DBnary: Wiktionary as Linked Data for 12 Language Editions

… with Enhanced Translation Relations

Gilles SÉRASSET — GETALP - LIG - Université Grenoble 1
http://kaiko.getalp.org/about-dbnary/

MOTIVATIONS

Wiktionary is a large set of lexical data with many multilingual information.
However, its data is quite difficult to extract:

- each language edition uses its own templates/guidelines
- many entries are incoherent, inconsistent or erroneous

Several work has been done to extract this data:

- WiktionaryX (Sajous & al 2010) provides an XML version of the data (French and English language editions),
- JWKTL (Zesch et al., 2008) is a free API to access wiktionary data,
- dbpedia.org/wiktionary (Hellmann et al., 2013) provides an alternative way, inspired by dbpedia tools,
- Many others extracted data for their own purpose… but mainly from the English language edition

Advantages of DBnary approach:

- Based on LEMON standard
- Available as Linked Data
- Extractors/Data evolve with Wiktionary
- programs are open source (extensions are welcome)

RESULTS

Open source extraction program (LGPL) written in Java (using maven, so it's easy to use)

- http://forge.imag.fr/projects/dbnary
  Extractor currently supports
- Bulgarian, English, Finnish, French, German, Greek, Italian, Japanese, Portuguese, Russian, Spanish and Turkish
- Polish (in progress)

Extraction features

- allows extraction from a dump or from a single entry (no DB setup needed)
- Full extraction of English (4.1Gb) takes ~4 min on a laptop

Extracted data is available

- as RDF (n-triple) dump files
- As linked data
- http://kaiko.getalp.org/dbnary/fra/exemple

DIVERSITY AND INCOHERENCE…

```
=====Translations=====
{{trans-top|domestic species}}
* Abkhaz: {{t|ab|ацгə|sc=Cyrl}}
* Acehnese: {{t|ace|mië}}
* Adyghe: {{t|ady|кIэтыу|sc=Cyrl}}
* Afrikaans: {{t+|af|kat}}
* Ainu: {{t|ain|チャペ|tr=cape}}
* Akan: {{t|ak|agyinamoa|n}}
* Albanian: {{t+|sq|mace|f}}
* Alemannic German: {{t|gsw|Chàtz}}
* Amharic: {{t|am|ድመት|sc=Ethi}}
```

```
====Käännökset====
{{kohta|1|eteläkiinalainen puu (''Litchi chinensis''); sen hedelmä, joka
*englanti: {{käännös|en|lychee}}
*espanja: {{käännös|es|lichi|suku=m}}
*ruotsi: {{käännös|sv|litchi}}
*viro: {{käännös|et|litši}}
|loppu}}
```

```
===Tradução===
{{tradini|De 1 (animal)}}
* {{trad|af|kat}}
* {{trad|ay|michi|phisi}}
* {{trad|sq|macja|maçok}}
* {{trad|gsw|Hauskatze}}
* {{trad|de|Hauskatze|Katze|Kater|Mieze|Miezekatze}}
```

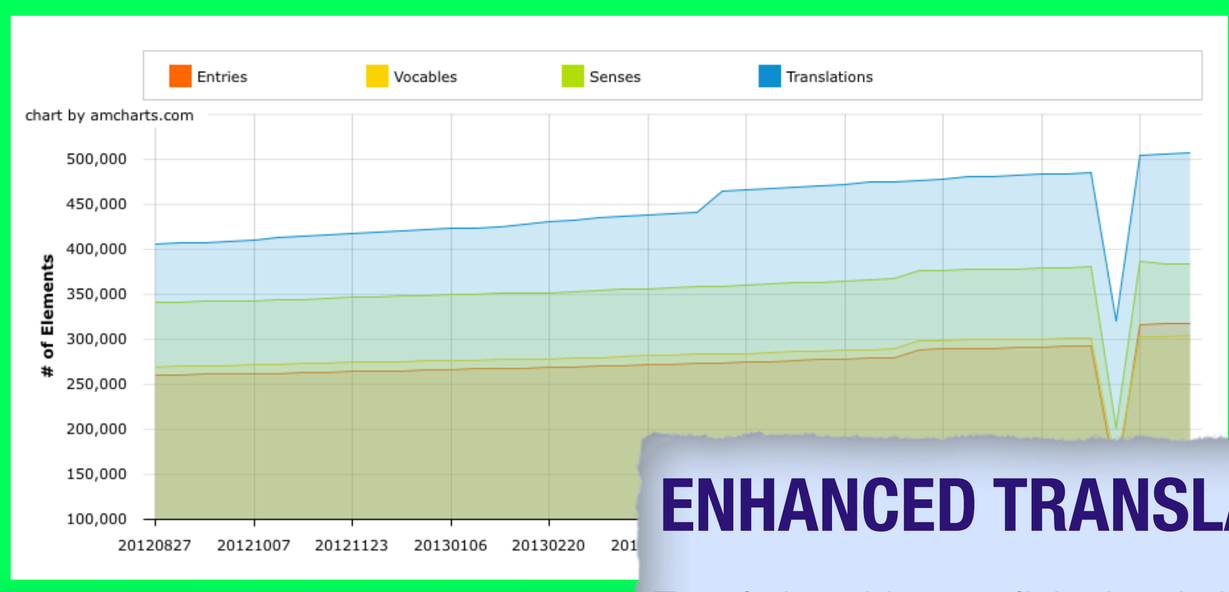
```
| ПРЕВОД1 =

* английски: [[cat]] [[:en:cat|(en)]]
* арабски: [[]] [[:ar:|(ar)]]
* арменски: [[]] [[:hy:|(hy)]]
* африкаанс: [[]] [[:af:|(af)]]
* белоруски: [[]] [[:be:|(be)]]
* гръцки: (мъжки) [[γάτος]] м., (всеобщ, ж.) [[γάτα]] ж. [[:el:γάτ
```

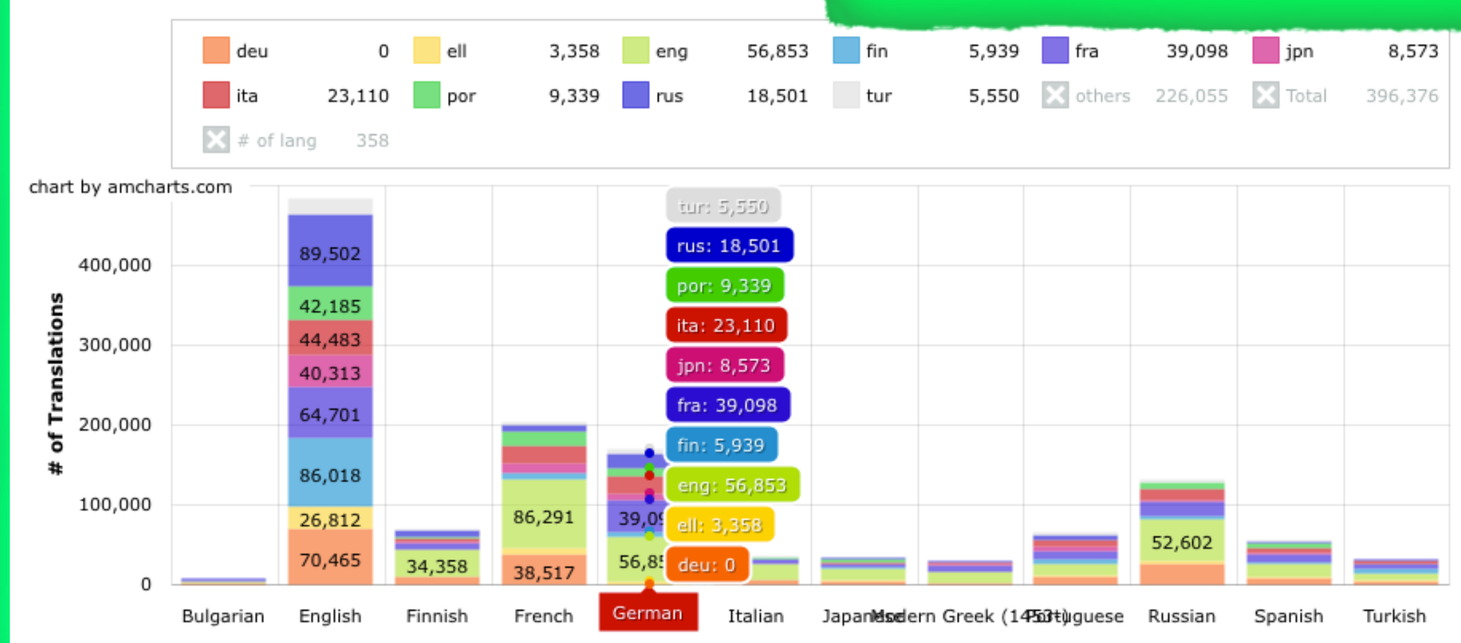
```
{{μτφ-αρχή|κατοικίδιο ζώο}}
* {{en}} : {{τ|en|cat}}
* {{sq}} : {{τ|sq|mace}}
* {{ar}} : {{τ|ar|قط|tr=qéṭṭ}}, {{τ|ar|قطة|tr=qéṭṭa}}
* {{vi}} : {{τ|vi|mèo}}
* {{bg}} : {{τ|bg|котка|tr=kotka}}, {{τ|bg|котак|noentry=1|tr=kotak}}, {{
* {{fr}} : {{τ|fr|chat}}, {{τ|fr|chatte}}
* {{de}} : {{τ|de|Katze}}
* {{da}} : {{τ|da|huskat|noentry=1}}, {{τ|da|kat}}
* {{he}} : {{τ|he|חָתוּל|noentry=1|tr=khatul}}, {{τ|he|חתולה|noentry=1|tr
```

```
=== Перевод ===
{{перев-блок|домашняя кошка|
|az=[[pişik]]
|sq=[[macja]]
|gsw=[[Chatze]]
|en=[[cat]]
|ar=[[قط]] (qiṭṭ) {{m}}, قطط (qíṭaṭ) {{мн}}
|an=[[gato]]
|hy=[[կատու]] (katu)
|ast=[[gatu]]
```

```
==Traducciones==
{{trad-arriba}}
{{t+|af|1|kat}}
{{t+|de|1, 2|Katze|f|,|1|Kater|m|nota|gato macho|,|7|Wagenheber|m|,|7|
{{t+|ay|1|phisi}}
{{t+|hy|1|կատու|tr|katu}}
{{t+|ast|1|gatu}}                                {{t+|az|1|pişik}}
{{t+|br|1|kazh}}
{{t+|bg|котарак|m|,|котка|f|}}
{{t+|ca|1|gat|,|8|tres en ratlla|,|1|móx|nota|Islas Beleares}}
```

EVALUATING QUALITY/MAINTAINING EXTRACTORS





ENHANCED TRANSLATIONS

Translation object are linked to their source Lexical Sense (in addition to their source Lexical entry).

BG	10068	13925
DE	388988	396376
EL	8506	58952
ES	61079	116912
FI	121660	125615
FR	136685	514525
IT	0	63711
JA	22229	93161
PT	74426	269516
RU	153485	381641
TR	51791	68416