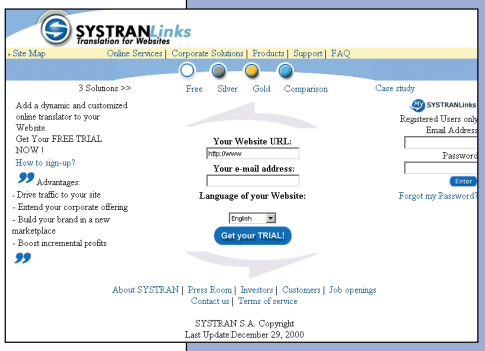
Linuxos feljlesztések a számítógépes fordításhoz

A Systran Internet Translation Technologies a hidegháború alatt született, amikor az Egyesült Államok kormánya nagy mennyiségű orosz szöveget akart gyorsan lefordíttatni. A hatvanas évek végén magánkézbe került a cég és Systranra módosították a nevét, székhelyéül pedig a kaliforniai La Jollát választották.



A kilencvenes években a Systrannál úgy döntöttek, hogy kidobják az MVS-t futtató OS/390-et és az egész rendszert átültetik Unixra. Ekkorra a PC-k már elegendő teljesítménnyel rendelkeztek ahhoz, hogy futtassák a fordítómotorokat. Az Assembly-kód nagy részét önműködő program segítségével fordítottuk le C-re. Elsőként Solarisra ültettük át, de hamarosan olcsóbb gépekre váltottunk, így jött a PC és a Slackware kombinációja (azóta már RedHatet használunk). A Linux választásakor a következő szempontokat vettük figyelembe: többféleféle géptípuson is futtatható legyen; minden programot elérhessenek a fejlesztők, amire csak szükségük lehet; a természetes nyelvek feldolgozása nagyméretű szöveges állományok kezelésével jár, ehhez pedig hatékony programok szükségesek; a fordítómotor összetett szabályrendszert használ, ezért a kód átültetése nagy C, illetve C++ programokat eredményez, ehhez szükség van az olyan hatékony eszközökre, mint a gcc, illetve a g++, valamint a GNU make; a széles közönséget ellátó, például az AltaVistához hasonló méretű ügyfelek számára a legfontosabb az alkalmazás biztonsága, illetve a rendszer megbízhatósága; és ne feledkezzünk el arról sem, hogy minden felhasználó költségkímélő megoldást kíván.

Ezekhez még az is hozzáadódik, hogy az új alkatrészekhez a meghajtóprogramok leggyorsabban Linuxra jelennek meg, továbbá más rendszerekhez képest sokkal kisebb az erőforrásigénye. A Linux beállításai egységesek és könnyen másolhatók további gépekre, valamint a rendszer igen jól méretezhető. A Linuxhoz tűzfal, sendmail, Apache, mod_perl és PostgreSQL is tartozik, ezeket mind használják a cég webes szolgáltatásaihoz (➲ http://www.systranlinks.com/, ➲ http://www.systranet.com/). Emellett a KDE és a GNOME lehetővé teszi, hogy a cég nem programozó alkalmazottai szintén használhassák a Linuxot. Ez azért olyan fontos, mert a Systrannál sokan inkább nyelvészek, mint programozók. Végül a POSIX-szabványhoz való igazodás lehetővé teszi, hogy igény esetén könnyen átültethessük a rendszert más Unixokra is.

A Systran programjai találhatók meg világszerte a legtöbb önműködő fordítórendszer mögött. Ügyfeleink közé nem csak az Egyesült Államok kormányügynökségei sorolhatók, de olyan cégek is, mint az AltaVista, a Microsoft, az Apple, a Lycos vagy az AOL.

A számítógépes fordítás a nyelvészet és a számítástechnika határterülete. A termék fejlesztése valójában az emberi nyelv szabályainak gépi nyelvre (kódra) történő átültetését jelenti. A nehézségek nagy része nyelvészeti, mivel először pontosan le kell írni a kérdéses nyelveket. A folyamat a következőképpen zajlik: elemezni kell a kiindulási nyelvet, majd leírni, azután létre kell hozni a célnyelvet. A kódkészítés négy részre bontható: 1. a kiindulási nyelv elemzése; 2. a célnyelven történő összeállítás; 3. az átviteli szabályok; és 4. a fordítómotorok által közösen használt eljárások (például tárkezelés, parancssor-értelmezés, szótárazó eljárások, szűrők, elő- és utófeldolgozók stb.).

A cégnél különleges szótárakat alkalmazunk. Ezekben nemcsak az adott kifejezés fordítása található meg (pl.: manger = to eat), hanem a kapcsolódó mondattani és lexikai adatok is (például: „ez az ige tárgyas, ebben az adott szövegkörnyezetben ezt és ezt jelenti"). Három különböző szótárat használunk. Az első kettő úgynevezett belső, az egyik az egyszerű szótöveket tartalmazza, a másik az összetett szavakat és kifejezéseket. A harmadik a „külső" szótár. Ez utóbbit az adott ügyfél igényeinek megfelelően állítjuk össze, meghatározott témakörökben.

A Systran olyan segédfájlokkal is rendelkezik, amelyek az igék, a főnevek és a melléknevek ragozásával kapcsolatos szabályokat írják le (ezek természetesen nyelvfüggők), továbbá meghatározzák az elsőbbségi szabályokat az ügyfél szótára számára, valamint megadják szövegsablonokat. Ezt mind C-ben valósítottuk meg, de az újabb programegységeket általában már C++-ban készítjük.

A szabályok előállításához a nyelvészek GTK-alapú grafikus programot használnak. Az adatokat ASCII fájlokban tárolják, ezek egy Perl programon folynak át, így készülnek el az adatokból a kód számára értelmezhető makróutasítások. A szótárak félig önműködő módon épülnek fel, azon szabályok felhasználásával, amelyeket a nyelv elemzése közben állítottunk fel. Először minden nyelvhez egynyelvű mesterszótár készül. A nyelvi rész feltöltése után a Systran-rendszer táblázatok alapján önműködően lényeges nyelvi adatokat ad a bejegyzésekhez. Például az „automatically" angol szót a program határozószónak ismeri fel, mert -ally végződésű. Ezután hozzák létre a kétnyelvű szótárakat, egyszerű kétbejegyzéses listaként, amely a lekérdezésnek megfelelő szintaktikai adatokat az egynyelvű mesterszótárból kéri le.

A szótárakat csak a legvégső lépésnél fordítjuk le bináris alakra, ezzel gyorsabbá tehető a szótárt használó program. Amikor az AltaVista weblapján valaki rákattint a *Translate* gombra, nem is gondol bele, milyen összetett folyamat áll a fordítás mögött.

A Systran tervei között szerepel egy szabad linuxos változat kiadása, amely tudását tekintve megegyezik a Systran Personal windowsos kiadásával.

Thunus F.
A Systran Luxembourg igazgatója

Csúcsfejek a tökfejek ellen

Ez egy jó év a vállalkozók számára. A kutya ott van elásva, hogy a befektetők nem a vállalkozásokat támogatják, hanem a valódi szakmai tudás nélküli bábokat.

Dave Winer