Literumilo por Esperanto

Bc. Marek BLAHUŠ <marek@ikso.net>

E@I, Nancio (Francio) /
Masaryk-Universitato, Brno (Ĉeĥio)

KAEST 2010, Modra, 2010-11-21

Origino de la literumilo

- E@I (“Esperanto@Interreto”)
- Subvencio de ESF por gramatika kontrolilo de esperanto (2007)
  - projekto “Lingvohelpilo”
  - antaŭ gramatika kontrolo necesas literuma (ortografia) kontrolo

Universitata laboraĵo

- Masaryk-Universitato, Brno (Ĉeĥio)
  - patrono: RNDr. Petr Sojka, Ph.D.
- A Spell Checker for Esperanto
  - bakalaŭra studlaboraĵo en aplikata informadiko (komputila prilaboro de natura lingvaĵo)
    - defendita: 2008-06-26
  - Trarigardo de ekzistantaj solvoj, ellaboro de prototipo de novspeca solvo
  - 29 paĝoj da angla teksto kaj 7 paĝoj da aldonaĵoj (+34 fontindikoj)
    - Elŝutebla de: http://is.muni.cz/th/172464/fi_b/?lang=en
- Stipendiata esplorprojekto (2008/12-2009/07)
  - studentaj esploraj kaj evoluaj projektoj (MUNI33/212008)
  - finpretigo de la literumilo kaj integrigo en OpenOffice.org
  - retpaĝaro: http://nlp.fi.muni.cz/projekty/esperanto_spell_checker/

Ekzistanta programaro

- Literumiloj dediĉitaj (por Esperanto)
  - Kontrolu Literumadon (1992) de Klivo Lendon
  - Ĉapelilo (1995) de Simono Pejno
  - Esperantilo (2003) de Artur Trzewik
- Literumiloj universalaj (lingve sendependaj):
  - Ispell (1971) por Unix
  - GNU Aspell (1998) por Projekto GNU
  - MySpell (2000) por OpenOffice.org
  - Hunspell (2005) por la hungara lingvo

Ekzistanta vortaro

- Sergio Pokrovskij (Ispell, Aspell) –> Dmitri Gabinski (MySpell, Hunspell)
- .dic: 19.342 eroj kun atributoj
- .aff: 58 afiksoj (34 prefiksoj kaj 24 sufiksoj) kun entute 2.426 rikordoj
- Ne estas dokumentitaj la signifoj de la afiksoj kaj ties kombinaĵoj

Necesaj datumoj (Hunspell)

- Por ĉiu lingvo unu „vortaro“
  = vortlisto + afiksaj reguloj
- Vortlista dosiero `lingvo.`**`dic`**
  - Listo de vortoj kaj vorteroj kun markoj (=atributoj)
- Afiksregula dosiero `lingvo.`**`aff`**
  - Prefiksoj kaj sufiksoj, kiuj korespondas kun la markoj
  - Reguloj priskribas kiel eblas uzi kaj kombini ilin

Klarigo pri afiksaj reguloj

```
SFX D Y 4
SFX D   0   d e          Ekz.: breathe > breathed
SFX D   y ied [^aeiou]y  Ekz.: cry > cr > cried
SFX D   0  ed [^ey]      Ekz.: want > wanted
SFX D   0  ed [aeiou]y   Ekz.: play > played
```

- Klarigo pri la kolumnoj:
  1. Mallongigo: SFX D = Sufikso „D“
  2. Forprenataj finaj literoj (0 = neniuj)
  3. Aldonataj finaj literoj
  4. Kondiĉo pri finaj literoj:
     [abc] = unu el tiuj, [^abc] = ĉiu krom tiuj

Ekzempla „vortareto“ por Esperanto

ekzemplo.aff:

```
SET UTF-8
TRY
   oaeinsrltkumvjdpbĉgfcĝAŭK
   MzNhŝSPLDUIVJOTEFHBĵĈRŜCĜ
   ZĥGĴĤŬ'-

PFX D Y 1
PFX D 0 dis .

SFX A Y 1
SFX A i a .

SFX O Y 1
SFX O i o .
```

ekzemplo.dic:

```
3
kaj
esti/OA
doni/OAD
```

rezultoj:

```
kaj
esti        doni     disdoni
esto        dono     disdono
esta        dona     disdona
```

Sugestado de korektoj

- Ĉu ne en vortlisto? Ĉu ne kunmetaĵo? → Proponu korekton
- Levenshtein-distanco
  - Operacioj: Aldoni, Forigi, Anstataŭigi
- Ekzemplo de malkorekta vorto *frao:
  - Levenshtein (frao, frato) = 1
  - Levenshtein (frao, fora) = 2
  - → „frato“ estas pli verŝajna korekto

Ideoj por la nova vortaro

- Atento al la nesenfina uzeblo de afiksoj
- Malaltigo de la suma nombro de afiksregulaj rikordoj profitante la avantaĝojn de Hunspell
- Komparo kun formalaj priskriboj de Esperanto-morfologio, kompletigo kaj precizigo
- Nova, pli ampleksa vortlisto
- Ĝeneraligo de la permesita vortofarado (afiksoj)
- Specifaj trajtoj de ortografia kontrolado por esperanto (tipaj mistajpaĵoj – c/ĉ;
  oftaj eraroj – delfino/delfeno, ...)

Metodologio

- Trovo de taŭga morfologia aliro al Esperanto
- Eltiro de radikoj el elektronika vortaro
- Semantika klasado de radikoj
  - Propono de klasado (sistemo)
  - Aŭtomata klasado (procezo)
- Elkreo de afiksaj reguloj
  - Analizo de jam-pretaj kunmetaĵoj
- Implemento en Hunspell

Konkreta realigo

- Ĉefaj fontoj de datumoj:
  - PIV (el. Versio) – listo de 16.780 radikoj
  - PMEG – gramatiko
  - Tekstaro de E. Bick – 18,5 milionoj da pozicioj
- Diagramo de proponita klasado
- Moore-maŝino por klasado de radikoj
- Afiksreguloj permane surbaze de la klasado
- Regulaj esprimoj en Hunspell

Elektita morfologia aliro

- Surbaze de PMEG (10 prefiksoj kaj 31 sufiksoj)
- Strukturo de rekonata vorto (kunmetaĵo):

  (afikso* radiko afikso* ligilo?)* afikso* radiko afikso* finaĵo?

- Ekzemploj:
  - `MAL - SAN - UL - EJ | DOM | O`
    `afks radik afks afks radik finaĵ`
  - `ŝAJN | MULT - E | KOST | A`
    `afks  radik ligil radik finaĵ`

Semantika klasado de radikoj

- Ne ĉiuj kombinaĵoj radiko + afiksoj estas validaj
- Ekzemplaj limigitaj afiksoj: “bo” (familio), “pra” (du signifoj)
- PMEG
  - Ĉap. 38: Afiksoj (ekz. “objektaj agaj vortoj”)
  - Ĉap. 37: Vortelementoj kaj signifoj (“ĉiu radiko havas jam per si mem signifon ... Ekzistas multaj diversaj grupoj kaj kategorioj...”; I-vortoj, A-vortoj, O-vortoj)
    - Homoj (AMIK, TAJLOR, INFAN, PATR, SINJOR, VIR...)
    - Bestoj (ĈEVAL, AZEN, HUND, BOV, FIŜ, KOK, PORK...)
    - Plantoj (ARB, FLOR, ROZ, HERB, ABI, TRITIK...)
    - Iloj (KRAJON, BROS, FORK, MAŜIN, PINGL, TELEFON...)
    - Agoj (DIR, FAR, LABOR, MOV, VEN, FRAP, LUD...)
    - Trajtoj kaj ecoj (BEL, BON, GRAV, RUĜ, VARM, ĜUST, PRET...)

Rekoneblaj klasoj surbaze de PMEG

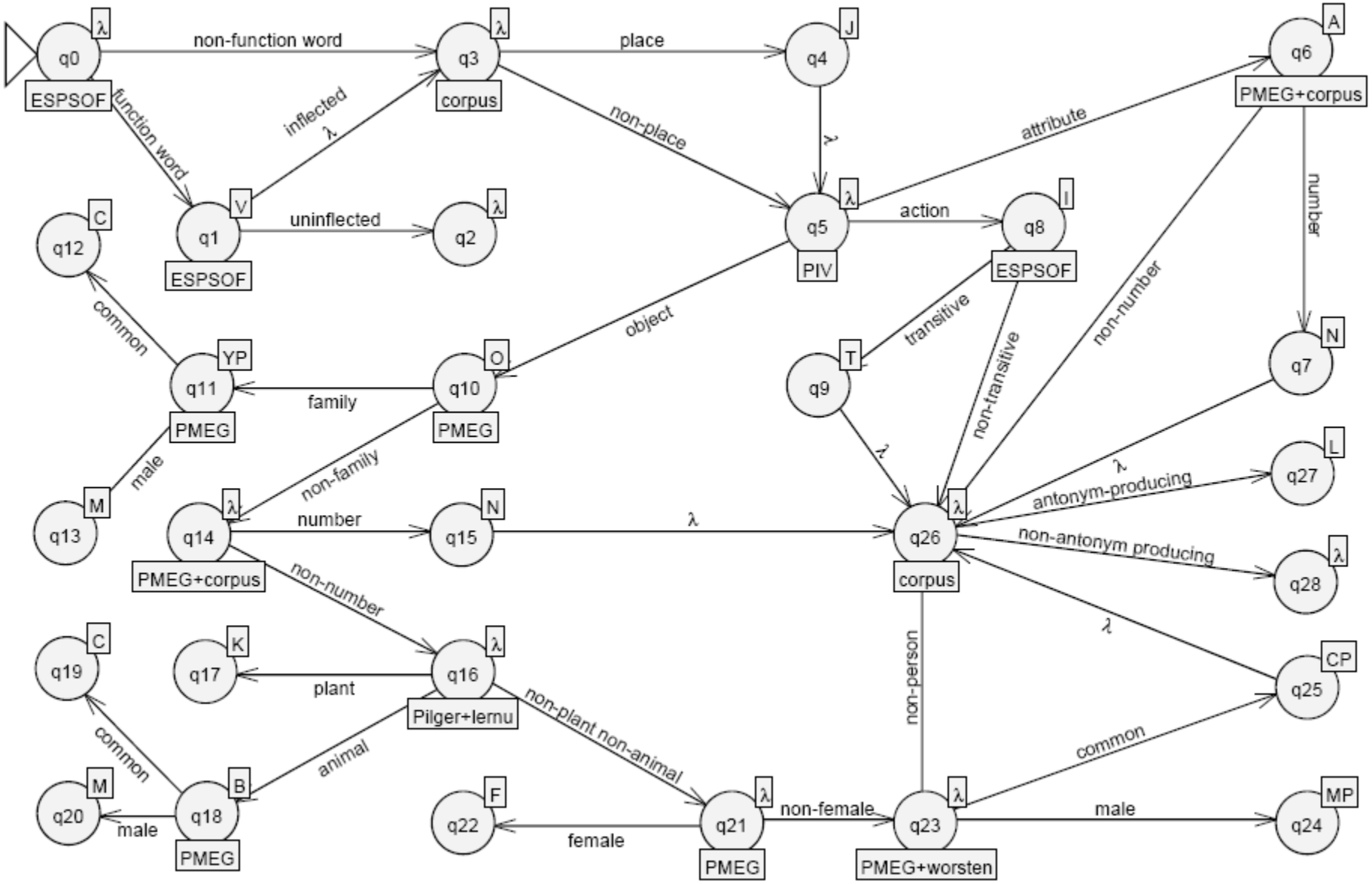
- **A** atributaj radikoj, havas A-finaĵon en la baza formo
- **B** bestoj
- **C** komuna ĝenro por estaĵoj
- **F** femala ĝenro por estaĵoj
- **I** agaj radikoj, havas I-finaĵon en la baza formo
- **J** lokaj radikoj, kreas adverbojn de spaca signifo (ĉambr-e = en ĉambro)
- **K** kreskaĵoj
- **L** antonim-kreivaj radikoj, kiuj akceptas la prefikson “mal“
- **M** maskla ĝenro por estaĵoj
- **N** numeroj (ciferoj kaj kelkaj aliaj nombraj radikoj)
- **O** objektaj radikoj, havas O-finaĵon en la baza formo
- **P** personoj
- **T** transitivaj radikoj, kreas transitivajn verbojn
- **V** funkcivortoj, kiuj povas aperi sen finaĵo
- **Y** familiaj rilatoj

Ekzemple ***PATR*:**
O (objektaj vortoj)
P (personoj)
M (maskloj)
Y (familio)

Fontoj por semantika klasado

- Vortostoko
  - Vortaraj kapvortoj (PIV)
  - Funkcivortoj (ESPSOF de Toon Witkam)
- Semantiko
  - PIV (A-, I-, O-vortoj)
  - “Fermitaj” listoj el PMEG (familio: BO-)
  - Fakaj vortaroj (personoj: REĜISOR)
  - Tekstaro de E. Bick (MAL-)

Aŭtomata semantika klasado



Rezulto de la semantika klasado

- Klasita vortlisto (**.dic**)
- Ekzemploj:
  - patr/MOPY
  - taŭr/BMO
  - dom/JO
- Sume:
  - 15.914 eroj en la vortlisto
  - 15 bazaj semantikaj klasoj

Elkreo de afiksaj reguloj

- Neceso observi realan uzon
- Listo de tuj-pretaj kunmetaĵoj (ESPSOF)
- Kunmetaĵa ŝablono = afiksoj plus radik-lokoj
  - Ŝablono por *MALSANULEJ* estas *mal/*/ul/ej*
- Ŝablonoj donas informojn pri efektiva kunmeteblo de afiksoj kun diversaj radikoj
- En 33.000 estis 632 malsamaj, plej oftaj:
  * (39%), */* (18%), **/o/** (4%), **/aĵ* (2%)
  **/ad*, **/ist*, **/ig*, **/ec*, ..., **/*/ig* (1%), ...

Elkreo de afiksaj reguloj

- Permana determino de permesataj radikaj klasoj por ĉiu ŝablono, ekz.:
  - Ŝablono: **/id*
  - Aperoj: *kat/id*, *kverk/id*, *hom/id*, ...
  - Decido: permesitaj estas radikoj markitaj P, B, K (personoj, bestoj, kreskaĵoj)
- Regulo por ĉiu el la 632 ŝablonoj

Ekzempla vortareto por Esperanto

```
a/z
i/z
o/z
ebl/e
est/I
hav/ITx
```

- Reguloj:

  ```
  COMPOUNDRULE Iz
  COMPOUNDRULE xez
  ```

- Produkteblaj vortoj:

  ```
  esta, esti, esto,
  hava, havi, havo,
  havebla, havebli, haveblo
  ```

Rezulto de la regulkreado

- Aldone al la 15 bazaj klasaj markoj ankoraŭ 81 por esprimi kombinojn
- Pro limigo je 3 kunmetaĵpartoj ekzistas 37.155 afiksaj reguloj (varieblaj per radikoj laŭ klasoj)
- Rekoneblas sume $2{,}8\times10^{18}$ da „sencohavaj“ vortformoj

Integrigo en aliaj projektoj

- Funkcianta aldonaĵo por:
  - OpenOffice.org 2.0.2+
  - Firefox 3.0+ (ekde 2008-06-17)
  - Hunspell mem (komandlinio)

http://extensions.services.openoffice.org/project/literumilo

Pritakso de la rezultinta vortaro

- Švandrlík, M.: Edzinigebla knabino
  - Transskribo de manuskripto, 52.700 vortoj (el tio 10.400 unikaj)
- Falsaj pozitivoj:
  - Pokrovskij: 3.823 / 7 % (1.565 / 15 %)
  - Blahuš: 2.140 / 4 % (546 / 5 %)
- Detekto:
  - Pokrovskij: 206 (148) eraroj
  - Blahuš: 167 (113) eraroj (-19 %)

(Malfermitaj) problemoj

- ~~Sugestoj nur por maksimume dupartaj Hunspell-kunmetaĵoj~~
- Interpunkcio kaj nealfabetaj signoj (ekz. dividstreko)
- Propraj nomoj (majusklo, diversaj formoj, sufiksoj -nj-, -ĉj-)
- Plibonigo de la klasado (ekz. personaj radikoj)
- Pliriĉigo de la korpuso de kunmetaĵaj ŝablonoj

Dankon pro via atento.

Bc. Marek BLAHUŠ <marek@ikso.net>