Hunmorph-foma morfológiai elemző és generátor a magyar nyelvre.

Az elemző és generátor a foma eszközön alapul (http://code.google.com/p/foma), amely a Xerox laboratórium lexc/xfst nyelvtani analizátor eszköz nyílt forráskódú implementációja. A Xerox laboratórium eszközeinek kézikönyve (fsmbook néven is ismert) a https://victorio.uit.no/langtech/tags/DIVVUN-NO_1_0_RELEASE/gt/doc/book.pdf_1.pdf címen található meg.

Ismert nyílt forráskódú, szabadon használható eszközök a magyar nyelvhez:

A *hunmorph* eszköz (http://mokk.bme.hu/resources/hunmorph/) több komponensű eszköz. Csak analízis irányban működik, azaz az „ablakot” szó alapján megtalálja a szógyököt és a tárgyragot, de generálás irányba nem működik, azaz nem képes az „ablak ACC” utasításra az ablakot szót előállítani. Szótára nem módosítható egyszerűen, mert az ahhoz szükséges eszközök nem dokumentáltak. Szókészlete töredéke a Hunmorph-fomának. Automatizálása nincs megoldva, az esetleges felhasználónak magának kell a kiszolgáló, felhasználó modellt elkészítenie. Nekem lassúnak és memóriaigényesnek tűnt a működése, ami általában nem előnyös semmilyen alkalmazásnál.

A *hunspell* (http://sourceforge.net/projects/hunspell/)a lehetséges ragozási alakok kb. 60%-ával képes megbirkózni, a többit nem ismeri. Szókincse szegényes, sok helyes magyar szó hiányzik belőle. Működése nyelvtani analízis módban instabil, pontosítva ebben a módban gyakran rohad le. Támogatása ebben a módban nem létező, az erre vonatkozó hibajelentések általában válaszolatlanul maradnak vagy törlésre kerülnek. A http://bug.openscope.org hibakövető rendszeren megtekinthetők a hunspell hibajelentései, melyek a problémákat példákkal illusztrálják.

A fenti hátrányok kiküszöbölése indokolja a Hunmorph-Foma előállításának értelmét és szükségességét.

Hunmorph-foma használatának előfeltétele foma (http://code.google.com/p/foma) lehozatala és sikeres installálása.

Használata:

1. Hozza le a fát és bontsa ki.
2. Ahol a "Makefile" található,
   sh tools/chksys.sh

Ha a válasz pozitív (foma és make installálva vannak):
   make
A make parancs kiadása Hunmorph-foma lefordítását eredményezi, a fordítási idő nálam 5-6 perc.

Az első tesztelés:

$ foma -l hfnnum.foma
27.3 MB. 599957 states, 1788091 arcs, Cyclic.
Foma, version 0.9.16alpha
Copyright © 2008-2011 Mans Hulden
This is free software; see the source code for copying conditions.
There is ABSOLUTELY NO WARRANTY; for details, type "help license"

Type "help" to list all commands available.
Type "help <topic>" or help "<operator>" for further help.

foma[1]: up
apply up> kilincs
kilincs+Noun+Nom
apply up> hajókra
hajó+Noun+Plur+Sub
apply up> kellet
kellet+Verb+IndefSg3
apply up> hibás
hibás+Noun+Nom
hibás+Adj+Nom
apply up> és
és+Con
apply up> (Ctrl-D lenyomása)
foma[1]: down
apply down> kilincs+Noun+Posss3s+Dat
Kilincsének
kilincsének
apply down> (Ctrl-C lenyomása)

A Hunmorph-fomával természetesen program segítségével is lehet társalogni, azaz működése teljes mértékben automatizálható, részletesebben erről: http://code.google.com/p/foma/wiki/FlookupDocumentation, és Kliens-szerver mód hátrébb.

Hunmorph-foma alkalmazási lehetőségei sokoldalúak, többek között:

- helyesírás ellenőrzés
- fordító támogatás
- szótövezés
- korpusz analizálási és előállítási segédeszköz
- szövegek átformálása
- ...

Hunmorph-foma szófajai:

- főnév
- melléknév
- ige
- számnév
- névmás
- határozó, kötőszó, névutó, egyéb
- ...

Ragozás

Főnevek ragozása így zajlik le a magyarban:

					Példa
szó	birtokrag	családias rag	birtokosi rag	ragok	sógoromékéra
szó	többesszám ragja		birtokosi rag	ragok	sógorokénak
szó		családias rag	birtokosi rag	ragok	sógorékéitól

Hunmorph-foma jelölései

					Példa
szó	birtokrag (Poss)	családias rag (Fam)	birtokosi rag (Gen)	ragok	sógoromékéra
szó	többesszám ragja (Plur)		birtokosi rag (Gen)	ragok	sógorokénak
szó		családias rag (Fam)	birtokosi rag (Gen)	ragok	sógoréktól

A ragok jelölésére Hunmorph-foma a következő rövidítéseket használja, hasonlóan a Szeged-Korpuszhoz (http://www.inf.u-szeged.hu/projectdirs/hlt/), a Szószablyához vagy a Hunspellhez:

Használt név	**Rag**
Abl	tól, től
Acc	t, ot, et, öt
Ade	nál, nél
All	hoz, hez, höz
Cau	ért
Dat	nak, nek
Ela	ból, ből
Ess	ul, ül
Fac	vá, vé
For	ként
Forp	képp
Forpen	képpen
Ill	ba, be
Ine	ban, ben
Ins	val, vel
Nom	
Sub	ra, re
Sup	on, en, ön, n

Használt név	Rag
Ter	ig
Tem	kor
Dis	anként, enként
Soc	astul, estül

Dis és Soc-t csak közvetlenül a szóra alkalmazzuk, nem Poss, Plur vagy Fam után.

Gen alkalmazása csak egyszer történik, noha elméletileg tetszőleges mélységig alkalmazható (kutyáéé..), az alkalmazott korpuszok (Szószablya korpusz és Szeged korpusz) ilyenre nem szolgáltak példával.

Melléknevek

A melléknevek a főnevekkel analóg módon ragozódnak. Mellékneveket ezen kívül fokozni is lehet, amikor felső és túlzófokban előtagot is kapnak.

Melléknév fokozás:

Alapfok	Középfok	Felsőfok	Túlzófok
Szó (fnragozás)	abb, ebb, obb, bb (fnragozás)	leg .. abb, ebb, obb, bb (fnragozás)	legesleg .. abb, ebb, obb, bb (fnragozás)

A fokozás tényét a "Mid" szóval jelzi Hunmorph-foma, a felső és túlzófok jelenlétét a leg vagy legesleg előtag megadása mutatja.

Példák:
szebbek: szebbek szép+Adj+Mid+Plur+Nom
legokosabbak: legokosabbak legokos+Adj+Mid+Plur+Nom
legeslegcsúnyábbak: legeslegcsúnyábbak legeslegcsúnya+Adj+Mid+Plur+Nom

Igék

Igék ragjai egyes szám első személytől harmadik személyig, és többes szám első személytől harmadik személyig tartanak.
Ige állhat főnévi igenévi alakban is, amely lehet egyes szám első-harmadik személyű vagy többes szám első-harmadik személyű.
A személyt és számot kifejező rag előtt még állhat a felszólítás (tat,tet) illetve az ismétlés (gat,get) képzője.
Az ige lehet tárgyas vagy tárgyatlan ragozású, jelen vagy múlt idejű, feltételes ragozású vagy kötő módban levő.

Példa:

						Példa
Szó	ismétlés	felszólítás	ismétlés	felszólítás	szám, személy, tárgyatlan	írogattatgattatok
Szó	ismétlés	felszólítás	ismétlés		szám, személy, tárgyatlan	írogattatgatok
Szó	ismétlés	felszólítás			szám, személy, tárgyatlan	írogattatok
Szó	ismétlés				szám, személy, tárgyatlan	írogatok
Szó				felszólítás	szám, személy, tárgyatlan	íratok
Szó		felszólítás	ismétlés		szám, személy, tárgyatlan	íratgatok
Szó					szám, személy, tárgyatlan	írok
Szó					szám, személy, tárgyas	írom
Szó					szám, személy, tárgyas, feltételes	írnám
Szó					szám, személy, tárgyatlan, feltételes	írnék
Szó					szám, személy, tárgyas, kötőmód	írjam
Szó					szám, személy, tárgyas, kötőmód	írjak
Szó	ismétlés				Szám, személy, tárgyas, múlt	írogattam
Szó		felszólítás			Szám, személy, tárgyatlan, múlt	írattál

Hunmorph-foma jelölései

- Az ismétlés jele a Rep,
- A felszólítás jele Imper.
- Tárgyas: Def Tárgyatlan: Indef
- Feltételes: Cond
- Kötőmód: Conj
- Múlt idő: Past

Számnevek
A számnevek ragozása teljesen azonos a főnevekével

Névmások

A névmások egyes szám első..harmadik és többes szám első.. harmadik személyben lehetnek.

Névmások kapcsolódnak szinte minden főnévi raghoz (rám, rád, ...), és esetenként névutókhoz is (alattam, alattad, ...).

Példák:

- enyém
- mellettünk
- tőletek
- ....

Határozók, kötőszavak, stb..
Határozók, kötőszavak és egyéb szófajok ragozatlanok.

Rendhagyó főnevek

A rendhagyó főnevek osztályai:

1	bokor, bokrot, bokrok	hangzókihagyás
2	sátor, sátrat, sátrak	hangzókihagyás, o-a hangzóváltás
3	hal, halat, halak	o-a hangzóváltás
4	csík csíkot csíkok	mély végződést kapó magas hangrendű vagy magas utolsó szótagú szavak
5	írt írtat írtak	mély végződést kapó magas hangrendű vagy utolsó szótagú szavak o-a hangzóváltással
6	levél, levelet,levelek	ragozásnál ékezetelhagyó szavak
7	derék, derekat,derekak	ragozásnál ékezetelhagyó szavak mély végződéseket követelve, noha a szó magas hangrendű
8	team, teamet teamek	magas hangrendű idegen szavak, melyeket ragozunk
9	hó, havat, havak	szótőváltó szavak
10	ló lovat lovak	szótő módosulással ragozandó szavak
11	erő, erejét erők	poss ragoknál szóvégi hangváltó szavak
12	fi fiat fiak	fi típusú szavak
13	ín inat inak	ín típusú szavak
14	bakter, bakterek, bakterok	ingadozó ragozású szavak, mély és magas ragozás is jó
15	öcs, báty néne	öcs, báty, néne szavak
16	szülő szüleim szülőim	szülő szó
17	tűz tüzet tüzek	tűz, fűz, szűz szavak, ékezetelhagyóak
18	y-i	bizonyos neveknél a szóvégi y i-nek tekintendő
19	y-y	bizonyos neveknél szóban szereplő y i-nek tekintendő, pl Ghyczy
20	ae	általában poss jelölésére ae és jaje is jó. Bizonyos szavaknál csak ae használható
21	jaje	általában poss jelölésére ae és jaje is jó. Bizonyos szavaknál csak jaje használható
22	névmások	enyém.. övéik magam.. maguk szavak

Rendhagyó főnevek esetén az összes összetett szó, melyek utótagja a rendhagyó szó, a rendhagyó csoportba sorolandó be. Például a hintaló a lóval azonos .lexc fájlba sorolandó be.

Rendhagyó igék

Az igéknél kb. 25 féle ragozású ige létezik. Az ikes igék külön csoportban vannak az egyszerűség kedvéért, pl enhuige7.lexc és enhuige7ik.lexc. A nem rendhagyó igék az enhuige.lexc fájlban vannak.

Rendhagyó melléknevek

1	bátor, bátrabb
2	sok-több, lassú-lassabb, nehéz-nehezebb, könnyű-könnyebb, szép-szebb, hosszú-hosszabb, bő-bővebb, ifjú-ifjabb

Rendhagyó számnevek

1	Ezer, ezrek	hangzókihagyás

Igekötők

Igekötők nemcsak igékhez járulhatnak, hanem főnevekhez és melléknevekhez is (kinézés, benézés, kilátó, belátó, stb...)

Az igekötők szavakhoz adása a LEXICON Prefigekoto blokknak a LEXICON Verb blokk elé kapcsolásával valósítjuk meg.

Azaz:

```
LEXICON Root
Verb;
Prefigekoto;

LEXICON Prefigekoto
+Pref+:keresztül Verb;
...
+Pref+:be   Verb;

LEXICON Verb
abajgatásoz AddVerbmelyik;
...
```

Ez a szerkezet garantálja, hogy minden igekötő minden igével kombinálva lesz. Ilyen szerkezet van még az enhuadjigem.lexc, enhuadjmelyige.lexc, enhuadjfnmn.lexc, enhufnige.lexc, az összes igével kapcsolatos és más fájlokban is, ahol gyakori az igekötővel való kombináció.

Új szavak beadása.

Meg kell állapítani, hogy a beadandó szó főnév, melléknév, melléknév és főnév, ige, számnév vagy egyéb-e.

Meg kell nézni, hogy hasonló végződéssel van-e már szó a szókészletben.

A hozzá leghasonlóbb szó mellé betenni, ugyanolyan ragozási utasítással, mint amilyen a hasonló szóé.

Ha a szóhoz igekötők is tartoznak, olyan helyre kell betenni, ahol az igekötőket is hozzáadjuk.

Példa:
Tegyük fel, a facsart szó hiányzik. A szó melléknév. Igekötők tartozhatnak hozzá (kifacsart, stb..)

grep "art " lexc/adj/*
enhuadjigem.lexc jó jelölt, mivel ez az igekötőket is kezeli

a csavart szó így van benne:
csavart AddAdjboeet;

tehát ez után a szó után berakjuk a következő sort:
facsart AddAdjboeet;

és újra kompilálás után a szó része a szókészletnek.

Képzős ragozási utasítások

Az adj fában előfordulnak képzős ragozási utasítások, pl. enhuadjfnmn.lexc-ben:
AddAdjeEnd
AddAdjeEndat
AddAdjeEndnyi
AddAdjeEndnyiat

ragozási utasítások, melyek kizárólag képzőket állítanak elő (szerű, féle, fajta, nyi, i); Ezzel sok helyet lehet megtakarítani, viszont az analízis kissé áttekinthetetlenebb lesz, mert pl. a próba szó esetén a próbaszerű illetve próbaféle szavak esetén az alapszót próba-nak adja meg és nem próbaszerűnek.

Példa:
apply up> próbaszerűt
próba+Adj+Acc
apply up>

Összetett szavak kezelése

Az összetett szavak kezelését eredetileg úgy terveztem, hogy egyszerű szavak listáját ragozási utasítással "fésülöm össze". Az egyszerű szavak listája kb. 30 ezer szót tartalmazott volna. Ez gyakorlatilag nem volt járható út, mert a .lexc fájl lefordítása több, mint 30 percet vett volna igénybe.

Ezért saját gyűjtésű korpuszból vett listát használok. A korpuszból vett szógyűjtemény a noun/enhuossz3.lexc fájlban található meg. A szógyűjtemény nem szabályos szavait ezután kézzel átraktam a megfelelő .lexc fájlokba. A tesztelést és a szógyűjtemény javítását ezután magával a szógyűjteménnyel, a Szeged korpusz szókészletével majd a Szószablya projekt szókészletének felhasználásával végeztem. Minden esetben a Hunspellt használtam szűrőként.

Mivel az enhuossz3.lexc fordítása kb. 3 percet vesz igénybe, ez lelassítja a fordítást. Esetleges szókészleti munka megkönnyítésére bevezettem a "nagy" és "kis" szótár használatának lehetőségét. A "nagy" szótár a végleges forma, a "kis" szótár a szabályos összetett főnevek nélküli forma. Az átkapcsolást a nagy és kis szótár között automatizáltam az smake.sh szkript segítségével.
használata:
./smake.sh k
kicsi

vagy:
.smake.sh n
nagy

parancsok segítségével lehetséges. Ezután a make a kiválasztott változatot fogja lefordítani.

A módszer előnye a rendkívül egyszerű átkapcsolás, és ennek során esetleges hibák elkerülése.
A módszer hátránya az, hogy létezik egy enhu2.foma.kicsi és egy enhu2.foma.nagy fájl, és ha az enhu2.foma fájlt változtatjuk a változást az enhu2.foma.kicsi és az enhu2.foma.nagy fájlon is végre kell hajtani.
Ugyanígy létezik egy Makefile.kicsi és Makefile nagy fájl. Minden változást a Makefile fájlon ezekre is alkalmazni kell.

Továbbfejlesztési ötletek:

Analízis esetén célszerű, ha az eszköz lehetőleg minden lehetséges alakot fölismer. Generálás esetén viszont az kívánatos, hogy a legszokásosabb alakot generálja az eszköz. A jelenlegi forma analízisre jó, de generálásra nem optimális, mivel sokféle formát generálhat. A "legszokásosabb" forma eléréséhez a lexikont célszerű redukálni és a lexc fájlokat leegyszerűsíteni, hogy ne generálják a ritkán előforduló formákat. Erre példa az írnánk/írnók pár, ahol kétségtelenül az írnánk forma lényegesen gyakoribb, de az írnók formát sem lehet kizárni, mert az a analízis esetén a megértési pontosságot fölöslegesen csorbítaná.

A Hunmorph-foma megírásának története

A Hunmorph-foma készítése az SFST nevű eszközzel indult, (http://www.ims.uni-stuttgart.de/projekte/gramotron/SOFTWARE/SFST.html) mivel az SFST áttekinthető analizátor/generátor, és segítségével már elkészült egy török morfológia, trmorph. http://www.let.rug.nl/~coltekin/trmorph/. A török nyelv meglehetősen hasonlít a magyarra, kb. német:angol megfelel a magyar:török viszonynak. A török morfológiát lényegesen leegyszerűsíti, hogy a török ragozás kicsit egyszerűbb, bár szinte minden szabálytalanság megvan benne mint a magyarban. A török igeragozás ezzel szemben kissé komplikáltabb, mint a magyar, ez viszont nem nehezíti meg jelentősen a morfológiai eszköz írását. Viszont a török nyelv nem használ sem igekötőket, sem összetett főneveket, ami egy morfológiai eszköz írását nagyban leegyszerűsíti.

Az SFST használata során ugyan sikerült kitalálni a csíziót, azaz megtalálni azt a módozatot, ahogy

elfogadható memória használattal működjön az eszköz. A probléma az olvashatóság lett. Az SFST fájlok olyan nehezen olvashatók voltak, hogy kénytelen voltam új eszközt keresni, és ez a foma volt. A foma tökéletesen alkalmas a magyar nyelv komplexitásának a teljes körű kezelésére, és azt nagyon jó hatásfokkal teszi.

Segédeszközök

LEXC ellenőrzés:

A lexc fájlok ellenőrzésére írtam egy kis programot, tools/fomachk/fomasyn2.pl, amely gyorsan megvizsgálja a lexc fájlok integritását. A program, mint minden segédprogram perl-ben íródott, és működéséhez működő perl interpreter jelenléte szükséges, ami modern rendszereken magától értetődő.

Első használat előtt a fomasyn2.pl programban levő $root változót, úgy kell beállítani, hogy az arra a mappára mutasson, ahol a Makefile van.

Egy paramétert kell megadni fomasyn2.pl-nek, ami a lexc-ben levő szófaj neve, azaz noun, adj, verb, num, fxpp, vagy misc.

Használata ezután a fő mappából:
```
 $ perl tools/fomachk/fomasyn2.pl verb
 $
```
fxpp paraméter esetén panaszkodik nem használt elemekről, ami ártalmatlan. Esetleges más panaszait viszont érdemes komolyan venni és utánuk nézni.

Kliens-szerver mód

Egy példa a kliens-szerver mód használatára a tools/chkwds mappában van. Használata:

1. a do_testup.sh fájlban a gyökérmappa beállítása (nálam a gyökérmappa: /home/en/program/foma/tktest1, ezt kell beállítani).
2. sh do_testup.sh
a rendszer válasza: Started flookup server on 127.0.0.1 port 6062
3. perl chkwdlistup.pl szolista.txt
```
kapunak: kapunak      kapu+Noun+Dat
ajtótól: ajtótól ajtó+Noun+Abl
megírják: megírják     +Pref+ír+Verb+DefPl3
megírják         +Pref+ír+Verb+ConjDefPl3
szebbek: szebbek       szép+Adj+Mid+Plur+Nom
legokosabbak: legokosabbak legokos+Adj+Mid+Plur+Nom
legeslegcsúnyábbak: legeslegcsúnyábbak     legeslegcsúnya+Adj+Mid+Plur+Nom
jó: jó    jó+Adj+Nom
laskagombáért: laskagombáért          laskagomba+Noun+Cau
send:  at chkwdlistup.pl line 28, <FILE> line 9.
```

Segédprogramok

A tools/perl mappában levő programok:

Programok a szólisták karbantartására

kivon.pl - hívása egy paraméterrel történik, a paraméter egy fájl neve
kivon.pl beolvassa a megadott fájlt.
Beolvassa stdin-t, a fájlban szereplő szavak kivételével listázza stdin-t.

kivon1.pl kivon.pl megfordítottja:
kivon1.pl - hívása egy paraméterrel történik, a paraméter egy fájl neve
kivon.pl beolvassa a megadott fájlt.
Beolvassa stdin-t, listázza stdin azon szavait, melyek a megadott fájlban szerepelnek..

Hivatkozások jegyzéke

- http://code.google.com/p/foma
- https://victorio.uit.no/langtech/tags/DIVVUN-NO_1_0_RELEASE/gt/doc/book.pdf_1.pdf
- http://mokk.bme.hu/resources/hunmorph/
- http://sourceforge.net/projects/hunspell/
- http://bug.openscope.org
- http://code.google.com/p/foma/wiki/FlookupDocumentation
- http://www.ims.uni-stuttgart.de/projekte/gramotron/SOFTWARE/SFST.html
- http://www.inf.u-szeged.hu/projectdirs/hlt/
- http://www.let.rug.nl/~coltekin/trmorph/