

Loka Tarjamah Otomatis Indonésia-Sunda (LTIS)

Dian Tresna Nugraha <dian.nugraha@gmail.com>

(Versi dokumen: 1.0)

1 Bubuka

Program “Loka Tarjamah Otomatis Indonésia-Sunda” (LTIS) nyaéta hiji program anu ditulis dina basa PHP, anu disimpen dina *webserver*¹ pikeun ngalayanan paménta tarjamahan sacara otomatis tina basa Indonesia kana basa Sunda, boh paménta nu mangrupa téks, atawa paménta nu mangrupa URI.

2 Daptar Singgetan

Istilah: **Maksudna:**

DOM	<i>Document Object Model</i> , tingali dina: http://en.wikipedia.org/wiki/Document_Object_Model
HTML	<i>Hypertext Markup Language</i> , tingali dina: http://en.wikipedia.org/wiki/HTML
XML	<i>eXtensible Markup Language</i> , tingali dina: http://en.wikipedia.org/wiki/XML
URI	<i>Unique Resource Identifier</i> , tingali dina: http://en.wikipedia.org/wiki/URI
CSV	<i>Comma-Separated Values</i> , tingali dina: http://en.wikipedia.org/wiki/Comma-separated_values
jrrd.	jeung réa-réa deui
jsté.	jeung sajaba ti éta
jst.	jeung saterusna

¹ program dina server anu husus ngalayanan paménta informasi ku cara maké protokol HTTP atawa HTTPS.

3 Tata Diréktori

Kode sumber program LTSI dina komputer disusun ku cara kieu:

Diréktori:	Pikeun:
./	nyimpen kode sumber anu mimiti ngalayanan paménta pamaké program (dina wéb)
./inc	nyimpen kode sumber program inti
./data	nyimpen kumpulan kecap-kecap atawa data kamus

File:	Diréktori:	Pikeun:
index.php	./	némbongkeun kaca wéb utama
loka.php	./	ngalayanan paménta mangrupa URI
teks.php	./	ngalayanan paménta mangrupa téks
NL_Translator.class.inc	./inc	kode fungsi-fungsi dasar program tarjamahan jeung analisis téks
NL_Translator_DOM.class.inc	./inc	kode fungsi-fungsi dasar pikeun parser HTML atawa XML, ngarombak jadi tangkal DOM, terus malikkeun deui jadi HTML/XML sanggeus ditarjamahkeun
NL_Translator_ID2SU.class.inc	./inc	kode fungsi-fungsi pikeun ngarombak tata basa Indonésia kana tata basa Sunda
NL_InputBox.class.inc	./inc	kode pikeun <i>user-interface (UI)</i>
id_su_main.csv	./data	data kecap-kecap utama basa.
id_su_names.csv	./data	data kangaranan jalma, tempat, waktu, jsb.
id_su_entries.csv	./data	data kecap-kecap lainna
id_su_phrases.csv	./data	data kecap-kecap kantétan
id_su_foreign.csv	./data	data kecap-kecap tina basa deungeun atawa anu teu kudu ditarjamahkeun

4 Ngajalankeun Program

Sakumaha waé ilaharna program anu dijalankeun dina server wéb, program LTIS dimimitan ku cara ngasupkeun URI dina *browser*². Contona, lamun program ieu disimpen di server <http://tarjamah.sabilulungan.org>, sacara otomatis `index.php` bakal dijalankeun. Mimitina katémbong dina *browser* siga kieu:



Gambar 1: Témbongan awal Loka Tarjamah

Dina gambar di luhur katémbong aya sababaraha kotak keur ngasupkeun téks.

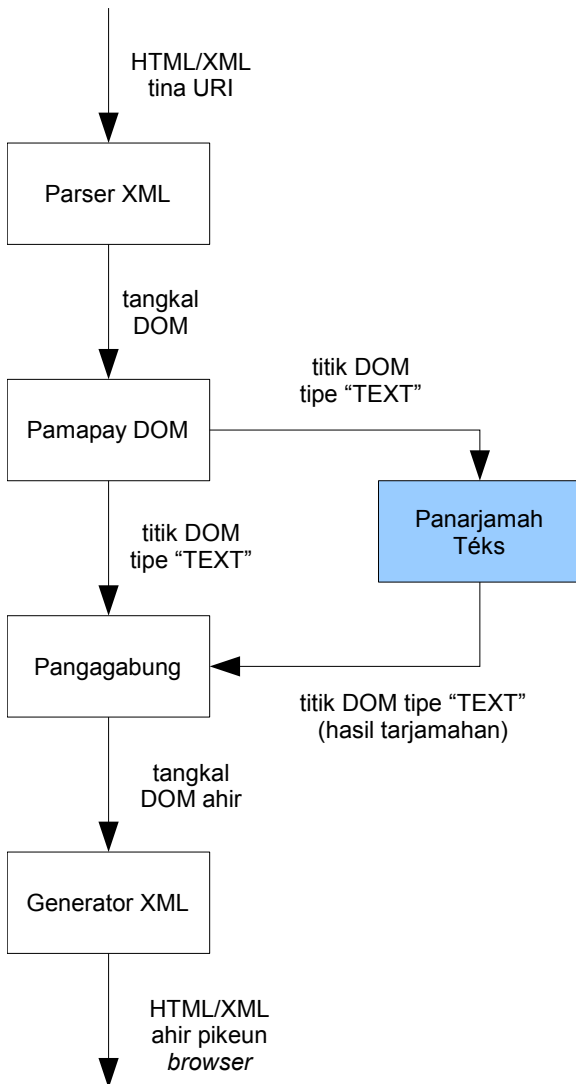
1. Anu luhur, kotak bobogaanana *browser*, pikeun ngasupkeun URI anyar atawa pindah ka ramatloka séjén. Tingali anu aya tulisan "<http://tarjamah.sabilulungan.org>".
2. Kotak pikeun narjamahkeun URI. Tingali anu aya tulisan "`http://`" wungkul.
3. Kotak nu badag panghadapna, pikeun narjamahkeun téks basa Indonésia anu diketikkeun ka dinya.

2 program pikeun urang ngambahan internét, contona: Internet Explorer, Mozilla Firefox, jrrd.

5 Galur Program

5.1 Narjamahkeun Loka tina URI

Nalika program “loka.php” narima URI pikeun ditarjamahkeun eusina, lalampahan program bisa ditengenan dina Gambar 2.

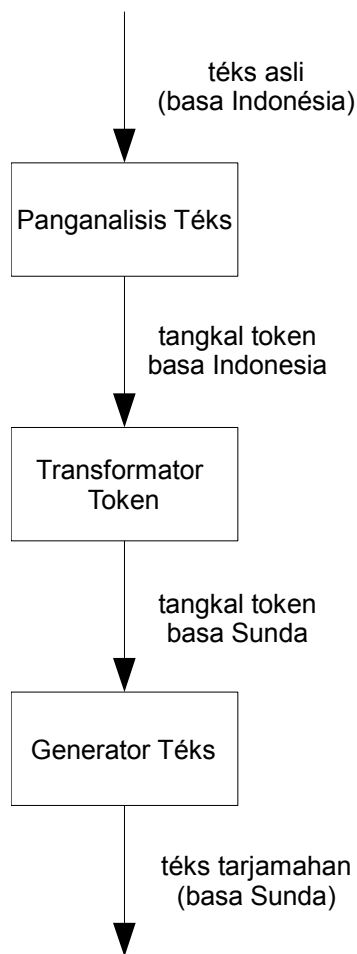


Pagawéan unggal blok anu aya dina gambar kasebut bisa dijéntrékeun siga kieu:

1. **XML Parser:** maca data HTML/XML anu dituduhkeun ku URI, tuluy ngabentuk wangun “tangkal DOM”-na.
2. **Pamapay DOM:** mapayan tangkal DOM, milihan unggal titik dina tangkal DOM mana anu ngabogaan tipe “TEXT” atawa anu séjénna.
3. **Panarjamah Téks:** nampa téks ti Pamapay DOM, terus narjamahkeun kana téks basa Sunda.
Cara gawé blok ieu dijelaskeun leuwih daria dina bab 5.2.
4. **Pangagabung:** ngahijikeun téks tarjamahan kana tangkal DOM indungna.
5. **Generator XML:** ngarobah tangkal DOM jadi data HTML/XML deui, pikeun ditémbongkeun dina *browser*.

Gambar 2: Galur utama narjamahkeun loka tina URI

5.2 Narjamahkeun Téks



Pagawéan unggal blok anu aya dina gambar kasebut bisa dijéntrékeun siga kieu:

1. **Panganalisis Téks:** maca téks asli (basa Indonésia), ngawilah-wilah jeung nandaan unggal token, bisa mangrupa kecap, angka, jeung tanda baca; saterusna nyimpen kana mémori, mangrupa tangkal token, atawa token anu silih tumbu.
2. **Transformator Token:** ngarombak susunan token numutkeun “Aturan Narjamahkeun” anu dijelaskeun dina bab 5.3.
3. **Generator Téks:** nyusun téks (basa Sunda) tina tangkal token hasil ngarombak tadi.

Gambar 3: Galur program narjamahkeun téks

5.3 Aturan Narjamahkeun

Aturan ieu dipaké ku “Transformator Token” pikeun ngarombak “tangkal token basa Indonésia” jadi “tangkal token basa Sunda”. Aturan ieu dipapay ti nomer hiji heula. Tingali ogé katerangan dina catetan suku.

Aturan narjamahkeun:

1. Frase ditarjamahkeun sacérwéléna jadi Kecap Kantétan numutkeun data dina kamus `id_su_frase.csv`.
2. Nama atawa Kata Asing ditarjamahkeun numutkeun data dina kamus `id_su_names.csv` atawa `id_su_foreign.csv`.
3. [Éksésif]: “terlalu” + Kata Sifat dirombak jadi Kecap Sipat + “teuing”
4. [Superlatif]: “paling” + Kata Sifat dirombak jadi “pang-” + Kecap Sipat + “-na”
5. [Kontinyu]: “masih” + Kata Sifat dirombak jadi Kecap Sipat + “kénéh”
6. “kembali” + Kata Kerja dirombak jadi Kecap Pagawéan + “deui”
7. “orang” ...

- a. dituturkeun ku Kata Benda ditarjamahkeun jadi "urang"
 - b. lainna, ditarjamahkeun jadi "jelema"
8. "baru" ...
- a. dituturkeun ku (Kata Kerja | Kata Benda | Modal | Waktu | Angka) ditarjamahkeun jadi "kakara"³
 - b. nuturkeun Terminasi^{4 5}, ditarjamahkeun jadi "kakara"
 - c. lainna, ditarjamahkeun jadi "anyar"
9. "dengan" + ...
- a. dituturkeun ku Kata Benda Perkakas, ditarjamahkeun jadi "maké"
 - b. dituturkeun ku (Kata Benda | Kata Ganti⁶ | Kata Bilangan | Modal), ditarjamahkeun jadi "jeung"
 - c. lainna, ditarjamahkeun jadi "kalawan"
10. "baik" + A + ("atau" | "maupun") + B, ditarjamahkeun jadi "boh" + A + "boh" + B
11. "pukul" + (Kata Bilangan | Angka), ditarjamahkeun jadi "jam" + (Kecap Bilangan | Angka)
12. "hari" ...
- a. dituturkeun ku (Kata Sifat | Kata Keterangan), ditarjamahkeun jadi "poé"
 - b. lainna, henteu ditarjamahkeun, tetep "hari"
13. "masing-masing" ...
- a. dituturkeun ku Kata Benda, ditarjamahkeun jadi "séwang-séwangna"
 - b. dituturkeun ku Terminasi, ditarjamahkeun jadi "séwang-séwangan"
 - c. lainna, ditarjamahkeun jadi "unggal"
14. "sekali" ...
- a. dituturkeun ku Terminasi⁷, ditarjamahkeun jadi "pisan"
 - b. lainna, ditarjamahkeun jadi "sakali"
15. Upama kapanggih, kecap-kecap séjén ditarjamahkeun numutkeun data dina kamus `id_su_main.csv` atawa `id_su_entries.csv`
16. [Sufiks]:
- a. Kata + "-ku", ditarjamahkeun jadi Kecap + "kuring"
 - b. Kata + "-mu", ditarjamahkeun jadi Kecap + "kuring"
 - c. Kata + "-nya" ... ditarjamahkeun jadi (Kecap + "-na" | Kecap + "-ana")
17. [Afiks]:
- a. "se-" + Kata, ditarjamahkeun jadi "sa-" + Kecap
 - b. "ke-" + Kata, ditarjamahkeun jadi "ka-" + Kecap
 - c. "ter-" + Kata Sifat, ditarjamahkeun jadi "pang-" + Kecap Sipat + "-na"

3 Tanda pipa ajeg '|', dibacaana 'atawa'. Conto: (Kata Bilangan | Angka), dibaca: Kata Bilangan atawa Angka.

4 ciri yén aya di awal klausa atawa awal kalimah

5 Terminasi, nyaéta rupa-rupa tanda baca: pananya, panyeluk, titik, koma, jsté.

6 *Pronomina*

7 ciri yén aya di tungtung klausa atawa tungtung kalimah

6 Kecap dina Kamus

Sakabéh data ngeunaan kecap-kecap disimpen dina “kamus”. Kamus ieu mangrupa *file*anu formatna CSV. Contona:

```
"ID";"Valid";"SU";"Fn";"Ctx";"Form";"Root"
"--";1;"maenya";"q";"--";"--";"maenya"
"--";1;"kapalang";"ad";"--";"--";"kapalang"
"adalah";1;"nyaéta";"ad";"--";"--";"éta"
"adapun";1;"sedengkeun";"c";"--";"--";"sedeng"
"adik";1;"adi";"pn";"pancakaki";"--";"adi"
"abang";1;"akang";"pn";"pancakaki";"--";"akang"
.....
```

Dina kamus, unggal kolom dipisahkeunana ku titik-koma. Baris kahiji minangka ngaran kolom, nyaéta:

- **ID:** kolom kecap basa Indonésia. Saupama eusi kolom ieu mangrupa '--', data nu aya dina sabaris moal dipaké dina program.
- **Valid = '1':** kecap bakal dipaké nalika narjamahkeun.
- **Valid = '0':** kecap moal dipaké nalika narjamahkeun.
- **SU:** kolom kecap basa Sunda, tarjamahan tina kolom ID.
- **Fn:** fungsi kecap, bisa mangrupa:
 - 'q' = Kata Tanya
 - 'ad' = Katerangan
 - 'c' = Kata Sambung
 - 'pn' = Kata Ganti
 - 'vi' = Kata Kerja Dasar
 - 'v' = Kata Kerja Aktif
 - 'vp' = Kata Kerja Pasif
 - 'r' = Kata Seru
 - 'm' = Modal **atau** Modalitas
 - 'b' = Kata Bilangan
- **Ctx:** Kontéks (larapan) kecap, contona:
 - 'alat'
 - 'awak' = bagian awak
 - 'waktu'
 - 'warna'
 - 'sato'
 - 'tempat'

7 Panutup

Sakieu heula dokuméntasi téknis program LTSI. Bongbolongan, pamundut, jsb. tiasa dikintunkeun ka surélék sim kuring: dian.nugraha@gmail.com

Hatur nuhun kana perhatosanana.

München, 29. November 2007

Dian Tresna Nugraha