

Language Technology Ecosystem

Richard L. Sites

Google, Inc

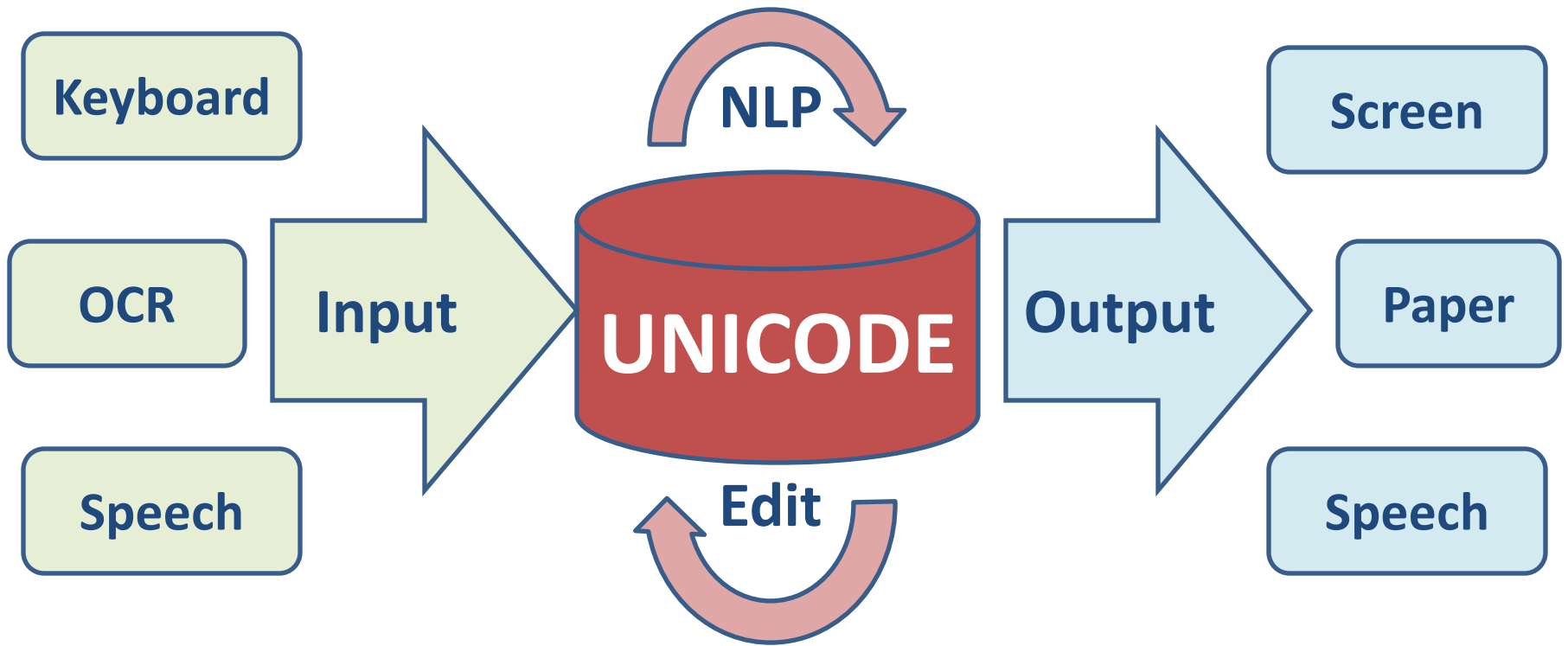
May 3, 2011

Abstract

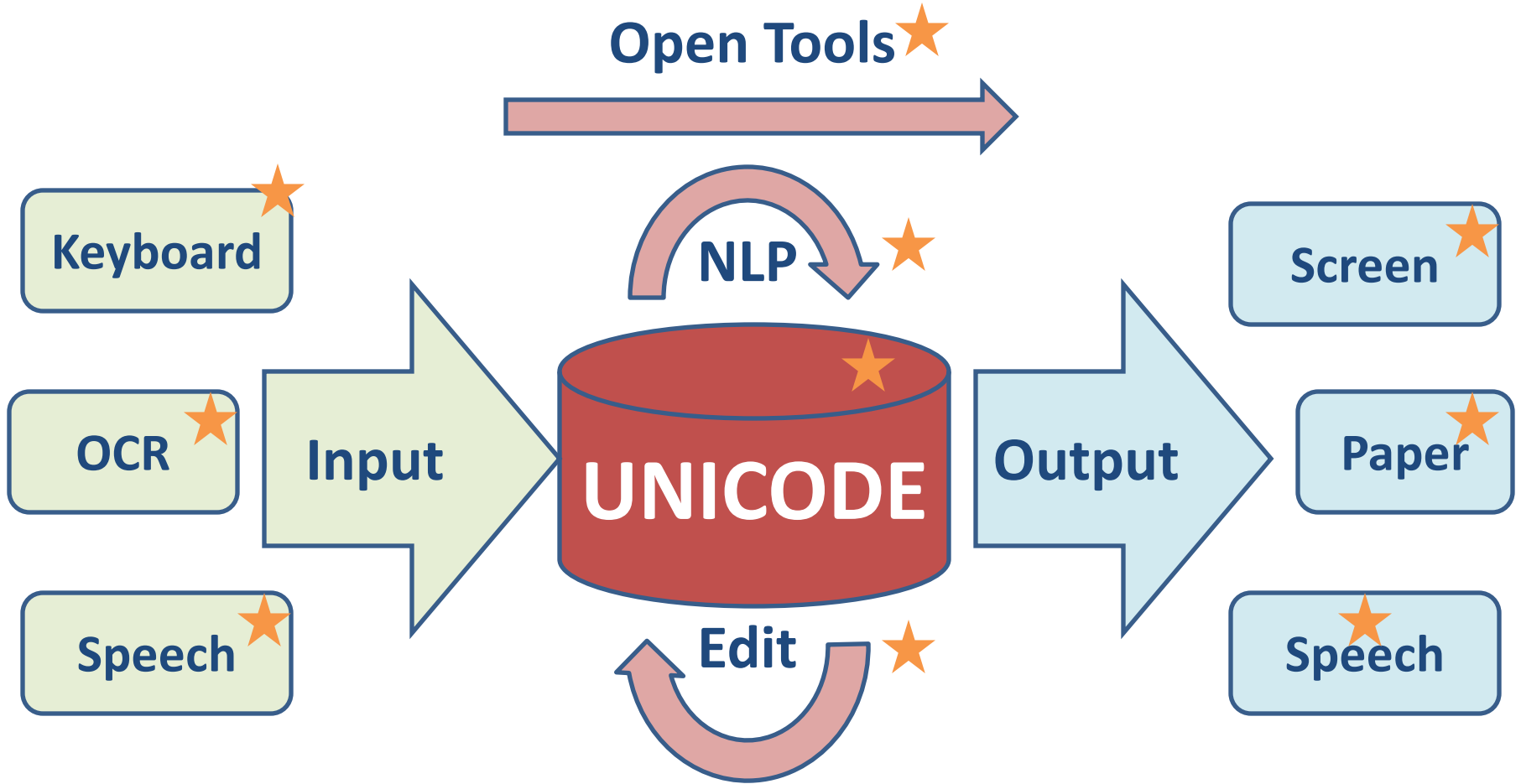
As information dissemination moves from traditional speech and written media to computers and the Web, language technology is a key to improved education, improved social contact, and improved commerce. Populations that cannot access information in their own languages are left out.

Conversely, Internet companies that invest in the **language ecosystem** enable development and also increase their user population by millions of people.

Language Ecosystem

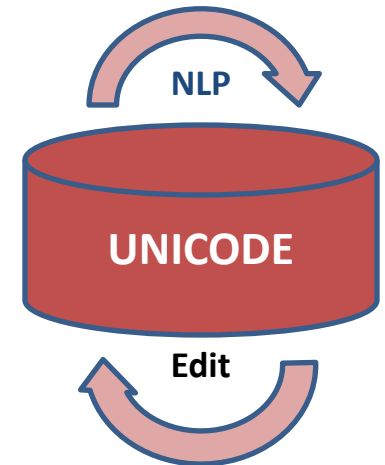


Google Language Ecosystem

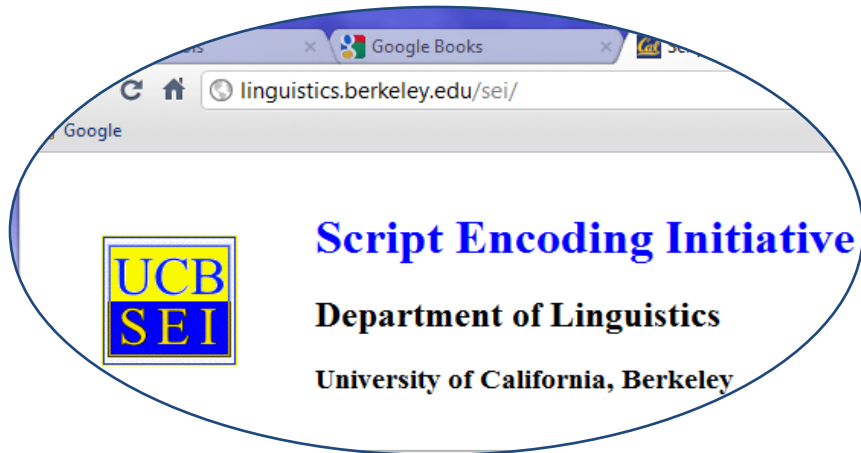
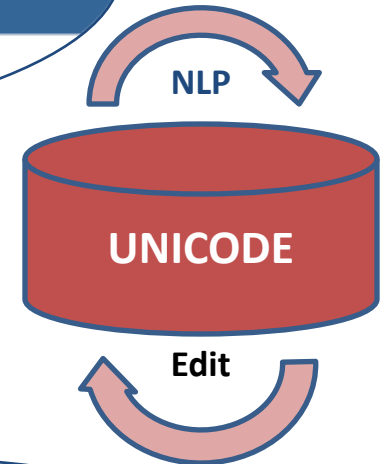
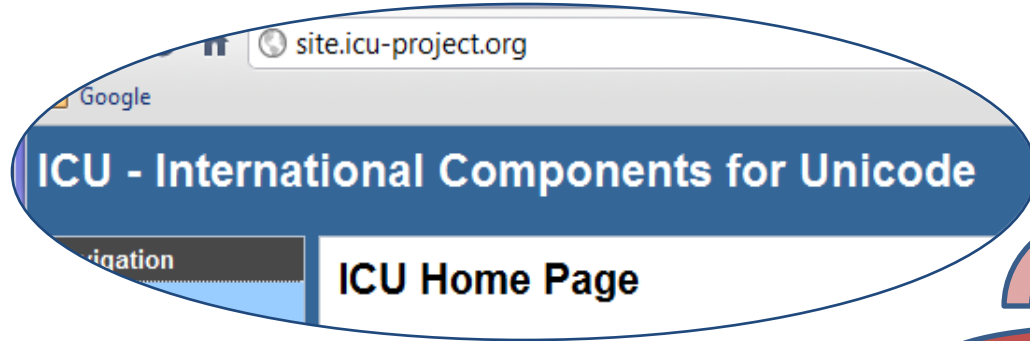
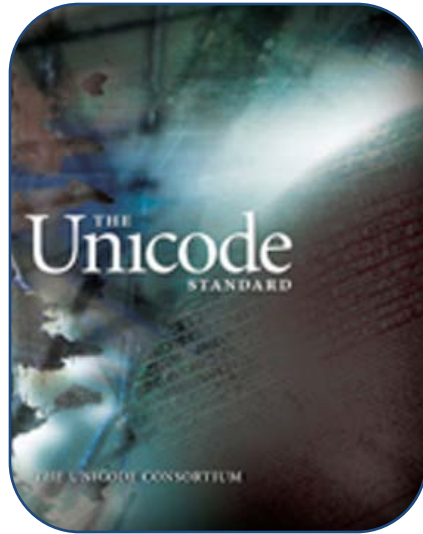


Unicode

- All incoming text is converted to Unicode
- All processing is done on Unicode
- **No Unicode = No search**

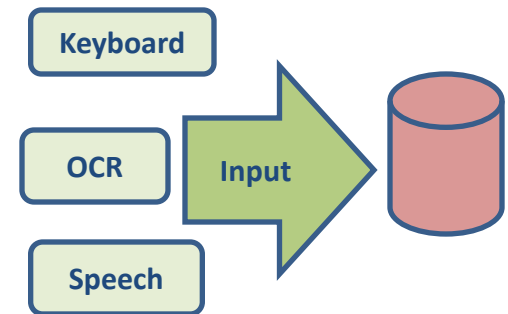


Supporting Unicode



Input

- Virtual keyboards
- OCR to text
- Speech to text
- Google IMEs, transliteration
 - Arabic, Bengali, Farsi (Persian), Greek, Gujarati, Hindi, Kannada, Malayalam, Marathi, Nepali, Punjabi, Tamil, Telugu and Urdu
 - Pinyin
- Chrome browser

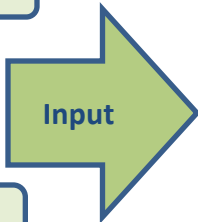


Input

Keyboard

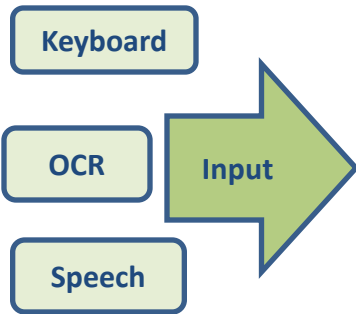
OCR

Speech



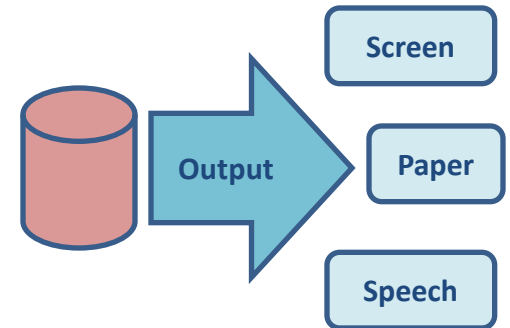
GERONIMO
sacrifice was **deemed necessary**. Sometimes
the offending one was punished.
If an **Apache** had allowed his age
suffer for food or shelter

Input

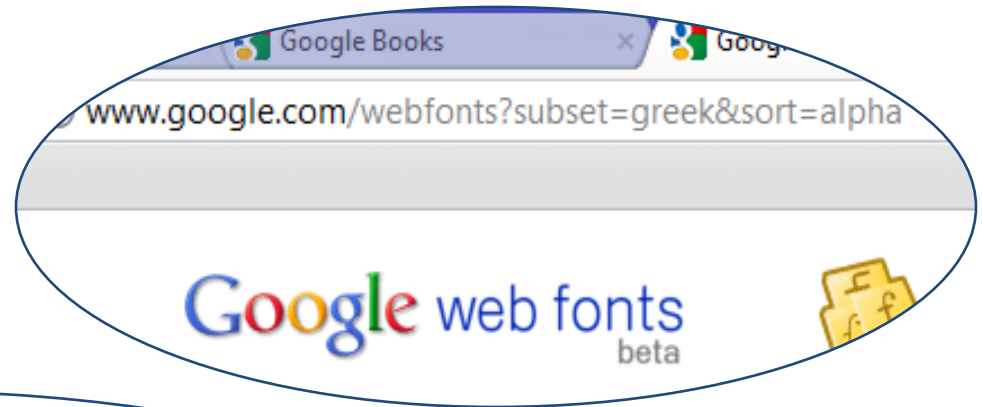
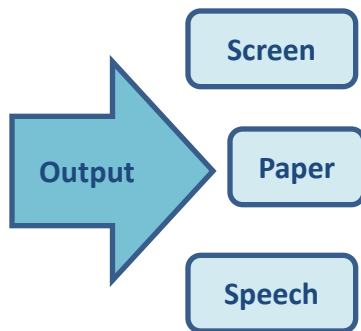


Output

- Fonts (Google funding)
- Text to speech
- 1500+ local-language versions of Google various services
- 40-language initiative, forcing issues up front:
 - Keyboards, fonts, BIDI, form fill-in, text width, names, addresses, phone numbers, dates, currency, advertisements
- Chrome browser



Output



Hitting 40 languages

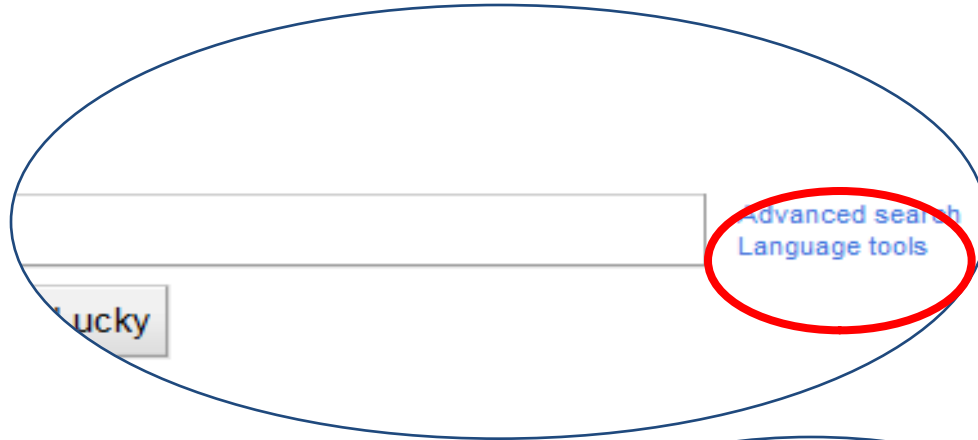
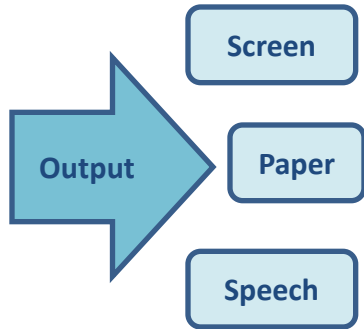
7/18/2008 07:01:00 AM

One of our goals is to give everyone using Google the information they want, wherever they are, in whatever language they speak, and through whatever device they're using. Part of that goal is making our services available in as many languages as possible. It's not always sure you can imagine, that isn't as easy as simply as translating a few lines

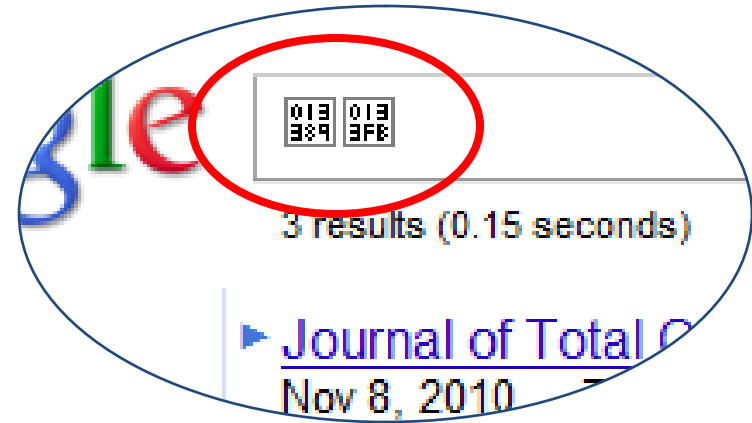
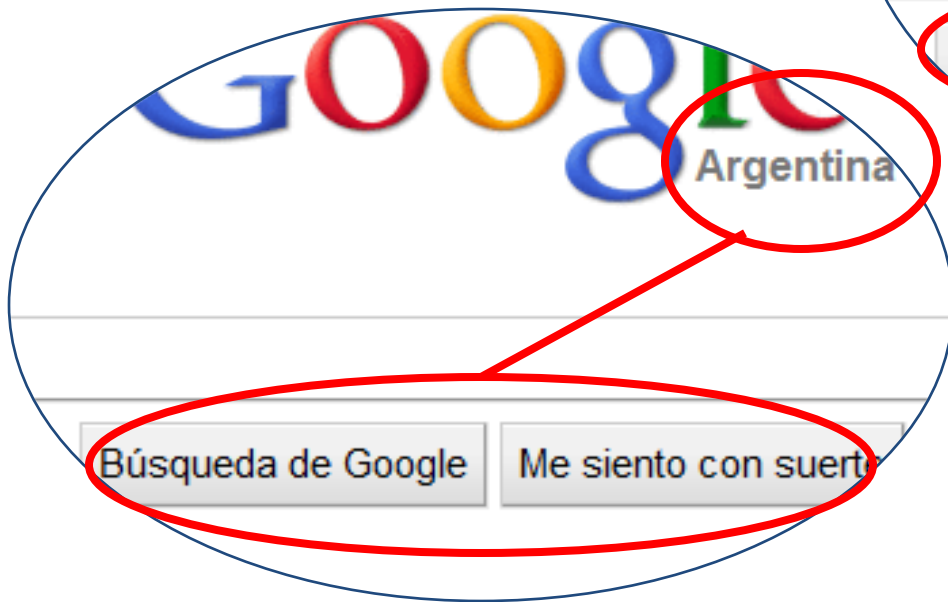
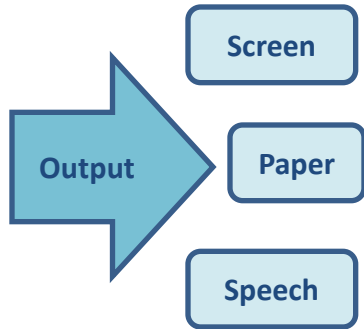
Hebrew or Arabic, which are written from right to left. An Arabi

[Football 2008] [كأس العالم 2008 لكرة القدم]. Part of the

User Interface

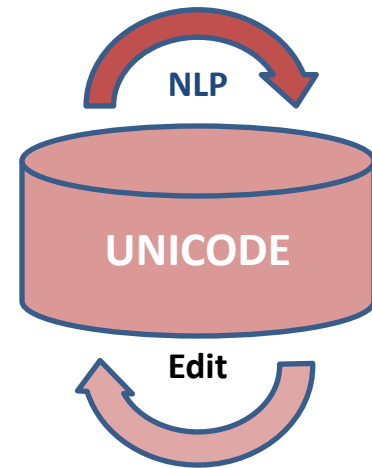


User Interface

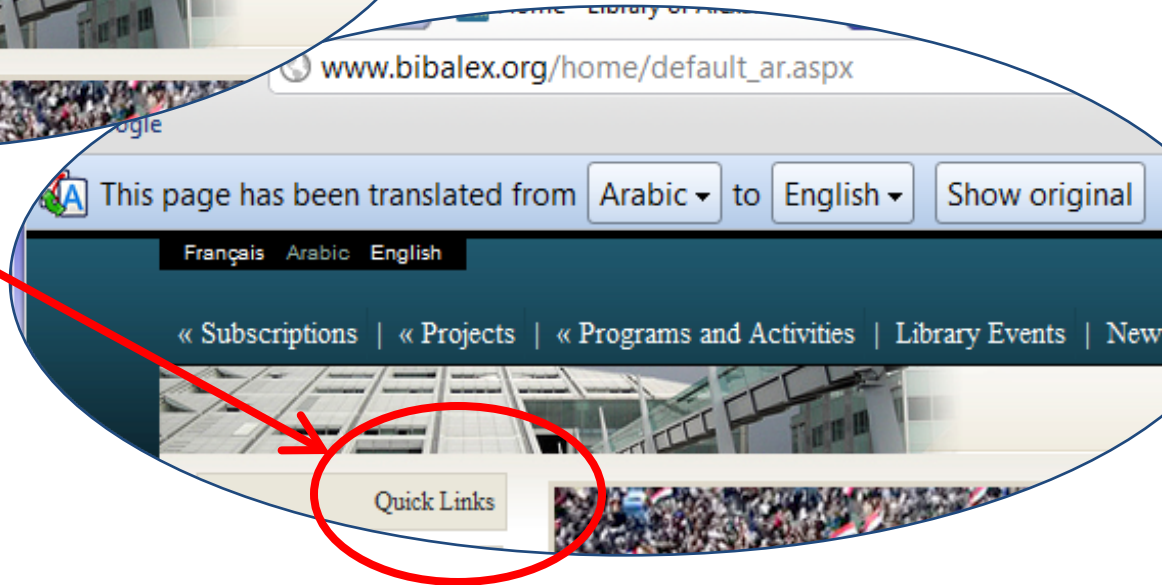
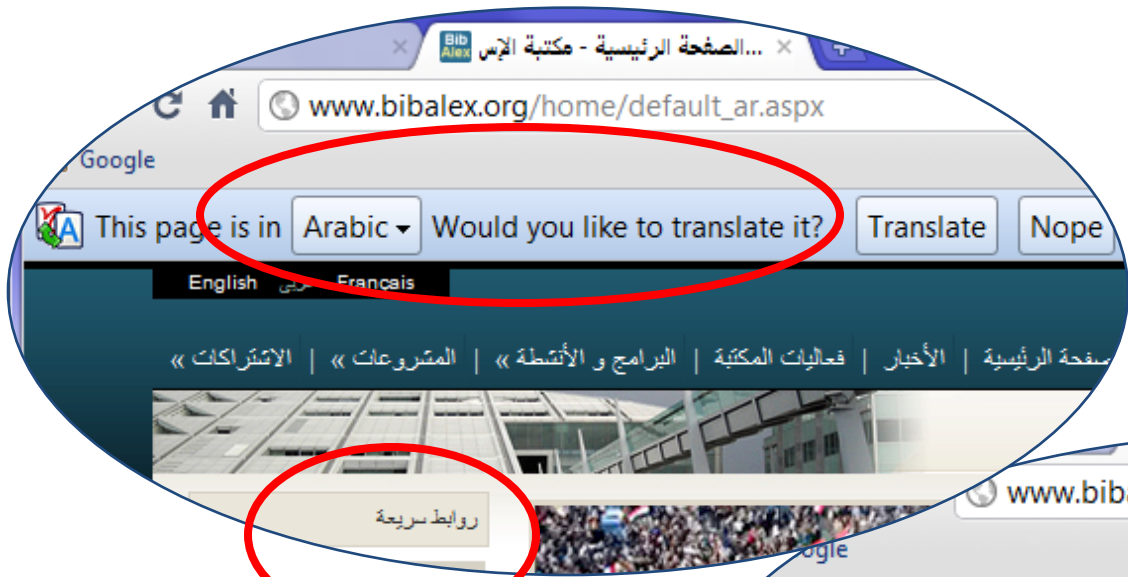


Language Processing

- Language detect
- Spelling correct/suggest
- Translation

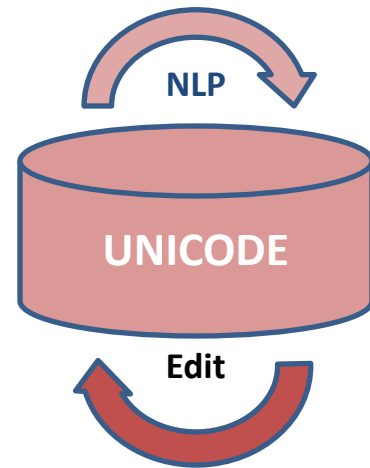


NLP: Detect, Translate



Editing

- URL, user name, address, date, currency, numbers, telephone numbers, forms of address, place names
- Parse: whitespace, quotes, delimiters, line breaks
- Singular/dual/plural, male/female, bidi



Editing

25.7 km

2. Matar Al Kahera Al Dawelli/مطار القاهرة الدولي dakika 51

23.0 km

A Cairo International Airport
القاهرة
مصر

1. Elekea kaskazini mashariki kwenye Airstrip Of International Airport/مهبط مطار القاهرة الدولي kuelekea Matar Al Kahera Al Dawelli/مطار القاهرة الدولي

400 m

2. Chukua njira pili kashoto kuingia Matar Al Kahera Al Dawelli/مطار القاهرة الدولي



Caire Visite

Les choses à ne pas manquer tout au Caire - épargne en ligne

Camera dome ip panasonic interieure 30x

References : | 049104 | Camera dome ip panasonic interieure 30x | Caméra Dôme réseau avec zoom optique 30x et technologie SDIII Sensibilité élevée avec ...

3 183,75 € - Neuf

Livraison gratuite
Maison-Des-Cameras

Open Tools

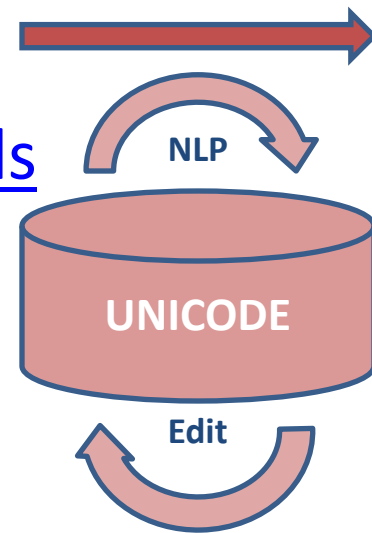
- Open source

http://translate.google.com/translate_tools

<http://src.chromium.org/>

<http://chrome.google.com/>

- E.G. auto-translate button on web sites



Open Tools

Google Books

code.google.com/apis/language/

拡散予測: [日本気象庁](#) | [ドイツ気象局](#) | [オーストリア気象局](#) | [イギリス気象局](#) | [ノルウェー気象局](#)

その他: [原子炉保安院による報道資料](#) | [茨城原発周辺](#) | [宮城県全域](#)

お知らせ 4/18 [福島県内の小中学校等の放射線量マップ](#)を追加しました。

Language: [Select Language](#) [B!](#) [tweet](#) 13.8K [Like](#) 7K

[Select Language](#) [Czech](#) [Hebrew](#) [Malay](#) [Spanish](#)

[English](#) [Danish](#) [Hindi](#)

[Afrikaans](#) [Dutch](#) [Hungari](#)

diffusie: [Japan Meteorological Agency](#) | [Weer lessenaar Duitsland](#) | [Oostenrijk Weertburg](#)

Ander: [Press Release van Nuclear Safety Agency](#) | [Area Primère Itaraki](#) | [Miyagi hele](#)

Original text: [Google™](#) [X](#)

文科省が公表している 情報 を元に日本全国の放射能値をグラフ化しています。

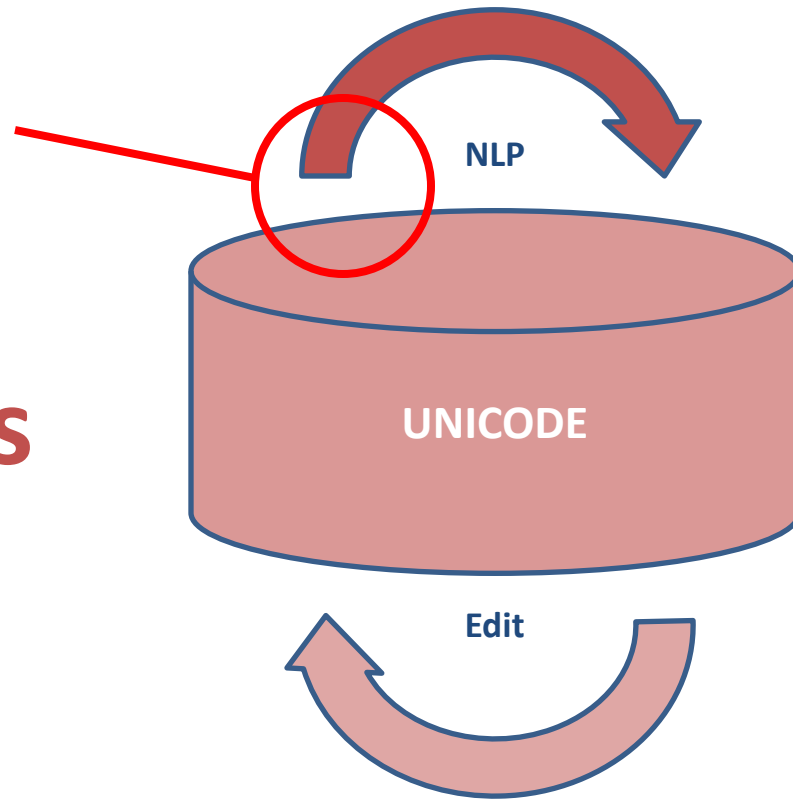
Die bediening het gepubliseer inligting in Japan te grafiek om die waarde van die oorspronklik Sodra die data is gepubliseer en opgedateer as die grafiek.

src.chromium.org/viewvc/chrome/trunk/src/third_party/cld

index of /trunk/src/third_party/cld

(next)

Statistical Language Detection in Web Pages



The problem

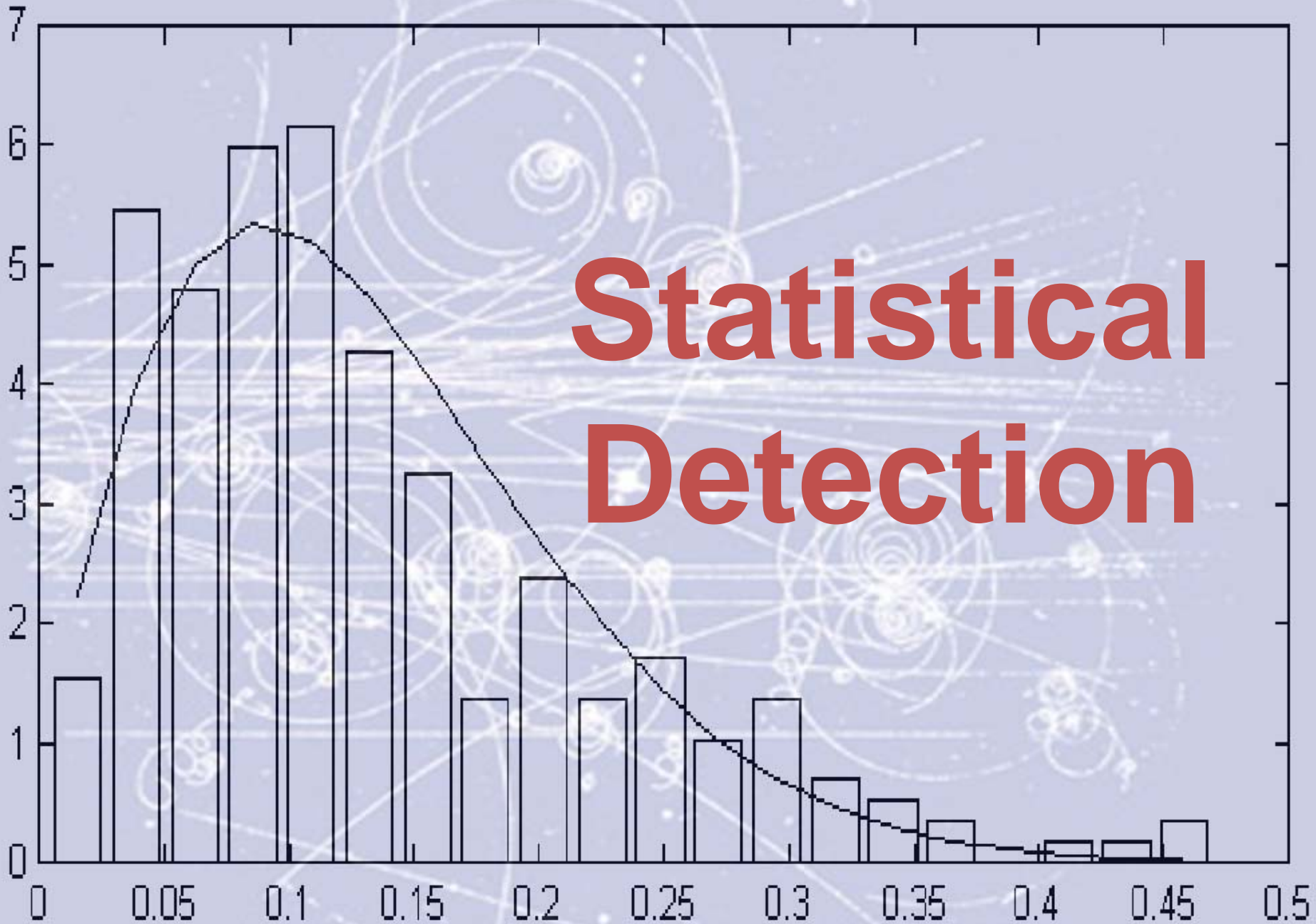
- Given a web page, determine the (human) languages it contains
- **Why?**
 - A French user preferentially sees French pages
 - Your browser and email can offer to translate text

Ce blog est une invitation au voyage, à la rencontre d'hommes et de femmes qui font de beaux vins, des vins de terroir, sans sacrifier aux modes. Un lieu de débat aussi, sans chanelle ni

A première vue, on pourrait croire à un incendie. Un feu de plus dans la dernière longueur d'un été meurtrier. Comme dans les Corbières il y a dix jours, on

⇒ **French**

Statistical Detection



Statistical Detection Overview

1. Training data to tokens and counts (offline)

French
training data



```
_des_ fr-Latn 21216547  
_les_ fr-Latn 26148645  
_une_ fr-Latn 11695306  
_dans_ fr-Latn 9181286  
_pas_ fr-Latn 7472078  
_vous_ fr-Latn 6469354  
_être_ fr-Latn 2382920  
  
...  
TOTAL fr-Latn 1448745452
```

English
training data



```
_now_ en-Latn 2978526  
_is_ en-Latn 59608675  
_the_ en-Latn 283476990  
_time_ en-Latn 6300667  
_for_ en-Latn 45836482  
_all_ en-Latn 8928935  
_good_ en-Latn 2591014  
  
...
```

Statistical Detection Overview

2. Token counts to **relative occurrence** log probabilities (offline)

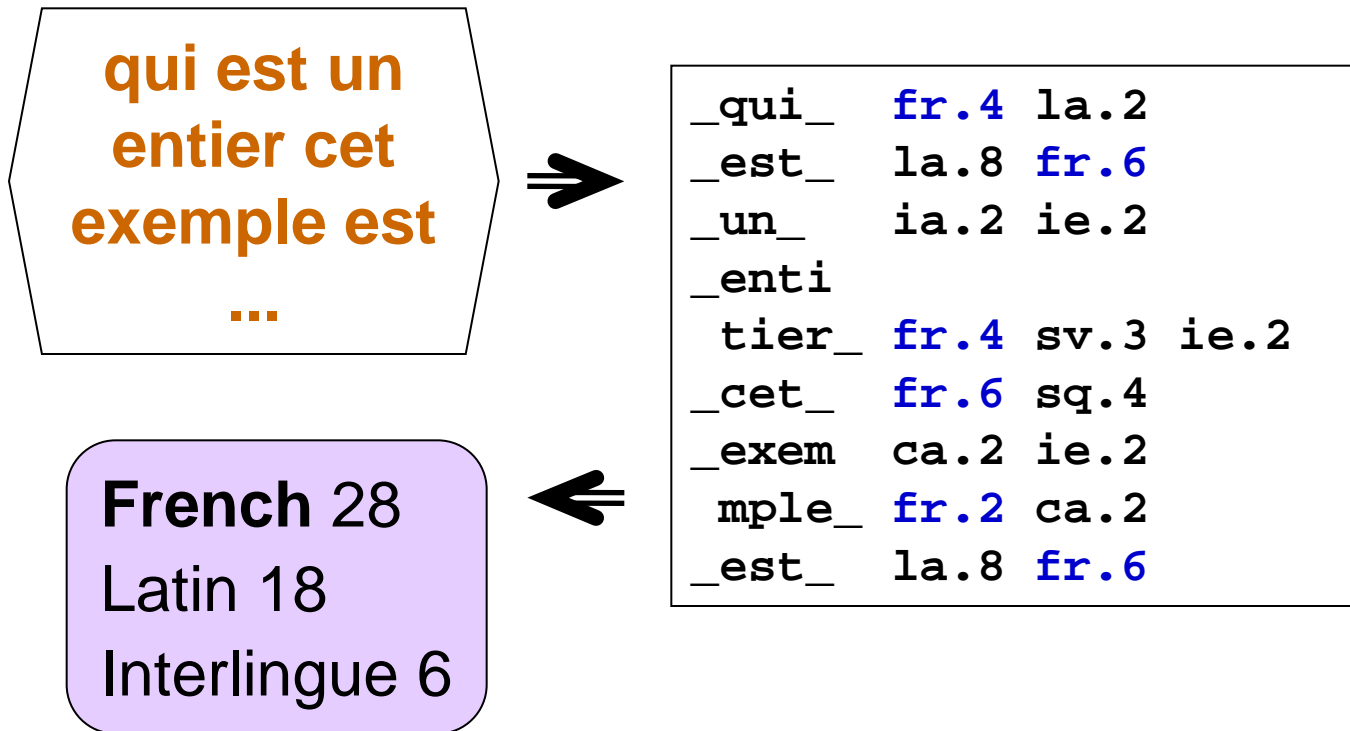
```
_camp ca-Latn 38022 / 20370777 = 1/536    lg = -9.1  
_camp ia-Latn  4270 /  4301970 = 1/1007   lg = -10.0  
_camp pt-Latn 66353 / 55361916 = 1/834    lg = -9.8  
_camp ro-Latn 49589 / 33645079 = 1/678    lg = -9.5
```

```
_camp ca 091  ro 095  pt 098  ia 100
```

Quantize probabilities and put tokens into a large hash table for detect-time lookup.

Statistical Detection Overview

3. Unknown page to tokens, to quantized language log probabilities, to **sum** (detect)



Token = Aligned Quadgram+

(include pre/post space, shown as underscore)

The quick brown fox jumped ...

The

_quic

ck_

_brow

wn_

fox

_jump

mped_

Significant information in word begin/end

Example

```
<HTML> <BODY>  
OPAC Online Public  
Access Catalog<BR>  
©1999 Perpustakaan STT  
Telkom <BR>  
The old saying "you  
need the right tool for  
the right job" applies  
even to your Internet  
business.  
<P>keyword <BR>  
untitled document  
...  
</BODY> </HTML>
```



Lowercase, just letters:

opac online public access
catalog perpustakaan stt
telkom the old saying you
need the right tool for the
right job applies even to
your internet business
keyword untitled document
...

Example

Lowercase, just letters:

opac online public access
catalog perpustakaan stt
telkom the old saying you
need the right tool for the
right job applies even to
your internet business
keyword untitled document
...

**ENGLISH(100%),
other(0%)**



opac	ia.4
_onli	xxx.4 ms.2
line_	xxx.4 et.3 ms.2
_publ	wo.2 ca.2
blic_	en.2
_acce	it.2 ro.2
cess_	en.2 rm.2
_cata	oc.2 vo.2
talo	eo.2
log_	xxx.2 vo.2
_perp	la.2 sq.2
rpus	id.2 ms.2
stak	
kaan_	om.4 fi.2
stt	
_telk	
lkom_	nl.4 fy.2
the	en.8
old	en.6 ie.4
...	

Mixed-Language Pages

- Break text into chunks of a dozen words or so
- Score top language of each chunk separately

[LUXEMBOURGISH] diskussioun haaptsäit wikipedia déi fräi enzyklopedie disk
[lb 24/de* 16] wikipedia der fräier enzyklopedie wiesselen op navigatioun sich h

[ENGLISH] tulations with the new wiki because lëtzebuergesch is also spoken i
[] added it to the country portal http www wikipedia be if you have users form
[] belgium consider using that url for publicity in belgium for this wiki greetings
[] kipedias do like this it would make it easier for me each time i update the mul

[GERMAN] al statistics pages thanks jul utc halli hallo viele grüße von der deu
[] tschen wikipedia noch nicht viel los hier oder trotzdem viel mut für die nä
[] chsten jahre auch wenn es noch lange dauern wird bis diese enzyklopädie br
[de 43/lb* 36] ckhaus qualität erreicht aug utc inhaltsverzeechnis wei starte mer

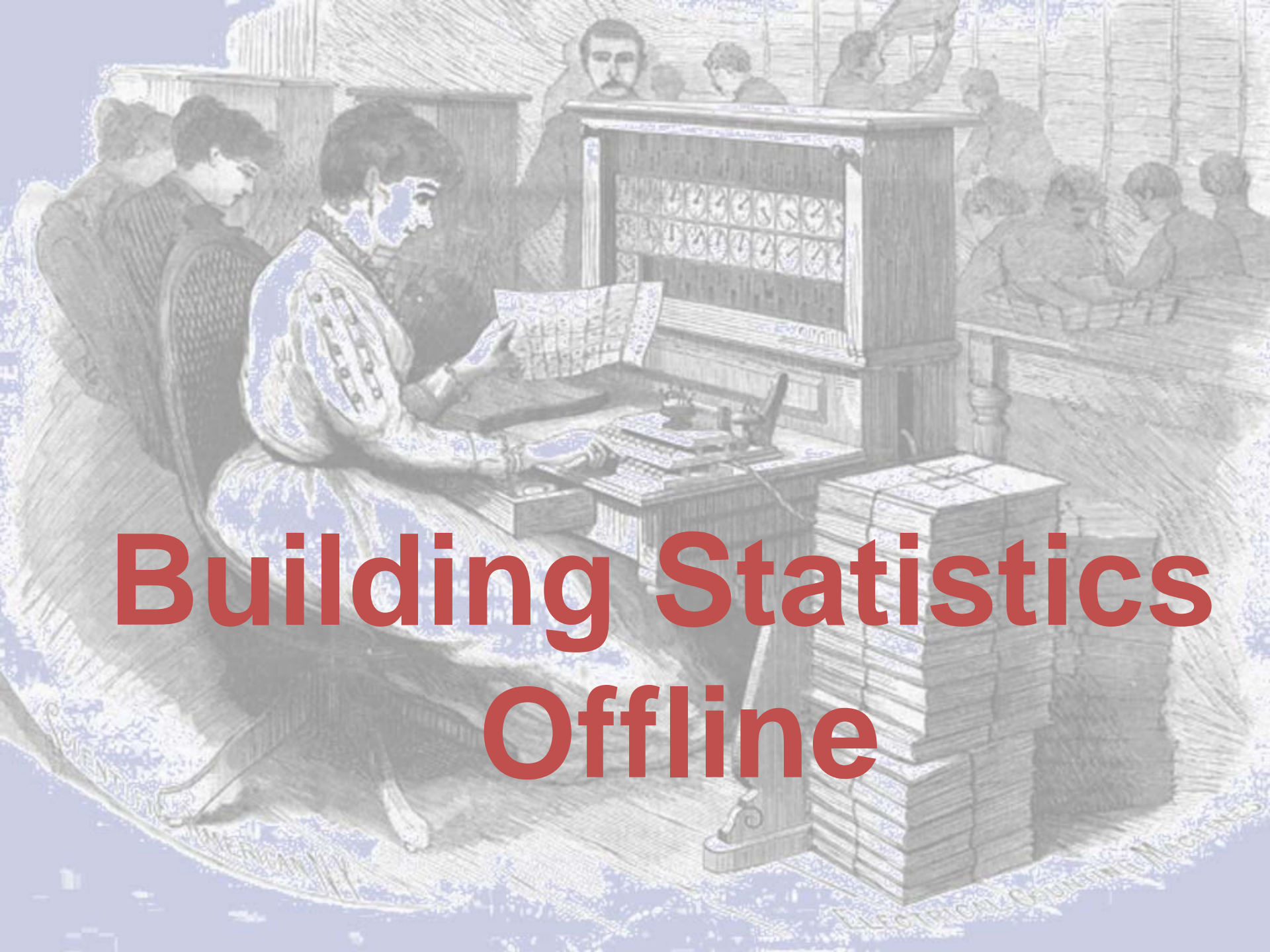
[LUXEMBOURGISH] chen artikel vun der woch reportage gréngs

[TAMIL] தமிழ்

[LUXEMBOURGISH] traut de kalenner sech net sorry eng

[] banal fro läänner bäiflécken startsäit neie modell vun haaptsäit fir d éischt sch

[FRENCH] iwen an da verbessern création d un portail supportant les wikipédi



Building Statistics Offline

Building Statistics Offline

- Training data to token counts (mapreduce)

power kiat mencegah
kejenuhan berpikir pengarang
brandreth gyles penerbit
semarang dahara no induk
jumlah kelas judul brainware
management generasi kelima
manajemen manusia
pengarang bahaudin t
penerbit jakarta elex media
komputindo no induk jumlah
kelas judul branding in asia
membangun merek di



	Quadgram	Count
QG id-Latn	_powe	1
QG id-Latn	wer_	1
QG id-Latn	_kiat_	1
QG id-Latn	_menc	1
QG id-Latn	nceg	1
QG id-Latn	gah_	1
QG id-Latn	_keje	1
QG id-Latn	jenu	1
QG id-Latn	nuha	1
QG id-Latn	han_	1
QG id-Latn	_berp	1
	...	

Building Statistics Offline

Token counts to languages and log probabilities

a	ha-Latn 046 gd-Latn 047 hu-Latn 050 to-Latn 050 sco-Latn 052 co-Latn 056
aa	sco-Latn 091 om-Latn 095 ik-Latn 110 sit-NP 125 tn-Latn 126 su-Latn 136 ..
aaa	om-Latn 152 gl-Latn 171 xxx-Latn 177 ru-Latn 188 cy-Latn 190 en-Latn 191 ..
_aaaa	xxx-Latn 170 cy-Latn 184 fy-Latn 193 ru-Latn 194 fo-Latn 196 az-Latn 213 ...
aaaa	ht-Latn 145 fo-Latn 163 oc-Latn 165 la-Latn 171 ia-Latn 178 gl-Latn 180 ...
_aaab	en-Latn 243
_zábo	sk-Latn 196 cs-Latn 211
_zábr	cs-Latn 194 sk-Latn 205
_zábě	cs-Latn 167
ώχου	el-Grek 221
ώχρα	el-Grek 215
_асуд	tk-Cyrl 131
_асқа	kk-Cyrl 239
لاقتا	kk-Arab 121
لالما	ug-Arab 167
لاماء	kk-Arab 121 ...

Building Statistics Offline

- Prune vertically (frequent quadgrams) and horizontally (top 3 languages only)
- Put into a large hash table

_aaab	en-Latn 243	
_zábo	sk-Latn 196	cs-Latn 211
_zábr	cs-Latn 194	sk-Latn 205
_zábě	cs-Latn 167	
ώχου	el-Grek 221	
ώχρα	el-Grek 215	
_асуд	tk-Cyrl 131	
_асқа	kk-Cyrl 239	
لاقتا	kk-Arab 121	
الاما	ug-Arab 167	
لاماء	kk-Arab 121	...



Bootstrapping

Bootstrapping Training Data

- Need training text to build language detector
- Need language detection to select training text

Oops!

- Start with hand-selected known-language text
- Use detector version N to find text for N+1
- Cross-check selected text

Bootstrapping Training Data

- What web pages to select for training data?
 - Need starting data for **each** of ~180 language-script pairs (of which I read two)
- fr.wikipedia.org [not so reliable, it turns out]
- news.bbc.co.uk/arabic
- www.voanews.com/persian/
- UDHR, Watchtower.org, etc.
- URLs containing language name/code:
[www.bhutan2008.bt/dz/...](http://www.bhutan2008.bt/dz/)

Bootstrapping Cross-checking

- Use pages from specified web-site patterns
- Use pages with a good amount of real text, and in only one language
- Cross-check
 - Top-level domain (e.g. [www.foo.ee](#) for Estonian)
 - Encoding (e.g. [GB2312](#) for Simplified Chinese)
 - Right letters for each language ("exemplars")
 - Common words

What Goes Wrong?



Training Data

Jumbled Training Data:

lb.wikipedia.org

<http://lb.wikipedia.org/wiki/Diskussioun:Haaptsäit>

Diskussioun:Haaptsäit

LUXEMBOURGISH
except as noted

Vu Wikipedia, der fräier Enzyklopedie.

Wiesselen op: [Navigatioun](#), [Sich](#)

ENGLISH

Hi, congratulations with the new Wiki! Because Lëtzebuergesch is also spoken in Belgium I have added it to the country portal <http://www.wikipedia.be> If you have users from Belgium consider using that url for publicity in Belgium for this Wiki. Greetings,
w:nl:gebruiker:walter

What? I'm absolutely flabbergasted. In BELGIUM, for real? Well, just recently (and I'm over 30), a Belgian girl told me in chat that she was German. Huh...Belgian-German? Yes! From Eupen, and that's indeed a part of Belgium hardly known about abroad: the German-speaking part! Flanders, Wallonia ... of course. But that a German minority is officially recognized, is some information arousing surprised looks outside Belgium. -andy [217.91.47.231](#) 08:36, 3 Februar 2006 (UTC)
I don't know about this girl in particular, but in principle people living around Eupen, St. Vith etc. (i.e. the East Cantons) are Belgians and not Germans, though they are German-speaking. They form an own federal community within Belgium. --Otets [★](#) 10:57, 3 Februar 2006 (UTC)
Otets, the German-speaking part of Belgium does not form its own federal comunity. It is a part of Wallonia. The German-speaking part of Belgium only forms some sort of "language community" which is used for example for statistical reasons.

How about including the {{NUMBEROFARTICLES}} link on the main page, as most Wikipedias do, like this: [24.375](#). It would make it easier for me each time I update the multilingual statistics pages! Thanks! [210.55.230.18](#) 13:21, 22 Jul 2004 (UTC)

GERMAN

Halli, hallo, viele Grüße von der deutschen Wikipedia!!! Noch nicht viel los hier, oder??? Trotzdem viel Mut für die nächsten Jahre, auch wenn es noch lange dauern wird, bis diese Enzyklopädie Brockhaus-Qualität erreicht--[217.2.42.13](#) 16:02, 5 Aug 2004 (UTC)

Inhaltsverzechnis

- [1](#) Wei starte mer???
- [2](#) ASBL & sou Saachen
- [3](#) Artikel vun der Woch
- [4](#) Reportage
- [5](#) Gréngs
- [6](#) தமிழ்
- [7](#) Traut de Kalenner sech net??
- [8](#) Sorry
- [9](#) Eng banal Fro
- [10](#) Länner bäiflücken
- [11](#) Startsäit
- [12](#) Neie Modell vun Haaptsäit
- [13](#) Fir d'éischt schreiwen an da verbesseren
- [14](#) Création d'un portail supportant les wikipédias dans les langues sont "traditionnellement présentes sur l'espace étatique français" ?
- [15](#) Datum

TAMIL

FRENCH

Omnipresent English

Afrikaans **camping in** frankryk skryf jou eie seks

Catalan ubicació exacta dels camps **camping** puigcerdà t interessa

Corsican ozzies cam pin ozzies **camping** ozzies**camping** ozzies **camping center**

Czech pláž kosmetické potřeby manikúry **camping** a grilovací potřeby

Welsh ymholiad hyn **hotel and camping** a **hotel near camping** yn edrych

Danish nået frem ferieboligmarked **camping** rejs med fdm travel

German mitfahrerinnen auf einem **camping**platz mit dusche morgens bekommen

English **advanced search map list all camping and caravans overview**

Spanish edad posts ese **camping** se deberia de llamar

Basque agroturismo kanping **camping** hostel kirol ekintzak actividades deportivas

Finnish båtslip kb vierasvenesatama **camping** alue talvisäilytys venehuolto

French les randonnées ou le **camping** notamment les radiobalises

Hungarian megtiltotta a mindenki által történő hozzászólást **camping**

Latin secundamano **camping** minibus para comprar minibus con chofer vigo

Luxembourgish verstéiss géint d bauteréglement um **camping** ugeet sou muss een

Dutch vakantie tafka mag niet mee naar de **camping** hij moet naar een pension

Norwegian velkommen til det nordiske **camping** treff les mer **the tall ships races** måløy

Swedish besöktes av flera invigning av **camping** så har då äntligen

Swahili wa kwanza uitwao bongo **camping** ni vyama viwili rafiki sisi wanachama

Uzbek kam o zbekistonda ham **camping rafting mountain tours** shaklida turli turlarni

Weird Web Pages

Remove Highly Compressible Text

booking for all major worldwide cities and travel destinations as Europa, Asia, Australia, Africa, United States, France, Spain United Kingdom, PULAU LANGKAWI, MY and much PULAU LANGKAWI SHERATON LANGKAWI BEACH RESORT hotel and more... PULAU LANGKAWI hotels, Bringing you some of the cheap vacation deals on the web. We offer a quick, easy and secure way to make your reservations online with instant availability checks and confirmations. Not only will we save you time and money but with our extensive destination guides we'll also help you plan your trips. Our new accommodation search engine will give you quick and easy access to thousands of (5 stars, 4 stars, 3 stars), hostels, motels, PULAU LANGKAWI best western hotels, pensions and apartments., , all eorld hotels world wide hotels welcomes you to the home of PULAU LANGKAWI car rental online where you can book your car hire or rent a car at any of our worldwide locations. , , We also offer PULAU LANGKAWI sightseeing tours, travel tours, PULAU LANGKAWI flight ticket and PULAU LANGKAWI vacation packages at around the world. Hotels and accommodation in PULAU LANGKAWI, MY, SHERATON LANGKAWI BEACH RESORT PULAU LANGKAWI **CN - WORLD HOTELS**, , All Right Reserved 2005 © SHERATON LANGKAWI BEACH RESORT PULAU LANGKAWI , PULAU LANGKAWI hotels, hotel 07.01.2008 14:32:14 : SHERATON LANGKAWI BEACH RESORT PULAU LANGKAWI Online hotels reservations. SHERATON LANGKAWI BEACH RESORT, aSHERATON LANGKAWI BEACH RESORT, bSHERATON LANGKAWI BEACH RESORT, cSHERATON LANGKAWI BEACH RESORT, dSHERATON LANGKAWI BEACH RESORT, eSHERATON LANGKAWI BEACH RESORT, fSHERATON LANGKAWI BEACH RESORT, gSHERATON LANGKAWI BEACH RESORT, hSHERATON LANGKAWI BEACH RESORT, iSHERATON LANGKAWI BEACH RESORT, jSHERATON LANGKAWI BEACH RESORT, kSHERATON LANGKAWI BEACH RESORT, lSHERATON LANGKAWI BEACH RESORT, mSHERATON LANGKAWI BEACH RESORT, nSHERATON LANGKAWI BEACH RESORT, oSHERATON LANGKAWI BEACH RESORT, pSHERATON LANGKAWI BEACH RESORT, rSHERATON LANGKAWI BEACH RESORT, sSHERATON LANGKAWI BEACH RESORT, tSHERATON LANGKAWI BEACH RESORT, uSHERATON LANGKAWI BEACH RESORT, vSHERATON LANGKAWI BEACH RESORT, wSHERATON LANGKAWI BEACH RESORT, ySHERATON LANGKAWI BEACH

DNA Sequences



Ralstonia metallidurans: Rmet_1390

Help

Entry	Rmet_1390	CDS	R.metallidurans
Definition	malate synthase (EC:2.3.3.9)		
Orthology	KO: K01638 malate synthase		
Pathway	PATH: rme00620 Pyruvate met PATH: rme00630 Glyoxylate a		
Class	Metabolism; Carbohydrate Met [PATH:rme00620] Metabolism; Carbohydrate Met metabolism [PATH:rme00630] BRITE hierarchy		
SSDB	Ortholog Paralog Gene cluster		
Motif	Pfam: Malate_synthase PROSITE: MALATE_SYNTHASE Motif		
Other DBs	JGI_1: Rmet1390 NCBI-GI: 94310332 NCBI-GeneID: 4038193 UniProt: Q1LNK3		
LinkDB	All DBs		
Position	1:1499337..1500920 Genome map		
AA seq	527 aa AA seq DB search MAITLPAGMKITGEILPAYEDILTPEAL LPDFLPETKSIREGDWKVAVVPKALECR		

NT seq	LIPEELAKVREVVGAGATYDRAAQIFEQ MSTSEDFAEFLTLPLYEEV
1584 nt	NT seq +upstream <input type="text" value="0"/> nt +downstream <input type="text" value="0"/> nt
	atggctatcacgctgcccgcgggcatgaagattaccggcgagattctgcccggcttatgaa gacatcctgacgcgggaagccctggccctggctcgacaagctgcaccgtgccttcgaggcc cgccgtcaggaactgctggccgcgcgcgtggcgcgaccaagcgctcgacgcggcgaa ctgcccgaacttctgcccgaaccaagagcatccgcgaaggcgactggaaggttgccgcg gtgcccgaaggcgctggaatgccgcgcgcgtggaatcaccggctccggctcgaagccaagatg gtgatcaacgccttcaactcggggcgtgacagctacatgaccgacttcgaggactccaac acccgaaactggcacaaccagatgcagggctcaggtgaacctgaagtccgcctgcccgcg acgctgacgctggacagcaacggaaagcactacaaactcaacgacaagatcgccacgctg caggtgctccgcgcggctggcacctagacgagaagcagctcacgatcgacggcaagcgc gtctcggggcggcatcttcgacttcgcgctgttctctgttccacaacgccaaggagcagatt gcccgcggcgccggcccgttcttctatctgccgaagatggaaagccatctggaagcgcgt ctgtggaacgacatcttctgtgatggcgcagaaggaagtcgggtctgcccgaaggcacggtc aaggccacgggtgctgatcgagacgatcctggccgcgttcgagatggaagaaatcctgtac gaactgcccgcgacagcgcggcctgaacgctggccgctgggactacatcttctcgtgc atcaagaagttcaaggtcgacaagaacttctgccttgccgaccgcgccaaggtgacgatg acctgcgcttcatgcccgcctacgcgctgctgctgctgaagacctgccacaagcgcggc gcccctgcatcggcggcatgagcgcgctgatcccgatcaagaacgatccggagaagaac gctatgccatgcagggcatcatcggcgacaagcgcgctgacgccaccgacggctacgac ggcggctgggtggcccaccgggtctggtcgagccggcaatgaaggaattcgtggacgtg ctgggcgaccgcccgaaccagttcgacaagcaacgtccggacgtcgaagtgaagggtgcc gacctgctgaacttccagccggaagcgcggatcaccgaagccggcctgcgcatgaacatc aacgtcggcatccaactacctgggcgctggctggctggcaatggctgctgcccgatccac aacctgatggaagacgcggccaccgcccagatctcgcgctcccaggtgtggcagtgatc cgctgcgcaagggcaagctggaagacggccgcaaggtgacggccgagatggtgcccgcg ctgattccggaagaactggccaaggtccgcgaagtggtgggcgaggcggccacgtacgac cgtgccgcgagatcttcgagcagatgtcgacttcggaagactttgccgaattcctgacg

DNA Sequences

[en/tl*] dbget result r metallidurans rmet ralstonia metallidurans rmet entry rmet cds r metallidurans
definition malate synthase a ec orthology ko k malate synthase pathway path rme pyruvate metab
path rme glyoxylate and dicarboxylate metabolism class ssdb motif pfam malate synthase [] pros
synthase other dbs jgi rmet ncbi gi ncbi geneid uniprot q lnk linkdb position aa seq aa
maitlpagmkitgeilpayeditpealalvdklhrafearrqellaarvartkrlidage lpdfipetksiregdwkvapvpkale
crrveitgpveakmvinafnsgadsymtdfedsnpnwhnqmqqvnlksavrrtltldsngkhy [eu/siN*]klndkiatlqvrprgw
dgkr vsggifdfalffhnakeqiargagpffylpkmeshlearlwndifvmaqkevglpqgtv katvlietilaafemeeilyelrehsaglr
agrwdyifscikkfkvdknfcladrakvtm tspfmrayalllktchkrgapaiggmsalipikndpeknaiamqqiigdkrrdatdgyd
ggwvahpglvepamkefvdlgdrpnqfdkqrpdvkvgadllnfqpeapiteaglrnmi nvgihylgawlagngcvpihnlmedaa
eisrsqwqwirspkgkledgrkvtaemvta [so/en*] lipeelakvrevvgagatydraaqifeqmstsedfaefltlplyeev nt se
upstream nt downstream nt atggctatcaggctgcccgcgggcatgaagattaccggcgagattctgccggcttatgaa
gacatcctgacgccggaagccctggccctggtcgacaagctgcaccgtgccttcgaggcc
cgccgtcagga [] actgctggccgcgcgctggcgcgcaccaagcgctcgacgccggcgaa
ctgccgacttcctgccggaaccaagagcatccgcgaaggcgactggaaggttgcgccc

...

cgtgccgcgagatcttcgagcagatgtcgacttcggaagactttgccgaattcctgacg ctgccgctgtacgaagaagtctga origin
dbget integrated [SWEDISH] database retrieval system genomenet [] =Final=
Languages **SOMALI*(40%)** ENGLISH(16%) TAGALOG(12%) 2638 bytes

"IgnoreMe" Language

DNA [acgt]*, Link farm, Filename farm, HTML tags

[en/tl*] dbget result r metallidurans rmet ralstonia metallidurans rmet entry rmet cds r metallidurans
definition malate synthase a ec orthology ko k malate synthase pathway path rme pyruvate metab
path rme glyoxylate and dicarboxylate metabolism class ssdb motif pfam malate synthase [] pros
synthase other dbs jgi rmet ncbi gi ncbi geneid uniprot q lnk linkdb position aa seq aa
maitlpagmkitgeilpayeditpealalvdklhrafearrqellaarvartkrdage lpdfipetksiregdwkvapvpkale
crrveitgpveakmvinafnsgadsymtdfedsnpnwhnqmqqvnlksavrrtltldsngkhy [eu/siN*]klndkiatlqvrprgw
dgkr vsggifdfalffhnakeqiargagpffylpkmeshlearlwndifvmaqkevglpqgtv katvlietilaafemeeilyelrehsaglr
agrwdyifscikkfkvdknfcladrakvtm tspfmrayalllktchkrgapaiggmsalipikndpeknaiamqqiigdkrrdatdgyd
ggwvahpglvepamkefvdlgdrpnqfdkqrpdvkvgadllnfqpeapiteaglrnmi nvgihylgawlagngcvpihnlmedaa
eisrsqvwqwirspkgkledgrkvtaemvta [xxx/en*] lipeelakvrevvgagatydraaqifeqmstsedfaefltlplyeev nt s
upstream nt downstream nt atggctatcaccgctgcccgcgggcatgaagattaccggcgagattctgccggcttatgaa
gacatcctgacgccggaagccctggccctggtcgacaagctgcaccgtgccttcgaggcc
cgccgtcagga [] actgctggccgcgcgcgtggcgcgcaccaagcgcctcgacgccggcgaa
ctgccgacttctgccggaaccaagagcatccgcgaaggcgactggaaggttgcgcgcg

...

cgtgccgcgagatcttcgagcagatgtcgcacttcggaagactttgccgaattcctgacg ctgccgctgtacgaagaagtctga origin
dbget integrated [SWEDISH] database retrieval system genomnet [] =Final=
Languages **ENGLISH(16%)** TAGALOG(12%) **Ignore*(40%)** 2638 bytes

Difficult Choices

Malay-Indonesian



Mencari Cinta Mu

Age:26
Occupation:PhD student
Address:Loughborough, UK

<< February 2005 >>

Sun	Mon	Tue	Wed	Thu	Fri	Sat
		01	02	03	04	05
06	07	08	09	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28					

Rindu yang tiada penghujung

Ya Allah
Sekiranya dia adalah suami pilihan
Mu di Arash
Berilah aku kekuatan dan keyakinan

Monday, February 28, 2005

Jiwang sikit

Assalamu'alaikum wbt.

2 hari aku tak mampu mengurus diri, tak boleh makan, tak boleh bau makanan..Badan rasa lemah tahap maksima sekaligus nafsu menguasai akal..Perasaan nak balik Malaysia membuak2, teringat waktu2 sakit di bawah jagaan mak..Terasa susahny sakit diperantauan ni..Apatah lagi apabila teringin nak makan itu ini..Hubby aku manusia paling penyabar, hanya aku dan Allah saja yang tahu tahap kesabaran dia memang amat mengagumkan!

Kerja research berlambak2 memaksa aku kuatkan semangat...Apatah lagi hari hari ni ada appoinment dengan University Health and Safety Officer dan esok (uwaaa) ada meeting bulanan dengan both of my supervisors. Pagi ini dalam kereta menuju ke office;

Hubby: Kuatkan semangat..Nanti 'dia' pun kuat semangat juga.
Aku: -pause-
Hubby: Tarik nafas panjang2..sedut udara luar sikit.

Close Pair: Indonesian - Malay

Indistinguishable via quadgrams

Sample Indonesian - Malay pairs

bagian	bahagian
bahwa	bahawa
bisnis	bisnes
keuangan	kewangan
majelis	majoriti
maluku	melayu
ngak	nak
serikat	syarikat

Use Whole-words

Force to Majority Language

[ms 55/id* 44] mencari cinta mencari cinta mu age occupation phd student address

[id 185/ms* 182] hujung ya allah sekiranya dia adalah suami pilihan mu di arash k

[ms 180/id* 172] mi yang akan membimbing tanganku dititiammu kurniakanlah ak

[id 164/ms* 164] ku di jannah mu limpahkanlah aku dengan sifat tunduk dan tawa

[ms 184/id* 180] rbaik untukku di duniamu peliharalah tingkah laku serta kataku c

[id 66/ms* 61] menghadapi segala kerenah dan ragamnya http profiles blogdrive

[ms 123/id* 115] weblog enter your email here tuesday february alamat betul ke a

[id/ms*]seperti wanita muslim jangan harap anda akan disayangi sepe

[MALAY] rti kucing parsi jika anda adalah tikus longkang yang busuk lagi membu

{ClosePair 166: **move 4867 bytes id => ms**}

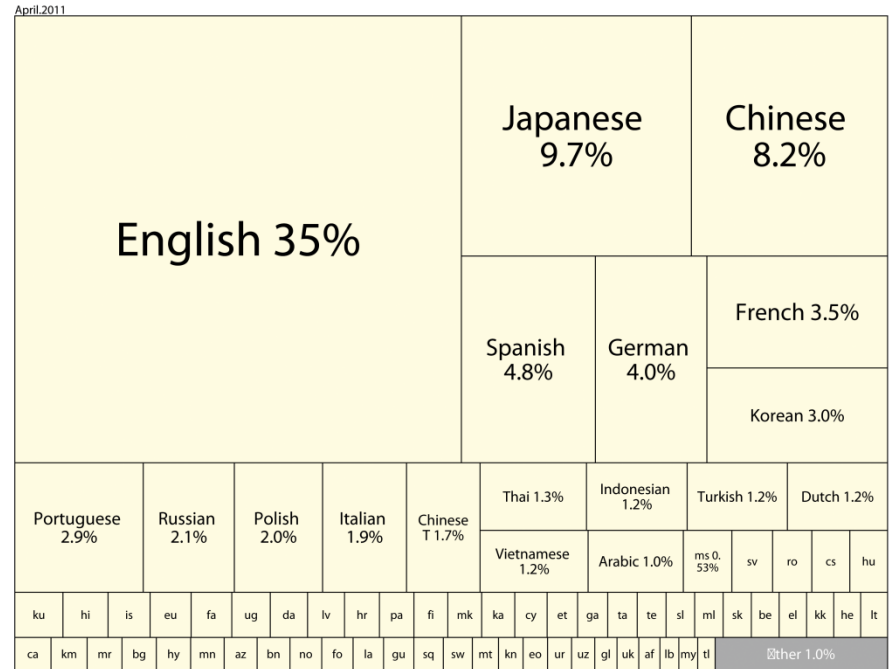
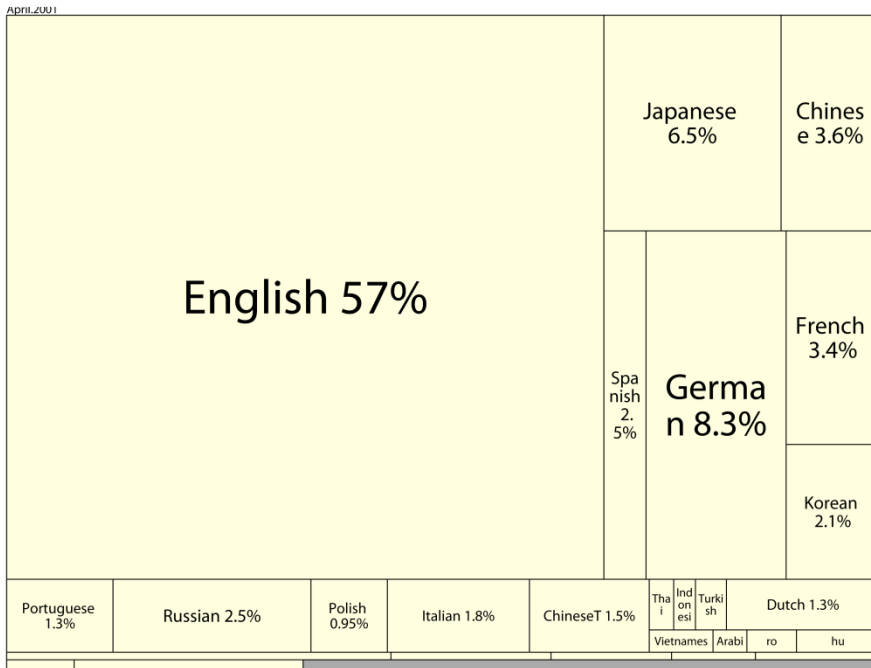
Languages **MALAY(100%),** **INDONESIAN(0%),** **other(0%),** 13/33 KB

Summary

- Text on the Web is unedited, intermixed, and sometimes not even text
- Difficult to find clean training data
- Quadgrams + Whole-words does pretty well
- There is a lot of garbage on the web

Webshare 2001 => 2011

(99% of active web: 32 =>74 languages)



English 35%

Japanese 9.7%

Chinese 8.2%

Spanish 4.8%

German 4.0%

French 3.5%

Korean 3.0%

Portuguese 2.9%

Russian 2.1%

Polish 2.0%

Italian 1.9%

Chinese T 1.7%

Thai 1.3%

Indonesian 1.2%

Turkish 1.2%

Dutch 1.2%

Vietnamese 1.2%

Arabic 1.0%

ms 0.53%

sv

ro

cs

hu

ku

hi

is

eu

fa

ug

da

lv

hr

pa

fi

mk

ka

cy

et

ga

ta

te

sl

ml

sk

be

el

kk

he

It

ca

km

mr

bg

hy

mn

az

bn

no

fo

la

gu

sq

sw

mt

kn

eo

ur

uz

gl

uk

af

lb

my

tl

Other 1.0%

Arabic up by 10x, Turkish by 5x. NEW-to-the-top-99% Kurdish, Hindi, Persian, Uighur, Pashto, Macedonian, Tamil, Telugu, Malayalam, Kazakh, Hebrew, Marathi, Armenian, Azerbaijani, Bengali, Gujarati, Swahili, Maltese, Kannada, Urdu, Uzbek, and Afrikaans

Questions?

