A MetaMorpho projekt 2007-ben – a sorozat vége

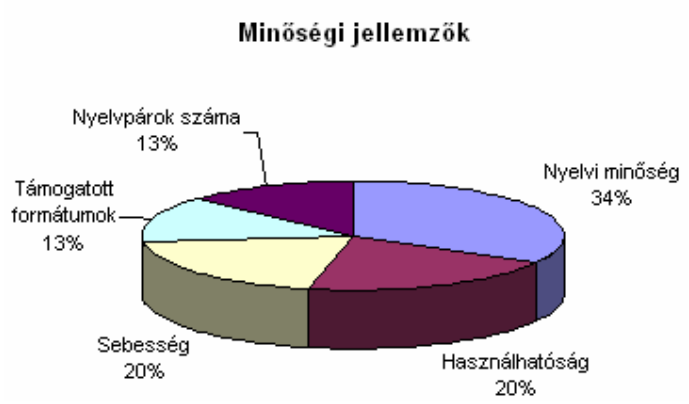
Tihanyi László

MorphoLogic, 1126 Budapest Orbánhegyi út 5
tihanyi@morphologic.hu

Kivonat: Ez a fordítóprogram fejlesztéséről szóló sorozat utolsó előadása. A 2000-ben indult fejlesztésről az első MSZNY konferencián 2003-ban számoltam be először. Az elért eredményeket az idén a programok minőségi jellemzőinek ismertetésével foglalom össze.

A fordítóprogramok minőségi jellemzői

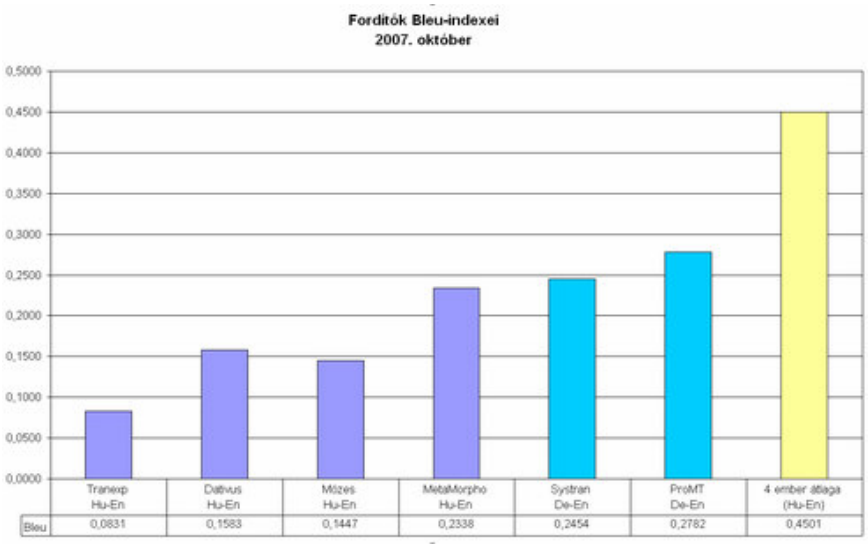
A MetaMorpho rendszert a fordítóprogramok minőségi jellemzőin keresztül vizsgáltam meg. A minőséget meghatározó szempontok között a nyelvi minőség a legfontosabb, de a működő rendszereknek további igényeknek is meg kell felelnie, amelyeket megfelelő súlyozással [5] szokás figyelembe venni.



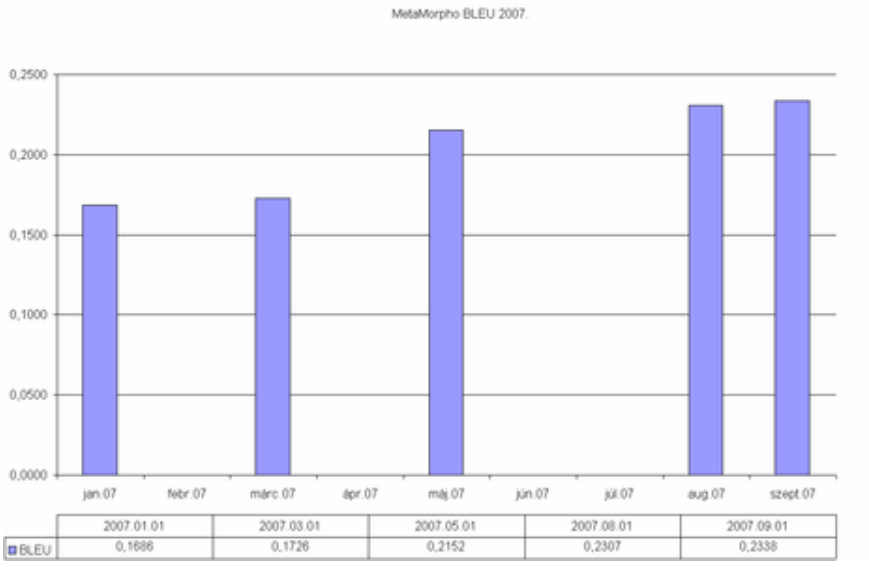
1.1 Nyelvi minőség

A fordítóprogramok minőségét és annak változását Bleu-értékkel mérik. A mérési eljárás lényege röviden, hogy a forrásszövegről több emberi fordítóval referenciafordítást készítenek, majd a gépi fordítást ezekkel a referenciákkal hasonlítják össze. A Blue-index a gépi fordításban lévő olyan 1-től 4-ig terjedő hosszúságú szósorozatok mértani közepe, amelyek megtalálhatók valamelyik referenciában.

Az összehasonlíthatóság érdekében a mérésekhez mi is a legtöbbek által használt NIST implementációt használtuk, és háromreferenciás méréseket végeztünk. A fordítás minőségének vizsgálatát a magyar–angol anyagon januárban kezdtük, és a mérést minden fejlesztés után elvégeztük. A Bleu-indexet elsősorban ellenőrzési céllal készítjük, de ezúttal háromféle összehasonlító mérést is elvégeztünk.
Elsőként összevetettük eredményeinket emberi fordítók teljesítményével. Ez gyakorlatban az adott szövegen elérhető maximális Bleu-érték meghatározását jelenti. A méréshez készítettünk egy negyedik referenciafordítást, majd mind a négy emberi fordítás Blue-indexét, a másik hármat referenciaként használva, meghatároztuk. A négy fordítás számtani középértéke 0,45-re adódott. Az emberi fordítások Blue-indexének eltérése igen alacsony volt, 0,021-es szórással.
Ugyancsak újdonság volt, hogy a magyar–angol Blue-indexet más nyelvű fordítók indexeivel is összevetettük. A kísérlethez a magyar forrásszöveget németre fordítottuk, majd ezt a német forrást különféle programokkal angolra. Azonos tartalmú szövegek, azonos nyelvű fordításait összemérhetőnek gondoljuk. A vizsgálatba a legismertebb Systrant, és a legjobb minőséget adó @promt fordítókat vontuk be. A fordítónk megközelítette a Systran színvonalát (MetaMorpho: 0,2338, Systran 0,2454). Fontos észrevenni, hogy a magyar–angol fordítót két germán nyelv közötti fordító teljesítményével vetettük egybe.
Megmértük más magyar-angol fordítók teljesítményét is: Tranexp: 0,0831, Dativus: 01583. Az eredmények alapján megállapítottuk, hogy rendszerünk jobb, és gyorsabban is fejlődik mint a többiek. Egészen friss az első statisztikai magyar-angol gépi fordító eredménye (Mózes: 0,1447). A program a Hunglish korpusz irodalmi részén lett betanítva.



Az ábrán a magyar-angol MetaMorpho modul ez évi minőségjavulása látható.



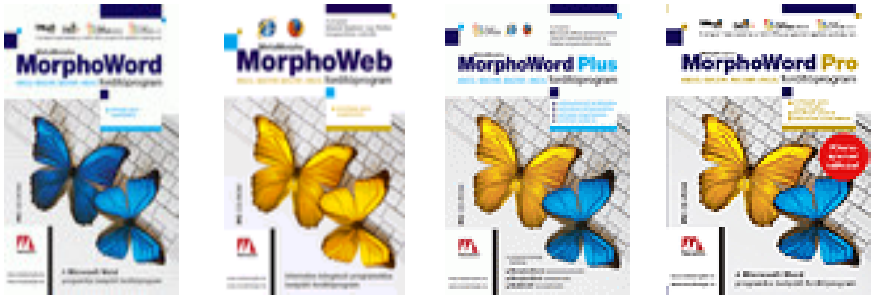
1.2 Használhatóság

A használhatósággal kapcsolatban két fontos kérdésre kell válaszolnunk: kik és mire tudják használni a fordítóprogramot? Az eszköz mai fejlettségi szintjén a megértés segítésére szolgál, elsősorban az anyanyelvre történő fordításkor használható.

Az egyéni, céges és internetes felhasználói köröknek különféle programtermékeket készítettünk. Mindamellett különféle megoldásokra van szükség az egyes alkalmazásokhoz és fájltípusokhoz, az internet böngészéséhez, a dokumentum és egyéb szövegtípusokhoz. Ezeknek kiszolgálására az alábbi termékszerkezetet alakítottuk ki:

	weblap	**dokumentum**	**egyéb szöveg**
egyének	MorphoWeb MoBiCAT	MorphoWord MoBiCAT	-
egyének, csomag		MorphoWord Plus	
cégek		MorphoWord Pro	
internet	webforditas/web	-	webforditas/szöveg

A cégeknek szánt **MorphoWord Pro** változat egy kliens/szerver megoldás, az egy felhasználós kombinált változat neve **MorphoWord Plus**. A céges változathoz egy terminológia kivonatoló programot is kifejlesztettünk. A **MoBiCAT** fordító kiegészítésként szolgál azok számára, akik csak egy-egy mondat fordítását szeretnék megkapni, de a szöveget eredeti formájában kívánják olvasni.



A program használhatóságának legfőbb bizonyítéka a használat. A különböző termékek értékesítési adatait nem közölhetjük, de a webforditas.hu oldal látogatottsági értékei is érdekesek. A webforditas.hu ingyenes szolgáltatás éppen a cikk írásakor (2007. október 22.) lett egyéves, így teljes évi adatokkal szolgálhatunk.

A webforditas.hu weboldalról egy év alatt 40 millió oldalt töltöttek le. 23 millió rövidebb szöveg-, és 2,4 millió weboldal-fordítási kérést szolgáltunk ki. A szótárból további 7,1 millió szó jelentését kérdezték le.

Ez a forgalom szövegmennyiségben is kifejezhető. Egy év alatt 6,3 GB szöveget, azaz 6, 3 milliárd karaktert fordítottak le. Ez 1800 karakteres oldalmérettel számolva 3,5 millió oldalnak felel meg. Ez többszöröse a teljes hazai emberi fordítás mennyiségének, amelyet az irodák forgalmából néhány százezer oldalra becsülhetünk.

A webforditas.hu 2007 szeptemberében több mint kétszázhúszezer látogatót szolgált ki, akik hétszázhúszezer látogatást tettek. Az induló forgalom az év végére megháromszorozódott. A napi látogatók száma októberben húszezer fölé nőtt, ezzel a webforditas.hu az ötven leglátogatottabb magyar weboldal közé került.

1.3 Sebesség

A fordítóprogramok sebességi és minőségi szempontjai ellentétesek. A nyelvi elemzésre annyi időt tervezhetünk amennyit az elfogadható sebességi érték megenged. Az egyre jobb nyelvészeti algoritmusok és egyre gyorsabb számítógépek a közeljövőben újabb minőségjavító eljárásokat tesznek majd lehetővé.

A sebességi adatok egyébként a számítógép memória méretétől, a processzor sebességétől és terheltségétől, valamint a mondatok bonyolultságától függően jelentősen eltérhetnek, ezért az adatok csak tájékoztató jellegűek: angol–magyar 400 karakter/s, magyar–angol 250 karakter/s. A számítógép paraméterei: P4 2,8 GHz, 1 GB RAM.

1.4 Nyelvpárok

A szakmai és laikus vélemények egybehangzóan állítják, hogy egy fordítóprogram annál jobb minél több nyelvet ismer. Mi szakítottunk ezzel a véleménnyel és kizárólag az angol-magyar nyelvpárral foglalkozunk. Az alábbiakban ismertetem miért tesszük ezt, és hogyan gondolom megoldani a soknyelvűségi problémát.

Mindnyájan ismerjük a nyelvi poligont. Ha *N* nyelv között akarunk fordítani, akkor az *N* oldalú poligon átlóinak és éleinek összegével azonos összeköttetést kell létesítenünk közöttük. Valójában az irányítottság miatt kétszerannyi, vagyis *(n(n-1)* megoldásra lesz szükségünk. Ha kiválasztunk egy közvetítő nyelvet, akkor elég azt a poligon csúcsaival irányítottan összekötnünk, vagyis 2**n* nyelvi modul szükséges.
Vizsgáljuk meg az Európai Uniót, melynek 2007. január elseje óta 27 tagországa, és 23 hivatalos nyelve van. Ez 23*22=506 közvetlen kapcsolatot jelentene. Ha csúcsokat a középponttal kötjük össze, és a közvetítő nyelves megoldást választjuk, a szükséges kapcsolatok száma 46 lesz. Csak az utóbbi megoldás megvalósítása tűnik reálisnak.
A probléma az, hogy bár a közvetítő nyelvre sok éve várunk, mégsem nem született meg. Ennek számos szakmai és egyéb oka is van. Mára az interlingva kérdése a gyakorlatban eldőlni látszik. A fordítóprogramok a természetes nyelvekhez hasonlóan természetes nyelveket választanak közvetítő nyelvül. Ez általában (bár nem kizárólagosan) az angol nyelv.
Értékeljük a kialakult helyzetet: A mesterséges közvetítőnyelvekkel kapcsolatban felmerülő, a nyelv kidolgozottságát érintő problémák kiküszöbölését egy világnyelv esetén a beszélők sokasága biztosítja. Az élő nyelvek területileg és időben változnak ugyan, de ez kezelhető mértékű. Nem jelentkezik a nyelvalkotóktól való függőség, hiszen a mű, embermilliók közös kincse, szabadon felhasználható szellemi termék. Ugyancsak nem probléma a fejlesztők előzetes nyelvi képzettsége, mert pl. angol nyelvtudással a fejlesztők rendelkeznek. Az általános érvény és az esélyegyenlőség szempontjai viszont nem teljesülnek. A közvetítő nyelv és a vele rokon nyelvek előnyösebb helyzetbe kerülnek, de ezek a szempontok alulmaradni látszanak.
Vizsgáljuk meg az angolt, mint számítógépes közvetítőnyelvet. Az angol nyelv ismeretét tekintve osszuk az embereket három csoportra: angol anyanyelvűek, angolul tudók és angolul nem tudók.

	Európai Unió (2006)	**Világ (2006)**
Angol anyanyelvűek	13%	5%
Angolt nem anyanyelvként beszélők	38%	8,2%
Angolul tudók összesen	51%	13,2%

Felhasznált adatok:: [6], [7]

Mit történik akkor, ha fordítóprogramok jönnek létre az angol és az Európai Unió nyelvei között? A 13% anyanyelvű számára az előny nyilvánvaló, mert közvetlenül érthetik meg az összes többi nyelvet. Az angolt második nyelvként beszélő 38% számára is elérhetővé válik a többi nyelv, mert az angol és a saját anyanyelvük közötti fordítást nyelvismeretük alapján elvégzik. Például, egy magyar–angol fordító nemcsak

az angoloknak jelent megoldást, de segítségével az angolul jól beszélő németek is el tudják olvasni a magyar weblapokat. A fordítások célja elsősorban a megértés, amelynél nincs szükség arra, hogy az anyanyelven írt szöveg valóban létre is jöjjön. És mit jelent a népesség felét kitevő angolul nem beszélő uniós polgárok számára? Számukra ez a megoldás az angol szövegek megértésének lehetőségét teremti meg. Azaz mindhárom csoport számára hasznos az angol központú nyelvi megoldás.
Várható, hogy az Európai Unió lakosságának nyelvtudása a jövőben intenzíven növekedni fog. 2002-ben elhatározás született arról, hogy minden uniós állampolgárnak lehetőség szerint két nyelvet kell majd megtanulni. Ez a felmérés mutatja, hogy melyek az első és második helyen választott nyelvek. Az angol vált a természetes közvetítő nyelvünkké, és a fenti szempontok miatt, most mint számítógépes közvetítőnyelv is egyre nagyobb tért hódít.

	English	French	German	Spanish	Russian	Italian	Swedish
EU25	**77%**	**33%**	**28%**	**19%**	**3%**	**2%**	**0%**
BE	88%	50%	7%	9%	0%	1%	-
CZ	89%	9%	66%	4%	9%	0%	-
DK	94%	13%	62%	13%	0%	0%	0%
DE	89%	45%	3%	16%	6%	2%	-
EE	94%	6%	22%	1%	47%	0%	1%
EL	96%	34%	50%	3%	0%	6%	-
ES	85%	44%	14%	4%	0%	1%	-
FR	91%	2%	24%	45%	0%	6%	-
IE	3%	64%	42%	35%	1%	4%	0%
IT	84%	34%	17%	17%	0%	0%	-
CY	98%	49%	19%	2%	4%	4%	0%
LV	94%	6%	28%	1%	42%	0%	0%
LT	93%	6%	34%	2%	43%	0%	0%
LU	59%	83%	43%	2%	0%	1%	-
HU	85%	4%	73%	3%	2%	2%	-
MT	90%	24%	13%	2%	-	61%	-
NL	90%	22%	40%	21%	0%	0%	-
AT	84%	29%	2%	10%	4%	11%	-
PL	90%	7%	69%	1%	10%	1%	-
PT	90%	60%	8%	7%	-	0%	-
SI	96%	6%	69%	3%	0%	12%	0%
SK	87%	7%	75%	3%	6%	1%	0%
FI	85%	10%	24%	3%	10%	0%	38%
SE	99%	17%	35%	31%	1%	0%	1%
UK	5%	71%	34%	39%	1%	3%	-
BG	87%	13%	49%	5%	14%	1%	-
HR	82%	5%	69%	2%	0%	14%	-
RO	64%	34%	17%	7%	2%	8%	-
TR	72%	12%	52%	1%	2%	1%	-

= First language = Second language

A gépek nyelvtudása azonban még az emberek nyelvtudásánál is gyorsabb ütemben fejlődik. Egy adott fejlettségi szint után a gépi fordítások összeláncolhatóvá vállnak,

mert a kétszeres fordítás is elfogadható minőséget fog adni. Ezt a lépést a minőségre kevesebbet adó on-line internetes fordítók már egy ideje meg is tették. Ekkor pedig az angolul nem tudók számára is megteremtődik az egyéb nyelveken való kommunikáció lehetősége.
A fentiek alapján, mi a többnyelvűségre való törekvés helyett a közeljövőben inkább az angol-magyar és magyar-angol rendszereink továbbfejlesztésén fogunk dolgozni.

A jövő

A fejlesztés következő, harmadik fázisának célja, hogy a fordítóprogram valóban a mindennapok eszközévé váljon. Ehhez a nyelvi minőség és a funkciók további tökéletesítésére lesz szükség. Ezekről azonban már nem ebben a sorozatban, hanem a problémával foglalkozó konkrét előadások formájában számolunk be.

Bibliográfia

1. Tihanyi László: A MetaMorpho projekt története. Alexin Zoltán; Csendes Dóra (szerk.) *Az 1. Magyar Számítógépes Nyelvészeti Konferencia előadásai*, 247–253. SZTE, Szeged (2003)
2. Tihanyi László: A MetaMorpho projekt 2004-ben. Alexin Zoltán; Csendes Dóra (szerk.) *A 2. Magyar Számítógépes Nyelvészeti Konferencia előadásai,* 85–87. SZTE, Szeged (2004)
3. Tihanyi László: A MetaMorpho fordítóprogram projekt 2005-ben. Alexin Zoltán; Csendes Dóra (szerk.) *A 3. Magyar Számítógépes Nyelvészeti Konferencia előadásai*, 99–107. SZTE, Szeged (2005)
4. Tihanyi László, Merényi Csaba: A MetaMorpho fordítóprogram projekt 2006-ban. Alexin Zoltán; Csendes Dóra (szerk.) *A 4. Magyar Számítógépes Nyelvészeti Konferencia előadásai*, SZTE, Szeged (2006)
5. Hans-Udo Stadler, Ursula Peter-Spörndli: The Quest for Machine Translation Quality at CLS Communication, Proceedings of MT Summit 2007, Coppenhagen

Adatok:

6. Az angol anyanyelvűek (2006): 326 millió, angolul tudók: 860 millió http://en.wikipedia.org/wiki/List_of_countries_by_English-speaking_population
7. A világ lakossága (2006): 6,5 milliárd: http://en.wikipedia.org/wiki/World_population
8. A beszélt nyelvek száma és lefedettségük: az 374 nyelv, amelyet 1 milliónál többen beszélnek (és amely az összes 6912 nyelv 5%-a) lefedi a lakosság 94%-át. http://www.ethnologue.com/ethno_docs/distribution.asp?by=size