

Kedves Szerkesztőség!

Megjegyzés a <http://www.nol.hu/kult/lap-20090204-20090204-30> cikkre.

Örömmel olvastam tegnapi számukban – a NOL-ban február elsején megjelent „előzetessel” – cikküket a Google fordítóról. Tettem ezt azért, mert szakmabeli vagyok, és véleményem szerint a téma figyelmet érdemel.

Ami nem teszi örömmé, az az, hogy számos tévedés szerepel a cikkben.

Mielőtt a tévedésekről beszélnék, némi adalékkal had szolgáljak:

Kipróbáltam a három „használató” fordítót az NOL idevágó cikkén (http://www.nol.hu/tud-tech/magyarra_forditjak_az_internetet). Amit hamar megállapíthattam, a Google nem tudott megbirkózni a NOL oldalának bonyolultságával, a fordításig el sem jutott. Fordítani azért lehetett, a kiemelt szöveget – mint „plain text” adtam be.

Az eredeti szöveg: http://www.nol.hu/tud-tech/magyarra_forditjak_az_internetet

Magyarra fordítják az internetet

Hamarosan magyarra, illetve magyarról is tud majd fordítani a Google "megértéstámogató" programja, a Google Translate. A változás egyfelől jó hír számunkra (hiszen így az idegen nyelvet nem ismerők számára is könnyebben hozzáférhetővé válnak a világháló angl, német stb. tartalmai), másfelől viszont rossz: kiszoríthatja a magyar fejlesztőket egy olyan ágazatból, ahol hagyományosan a főszereplők közé tartoztak.

És most nézzük a fordításokat.

www.webforditas.hu

The internet is translated to Hungarian

Soon onto Hungarian, Google comprehension assistant program, Google Translate can translate from Hungarian then concerned. The change on the one hand good news for us (since in this manner the foreign language not knowing they become accessible much more easily for him the web angl, German stb. his contents), on the other hand on the other hand bad: may supplant the Hungarian developers from a section like that, where traditionally between the main characters owed.

Datívus:

They are translating into Hungarian the Internet

To Hungarian one soon, or off Hungarian one the Google is also able to translate "a megértéstámogató" his program, the Google Translate. The change good for our number from the same direction good for news surely the foreign language not cognising for her also more easily available contents of an angl, German etc get a divorce) the web like this conversely moreover it is bad: it can squeeze the Hungarian developers out a such branch, where traditionally they owed tween the protagonists.

www.translate.google.com

Devoted to the Internet in English

Soon, English, Hungarian and Google also will be able to reverse the "megértéstámogató" program, the Google Translate. The first change is good news for us (because it is known in a foreign language will be more easily accessible on the World Wide Web, angl, German, etc.. Contents), on the other hand, poor: the Hungarian developers kiszoríthatja a sector which is traditionally among the stars.

Szemmel láthatóan egyik fordító sem tudott mit kezdeni a hibásan leírt angl (angol) szóval, de a Google egyéb, szavakat sem ismert fel. Ebben persze a Google módszere és a magyar nyelv bonyolultsága egyaránt ludas. Erre majd még kitérek.

A legérthetőbbnek, a Morphologic Webfordítás ingyenes szolgáltatása tűnik – de azt sem lehet kiadni hivatalos átültetésnek.

Ezek után különböző témájú cikkekből különböző bonyolultságú mondatokkal etettük a fordítókat, angol-magyar és magyar-angol viszonylattal, majd többen egymástól függetlenül értékeltük a fordítást. A teszt objektív, annak ellenére, hogy Morphologic-kal kapcsolatban levő személyek végezték a minősítést – de nem tudták, melyik fordítást melyik program adta. A jósaági sorrend mindkét irányban megegyezik azzal, amit a fenti fordítások is prezentálnak.

Az szembetűnő, hogy a most megjelent „világszenzáció” mennyire lemarad a másik kettőtől. Ennek persze oka van. A statisztikai elven történő fordításhoz nagy minta kell. Ha nagyobb a tér, amelyet le kell fednie, annál hatványozottan több példa, minta kell, hogy minimálisan megbízható eredményt szolgáltatson a statisztika, és a magyar nyelvben levő variabilitás sokkal nagyobb, mint például az indoeurópai nyelvekben. És itt van a Google tévedése. Míg pl. angol-francia viszonylatban garadával lehet találni kétnyelvű anyagot, addig a kis nyelvek, mint a galíciai (gallego), albán, magyar nyelvekre tized annyi anyagot sem lehet begyűjteni. Ezt tetézi nyelvünk nagy variációs lehetősége, amely már szókakok számában is több nagyságrenddel nehezebb feladatot követel a statisztikai feldolgozástól, hát még szavak együttes előfordulásainak gyűjtésétől.

A Google módszerében az a zseniális, hogy nyelvektől független olyan módszert használ, amely szinte emberi beavatkozás nélkül, csupán a felismert szöveghalmazok ismeretében építi ki fordítóját. Ennek minősége bizonyos nyelvpároknál meg is közelíti a sok munkával előállított nyelvtani alapokon nyugvó fordítók minőségét. Egyes nyelveknél, ilyen a finn, a török, a magyar, viszont matematikailag kimutatható, hogyha annyi lefordított anyaggal etetnék, mint amennyi a világon van, és várható az elkövetkezendő 50 évben keletkezik, statisztikai módszerekkel akkor sem nyerhetünk lényegesen jobb minőségű fordítást, a most prezentálnál. Hasonló minőségű fordításhoz magyarra, mint amilyent a google ad angol és francia között, több milliárdszor nagyobb kész fordítást kéne elemeznie, ami nincs, és nem is lesz. Azt a szintet nyilván el fogja érni, hogy a helyes szavak többségét felismeri – ma még ebben is erős hiányosságok vannak –, az olyan nyilvánvaló melléfordítások is el fognak tűnni, minthogy a „magyar” szót „English”-re teszi át statisztikai alapon, de meg sem közelítheti a több éven át kidolgozott szabály alapú fordítás minőségét. Ne higgyék, hogy ez magyar specialitás. Aki tud törökül, próbálja ki, és hasonló lesz a tapasztalata török angol fordításnál. Legfeljebb az a különbség, hogy török nyelvre sokkal gyengébb szabályalapú fordítót talál, mint magyarra, de még az is jobb a google-nál.

Ezek után had térjek át a cikk pontatlanságaira.

1. A Google ingyenes szolgáltatásánál némileg jobbat, korábban a magyar www.webforditas.hu honlapon már régóta elérhet bárki, tehát a magyarok számára annyira nem újdonság.
1. Az a „gyorsteszt” nagyon gyors lehetett, amelyik szerint a magyarról való fordítás jobb, mint a honi két létező eredménye. Ugyan a mi tesztünkben is előfordult olyan mondat, melynek google-fordítása jobban sikerült a másik kettőnél, de ezek oly ritkák voltak. A minőségben a másik irány a jellemző.
3. Hogy a Google fordítója a folyamatos tanulással fejlődni fog – kezdeti időszakban talán látványosan – az lehet, de a statisztika törvényeit át nem lépheti. A korábban írt okok miatt hiú remény, hogy a fejlődés töretlen. Egy bizonyos gyenge szintet nem lesz képes átlépni, hacsak a módszerekben nem találnak valamilyen lényeges újítást.
4. Bár minden nagyobb cég igyekszik a saját eszközeit kifejleszteni, nem várható ennél sokkal jobb eredmény sem az IBM-nél, sem a Microsoftnál. A ma divatos fordítóeszközök piacán – statisztikai fordítás, fordítómemória – a módszerek hasonlóak, a Google-nak viszont kellő anyaga és utólérhetetlen tapasztalata van, hogyan kell meglesni, kigyűjteni, feldolgozni a több nyelvű szövegeket.
5. Ilyen értelemben nem kell féltetni a magyar fejlesztőket, minőségben nem jelent konkurenciát a most megjelent eszköz. Konkurenciát jelentene, ha valóban kéttucatnyi magyar cég lenne, amelyik ezzel a témával foglalkozik, mint ahogy a cikk említi. Bár így lenne. Hogy ez az adat honnan született, nem tudom. Az igaz, hogy akár húsz helyen foglalkozhatnak számítógépes nyelvészeten ezen válfajával, de tudtommal csak egy olyan cég van – az sem akadémiai, amelyik a témával érdemlegesen, kitartóan, eredményeket elérve foglalkozik. A többi vagy velük társul részfeladatok elvégzésére, vagy annyira „akadémiai” a hozzáállásuk, hogy használható eredményre nem számíthatunk.