

# PanLex: The Schema

Jonathan Pool

Computational Linguistics Lab  
University of Washington  
9 October 2008

# Preface

PanLex aggregates lexical resources into a unified database.

What resources?

Dictionaries (including Wiktionaries)

Glossaries

Word lists

Thesauri

WordNets

Vocabulary databases

Standards

Subject heading systems

Some content is copied into TransGraph and used in Turing Center research:

PanDictionary

PanImages

Image Labeling Game

PanMail

Lematic Communication Studies

The data are available to you for research.

# Preface

The lexical resources can be fairly simple, like Kirill Panfilov's Basque-Russian dictionary.

Or a bit more complex, like this:

aDkanu	अड्कनु <sup>s</sup> (अक्री)	halti <sup>r</sup> , implikiği, fiksiği <sup>r</sup>
aDig	अडगि <sup>s</sup> (वी)	stabila <sup>r</sup> , firma <sup>r</sup>
aDkaaunu	अड्काउनु <sup>s</sup> (सक्री)	fiksi <sup>r</sup> , haltigi <sup>r</sup>

Or mildly ethnolinguistic, like this:

**bohupohi** [bohupohi] *See: kosa; koho'ya.*  
*n321.* mound, hump, mountain, mountainous terrain. *Note:* Compare: bohupohi seems to apply to the whole mountain and to rounded hills or humps; kosa seems to apply to sharp mountains, mountain peaks and ridges.

# Preface

Or intensely historical and speculative, like this:

**fouyapin** (également "foubap" selon le Dictionnaire de Pouillet et al., 1990), aussi "friyapen", avec la variante "penbwa" d'après R. Confiant ([dictionnaire en ligne](#)) : arbre à pain, fruit à pain.

Il est intéressant de signaler que si, à l'heure actuelle, ne sont répertoriées aux Antilles dans les dictionnaires courants que ces formes calquées sur le mot français, il existe à la Dominique (île voisine indépendante après avoir été colonie britannique au XIXe et XXe siècle) pour désigner la même réalité dans le créole local, le terme de "yanm-pen" (lit. igname-pain) qui tend à faire de "[yanm](#)" un terme générique pour "nourriture". Marcel Fontaine dans son dictionnaire cite également l'usage de "pen-pen" - sans indiquer toutefois s'il faut en rapporter l'usage à un groupe particulier.

De fait "penbwa" (pour arbre à pain) est attesté à Sainte-Lucie (île également indépendante après la colonisation britannique).

Ces mots composés créoles sont particulièrement significatifs et intéressants : ne peut-on pas penser que le créole de la Dominique et le créole de Ste-Lucie nous livrent là un usage "non-contaminé" par le français ? Resterait à chercher si ces formes ont été également attestées en créole au XIXe siècle en Guadeloupe et Martinique.

En Haïti semble attestée la forme (un peu étonnante : on s'interroge sur son origine) de "lam"/"lanm" ou "lam véritab" (cf. in Dictionnaire d'Albert Valdman et al., 1996, mais aussi in Wally R. Turnbull, 2003 : *Creole Made Easy*)

# Preface

Or complexly ambiguous, like this:

fei-zâwn, *n.* a measure; a conical heap of rice, etc, the apex of which will be level with the point of a spear held vertically at arm's length above the head of an ordinary-sized man when standing.

fek, *adj.* small, stunted or dwarfed in growth, undersized. *v.* to be small, etc. (Used of persons, trees, etc.)

fêk, *adj.* slender (as the waist), squeezed in. *v.* to be slender, etc, (as above). See **dul fêk, kâwng fêk, tai fêk.**

fel, *adj.* just, righteous, accurate, correct, proper, convenient, right, neat, tidy, orderly, virtuous, good, careful. *v.* to be just, righteous, etc, as above; to be settled (as a dispute). *adv.* justly, righteously, accurately correctly, properly, conveniently, rightly, aright, neatly, tidily, orderly, virtuously.

# Preface

Or they can defy automated processing, like this:

174

بَذَارَةٌ: see بَذَارَةٌ.

بَذَارَةٌ *Increase, redundance, exuberance, plenty, or abundance, in wheat, or food. (Lh, \* T, \* M, L, K. \*)* You say, طَعَامٌ كَثِيرُ الْبَذَارَةِ *Wheat, or food, in which is much increase, &c. (T, TA.)* — See also بَذْرٌ.

بَذَارَةٌ, and sometimes بَذَارَةٌ, (Lh, M, K,) and بَيِّذْرَةٌ, (AA,) and نَبَذْرَةٌ, with ن, (T, K,) i. q. تَبْذِيرٌ, (M, K,) *The dissipating, or squandering, of wealth, or property, in a way that is not right. (T, TA.)*

بَيِّذْرَةٌ: see what next precedes.

بَيِّذْرَانِي: see بَذْرٌ.

بَيِّذَارٌ: see بَذْرٌ.

A4- JJ[ I

;5lu: sec;lj..

;iJti *Increase, redundance, exuberance, plenty, or abundance, in wheat, or food. (Lh, T,0 M, L, K. \*)* You say, ;ij4JI ;,± *Wheat, or food, in nwhich is much increase, &c. (T, TA.)*

-See also .

6l~t, and sometimes t1.i, (Lb, M, g,) and

· \*it, (AA,) and \* t i, with ;, (T, g,) i. q.

seU, (M, ,) *The dissipating, or squanderiyv, of wealth, or property, in a way that is not right. (T, TA.)*

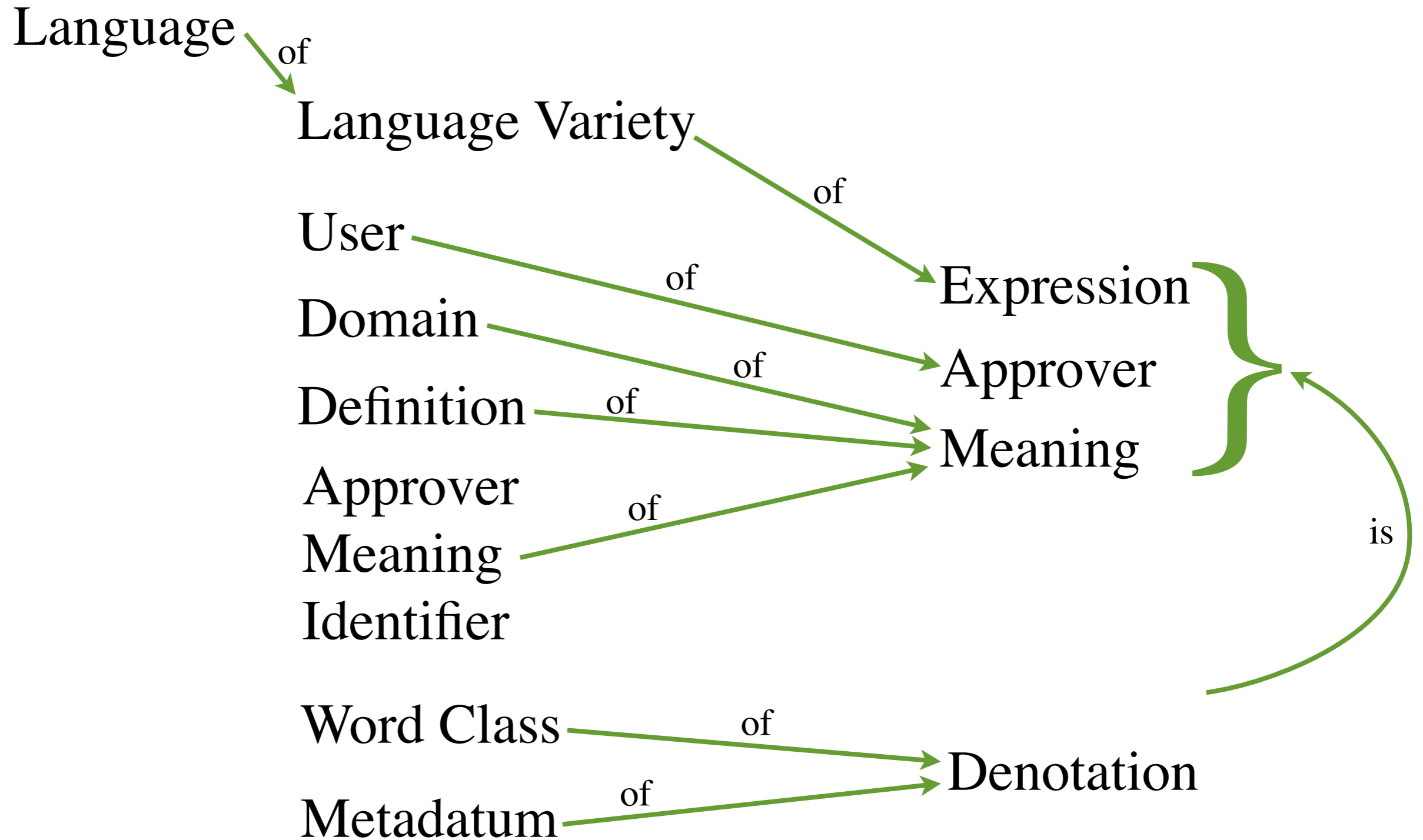
(T, TA.)

;jj~w: see what next precedes.

AlS: see

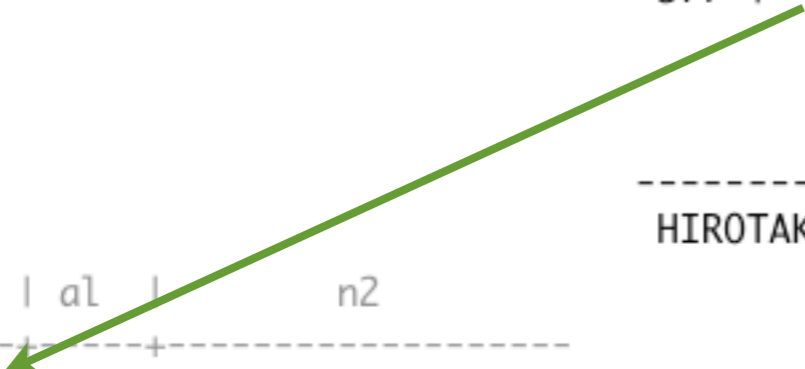
;lj,: see .ij.~

# Summary



# Approver

```
ap | us | tt | ur |
-----+-----+-----+-----+
377 | 3 | epo-jpn:PEJV | http://vastalto.com/pejv/ |
      |
      | au | ti | yr |
      |-----+-----+-----+
      | HIROTAKA Masaaki | Praktika Esperanto-Japana Vortareto | 2008
      |
us | al | n2
-----+-----+-----+
3 | smc | Susan M. Colowick
```



*Today: 586*

# Language

aaq | Alənaratəwewakan  
aar | Qafár af  
abe | Wôbanakiôdwawôgan  
abk | аҧсуа бызшәа  
abq | абаза бызшва  
abs | Malayu Ambon  
abx | Inabaknon  
acf | kwéyòl  
ach | Acoli  
ada | Adangme  
ady | адыгэбзэ  
aer | Iknɡerripenhe  
afr | Afrikaans  
agf | Arguni  
agt | Agta  
aia | Arosi  
aib | Äynú  
aie | Amara  
aii | ᱠᱟᱨᱱᱟᱲ  
ain | アイヌ イタク

yua | Maaya T'aan  
yuc | Yuchi  
yue | 廣東話  
zaj | Zaramo  
zak | Zanaki  
zap | diidzaj  
zay | Zayse-Zergulla  
zbc | Melawan  
zdj | Ngazidja Comorian  
zea | Zeêuws  
zga | Kinga  
zgr | Magori  
zha | Sawcuengh  
zin | Zinza  
ziw | Zigula  
zku | Kaurna  
zlm | Melayu  
zsm | Bahasa Melayu  
zul | isiZulu  
zun | Zuni



*Today: 7,766*

# Language Variety

lv	lc	vc	tt
396	lmo	0	lengua lumbarda
1253	lmo	1	bregagliotto
1254	lmo	2	milanese
1255	lmo	3	trentino

lmb | Merei  
lmg | Lamogai  
lml | Hano  
lmo | lengua lumbarda  
lmr | Lamalera  
lmy | Lamboya  
loj | Lou  
los | Loniu  
lou | Kréyol La Lwizyan

*Today: 1,256*

# Expression

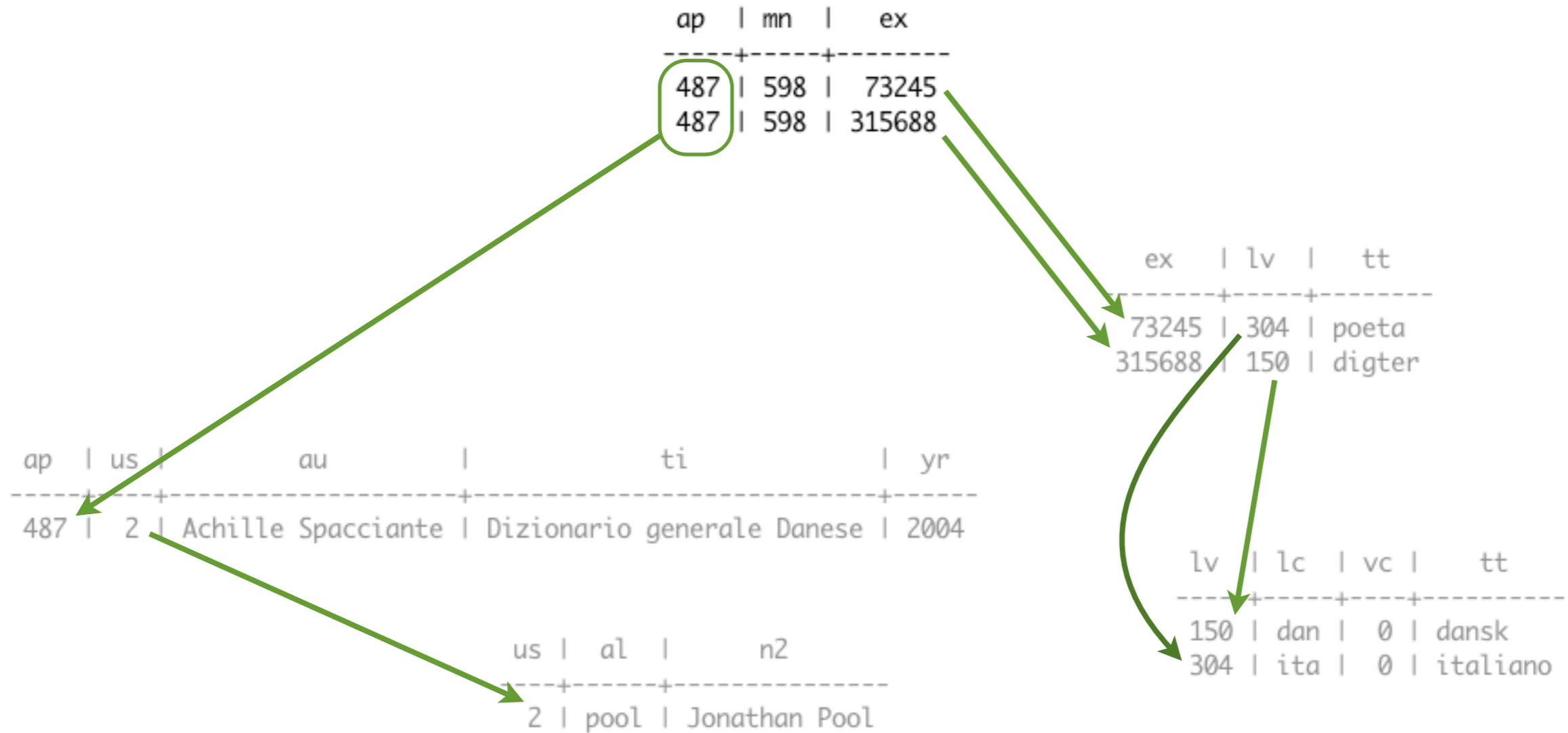
ex	lv	tt
3120024	34	إيقاف تشغيل إعادة توجيه الاتصال
3120025	93	Прехвърлянето на обажданията е изключено
3120026	106	Vypnout přesměrování volání
3120027	150	Slå Viderestilling af opkald fra
3120028	157	Anrufweiterleitung deaktivieren
3120029	184	Απενεργοποίηση προώθησης κλήσεων
3120023	187	Call Forwarding Off
3120030	190	Übersuunamine välja lülitatud
3120031	204	Soitonsiirto poissa käytöstä
3120033	271	העברת שיחות כבויה
3120034	275	कॉल फॉरवर्डिंग बंद
3120035	280	Prosljeđivanje poziva isključeno
3120036	283	Hívásátirányítás kikapcsolása
3120037	304	Inoltro di chiamata disattivato
3120021	304	Trasferimento chiamate
3120038	315	着信を転送しない
3120039	357	착신 전환 해제
3120040	383	Izslēgt zvanu pāradresēšanu

lv	lc	tt
32	apw	Ndee biyati'
33	apy	Apalaí
34	arb	إيقاف توجيه
35	arc	אומית
36	arg	luenga aragonesa
37	arn	Mapudungun

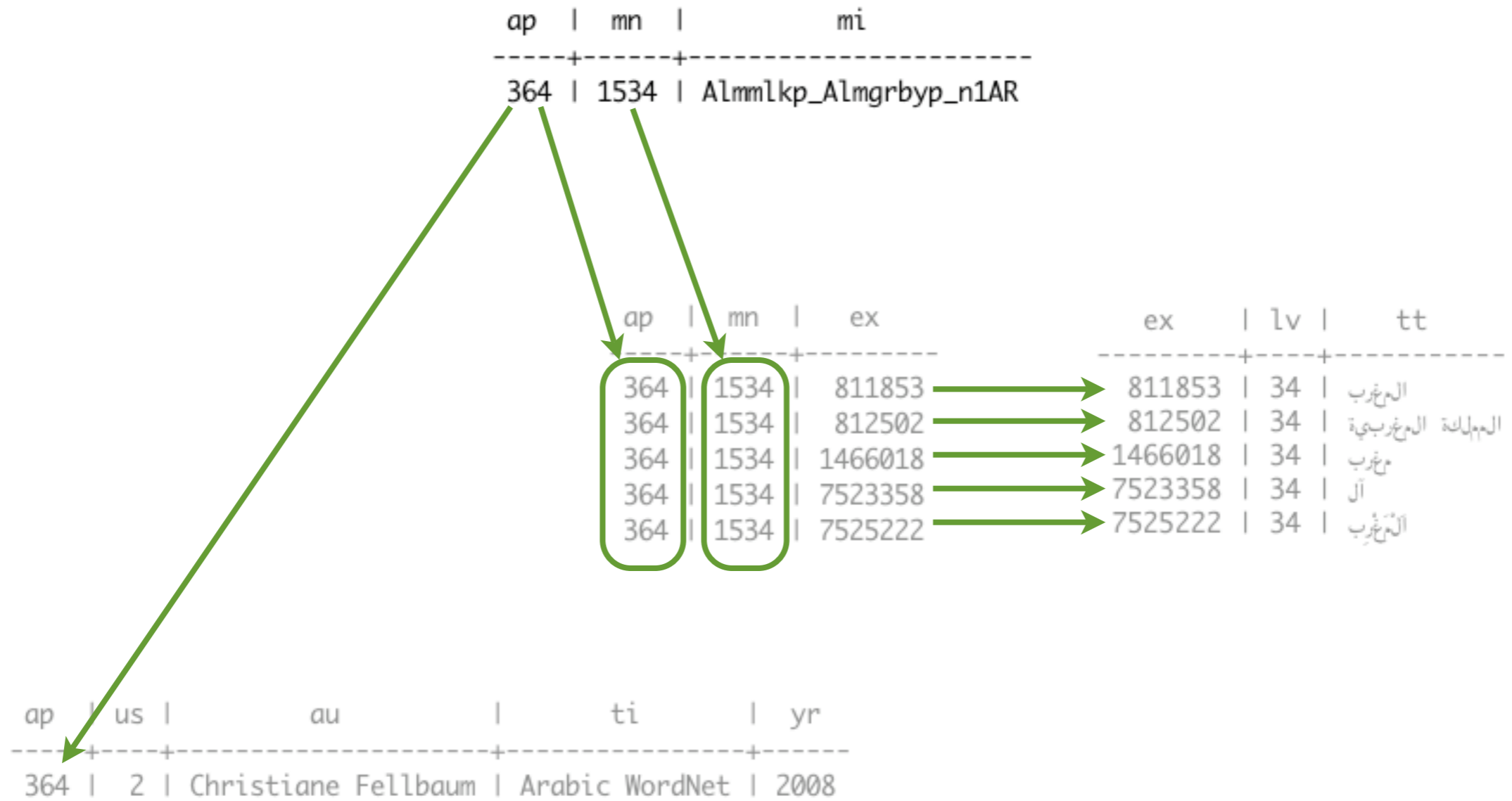
Today: 10,684,183

# Denotation



*Today: 24,804,636*  
*(Expression pairs: 78,895,306)*

# Approver Meaning Identifier



# Approver Meaning Identifier

```
ap | mn | mi  
-----+-----+-----  
364 | 1534 | Almmlkp_Almgrbyp_n1AR
```

```
<item itemid="Almmlkp_Almgrbyp_n1AR"  
offset="108413097" lexfile="" name="المملكة المغربية"  
type="synset" headword="" POS="n" source="NE file"  
gloss="Morocco." authorshipid="1980" />  
<authorship author="horacio" date="20070322" score=""  
comment="NE import from file couma.out."  
#Morocco#Kingdom of Morocco#" covering="0"  
authorshipid="1980" />
```

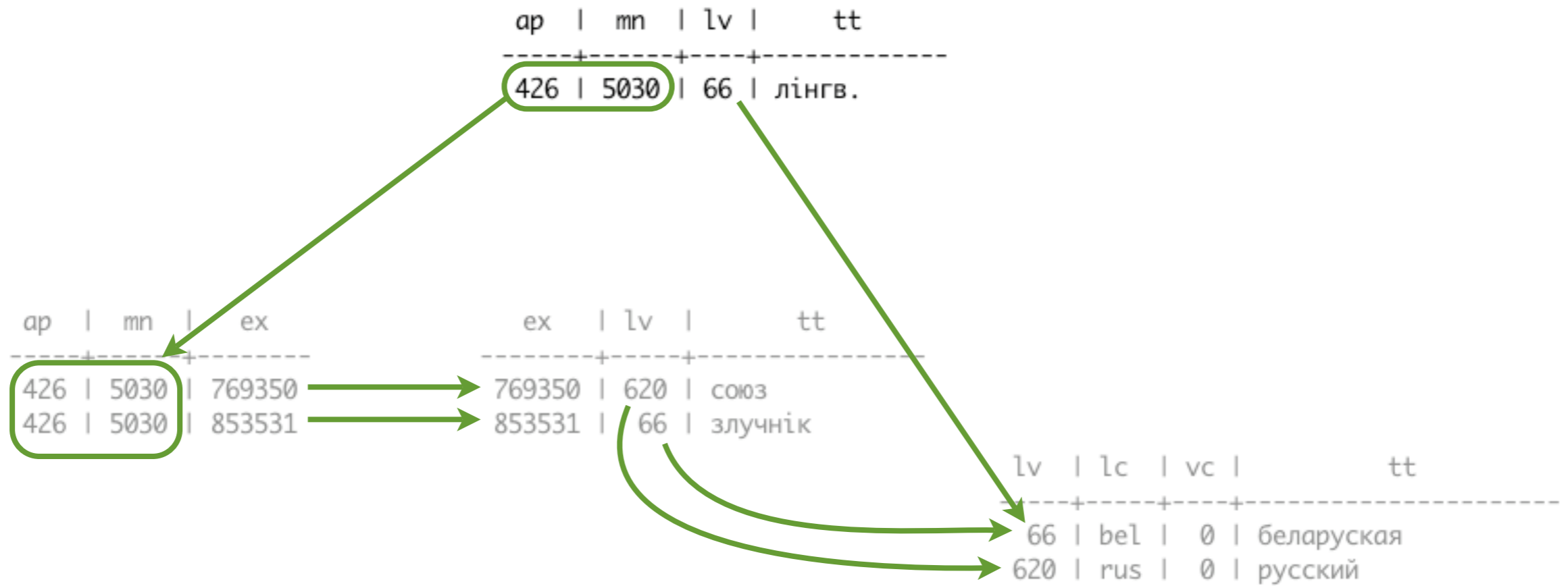
```
<word wordid="Almgrb_1" value="المغرب"  
synsetid="Almmlkp_Almgrbyp_n1AR" frequency=""  
corpus="" authorshipid="15880" />  
<authorship author="horacio" date="20070416" score=""  
comment="NE import from file stdin.Morocco#  
#Morocco#Morocco# #Morocco#Al# #Al#Al# #Al#"  
covering="0" authorshipid="15880" />
```

```
<link type="has_holo_member"  
link1="Almmlkp_Almgrbyp_n1AR"  
link2="jAmiEapu_n1AR" authorshipid="38311" />  
<authorship author="horacio" date="20080225"  
score="0" comment="from english WordNet 2.0"  
covering="0" authorshipid="38311" />
```

```
<link type="has_holo_part" link1="AlrbAT_n1AR"  
link2="Almmlkp_Almgrbyp_n1AR"  
authorshipid="38352" />  
<authorship author="horacio" date="20080225"  
score="0" comment="from english WordNet 2.0"  
covering="0" authorshipid="38352" />
```

```
<link type="has_instance" link1="balad_n2AR"  
link2="Almmlkp_Almgrbyp_n1AR"  
authorshipid="41539" />  
<authorship author="horacio" date="20070322" score=""  
comment="NE import from file couma.out."  
covering="0" authorshipid="41539" />
```

# Domain



Поиск [help](#)

Начинается  Содержит  Заканчивается  В описаниях

[Искать в Википедии](#) [Искать в Гугле](#)

[а](#) [б](#) [в](#) [г](#) [д](#) [е](#) [ж](#) [з](#) [и](#) [й](#) [к](#) [л](#) [м](#) [н](#) [о](#) [п](#) [р](#) [с](#) [т](#) [у](#) [ф](#) [х](#) [ц](#) [ч](#) [ш](#) [щ](#) [э](#) [ю](#) [я](#)

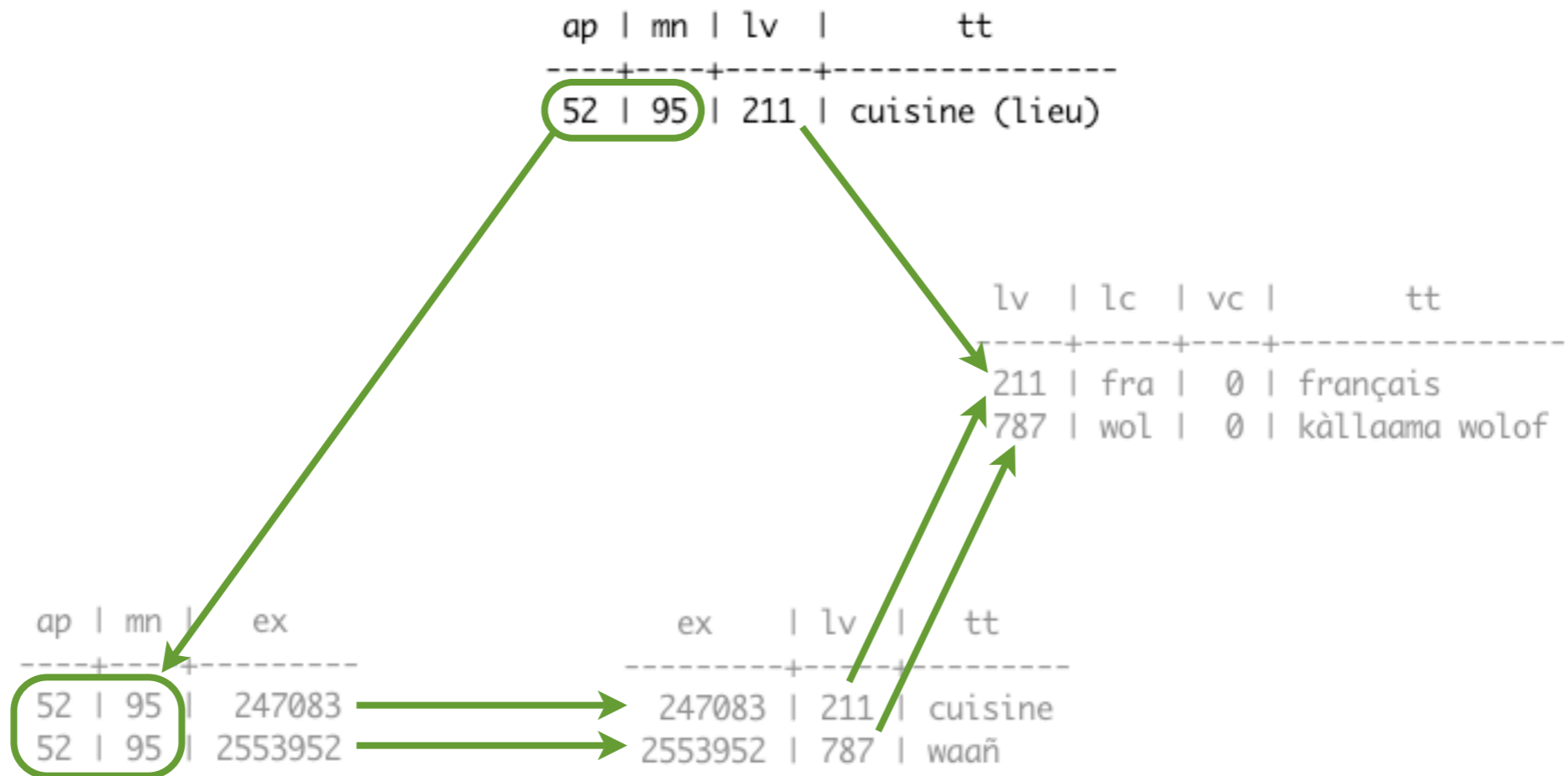
Количество найденных статей: 1185

Страница: [\[1\]<<](#) [90](#) [91](#) [92](#) [93](#) [94](#) [95](#) [96](#) [97](#) [98](#) [99](#) [100](#) [101](#) [102](#) [103](#) [>>\[119\]](#)

злучальны	соединительный
злучнік	лінгв. союз
злучок	лінгв. дефис

СМО

# Definition



Croire : v. gëm

Cuisine (lieu) : n. waañ

Cuisiner : v. togg

D

Dame : n. soxna

# Word Class

ap | mn | ex | tt  
-----+-----+-----  
385 | 74206 | 620174 | verb

ex | lv | tt  
-----+-----+-----  
620174 | 315 | 担う

noun	Common noun
name	Proper noun
pron	Pronoun
verb	Verb
vpar	Verb particle
auxv	Auxiliary verb
adjv	Adjective
detr	Determiner
advb	Adverb
prep	Preposition
post	Postposition
conj	Conjunction
ijec	Interjection
affx	Affix
punc	Punctuation
misc	Other

```

<entry>
<ent_seq>1599900</ent_seq>
<k_ele>
<keb>担う</keb>
<ke_pri>ichi1</ke_pri>
<ke_pri>news2</ke_pri>
<ke_pri>nf41</ke_pri>
</k_ele>
<k_ele>
<keb>荷う</keb>
</k_ele>
<k_ele>
<keb>荷なう</keb>
</k_ele>
<r_ele>
<reb>になう</reb>
<re_pri>ichi1</re_pri>
<re_pri>news2</re_pri>
<re_pri>nf41</re_pri>
</r_ele>
<sense>
<pos>&v5u;</pos>
<pos>&vt;</pos>
<gloss>to carry on shoulder</gloss>
<gloss>to bear (burden)</gloss>
<gloss>to shoulder (gun)</gloss>
<gloss xml:lang="fre">placer devant l'épaule (un fusils)</gloss>
<gloss xml:lang="fre">porter sur l'épaule</gloss>
<gloss xml:lang="fre">prendre sur soi (une faute)</gloss>
<gloss xml:lang="rus">нести на плече</gloss>
<gloss xml:lang="ger">auf der Schulter tragen</gloss>
<gloss xml:lang="ger">auf dem Rücken tragen</gloss>
</sense>
<sense>
<gloss xml:lang="ger">auf sich nehmen</gloss>
<gloss xml:lang="ger">übernehmen</gloss>
</sense>
</entry>

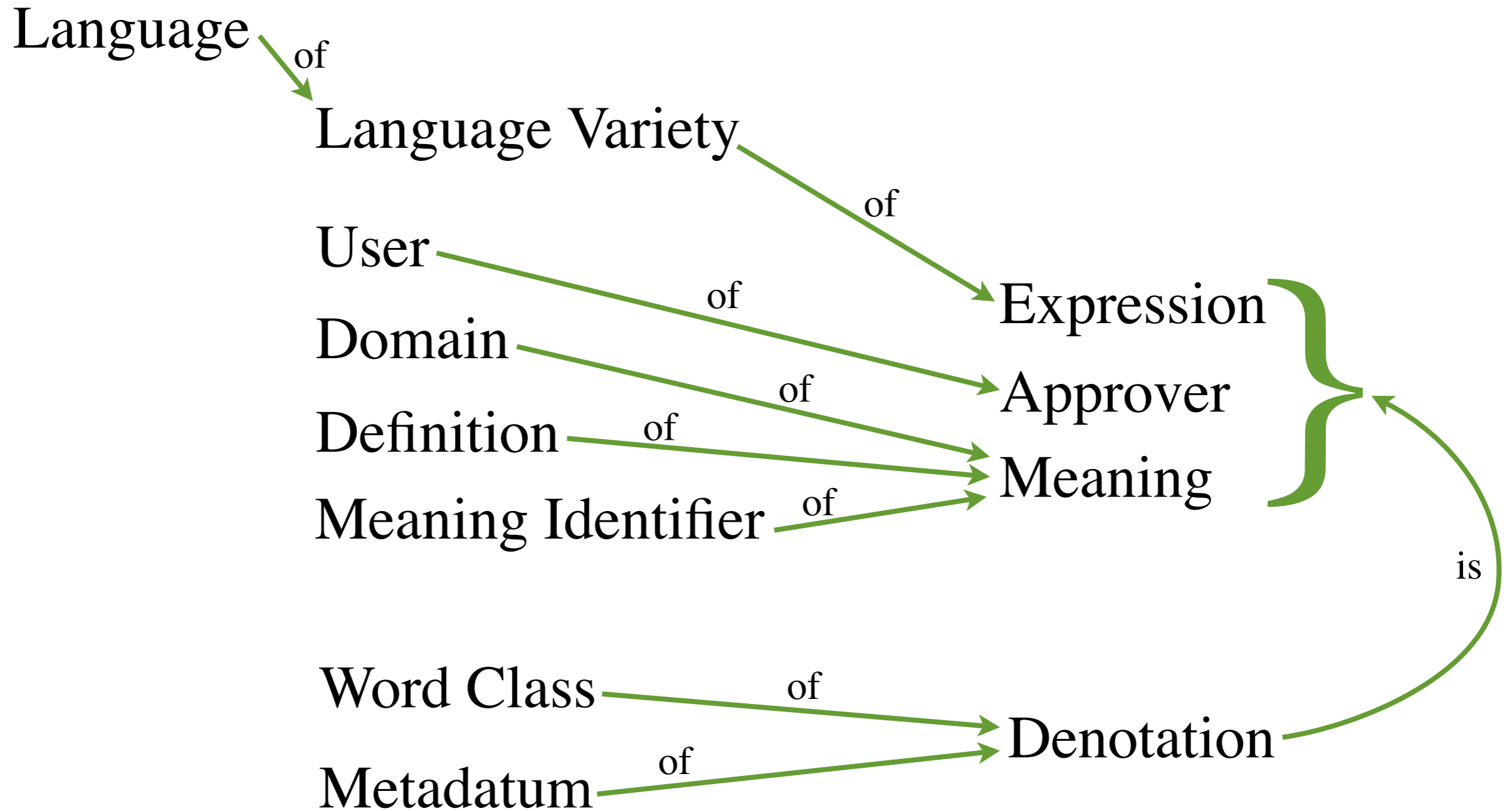
```

# Metadatum

ap	mn	ex	vb	vl	ex	lv	tt
454	3526	8165263	norm	depr	8165263	275	संकरणी करण

```
<skos:Concept rdf:about="http://www.fao.org/aims/aos/agrovoc#c_3706">
  <skos:prefLabel xml:lang="en">Hybridization</skos:prefLabel>
  <skos:prefLabel xml:lang="fr">Hybridation</skos:prefLabel>
  <skos:prefLabel xml:lang="es">Hibridación</skos:prefLabel>
  <skos:prefLabel xml:lang="ar">تهجين</skos:prefLabel>
  <skos:prefLabel xml:lang="zh">杂交</skos:prefLabel>
  <skos:prefLabel xml:lang="pt">Hibridação</skos:prefLabel>
  <skos:prefLabel xml:lang="th">การผสมพันธุ์</skos:prefLabel>
  <skos:prefLabel xml:lang="ja">雑種形成</skos:prefLabel>
  <skos:prefLabel xml:lang="sk">hybridizácia</skos:prefLabel>
  <skos:prefLabel xml:lang="de">HYBRIDISIERUNG</skos:prefLabel>
  <skos:prefLabel xml:lang="hu">hibridizáció</skos:prefLabel>
  <skos:prefLabel xml:lang="pl">Hybrydyzacja</skos:prefLabel>
  <skos:prefLabel xml:lang="fa">دورگه‌گی‌ری</skos:prefLabel>
  <skos:prefLabel xml:lang="it">Ibridazione</skos:prefLabel>
  <skos:prefLabel xml:lang="hi">वर्ण संकर उत्पन्न करना</skos:prefLabel>
  <skos:altLabel xml:lang="cs">tvorba hybridů</skos:altLabel>
  <skos:altLabel xml:lang="en">Hybridizing</skos:altLabel>
  <skos:altLabel xml:lang="fa">دورگه‌گی‌ری</skos:altLabel>
  <skos:altLabel xml:lang="hi">संकरणी करण</skos:altLabel>
  <skos:altLabel xml:lang="hu">hibridizálás</skos:altLabel>
  <skos:altLabel xml:lang="it"></skos:altLabel>
  <skos:altLabel xml:lang="ja">雑種形成</skos:altLabel>
  <skos:altLabel xml:lang="lo">ການເຮັດໃຫ້ເກີດລູກປະສົມ</skos:altLabel>
```

# Summary



# Future

(This page has been amended to include responses received during the presentation.)

Existing backlog will approximately triple the database size.

What other research could it benefit?

NSF-supported project on the collection of lexicons of endangered languages.  
Thai-language.com project.

What schema changes would help?

Integration of RDF triples into the schema.

Provision of 1-field primary keys for record types without them:

meaning	domain	word class
denotation	definition	metadatum
approver	meaning identifier	

What omitted data should be added?

Etymology.

Usage example.

Resource quality.

# Future

(This page has been amended to include responses received during the presentation.)

## What interface features should be added?

Tag-cloud-like graphical representation of relations among expressions.

Word-list authoring tool.

Graphical representation of the schema.

Application programming interface.

URI access to display of information about individual expressions.

Transaction logs.

## How should the content harvesting be improved?

Proceduralize resource conversion so it can be repeated with revised versions.

# Thanks

## Contributors to this project:

Kobi Reiter

Oren Etzioni

Marcus Sammer

Michael Schmitz

Mausam

Stephen Soderland

Susan Colowick

Michael Skinner

Chris Lim

Catherine Ono

Janara Christensen

Kate Everitt

Koshal Thirumalai

## and hundreds more, including:

ABBYY Software

Achille Spacciante

Adrian Otoiu

Adrian R.W. Room

AfroWeb

Ahmad Mumtaz

Aleandro Amadio, Pino Leo

Allen Hillel Merkrebs

Antonio Rapuano, Daniela Castrovillari

Aquilina Mawadza

Arnaud Le Floch

Attilio Farina

Ausseil

Bernard Vatant

Bernard Vivier

Besiki Sisauri

Carlo Minnaja

Christiane Fellbaum

Conor Quinn

Cristiano Screm

Cyril Babaev, Cristiano Screm

Cyril Babaev, Dario de Judicibus

Daniela Falessi

Dario de Judicibus

Dario de Judicibus, Claudio Porcellana

David M. Captain, Linda B. Captain

David Smyth

Dipti Misra Sharma et al.

Dévényi Károly

Edmondo Monti

Eduardo Sadier

Eesti Keele Instituut

El Meneghin

Emmanuel Rodary

Ergane

European Communities

European Schoolnet

Eurydice

Felice Pescatore

Food and Agriculture Organization of the United Nations

Franco Questa

Franco Questa, Dario de Judicibus

Franco Questa, Dinesh Prabhu

Franco Questa, Helena Ambroskiewicz

Franco Questa, Iqbal Wali

Franco Questa, Wang FuSheng

Frits van Zanten, Alain Rousseau, Arie Taal

Gabriele Brunini

Giampaolo Mazzola

Gildas Perrot

HIROTAKA Masaaki

Harold F. Schiffman

Heymans Institute of Pharmacology;

Mercator School, Department of Applied Linguistics

Ho Ngoc Duc et al.

Hope Studio

Horst Eyermann

Indian Institute of Technology and University of Hyderabad

International Monetary Fund

Jane Fajans

Jim Breen

Jitendra Chaudhary

Joan Francés Blanc

Joan-Francés Blanc

John Correll

Karl Hesse and Theo Aerts

Kay Williamson

Kisii Village Project

LEXiTRON

Language Technologies Research Centre and University of Hyderabad

Laura Dorst, Giancarlo Pocetta, Kerstin Karlström, et al.

Leon Kuperman

Les Mondes Polaires

Libor Sztemon

Linguistic Data Consortium

Luca Carrozzi

Marc Nery, Dario de Judicibus

Maria Amarillas

Marie-Christine Hazaël-Massieux

Mark R. Laws

Mary Kawena Pukui, Samuel H. Elbert

Michel Ditria, Bernard Lubin, Bruno Charpentier, Dario de Judicibus

Mickaël Estace, F. Ramaco

Microsoft Corporation

Mike Wright

Ministry of Water Resources, Government of India

Mélanie Coste-Bonnet, Dario de Judicibus

Nicola Selenu, Luca Ballore

Nino G. Barbieri

Nino Vessella

Northern Illinois University, Center for Southeast Asian Studies

Norwaydict

Nunavut Arctic College

Oren Etzioni, Kobi Reiter, Stephen Soderland, Marcus Sammer

Pamela Smith & Adebusola Onayemi

Paolo Castellina, Alberto Mardegan

Parahat, Akmuhammet, Sapar

Pascal Bayle, Michel Ditria, Dario de Judicibus

R. Hilgers; Yashovardhan

Razen Mandahar, Jacob Nordfalk, Narendra Bhattherai, Phillip Pierce, Poshraj

Robert Bassford

Robert Bassford, Fotis Gortsilas

Robert Bassford, Luca Carrozzi

Robert Young and Rosemary Young

SIL International

Sailendra Biswas

Simon Greenhill, Robert Blust, Russell Gray

Stuart Campbell, Chuan Shaweevongse

Sulayman Hayyim

Terminology and Reference Section, Documentation Division, DGACM

The Kamusi Project

The Library of Congress

Ufficio del Turismo di Tonga in Italia

Unicode Consortium

Università di Parigi, Dario de Judicibus

Universität Heidelberg, Seminar für Computerlinguistik

Vaman Shivaram Apte

Wayne Reid

Werner Fröhlich

Wikipedia

Yüksel Demir

Özer Ozankaya

Θεόφιλος Βαμβάκος

A. V. Александров

A.M. Касаев, Т.А. Гуриев

Б. С. Воскобойников

Б. С. Воскобойников, В. Л. Митровић

В. В. Быков, А. А. Поздняков

В. Д. Новиков

В. Л. Ривкин

В.И. Абаев

Владимир Савельев

Г. В. Чернов

Д.В. Поликанов

Е. А. Бокарёв, Юрий Финкель, Игорь Галицкий

Е. Г. Коваленко, А. И. Гриценко

Е. К. Масловский

Кирилл Панфилов

Корейско-Русский Словарь

М. А. Сторчевой

М. В. Тверитнев

Н. Ю. Борисова, М. А. Сторчевой

О. И. Чибисова, Н. Н. Смирнов

С. В. Глядков

С. М. Баринов

С. Н. Андрианов, А. С. Берсон

Т. Е. Апанасенко, М. А. Сторчевой

Ф.М. Таказов

Э. М. Пройдаков, Л. А. Теплицкий

Экономическая школа

Ю. Н. Маляревская, М. А.Сторчевой

刘芳