, tapasztalatunk szerint nem kritikus lépés. A komplex `huntoken` az illesztés szempontjából feleslegesen jelöl meg nyílt tokenosztályokat, címeket, kifejezéseket egy tokennek. Helyette egy egyszerű háromsoros programot használtunk, ami tulajdonképpen nem tesz mást, mint a szavak végéről leválasztja az írásjeleket.

A mondatszintű párhuzamosításhoz fejlesztettük ki a `hunalign` programot [11]. A `hunalign` legfőbb előnye, hogy ún. zajos szövegeken is megbízhatóan dolgozik. Mondatok kiesését és összeolvadását akár nagyobb számú mondat esetében is kezeli; ilyen például a csak az egyik oldalon jelen levő utószó, vagy a nagy számú lábjegyzet. A mondatok sorrendjének felcserélődését a `hunalign` nem kezeli. A tanítókorpuszainkban és szúrópróbaszerű korpuszelemzéseink során azt tapasztaltuk, hogy ez a jelenség igen ritka. Magyar–angol tesztkorpuszunkon a `hunalign` pontossága 99.34%, ami jelentősen meghaladja a standard statisztikus módszerekét, köszönhetően – többek között – a kétnyelvű szótár felhasználásának.

A `hunalign` működési folyamata nagyon vázlatosan a következő:

1. Elkészíti a nyers, tokenizált szöveg magyar mondatainak nyersfordítását: a bemeneti magyar–angol szótár alapján lecseréli a magyar szavakat a célszövegben leggyakoribb angol megfelelőjükre, a szótárban nem szereplő szavakat pedig meghagyja eredeti alakjukban (így például a számokat, jogszabályban a paragrafus-hivatkozásokat, email címeket stb. is).
2. A nyersfordítás és a célszöveg hasonlósága, valamint a mondathosszarány alapján hasonlósági mértéket számol a forrásszöveg és a célszöveg mondatai között. Ez alapján megkeresi a legjobbnak vélt illesztést egy dinamikus programozási feladat megoldásával.
3. A megtalált illeszkedő mondatpárok alapján statisztikai módszerekkel szótári tételeket azonosít, és ezekkel kiegészíti a kiinduló szótárat.
4. A kiegészített szótárat felhasználva újra elvégzi a szöveg illesztését az első két pont szerint. (Tapasztalataink szerint ennek a ciklusnak a további iterációja már nem javítja az illesztés minőségét.)

Az algoritmus leírásából látható, hogy a `hunalign`-nak az első nyersfordítás elkészítéséhez szüksége van egy kiinduló szótárra. Ehhez mi a Vonyó szótárat alkalmaztuk. Ahhoz viszont, hogy egy toldalékolt szót megtaláljunk a szótárban, szótövezést kell végeznünk. Ehhez a `hunmorph` programot [10] használtuk: a magyar szövegeknél a `morphdb.hu` [12], angol szövegeknél a szintén saját fejlesztésű `morphdb.en` nyelvi erőforrást alkalmaztuk. Nem foglalkoztunk azokkal az esetekkel, amikor a szótövezés nem egyértelműen adja meg a szó lemmáját. Ilyenkor egyszerűen a legkevesebb toldalékot leválasztó elemzést választjuk.

Vegyük észre azonban, hogy `hunalign` képes kiinduló szótár nélkül is működni. Ilyenkor az első lépésben a nyersfordítás nem változtat semmit a kiinduló mondaton, és a második lépésben használt hasonlósági függvény elsődlegesen a mondathosszak arányától fog függeni [4]. Ebben az esetben a `hunalign` működése hasonlít a sokak által használt vanilla [1] illesztőéhez. Az automatikus szótárépítési fázis után újból elvégzett második párhuzamosítás azonban már jóval nagyobb pontosságot ér el, mint az első.

A 2. táblázat a `hunalign` pontosságát és fedését mutatja be különböző erőforrásokkal a MULTEXT-East 1984 magyar–angol párhuzamos korpuszon [3] mérve.

Látható, hogy a kiinduló szótár növeli a pontosságot, különösen akkor, ha szótövezést is végzünk.

2. Táblázat: A hunalign pontossága és fedése az 1984 korpuszon különböző beállításokkal: szótár: kiinduló szótár használatával, tövez: tövező használatával, iter: automatikus szótárépítés, kiinduló szótár nélkül, id: szótár nélküli, csak azonos szótokeneket illesztő mód, hossz: illesztés a mondat karakterszámának alapján

módszer	**pontosság**	**Fedés**
hossz	97.58	97.55
hossz+id	97.65	97.42
szótár	97.30	97.08
hossz+szótár	98.86	98.88
hossz+szótár+tövez	99.34	99.34
hossz+tövez	98.63	98.74
hossz+iter+tövez	99.12	99.18

A 2. táblázatból az is kiolvasható, hogy az illesztő bármiféle nyelvi erőforrás nélkül is jó eredményt ér el. Megvizsgáltuk, hogy ez más nyelvpárokra is igaz-e. A 3. táblázat mutatja, hogy hogyan teljesít a `hunalign` nyelvi erőforrás nélkül a MULTEXT-East 1984 korpusz más nyelvű párhuzamos anyagain mérve. Megjegyezzük, hogy az SGML formátumú korpusz karakterkonverzióját nem minden esetben végeztük el, ami ront a mondathossz alapú heurisztika pontosságán, és feltehetőleg felelős a román–angol nyelvpáron elért kiugróan alacsony eredményért.

3. Táblázat: A hunalign pontossága és fedése a MULTEXT-East 1984 korpuszon különböző angol–X nyelvpárokra, szótári erőforrás használata nélkül.

nyelv	pontosság	fedés
észt	99.34	99.53
cseh	98.60	98.75
román	97.10	97.98
szlovén	99.44	99.61

4. Kézi illesztés

Az illesztőalgoritmus teszteléséhez szükségünk volt manuálisan illesztett párhuzamos szövegre. Ehhez a fent már tárgyalt MULTEXT-East 1984 párhuzamos korpusz mellett felhasználtuk John Steinbeck Egy marék arany című művének általunk elkészített manuális illesztését is.

A manuálisan végzett munka három részből állt:

1. automatikusan mondatokra bontott és mondatszinten automatikusan illesztett korpusz illesztésének kézi javítása
2. az eredeti automatikus mondatszegmentálás hibáinak kézi javítása
3. kézzel javított szegmentálású, automatikusan illesztett korpusz illesztésének kézi javítása

Az automatikus mondatszegmentálás a `huntoken`, az automatikus mondatszintű illesztés a `hunalign` programmal történt. Az első szakasz végén létrejött korpusz felhasználható párhuzamosító algoritmusok pontosságának kiértékelésére abban a tipikus helyzetben, amikor a bemeneti mondatra szegmentálás automatikus, tehát hibákat tartalmazhat. A második és harmadik munkafázis eredményeképpen hibátlannak tekinthető párhuzamos korpuszt kaptunk. Ez sokféle célra hasznosítható, de az elsődleges célja párhuzamosító algoritmusok pontosságának értékelése azon feltétel mellett, hogy a korpuszban a mondathatárokat hibátlanul ismerjük.

A korpuszt felhasználtuk a `hunalign` tesztelésére. Természetesen az algoritmus egy korábbi változatának kimenete befolyásolta a végeredményt, tehát az algoritmusunk ezen korpuszokon való értékelése megkérdőjelezhető. De egyrészt a végső manuális párhuzamosítás elegendően függetlennek tekinthető a gépi párhuzamosítási lépésektől, másrészt az algoritmus inkrementális paraméterbeállítására a korpusz mindenképpen alkalmazható.

A manuális mondatszintű illesztés munkaigényének becslését segítheti, ha közöljük a következő adatokat: A regény terjedelme 230 oldal, 57,000 szó. Ezen a szövegen a fent leírt manuális munkát négy ember végezte, a teljes ráfordított munkaidő körülbelül 240 órát, azaz 6 emberhetet tett ki.

5. Korpuszjavítás

Egy gépi tanulási szoftverimplementáció sebességét általában joggal tekintik kevésbé fontosnak a pontosságához képest. A Hunglish korpusz építése során azonban sokféle módon előnyünkre tudtuk fordítani azt a tényt, hogy a C++ nyelven írt `hunalign` legalább egy nagyságrenddel gyorsabb, mint más hasonló célú implementációk. A teljes, több tízezer dokumentumból álló korpuszunk párhuzamosítása így néhány nap helyett néhány órán belül elvégezhető volt.

Az algoritmus gyorsasága lehetővé tesz egy iteratív munkafolyamatot: A `hunalign` által legalacsonyabb konfidenciaszintűnek ítélt párhuzamosítások tipikusan valamilyen dokumentumszintű illeszkedési hibát, vagy súlyosabb szövegnormalizációs hibát tartalmaznak. (Példák az előbbire: több kötet illesztése egyhez, nagyszámú egynyelvű lábjegyzet, novelláskötet más novella-sorrenddel, vagy akár gyermekek számára átdolgozott kiadás.) Az alacsony konfidenciaszintű részek megvizsgálásával az ilyen problémák feltárhatók és orvosolhatóak, vagy a menthetetlen szöveg eliminálható. Előfordulhat, hogy a javítást a szövegkinyerő, tokenizáló, mondatra szegmentáló vagy párhuzamosító algoritmusainkon kell megtennünk. Egy-egy ilyen javítás a párhuzamosított dokumentumok ezreit érintheti, tehát a teljes korpuszépítési ciklus újbóli elvégzése után a legalacsonyabb konfidenciaszintű mondatok listája lényegesen megváltozhat. Ezt a folyamatot addig ismételtük, amíg már a legalacsonyabb konfidenciaszintű párhuzamosítások is elfogadhatóak voltak.

Egy másik példa iteratívan végezhető javításra a mondatra szegmentáló kivételszótárának bővítése. Ehhez felhasználtuk azon szavak listáját, amelyek a két mondatot egy mondathoz rendelő szegmentumokban nagy gyakorisággal az elválasztó írásjel előtt állnak.

6. Szótárépítés

A kiinduló kétnyelvű Vonyó szótárat a korpusz alapján javítottuk. Először a korpuszban nem megtalálható rekordokat törölve egy kisebb, de jobb minőségű szótárat kaptunk. Ezután új rekordokat vettünk be, amelyeket a párhuzamos mondatpárokon futó statisztikus alapú automatikus szótárépítő algoritmusunk azonosított.

A kétnyelvű szótárak általában rejtett, de fontos tulajdonsága, hogy a különböző jelentéseket, fordítási alternatívákat gyakoriság szerint súlyozva mutatják. Az elkészült Hunglish szótárba ezek a gyakorisági adatok a korpusz alapján kerültek be.

7. Keresőfelület

A korpuszhoz és szótárhoz készült kereső szolgáltatás kiegészítője lehet a jelenlegi webes szótár szolgáltatásoknak. A kereső találati listáján a párhuzamos mondatok jelennek meg. A beépített magyar és angol szótövezőnek, illetve a nyílt forráskódú Lucene programkönyvtár keresőalgoritmusának köszönhetően nem csak lemmák, hanem teljes kifejezések, idiómák is kényelmesen és hatékonyan kereshetők.

8. Köszönetnyilvánítás

A Hunglish projekt az Informatikai és Hírközlési Minisztérium ITEM pályázatán nyert támogatással vált lehetővé (IHM-ITEM 2003/76/6/2004). A projekthez való hozzájárulásukért köszönettel tartozunk Gyarmati Ágnesnek, Héja Enikőnek, Mészáros Ágnesnek, Balogh Attilának, Kornai Andrásnak és Trón Viktornak. Köszönetet mondunk a Magyar Telekom Rt.-nek a Sajtófigyelő korpusz publikálhatóvá tételéért és a projekt infrastrukturális támogatásáért.

Bibliográfia

[1] Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993.

[2] Jiang Chen and Jian-Yun Nie. Automatic construction of parallel english-chinese corpus for cross-language information retrieval. In *Proceedings of the sixth conference on Applied natural language processing*, pages 21–28, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[3] Ludmila Dimitrova, Tomaz Erjavec, Nancy Ide, Heiki Jaan Kaalep, Vladimir Petkevic, and Dan Tufis. Multext-east: Parallel and comparable corpora and lexicons for six central and eastern european languages. In Christian Boitet and Pete Whitelock, editors, *Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics*, pages 315–319, San Francisco, California, 1998. Morgan Kaufmann Publishers.

[4] William A. Gale and Kenneth Ward Church. A program for aligning sentences in bilingual corpora. In *Meeting of the Association for Computational Linguistics*, pages 177–184, 1991.

[5] Péter Halácsy, András Kornai, László Németh, András Rung, István Szakadát, Viktor Trón, and Dániel Varga. Hunglish: nyílt statisztikai magyar–angol gépi nyersfordító. In *II. Magyar Számítógépes Nyelvészeti Konferencia*, pages 81–84. Szegedi Tudományegyetem, 2004.

[6] András Miháczi, László Németh, and Miklós Rácz. Magyar szövegek természetes nyelvi előfeldolgozása. In *Proceedings of the 1st Hungarian Computational Linguistics Conference*. Szegedi Tudományegyetem, 2003.

[7] Philip Resnik. Parallel strands: A preliminary investigation into mining the web for bilingual text. In D. Farwell, L. Gerber, and E. Hovy, editors, *Machine Translation and the Information Soup: Third Conference of the Association for Machine Translation in the Americas*, Langhorne, PA, 1998. Springer.

[8] Philip Resnik and Noah Smith. The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380, 2003.

[9] Jörg Tiedemann and Lars Nygaard. The opus corpus - parallel and free. In *Proceedings of LREC'04*, volume IV, pages 1183–1186, Lisbon, 2004.

[10] Viktor Trón, György Gyepesi, Péter Halácsy, András Kornai, László Németh, and Dániel Varga. Hunmorph: open source word analysis. In *Proceeding of the ACL 2005 Workshop on Software*, 2005.

[11] Dániel Varga, László Németh, Péter Halácsy, András Kornai, Viktor Trón, and Viktor Nagy. Parallel corpora for medium density languages. In *Proceedings of the Recent Advances in Natural Language Processing 2005 Conference*, pages 590-596., 2005.

[12] Trón Viktor, Halácsy Péter, Rebrus Péter, Rung András, Simon Eszter, and Vajda Péter. morphdb.hu: magyar morfológiai nyelvtan és szótári adatbázis. In *III Magyar Számítógépes Nyelvészeti Konferencia*, 2005. megjelenés alatt.