A MetaMorpho projekt története

Tihanyi László

MorphoLogic Kft. 1118 Budapest Késmárki utca 8.
tihanyi@morphologic.hu

Absztrakt. A MetaMorpho fordítóprogram-család a fordítási feladatok teljes palettájára eszközöket kíván biztosítani. A nyelvet nem tudók számára megértéstámogató eszköz, az átlagos nyelvtudásúak számára fordítóprogram, és a profi fordítóknak intelligens fordítómemória készül.

1. Előzmények

A MorphoLogic 1991-es megalapítása óta foglalkozik egy gépi fordítási projekt beindításának gondolatával. A konkrét munka elkezdéséhez azonban meg kellett teremteni a szükséges anyagi és technikai hátteret. Sokáig úgy terveztük, hogy egy már meglévő fordítóprogramnak készítjük el a magyar modulját, és ezért korábban kapcsolatba léptünk több gépi fordító rendszer készítőjével (Logos [1], Systran [2] és LogoVista [3]). Végül – és ma már tudjuk: szerencsésen – a saját rendszer fejlesztése mellett döntöttünk.

2. 2000

A tulajdonképpeni munka 2000. júniusában kezdődött el. Először meghatároztuk az általunk kifejlesztendő rendszer működési elvét. [4, 5] Eszerint a MetaMorpho olyan szabály-alapú rendszer, ahol az elemzési szabályokhoz közvetlenül hozzárendelt fordítások tartoznak. Az így felírt mintapárok tetszőlegesen specifikált elemekből állhatnak, amelyeket ettől függően nyelvtani szabálynak, lexikális elemnek, vonzatkeretnek, vagy egyéb mintának nevezhetünk. Ha a teljes mondat minden szavát lexikálisan megszorítva tároljuk el, akkor a mintaalapú rendszerek működését valósítjuk meg, azaz fordítómemória-szerű működést nyerünk. A MetaMorpho ezek szerint egyszerre RBMT (Rule-Based Machine Translation) és EBMT (Example-Based Machine Translation), azaz egyesíti az MT (Machine Translation) és a TM (Translation Memory) rendszerek tulajdonságait.

Nem köteleztük el magunkat semmilyen ismert formális nyelvészeti elmélet mellett, környezetfüggetlen szabályokat használunk, praktikus kiegészítésekkel. Az elemzés balról jobbra halad, és bottom-up irányban építkezik.

A MetaMorpho kétnyelvű eszköz. Az elemzés eredménye a felcímkézett fa, a generálás a fa top-down bejárása az elemző szabályok generáló párjainak

kiértékelésével. A nyelvtan egyirányú, és bár a minták szimmetrikusak, ez a projekt kizárólag az angolról magyarra fordításra korlátozódik

Az egyszerű architektúra célja a hatékony bővíthetőség. A mintákra épülő rendszer képes arra, hogy tudását futási időben lehessen bővíteni, és a így szerzett ismeretet a sajátjával azonos módon kezelje.

Szeptemberben kialakítottuk az MMO-szabályok szintaxisát. Meghatároztuk az elemzési szimbólumokat, ezek tulajdonságait és értékkészletét. Az ötlet, hogy a szabályokat Clips-ben, egy Lisp-alapú szakértői rendszerben működtessük, Endrédy Istvántól származott. Kezdetben Tihanyi László végezte a rendszer tervezését és kialakítását, Vlaskovits Dóra a nyelvtani szabályok írását, Endrédy István pedig a szakértői rendszer programozását. A Clips azért bizonyult jó választásnak, mert saját elemzőrendszer írása nélkül meg lehetett bizonyosodni arról, hogy a tervezett rendszer működőképes. Ennek az volt az ára, hogy a szabályokat a Lisp által meghatározott formátumúra kellett konvertálni.

A találatok számának csökkentése érdekében még ez évben bevezetésre került egy mechanizmus, amely biztosította, hogy a lexikálisan kitöltöttebb, azaz specifikusabb minták felül tudják bírálni a velük azonos hosszon illeszkedő általánosabbakat. Ezzel el lehetett érni, hogy a rendszer gyakorlatilag mindig csak egy elemzési eredményt adjon.

Az első morfológiai generátornyelvtant – épp a MetaMorphóhoz – októberben készítette el Tihanyi László és Endrédy István. Az év végére kialakult a MetaMorpho laboratóriumi prototípusa.

3. 2001

Az első problémák akkor jelentkeztek, amikor a szabályaink bonyolódtak és a működés nyomkövetése a szakértői rendszer formalizmusában egyre nehézkesebbé vált. Emiatt a lehetővé tettük, hogy a szabályírás a már induláskor felvázolt MMO szintaxisban történjen. Ekkor, márciusban jött létre a szabályok szintaxisának specifikációja. [6]

A Clips-nyelvtanban írt mintáinkat Vlaskovits Dóra áttette saját formalizmusunkba. Tihanyi László konvertereket készített, amelyek a MMO formátumú mintákat először XML-re majd azt Clips-nyelvűre fordították. [7]

Májusban meghatároztuk a rendszer moduljait és a fordítási lépéseket: morfológiai elemzés, morfológiai elemek szintaktikai szimbólumokká konvertálása, szintaktikai elemzés és generálás, szintaktikai elemek morfológiaivá konvertálása, morfológiai generálás. A modulokat szakértői rendszerben is kialakítottuk, és a környezetbe bekötöttük a *Humor* [8] morfológiai elemzőt és generátort. A morfológiai és szintaktikai szimbólumok konverziója ekkor még Lisp nyelven történt.

Készítettünk egy – mintegy húszezer elemű – angol-magyar szótárat, amelynek fontos tulajdonsága volt, hogy minden szónak csak egy jelentése lehetett.

A rendszer 2001. májusára valódi fejlesztő környezetté nőtte ki magát, amelyből a szabályok, az akkor még különálló szótárak és a morfológia is bővíthetővé vált.

A saját formátum bevezetése még nem jelentett megoldást a nyomkövetésre. Kezdte éreztetni a hatását az eredetileg nyilván más célra készült szakértői rendszer

korlátoltsága. A szabályok számának növekedésével a rendszer lelassult. Kis Balázs megkezdte a kialakított adatszerkezetnek és működési elvnek megfelelő szintaktikai elemző, a *HumorESK* [9] fejlesztését.

Az év második felét a nyelvi adatok gyűjtésére fordítottuk. Ugray Gábor először szótárfejlesztési feladatokkal lett megbízva: több szótár összevetésével kiválasztotta az angol legvalószínűbb magyar jelentését. Közben MMO-formátumban elkészült mintegy tizenötezer angol ige vonzatkeret-leírása. Új kódokat kaptak továbbá a magyar névszók is, az eddig még nem kezelt „hol/honnan/hová" kérdésekre adott morfológiai viselkedésük alapján.

A nagytömegű Clips-minta előállítása az XSLT processzorral már kezdett lassúnak bizonyulni, ezért ezt a konverziót DOM-alapú C++ programmal kellett kiváltani. Még az év vége előtt összekötöttük a programot a *MoBiMouse* felhasználói felületével.

4. 2002

Először elvégeztük a morfológiai és szintaktikai modulok összekötését, melyből egy véges állapotú automatával működtetett konverter született. Időközben Endrédy István létrehozta a C nyelvű API-t, így ezzel a modulok önálló programmá váltak és megvalósulhatott a belső adatforgalom. A elemzőmodulok között ezután Tihanyi László XML interfészt definiált: a *format.dtd*-ket.

Az elhúzódó HumorESK-fejlesztés miatt további elemző-fejlesztések is elindultak a cégen belül. A versenyt Ugray Gábor *moose* elemzője „nyerte meg". Jelenleg is ez alkotja a MetaMorpho motorját: a *moose* egyben a köré épült nyelvtanfejlesztő környezet neve is. Nyílt elemzői felületünk lehetővé teszi, hogy más elemzők (pl. a HumorESK) a MetaMorpho rendszerbe illeszkedjenek. Az új elemző új lehetőségeket teremtett a nyelvtanban is, amely ennek megfelelően átdolgozásra került. Ilyenek a voltak: részfa-mozgatás, vagy a nem-terminális elemek beszúrása. Az új rendszer közvetlenül olvassa be az MMO-szabályokat: ezzel véget ért a Clips-korszak.

Ezután bevezettük a források kétszintűségét: létrejött az MMD-formátum. Ez egy MMO-val szintaktikailag azonos forma, mely az olvashatóság érdekében nem tartalmazza azokat a tulajdonságokat, amelyek triviálisan gépi úton is hozzájuk rendelhető (pl. az öröklődő tulajdonságok).

Endrédy István elkészítette a MetaMorpho első telepíthető változatát, a MoBiMouse-felületű *MoBiCAT* programot. Közben a projekt nyelvi- és program-forrásainak archiválására, illetve az egyidejű projektmunkára bevezettük a CVS-rendszert. Bevezettük a projekt belső ellenőrzését is: a szabályok működését a hozzájuk tartozó mintamondatokkal rendszeresen ellenőrizni kezdtük. Augusztusban elindítottuk a szabályok egymás közti és a morfológiai rendszerrel való konzisztenciájának ellenőrzését.

Az ismeretlen angol szavak magyar toldalékolására Novák Attila egy *guesser*-programot fejlesztett ki. [10]

Az igei vonzatkeretek fordítása májustól októberig, azaz fél évig tartott; tesztelése még ma is folyik. A lexikonok is fejlődtek: kifejezés-, névszóielőtag- és más tematikus tulajdonnév-lexikonok készültek.

5. 2003

A felmerült tennivalók kezelésére januárban bevezettünk egy intranetes célprogramot az mmotodo *bugz* rendszert. Ezzel egy időben Gröbler Tamás került a fordítómemória-projekt élére, mely ettől új lendületet kapott: először elkészítette a C++ alapú objektumorientált interfészt, amely a régi C-alapú XML API-t váltotta ki. Emellett több új modul is született, mint például a szövegfájlokon futtatható *documentTranslator,* vagy a szabályok bővítésére készített TM-interfész, Hodász Gábor pedig megoldotta a fordítómemória-bejegyzések tárolását mySQL adatbázisban.

Időközben Ugray Gábor elkészített egy új MMD–MMO konvertert, mely a szabálybővítéseket volt hivatott áttekintehetőbben kezelni. A nyelvtani leírás is folyamatosan bővült: Újvárosi Gábor végzett az időhatározós szerkezetek kódolásával [11]

Júniusban Kundráth Péter először a HTML oldalak fordítását oldotta meg, majd elkészítette a MetaMorpho szerverváltozatát, HTML- és WAP-kliensekkel. A nyáron tovább bővült a csoport Vlaskovits Dórával és Merényi Csabával, akik az alapnyelvtan fejlesztéséért felelősek.

Szeptemberben Újvárosi Gábor átfogó felmérést készített a nyelvtan állapotáról egy hagyományos angol-magyar nyelvtankönyv, a *Huron's Checkbook* [12] alapján. Ezt követően megindult a mostantól már folyamatosan folyó tesztelés, Vancsa László vezetésével. Ő a rendszeresített teszteljárások mellett bevezette az anyag Bleu-teszttel történő vizsgálatát. [13,14]

A Humor-rendszerben működő angol morfológia könnyebb belső kezelésére létrehoztunk egy listaformátumot, amellyel a karbantartás hatékonyabbá vált. Októberben Kundráth Péter és Endrédy István közreműködésével létrejött a második telepíthető MetaMorpho-alkalmazás: az *MmoServer*.

6. A jövő

Célunk, hogy a MetaMorpho olyan fordítóprogram-család legyen, amely a fordítási feladatok teljes palettáján használható eszközöket biztosít: a nyelvet nem tudók számára megértéstámogató eszköz, az átlagos nyelvtudásúak számára fordítóprogram, és a profi fordítóknak fordítómemória-változat készül. A program első, széles körben használható változata az uniós csatlakozásra megjelenő *MoBiCAT* mondatszintű megértéstámogató–fordító eszköz lesz.

A jövő májusig terjedő időszak főbb fejlesztései:

- Többszálú szintaktikai elemző
  A *moose* szintaktikai elemzőt Ugray Gábor alakítja át az internetes felhasználásnak is eleget tevő, többszálú üzemmódra.
- Kliens–szerver fordítórendszer
  A *Microsoft Word*-ön belüli fordítást támogató, szervermegoldásra épülő fordítóprogram első változatán Kundráth Péter dolgozik.

- Automatikus mintageneráló
  Vajda Kristóf szeptemberben megkezdte a szabálybővítő program és felhasználói felületet fejlesztését, melytől a mintabővítés hatékonyságának drasztikus javulását várjuk.
- Szabálykonverter
  Hegedűs Balázs végzi a szabályforrásaink bővítését végző konverter [15] újraírását.
- Szövegszinkronizáló
  Nyár óta folyik a szövegszinkronizáló program fejlesztése Pohl Gábor vezetésével. Októberre elkészült az eszköz motorja, és megindult a felhasználói felület fejlesztése. A modul segítségével fordításokból, párhuzamos szövegkorpuszokból hatékonyan tudunk majd mintatárat készíteni.
- Jelentés-egyérteműsítő
  Miháltz Márton kutatásai eredményeként a 2003-as év végére elkészül a szavak helyes jelentésének kiválasztását végző jelentés-egyértelműsítő eszköz.

7. Összegzés

A MorphoLogic cég 1991-es alapításától kezdve készített minden fontosabb nyelvtechnológiai modul felhasználásra került a MetaMorpho-projektben. A 2000. júniusa és 2003. decembere között zajló fejlesztési időszakban mintegy 30 ember működött közvetlenül közre, ami kb. 220 ember-hónapot jelent. A munka fele-fele arányban oszlik meg a programozás és a nyelvtanfejlesztés között: a program jelenleg mintegy 70 modulból (MSVC-projektből) áll, adatbázisa 110 ezer szabályt tartalmaz.

A projekten ma 13 fő dolgozik. Ezen kívül sok segítséget kap a MetaMorpho-csapat a cég többi munkatársától, illetve a kutatás-fejlesztéshez kapcsolódó egyetemistáktól és doktorandusz-hallgatóktól. A teljes MetaMorpho-projektet kezdektől fogva a MorphoLogic finanszírozza saját erőből.

Referenciák

1. Hawes, R.E.: Logos: The Intelligent Translation System. In: Lawson, V. (ed.) *Tools for the Trade. Proceedings of the Confetrence 'Translating and the Computer 5'.* London: Aslib, 131-139 (1985)
2. Toma, P.: Systran as a Multilingual Machine Translation System. In: *Overcoming the language barrier. Third European Congress on Information Systems and Networks,* Luxembourg,. München: Verlag Dokumentation, 569-581 (1977)
3. Akers, Glen: Logo Vista Conquers Japan *Language Industry Monitor.* 1994 Mar-Apr (1994)
4. Tihanyi László: A fordítóprogram működése *MorphoLogic belső dokumentáció* (2000. június)
5. Tihanyi László: A MetaMorpho formátum. *MorphoLogic belső dokumentáció* (2000. szeptember)

6. Tihanyi László: A MetaMorpho felépítése. *MorphoLogic belső dokumentáció* (2001. május)
7. Tihanyi László: A mmorpho2.dtd. *MorphoLogic belső dokumentáció* (2001. március)
8. Prószéky Gábor & Tihanyi László: A Fast Morphological Analyzer for Lemmatizing Corpora of Agglutinative Languages. *Papers in Computational Lexicography.* Linguistics Institute of H.A.S, 265–278 (1992)
9. Prószéky Gábor: Syntax As Meta-Morphology. *Proceedings of COLING-96,* Vol.2, 1123–1126. Copenhagen, Denmark (1996)
10. Novák Attila, Nagy Viktor, Oravecz Csaba: Magyar ismeretlenszó-elemző program fejlesztése. *I. Magyar Számítógépes Nyelvészeti Konferencia, Szeged* (2003)
11. Ugray, Gábor & Gábor Ujvárosi: English Adverbial NPs of Time in Machine Translation *Proceedings of RANLP*, Tzigov Chark, Bulgaria (2003)
12. Salamon Gábor, Zalótay Melinda (szerk..): *Huron's Checkbook 8000*. Műszaki Könyvkiadó (2001)
13. Papineni K., Roukos S., Ward T. & Zhu W.-J. BLEU: a Method for Automatic Evaluation of Machine Translation. *Research Report*, Computer Science IBM Research Division, T.J.Watson Research Center (2001)
14. Vancsa László: A „BLEU" automatikus kiértékelési eljárás alkalmazása angol-magyar fordítóprogram gyakori, folyamatos minősítésére. *I. Magyar Számítógépes Nyelvészeti Konferencia, Szeged* (2003)
15. Hegedűs Balázs: Szabálykonverzió a MetaMorpho rendszerben. *Szakdolgozat*. BME Villamosmérnöki Kar (2003)
16. Prószéky, Gábor & László Tihanyi: MetaMorpho: A Pattern-based Machine Translation Project. In: *Proceedings of the 24th 'Translating and the Computer' Conference.* London, United Kingdom, 19–24 (2002)

Függelék: A fejlesztésben közreműködő munkatársaink

Aggod Zsuzsa, Csordás Attila, Dominus Ákos, Endrédy István, Földes András, Grőbler Tamás, Hegedűs Balázs, Hodász Gábor, Keresztes Máté, Kis Balázs, Kiss Gabi, Kiss Márton, Kozma Andrea, Kundráth Péter, Légrádi Ágnes, Magyar Dóra, Merényi Csaba, Miháltz Márton, Novák Attila, Pál Miklós, Pohl Gábor, Prószéky Gábor, Tihanyi László, Tőkés Tamás, Ugray Gábor, Újvárosi Gábor, Vancsa László, Vajda Kristóf, Vlaskovits Dóra