Language Technology Ecosystem

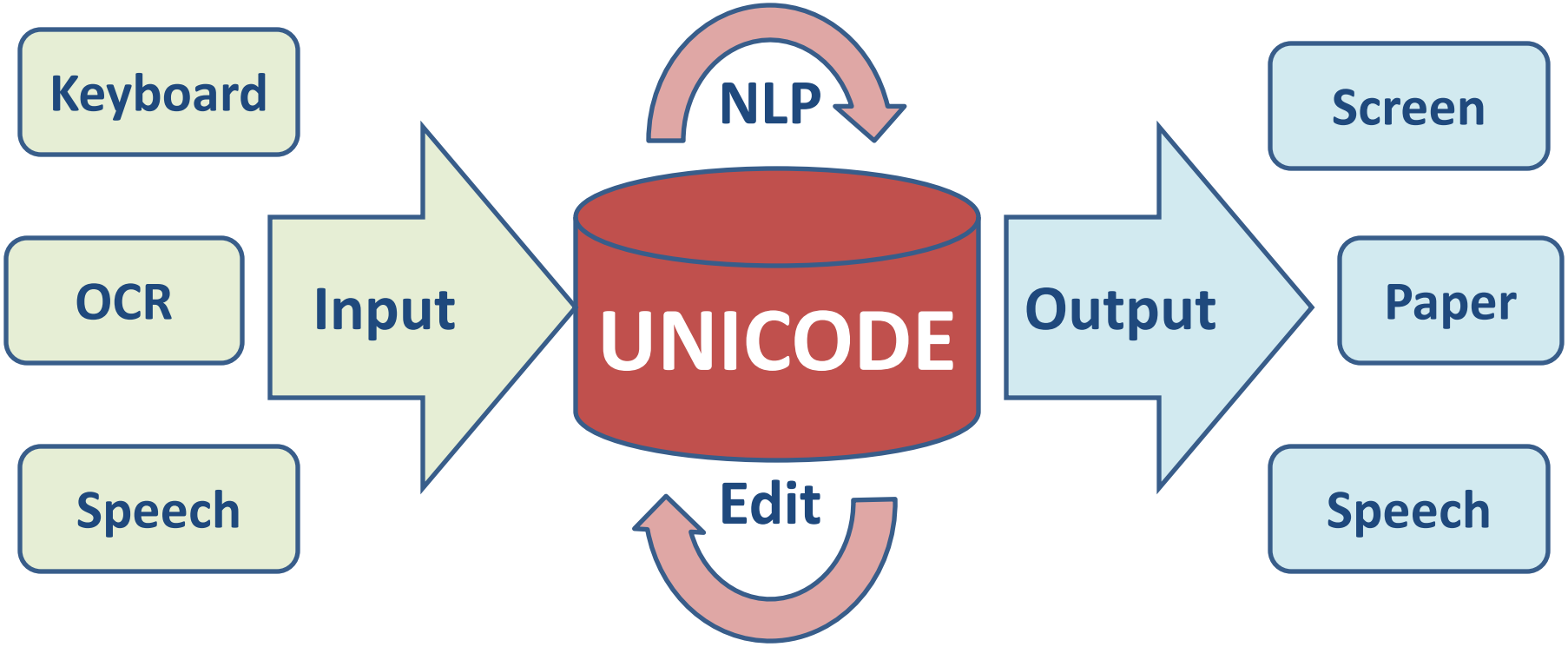
Richard L. Sites
Google, Inc
May 3, 2011

Abstract

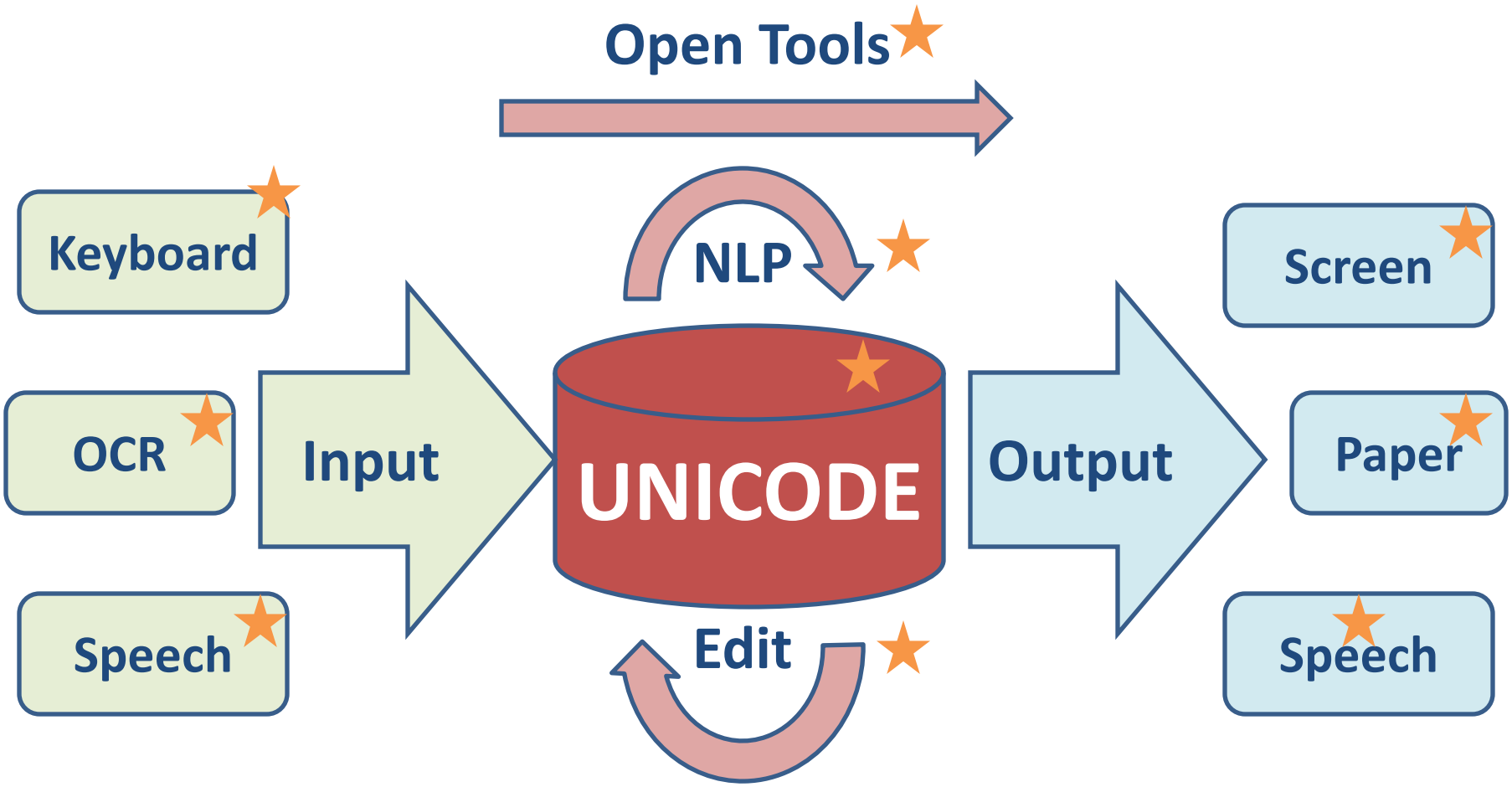
As information dissemination moves from traditional speech and written media to computers and the Web, language technology is a key to improved education, improved social contact, and improved commerce. Populations that cannot access information in their own languages are left out.

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

Language Ecosystem



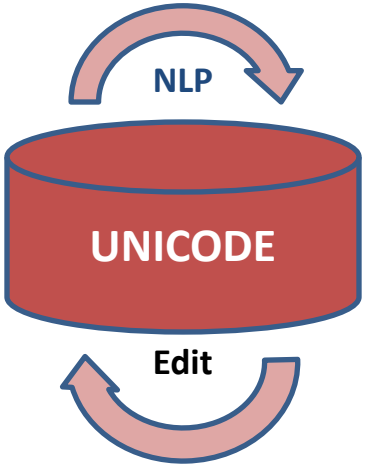
Google Language Ecosystem



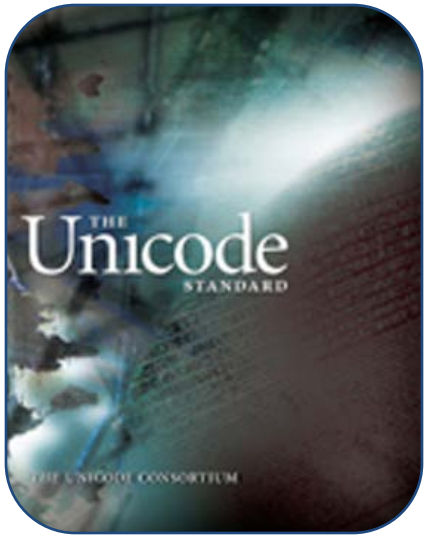
Unicode

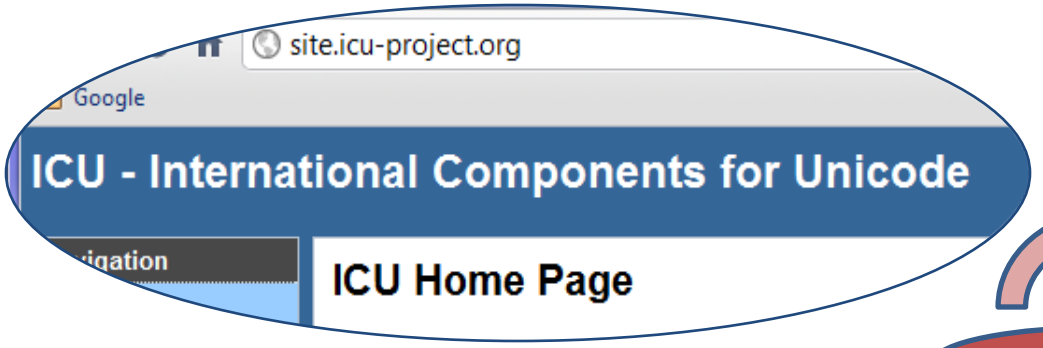
- All incoming text is converted to Unicode
- All processing is done on Unicode

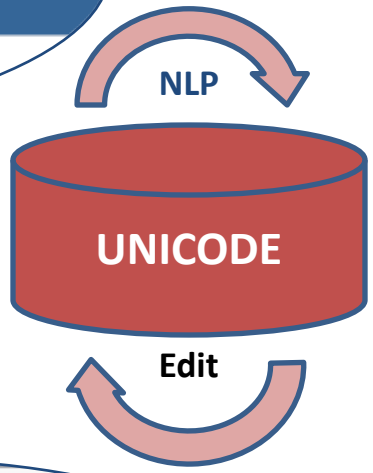
- **No Unicode = No search**

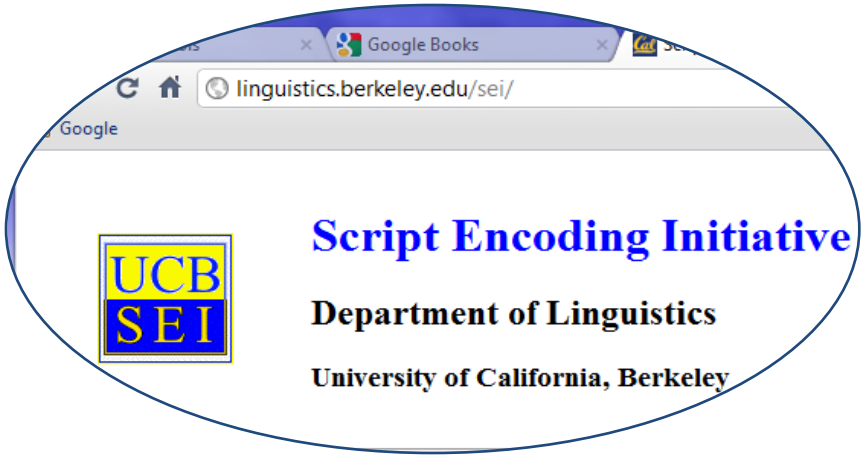


Supporting Unicode



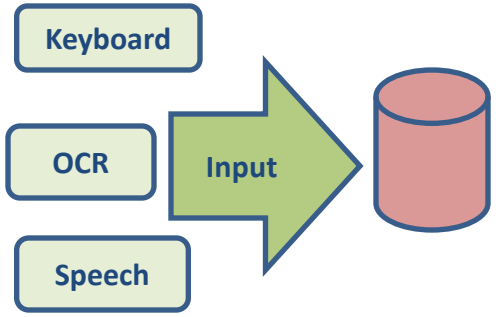






Input

- 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



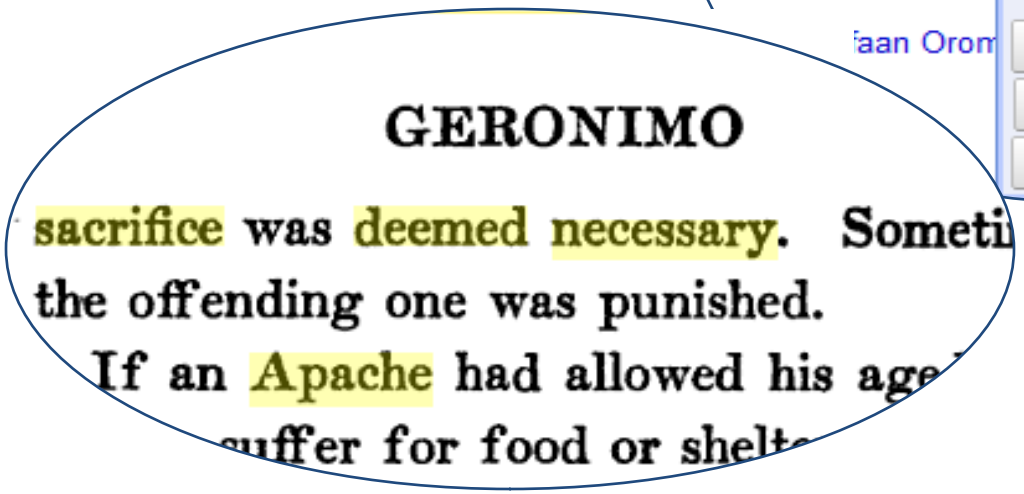
Input

Keyboard

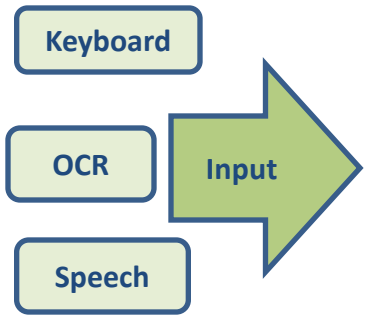
OCR

Input

Speech

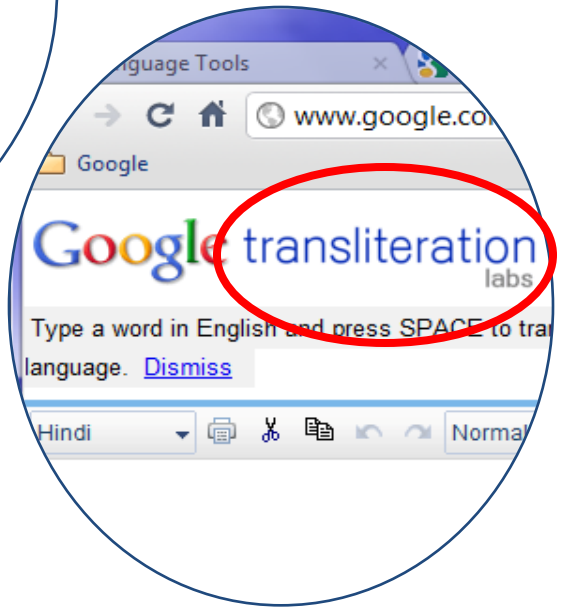


Input


Keyboard
OCR
Input
Speech

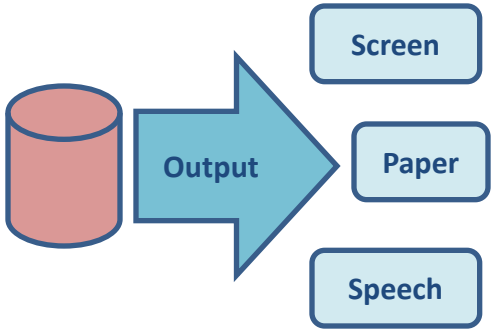

Google Books
www.google.com/ime/
Google IME
Please select an IME.
谷歌拼音输入法
Google Pinyin Input


5:51 AM
Google Translate
English
French
Type or paste text here to translate

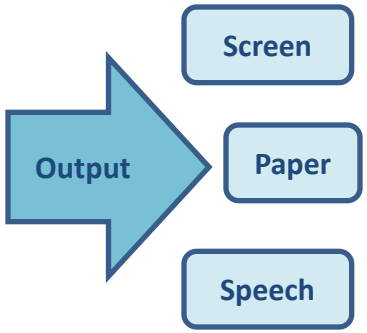

Google
Google transliteration
labs
Type a word in English
language. Dismiss
Hindi

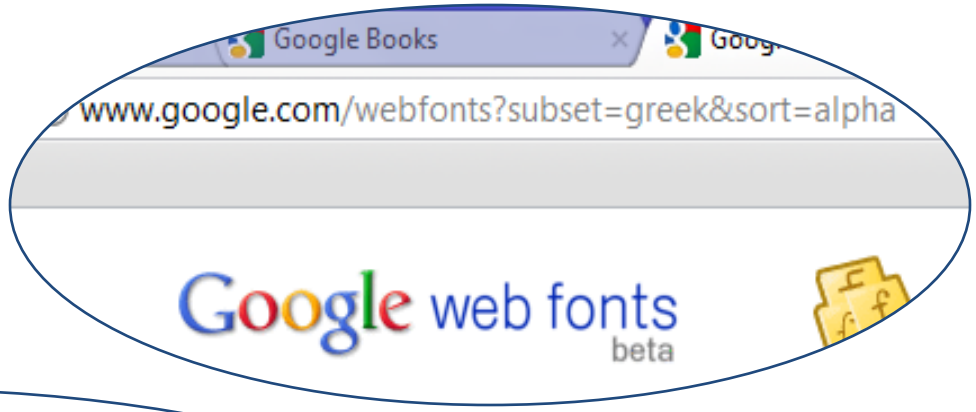
Output

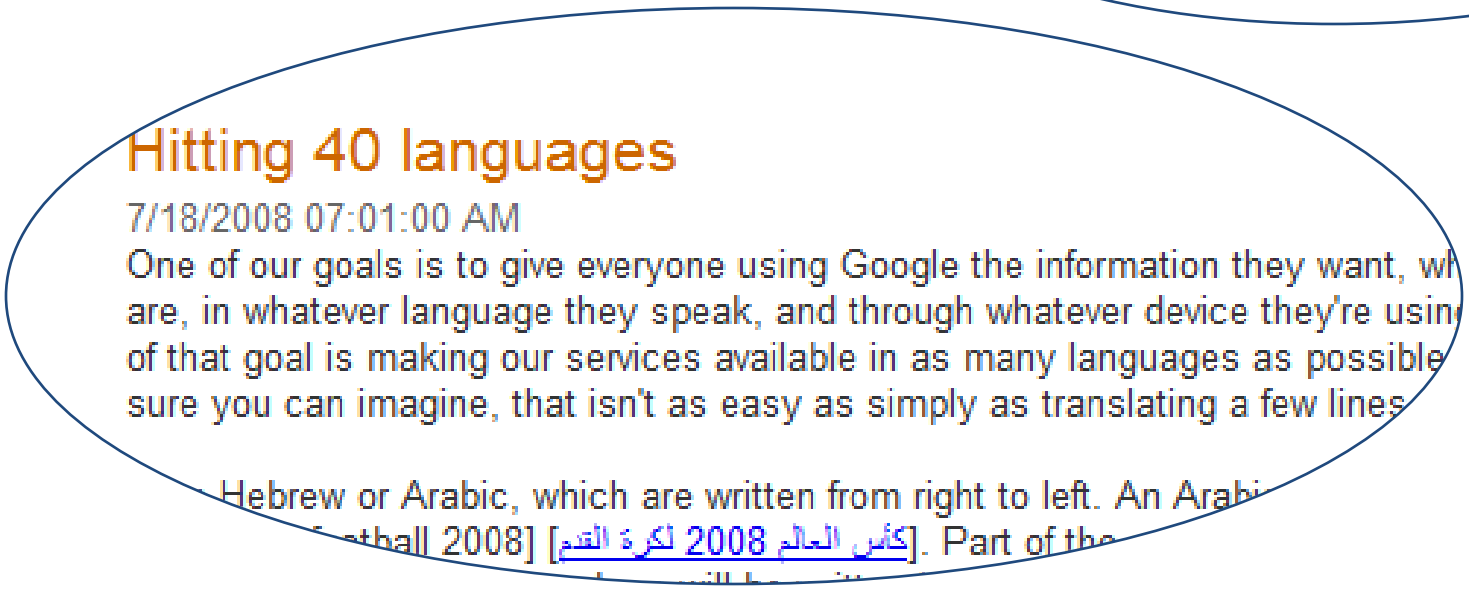
- Fonts (Google funding)
- Text to speech
- 1500+ local-language versions of Google various services
- 40-language initiative, forcing issues up front:
  - Keyboards, fonts, BIDI, form fill-in, text width, names, addresses, phone numbers, dates, currency, advertisements
- Chrome browser



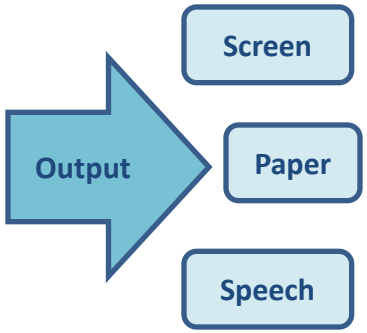
Output

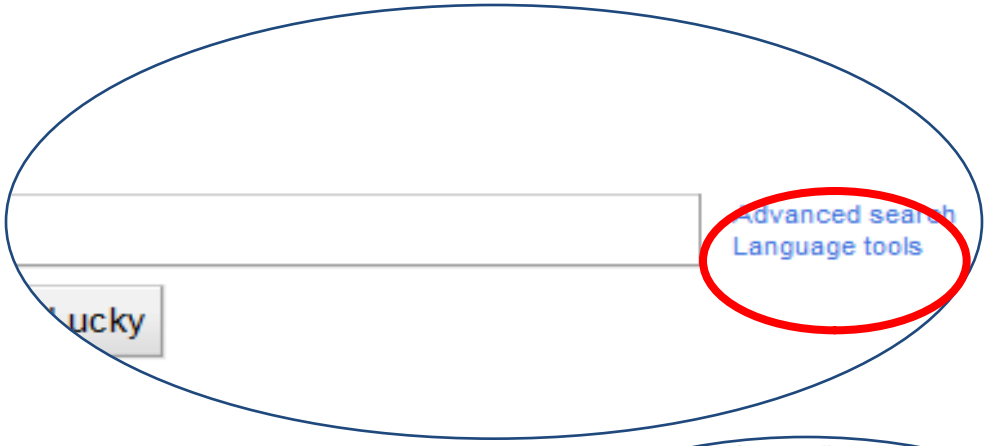

Output
Screen
Paper
Speech


Google Books
www.google.com/webfonts?subset=greek&sort=alpha
Google web fonts
beta

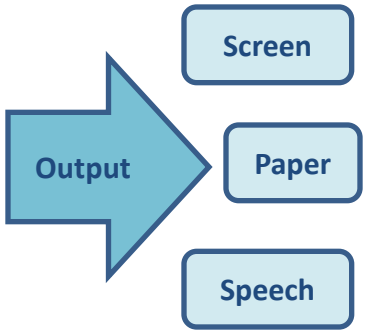

Hitting 40 languages
7/18/2008 07:01:00 AM
One of our goals is to give everyone using Google the information they want, wh
are, in whatever language they speak, and through whatever device they're usin
of that goal is making our services available in as many languages as possible
sure you can imagine, that isn't as easy as simply as translating a few lines
Hebrew or Arabic, which are written from right to left. An Arabi
tball 2008] [كأس العالم 2008 لكرة القدم]. Part of the

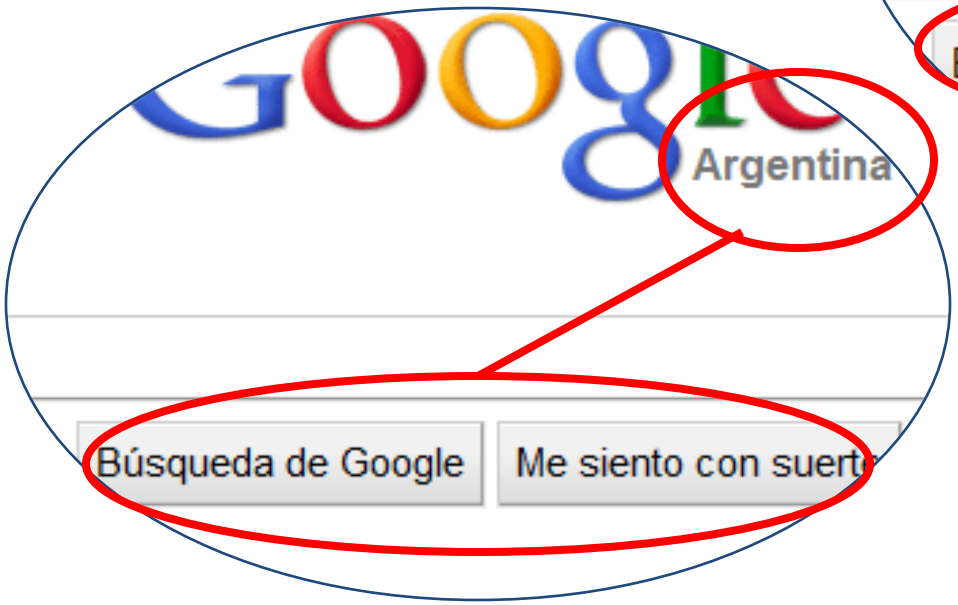
User Interface

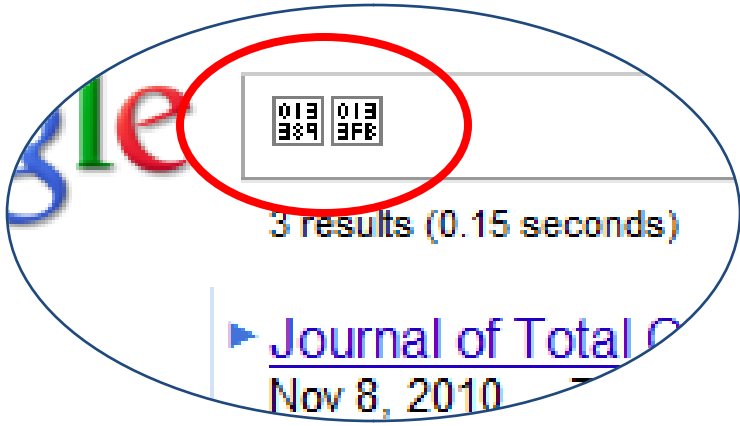




User Interface

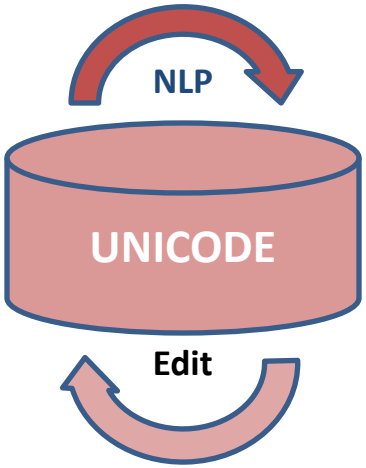




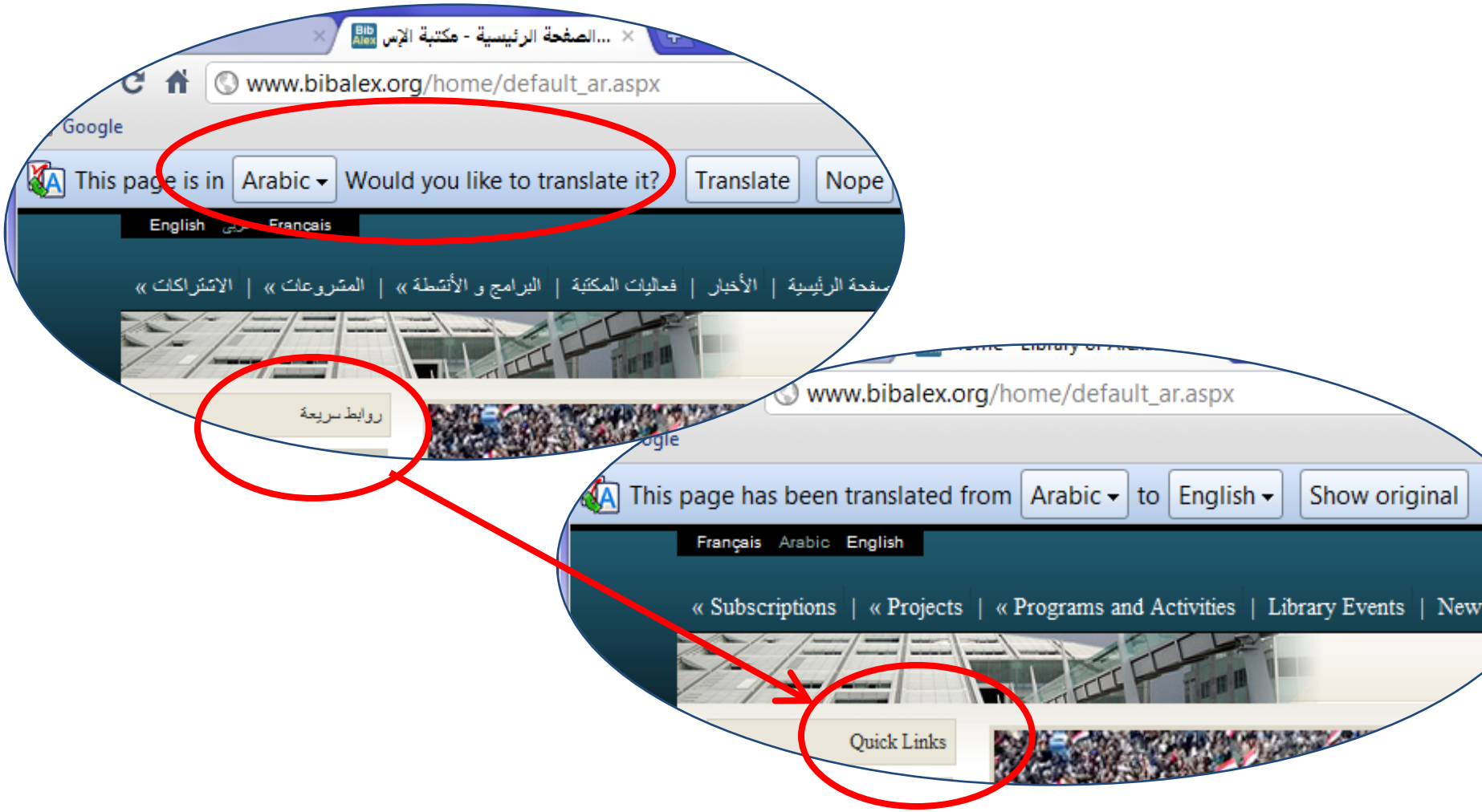


Language Processing

- Language detect
- Spelling correct/suggest
- Translation

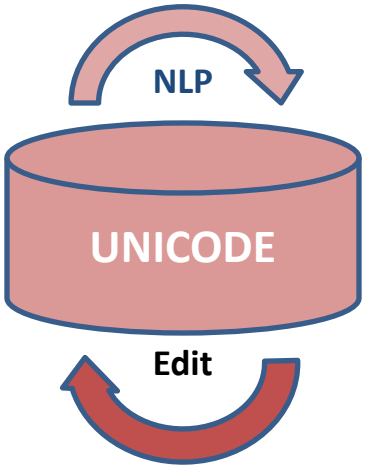


NLP: Detect, Translate



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

3. Mataar Al Kahera Al Dawelli/مطار القاهرة الدولي dakika 51

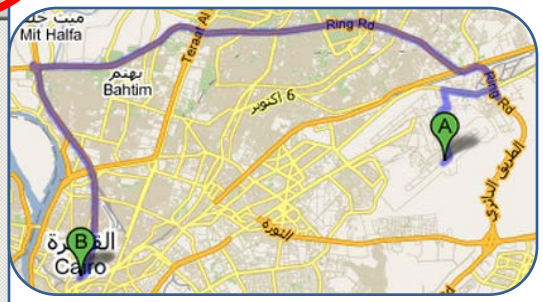
23.0 km

Cairo International Airport
القاهرة
مصر

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

400 m

2. Chukua njia pili kushoto kuingia **Mataar 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 SDIIISensibilité élevée avec ....

3 183,75 € - Neuf
Livraison gratuite
Maison-Des-Cameras

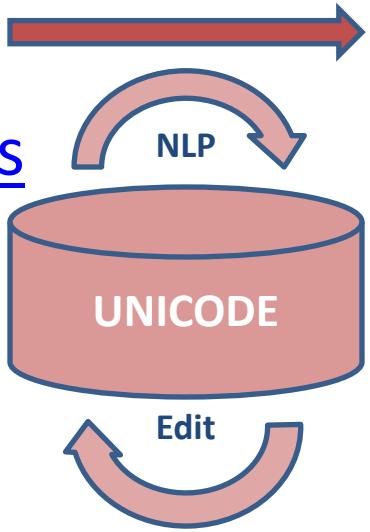
Open Tools

- Open source

  http://translate.google.com/translate_tools

  http://src.chromium.org/

  http://chrome.google.com/



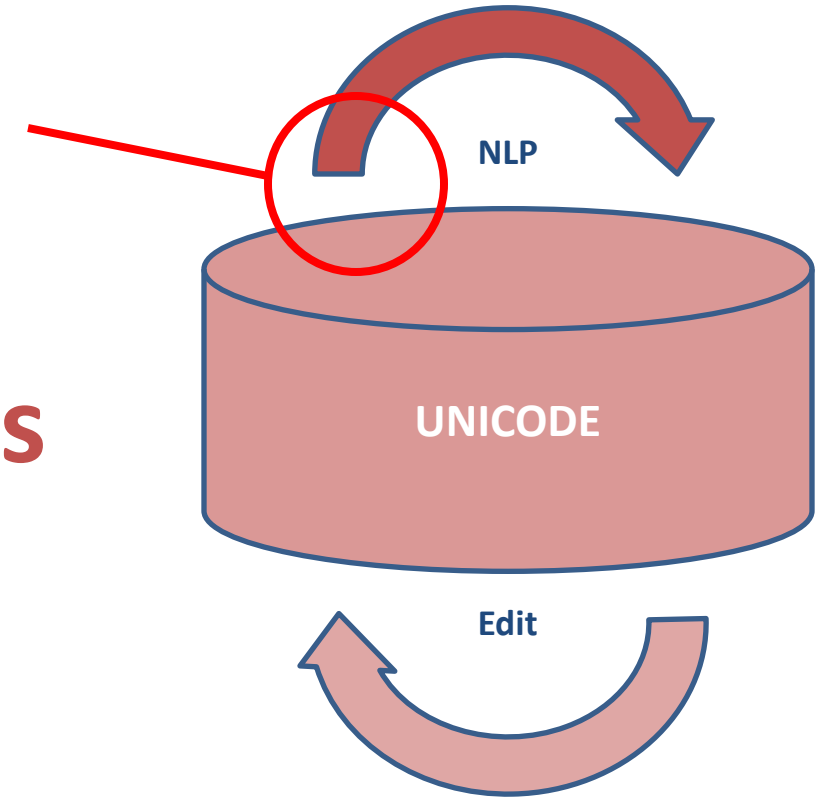
- E.G. auto-translate button on web sites

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
Tweet 13.8K
Like 7K
Select Language
Czech
Hebrew
Malay
Spanish
English
Danish
Hindi
Afrikaans
Dutch
Hungari
Diffusie:
Japan Meteorological Agency | Weer lessenaar Duitsland
Ander:
Press Release van Nuclear Safety Agency | Area Primêre Ibaraki | Miyagi hele
Original text:
Google
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.
Google Books
src.chromium.org/viewvc/chrome/trun
Google
chrome] / trunk / src / third_party / cld
dex of /trunk/src/third_party/cld

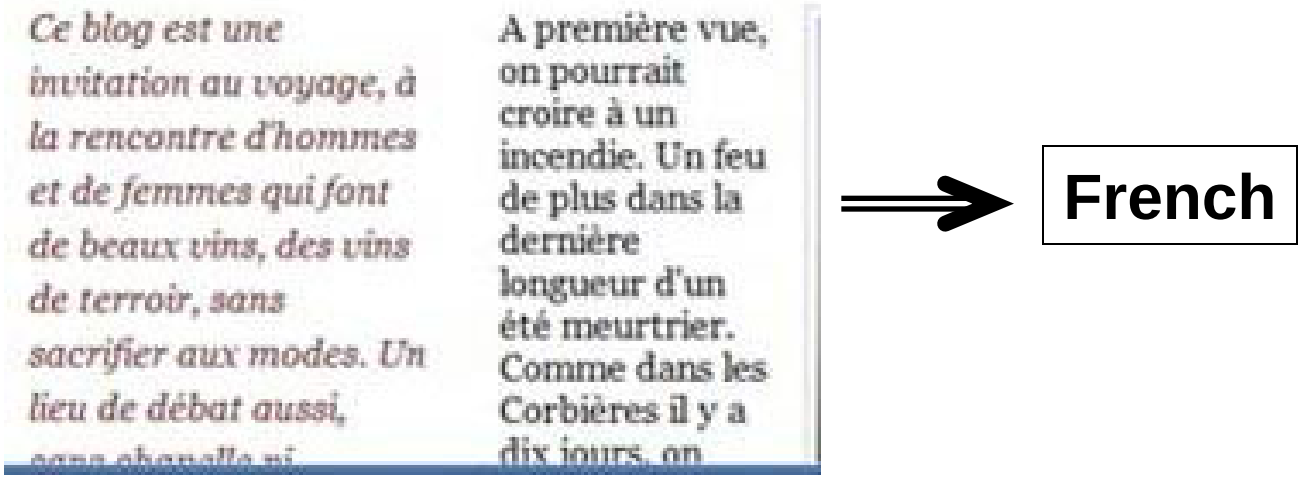
(next)

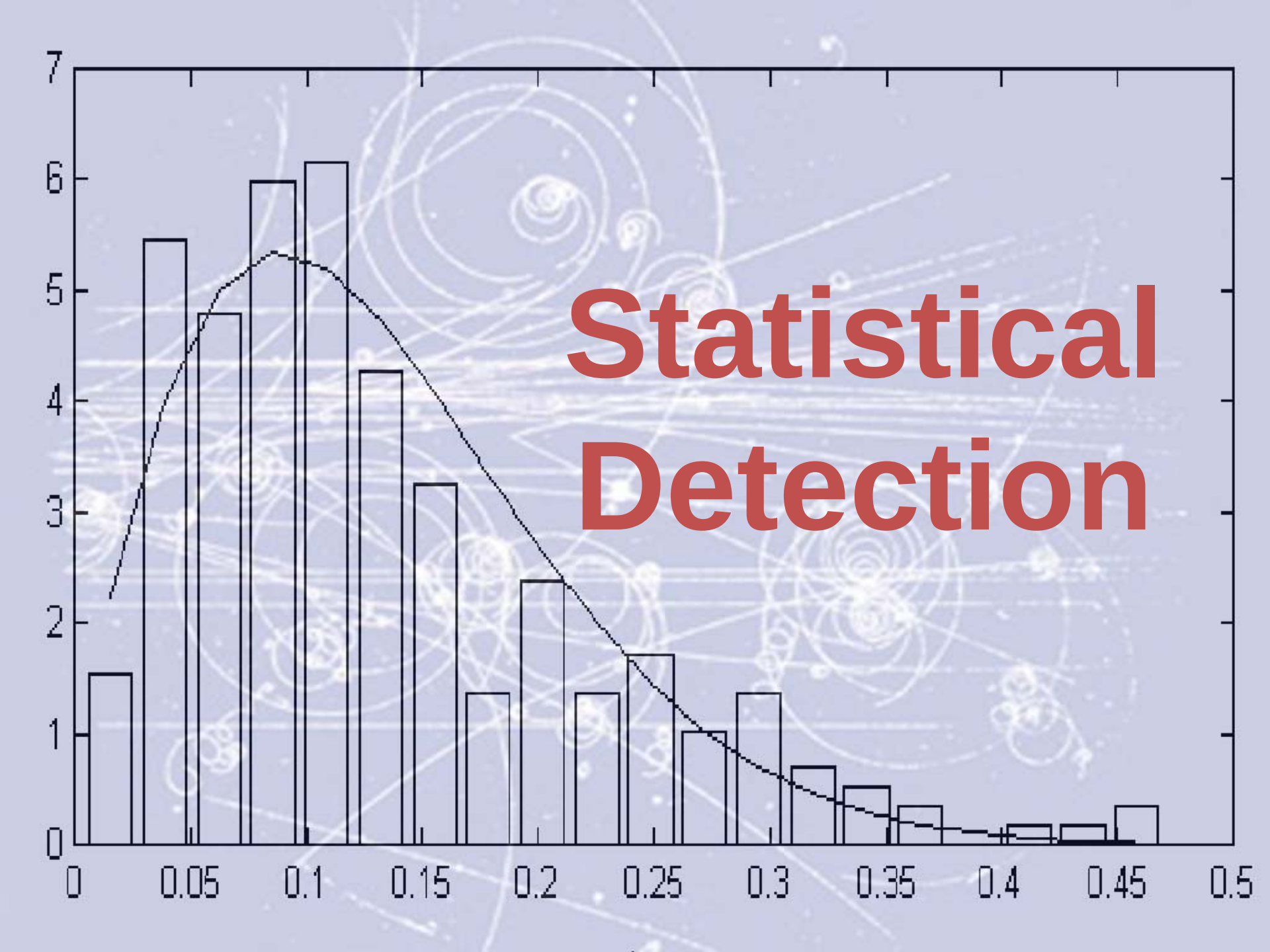
Statistical Language Detection in Web Pages



The problem

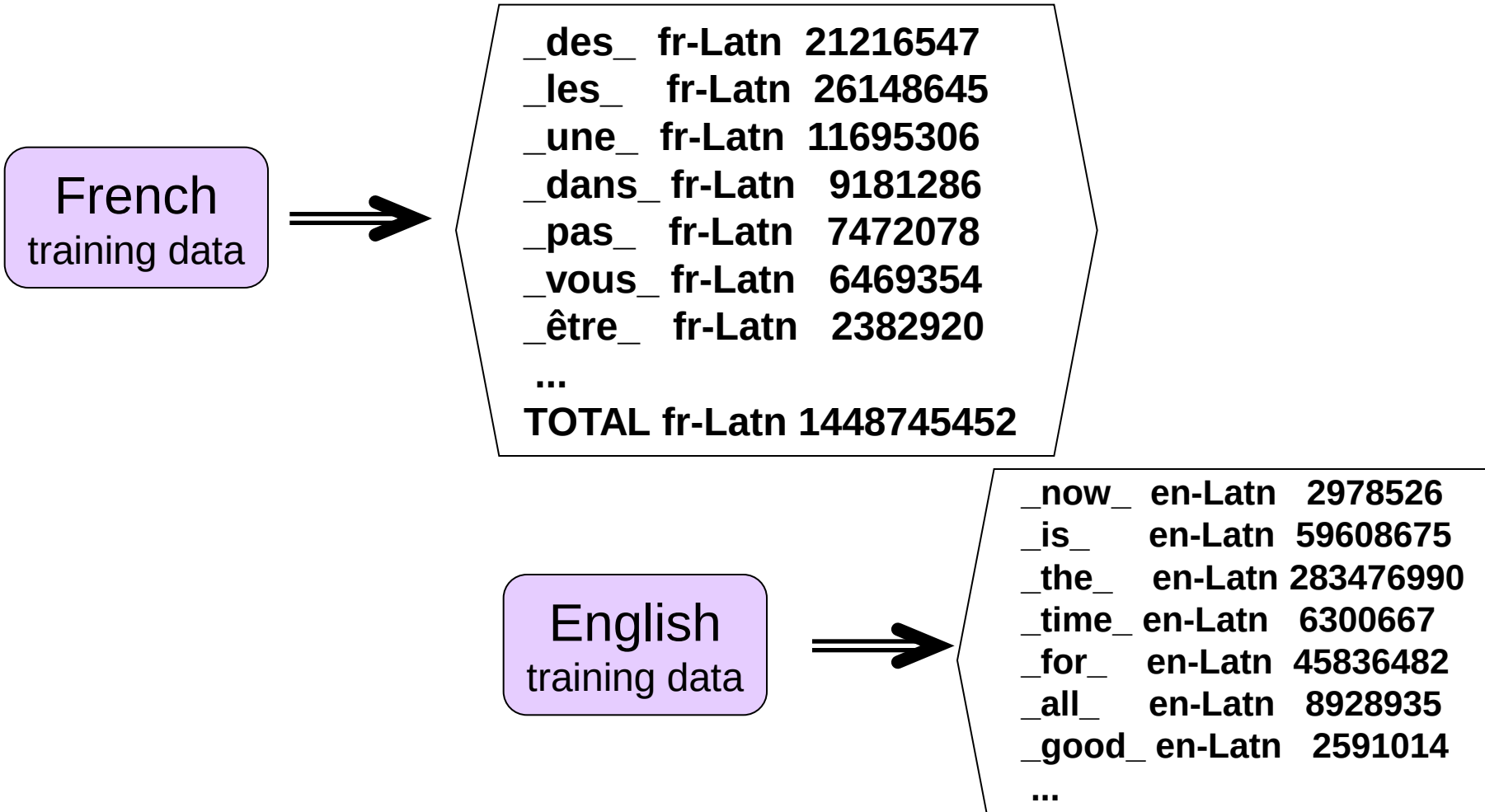
- Given a web page, determine the (human) languages it contains
- Why?
  - A French user preferentially sees French pages
  - Your browser and email can offer to translate text




Statistical
Detection
7
6
5
4
3
2
1
0
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5

Statistical Detection Overview

1. **Training data** to tokens and counts (offline)



Statistical Detection Overview

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

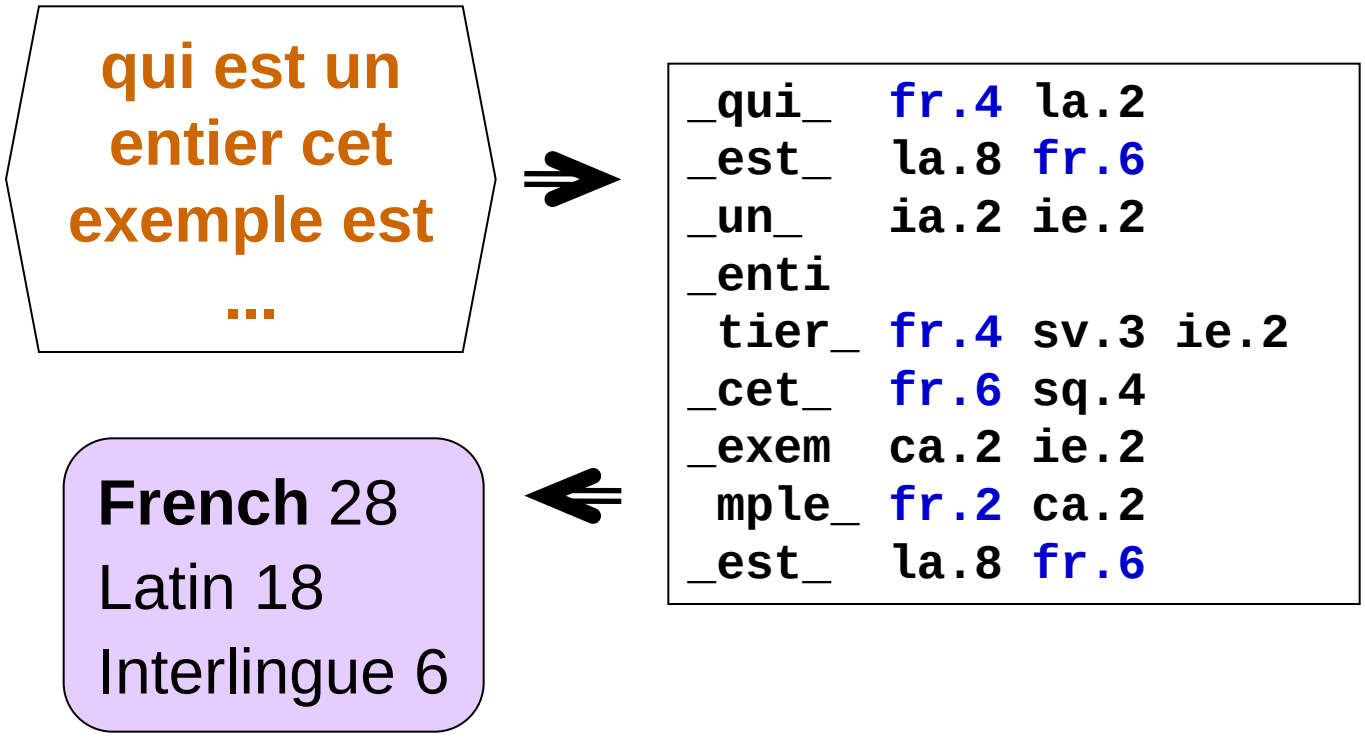
```
_camp ca-Latn 38022 / 20370777 = 1/536   lg = -9.1
_camp ia-Latn  4270 /  4301970 = 1/1007  lg = -10.0
_camp pt-Latn 66353 / 55361916 = 1/834   lg = -9.8
_camp ro-Latn 49589 / 33645079 = 1/678   lg = -9.5
```

```
_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
**

**©1999 Perpustakaan STT Telkom
**

The old saying "you need the right tool for the right job" applies even to your Internet business.

**<P>keyword
**

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 ...

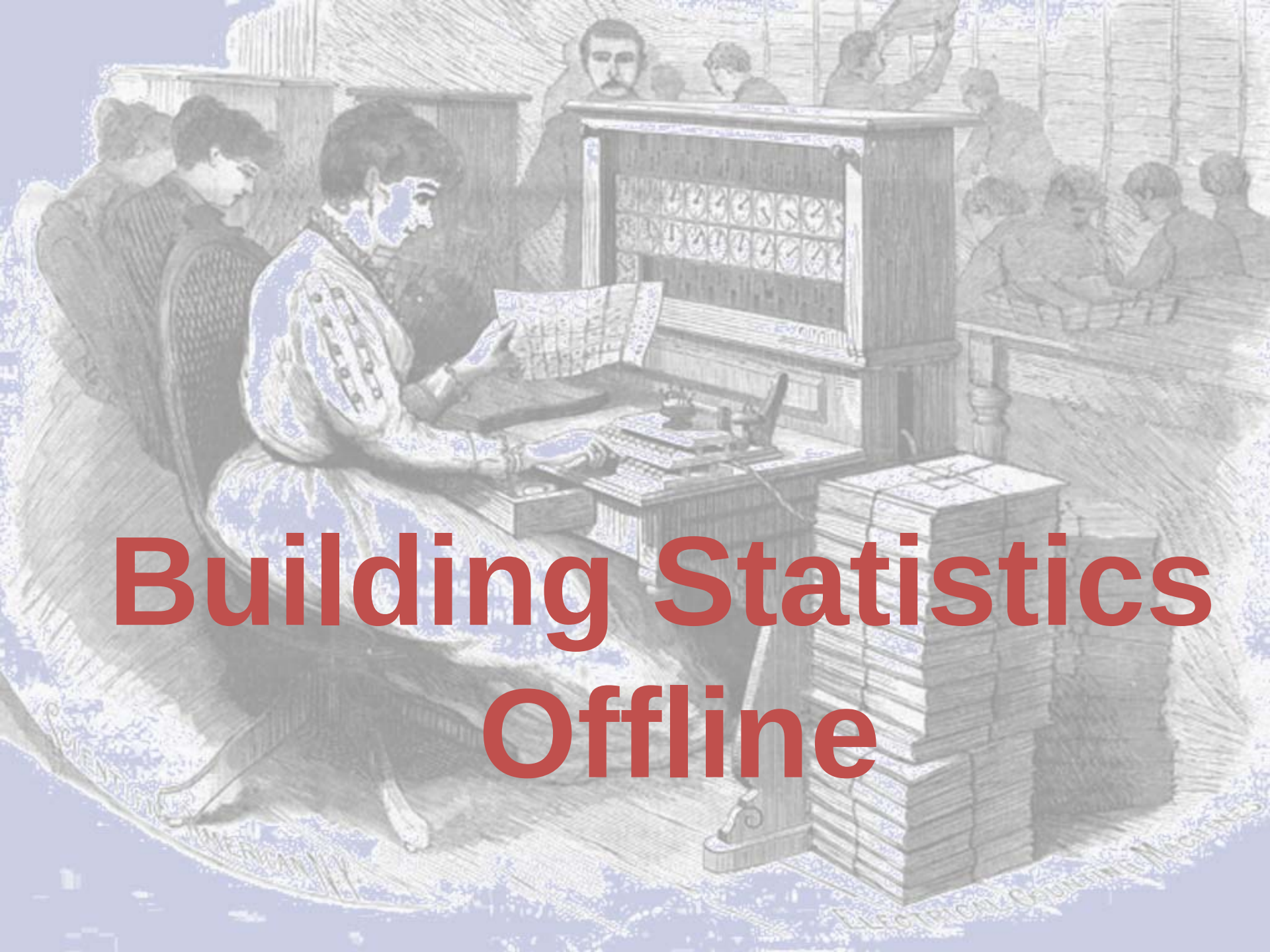
```
_opac_   ia.4
_onli    xxx.4 ms.2
 line_   xxx.4 et.3 ms.2
_publ    wo.2 ca.2
 blic_   en.2
_acce    it.2 ro.2
 cess_   en.2 rm.2
_cata    oc.2 vo.2
 talo    eo.2
 log_    xxx.2 vo.2
_perp    la.2 sq.2
 rpus    id.2 ms.2
 stak
 kaan_   om.4 fi.2
_stt_
_telk
 lkom_   nl.4 fy.2
_the_    en.8
_old_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

ENGLISH(100%), other(0%)

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 länner bäiflécken startsäit neie modell vun haaptsäit fir d éischt sch
[FRENCH] iwen an da verbesseren 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

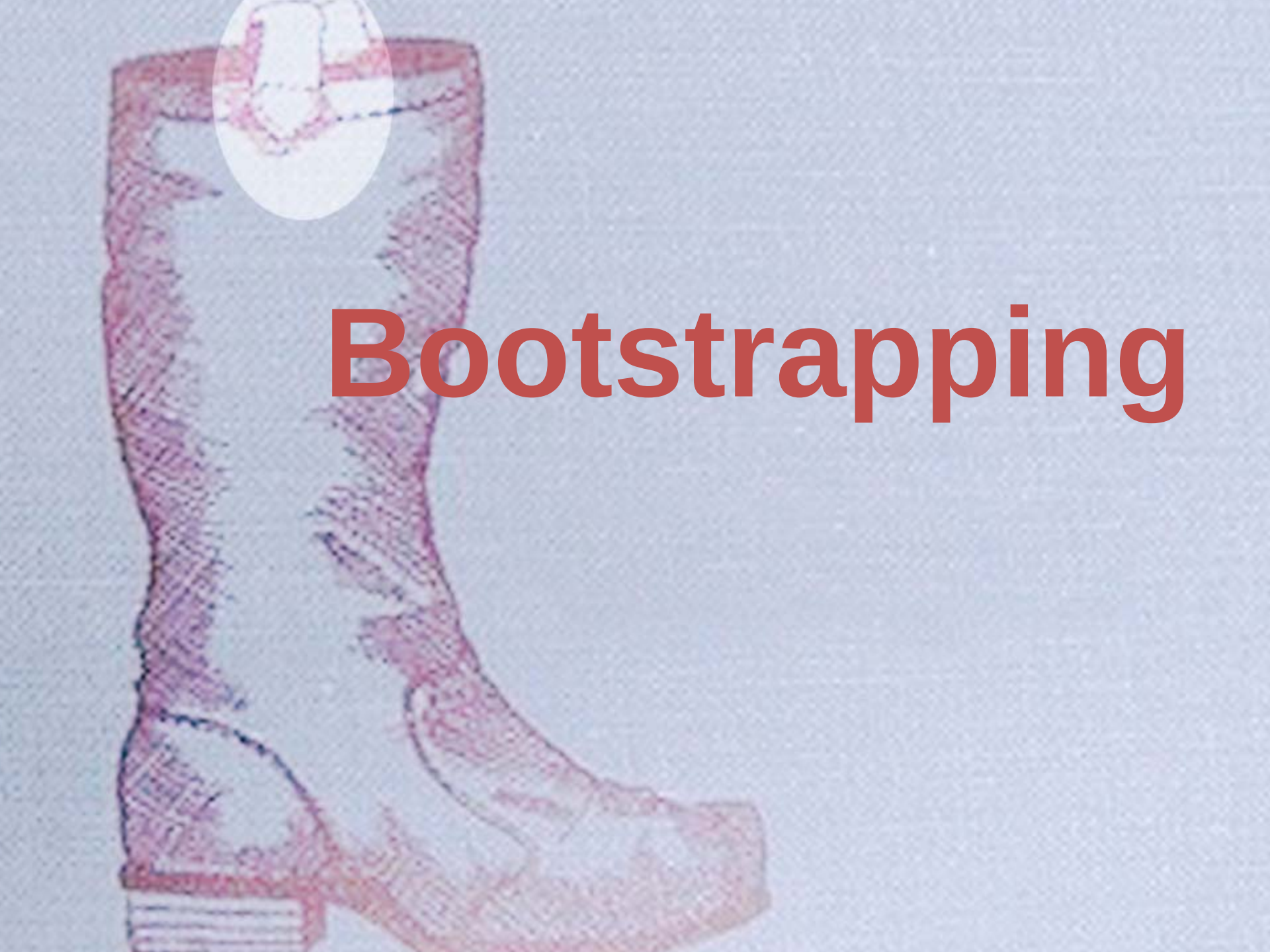
Token counts to languages and log probabilities

_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		
لالما	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/...

Bootstrapping Cross-checking

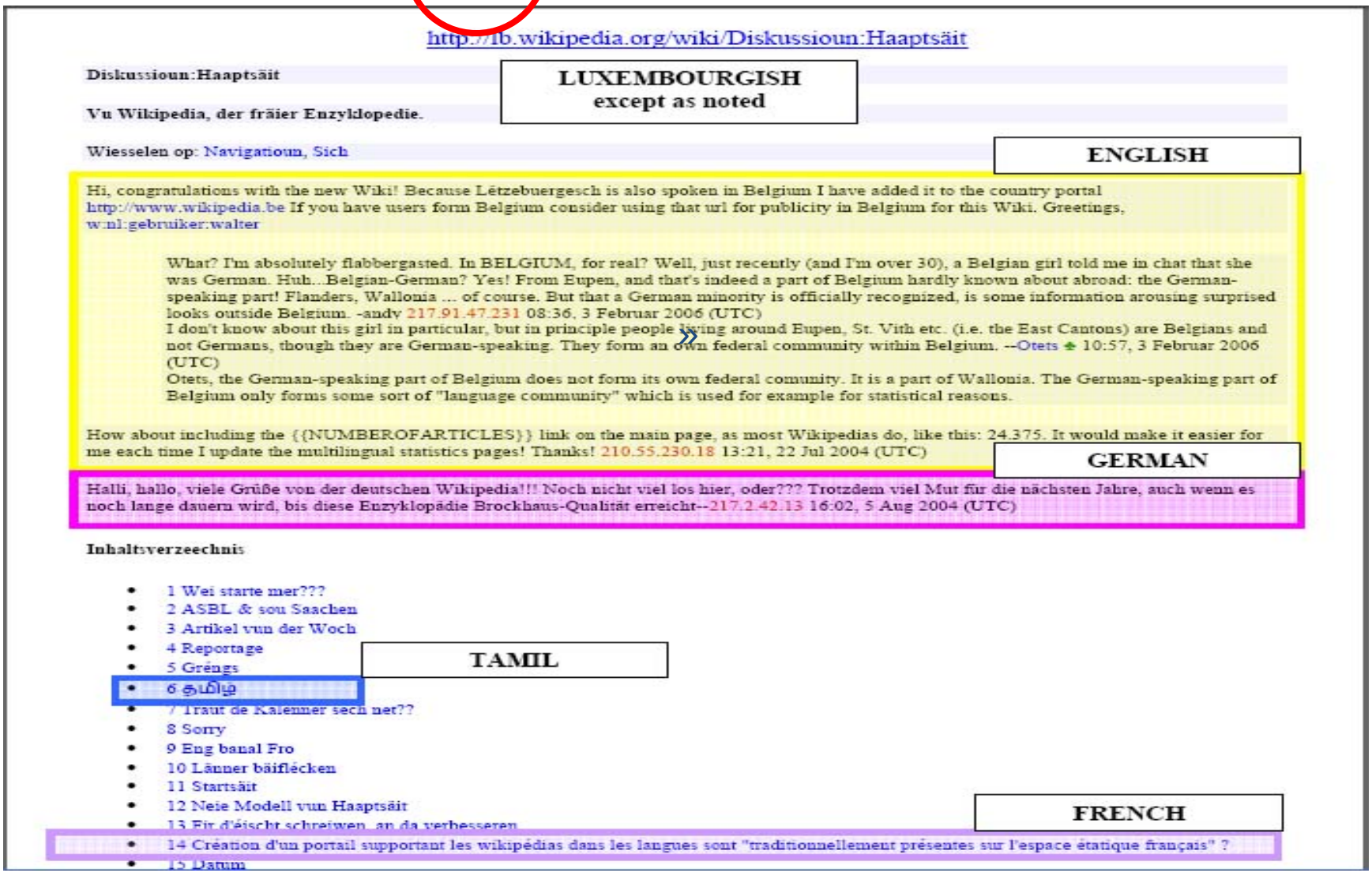
- Use pages from specified web-site patterns
- Use pages with a good amount of real text, and in only one language
- Cross-check
  - Top-level domain (e.g. www.foo.ee for Estonian)
  - Encoding (e.g. GB2312 for Simplified Chinese)
  - Right letters for each language ("exemplars")
  - Common words


What Goes
Wrong?

Training Data

Jumbled Training Data:

lb.wikipedia.org

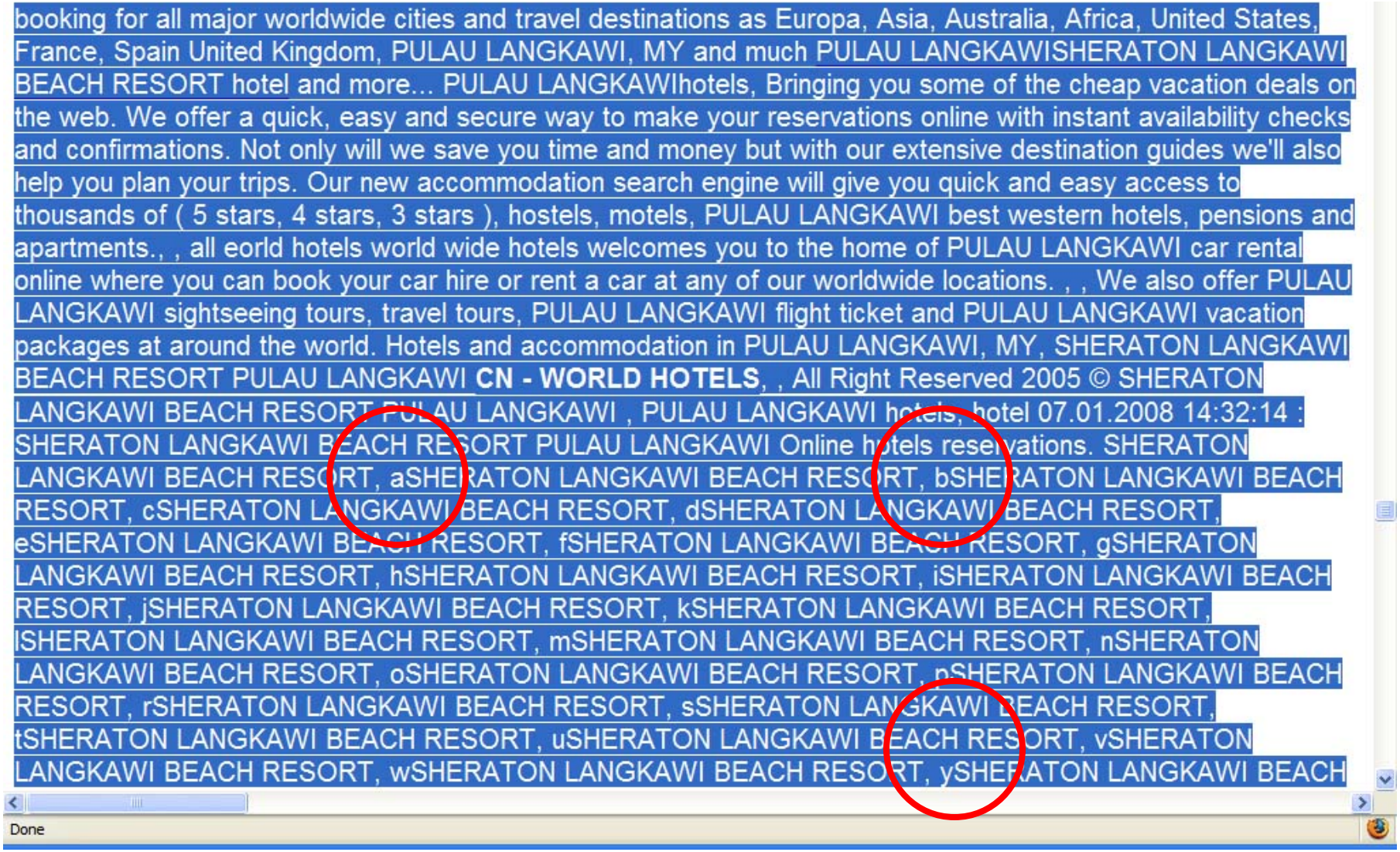


Omnipresent English

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** hostal 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 segundamano **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 tourlarni

Weird Web Pages

Remove Highly Compressible Text



DNA Sequences

KEGG **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 ma
AA seq	527 aa AA seq DB search MAITLPAGMKITGEILPAYEDILTPEAL LPDFLPETKSIREGDWKVAPVPKALECR

LIPEELAKVREVVGAGATYDRAAQIFEQMSTSEDFAEFLTLPLYEEV

NT seq

1584 nt NT seq +upstream 0 nt +downstream 0 nt

```
atggctatcacgctgcccgcgggcatgaagattaccggcgagattctgccggcttatgaa
gacatcctgacgccggaagccctggccctggtcgacaagctgcaccgtgccttcgaggcc
cgccgtcaggaactgctggccgcgcgcgtggcgcgcaccaagcgcctcgacgccggcgaa
ctgcccgacttcctgccggaaaccaagagcatccgcgaaggcgactggaaggttgcgccg
gtgccgaaggcgctggaatgccgccgcgtggaaatcaccggtccggtcgaagccaagatg
gtgatcaacgccttcaactcgggcgctgacagctacatgaccgacttcgaggactccaac
accccgaactggcacaaccagatgcagggtcaggtgaacctgaagtccgccgtgcgccgc
acgctgacgctggacagcaacggaaagcactacaaactcaacgacaagatcgccacgctg
caggtgcgtccgcgcggctggcacctagacgagaagcacgtcacgatcgacggcaagcgc
gtctcgggcggcatcttcgacttcgcgctgttcctgttccacaacgccaaggagcagatt
gcccgcggcgccggcccgttcttctatctgccgaagatggaaagccatctggaagcgcgt
ctgtggaacgacatcttcgtgatggcgcagaaggaagtcggtctgccgcaaggcacggtc
aaggccacggtgctgatcgagacgatcctggccgcgttcgagatggaagaaatcctgtac
gaactgcgcgagcacagcgccggcctgaacgctggccgttgggactacatcttctcgtgc
atcaagaagttcaaggtcgacaagaacttctgccttgccgaccgcgccaaggtgacgatg
acctcgccgttcatgcgcgcctacgcgctgctgctgctgaagacctgccacaagcgcggc
gcgcctgcgatcggcggcatgagcgcgctgatcccgatcaagaacgatccggagaagaac
gctatcgccatgcagggcatcatcggcgacaagcgccgtgacgccaccgacggctacgac
ggcggctgggtggcccacccgggtctggtcgagccggcaatgaaggaattcgtggacgtg
ctgggcgaccgcccgaaccagttcgacaagcaacgtccggacgtcgaagtgaagggtgcc
gacctgctgaacttccagccggaagcgccgatcaccgaagccggcctgcgcatgaacatc
aacgtcggcatccactacctgggcgcctggctggctggcaatggctgcgtgccgatccac
aacctgatggaagacgccgccaccgccgagatctcgcgctcccaggtgtggcagtggatc
cgctcgccgaagggcaagctggaagacggccgcaaggtgaccgccgagatggtgcgcgcg
ctgattccggaagaactggccaaggtccgcgaagtggtgggcgcaggcgccacgtacgac
cgtgccgcgcagatcttcgagcagatgtcgacttcggaagactttgccgaattcctgacg
```

DNA Sequences

[en/tl*] dbget result r metallidurans rmet ralstonia metallidurans rmet entry rmet cds r metallidurar
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 **[]** prosi
synthase other dbs jgi rmet ncbi gi ncbi geneid uniprot q lnk linkdb position aa seq aa
maitlpagmkitgeilpayediltpealalvdklhrafearrqellaarvartkrldage lpdflpetksiregdwkvapvpkale
crrveitgpveakmvinafnsgadsymtdfedsnpnwhnqmqgqvnlksavrrtltldsngkhy **[eu/siN*]**klndkiatlqvrprgw
dgkr vsggifdfalflfhnakeqiargagpffylpkmeshlearlwndifvmaqkevglpqgtv katvlietilaafemeeilyelrehsagl
agrwdyifscikkfkvdknfcladrakvtm tspfmrayallllktchkrgapaiggmsalipikndpeknaiamqgiigdkrrdatdgyd
ggwvahpglvepamkefvdvlgdrpnqfdkqrpdvevkgadllnfqpeapiteaglrmni nvgihylgawlagngcvpihnlmedaa
eisrsqvwqwirspkgkledgrkvtaemvra **[so/en*]** lipeelakvrevvgagatydraaqifeqmstsedfaefltlplyeev nt se
upstream nt downstream nt atggctatcacgctgcccgcgggcatgaagattaccggcgagattctgccggcttatgaa
gacatcctgacgccggaagccctggccctggtcgacaagctgcaccgtgccttcgaggcc
cgccgtcagga **[]** actgctggccgcgcgcgtggcgcgcaccaagcgcctcgacgccggcgaa
ctgcccgacttcctgccggaaaccaagagcatccgcgaaggcgactggaaggttgcgccg
...
cgtgccgcgcagatcttcgagcagatgtcgacttcggaagactttgccgaattcctgacg 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 metallidurar
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 **[]** prosi
synthase other dbs jgi rmet ncbi gi ncbi geneid uniprot q lnk linkdb position aa seq aa
maitlpagmkitgeilpayediltpealalvdklhrafearrqellaarvartkrldage lpdflpetksiregdwkvapvpkale
crrveitgpveakmvinafnsgadsymtdfedsnpnwhnqmqgqvnlksavrrtltldsngkhy **[eu/siN*]**klndkiatlqvrprgw
dgkr vsggifdfalflfhnakeqiargagpffylpkmeshlearlwndifvmaqkevglpqgtv katvlietilaafemeeilyelrehsagl
agrwdyifscikkfkvdknfcladrakvtm tspfmrayallllktchkrgapaiggmsalipikndpeknaiamqgiigdkrrdatdgyd
ggwvahpglvepamkefvdvlgdrpnqfdkqrpdvevkgadllnfqpeapiteaglrmni nvgihylgawlagngcvpihnlmedaa
eisrsqvwqwirspkgkledgrkvtaemvia **[xxx/en*]** lipeelakvrevvgagatydraaqifeqmstsedfaefltlplyeev nt s
upstream nt downstream nt atggctatcacgctgcccgcgggcatgaagattaccggcgagattctgccggcttatgaa
gacatcctgacgccggaagccctggccctggtcgacaagctgcaccgtgccttcgaggcc
cgccgtcagga [] actgctggccgcgcgcgtggcgcgcaccaagcgcctcgacgccggcgaa
ctgcccgacttcctgccggaaaccaagagcatccgcgaaggcgactggaaggttgcgccg
...
cgtgccgcgcagatcttcgagcagatgtcgacttcggaagactttgccgaattcctgacg ctgccgctgtacgaagaagtctga origin
dbget integrated **[SWEDISH]** database retrieval system genomenet **[]** =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 susahnya 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 addre
[id 185/ms* 182] hujung ya allah sekiranya dia adalah suami pilihan mu di arash
[ms 180/id* 172] mi yang akan membimbing tanganku dititianmu 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
[id 66/ms* 61] menghadapi segala kerenah dan ragamnya http profiles blogdrive
[ms 123/id* 115] weblog enter your email here tuesday february alamak 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

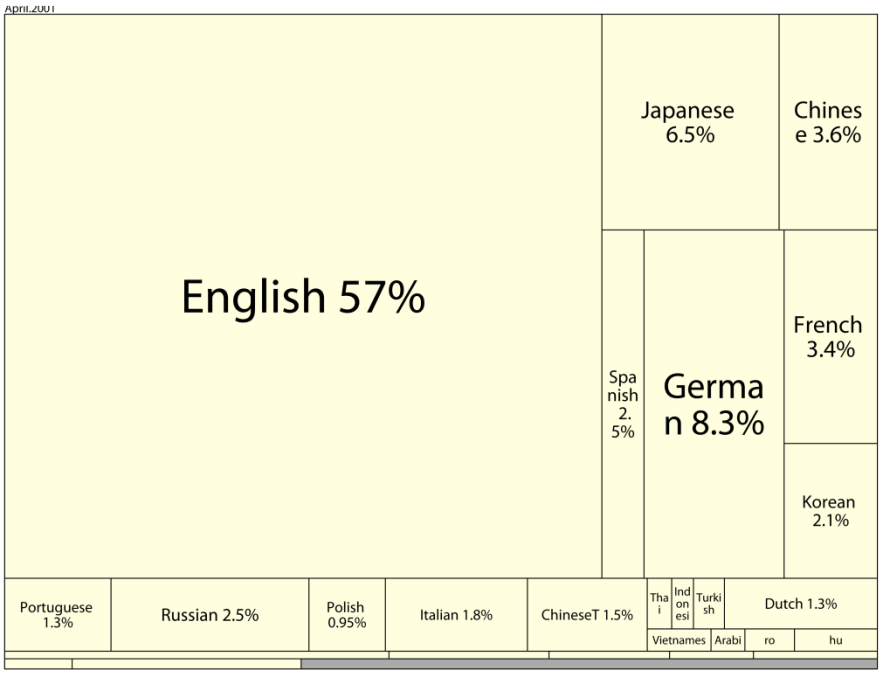
Summary

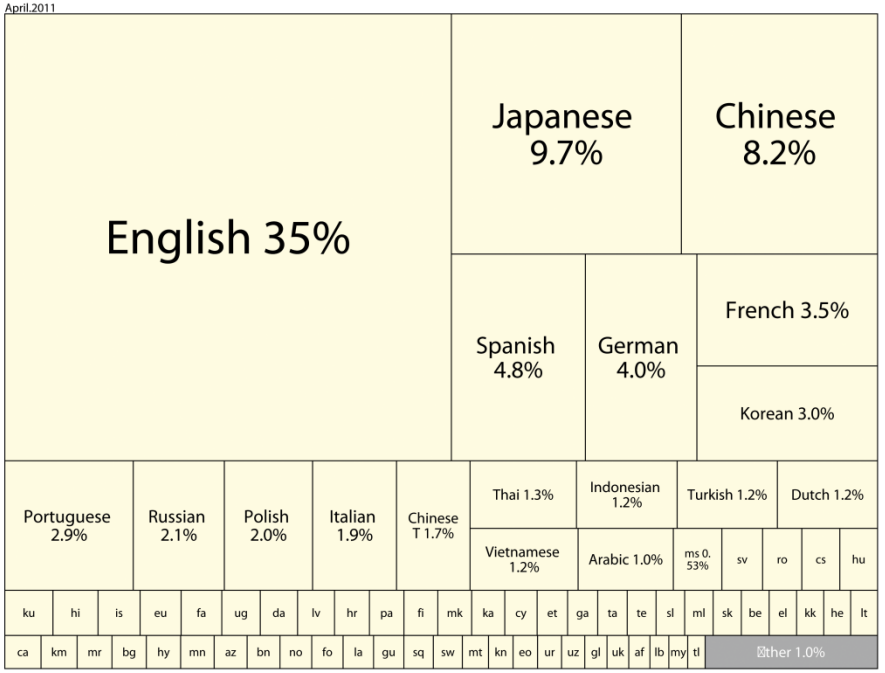
- Text on the Web is unedited, intermixed, and sometimes not even text
- Difficult to find clean training data
- Quadgrams + Whole-words does pretty well

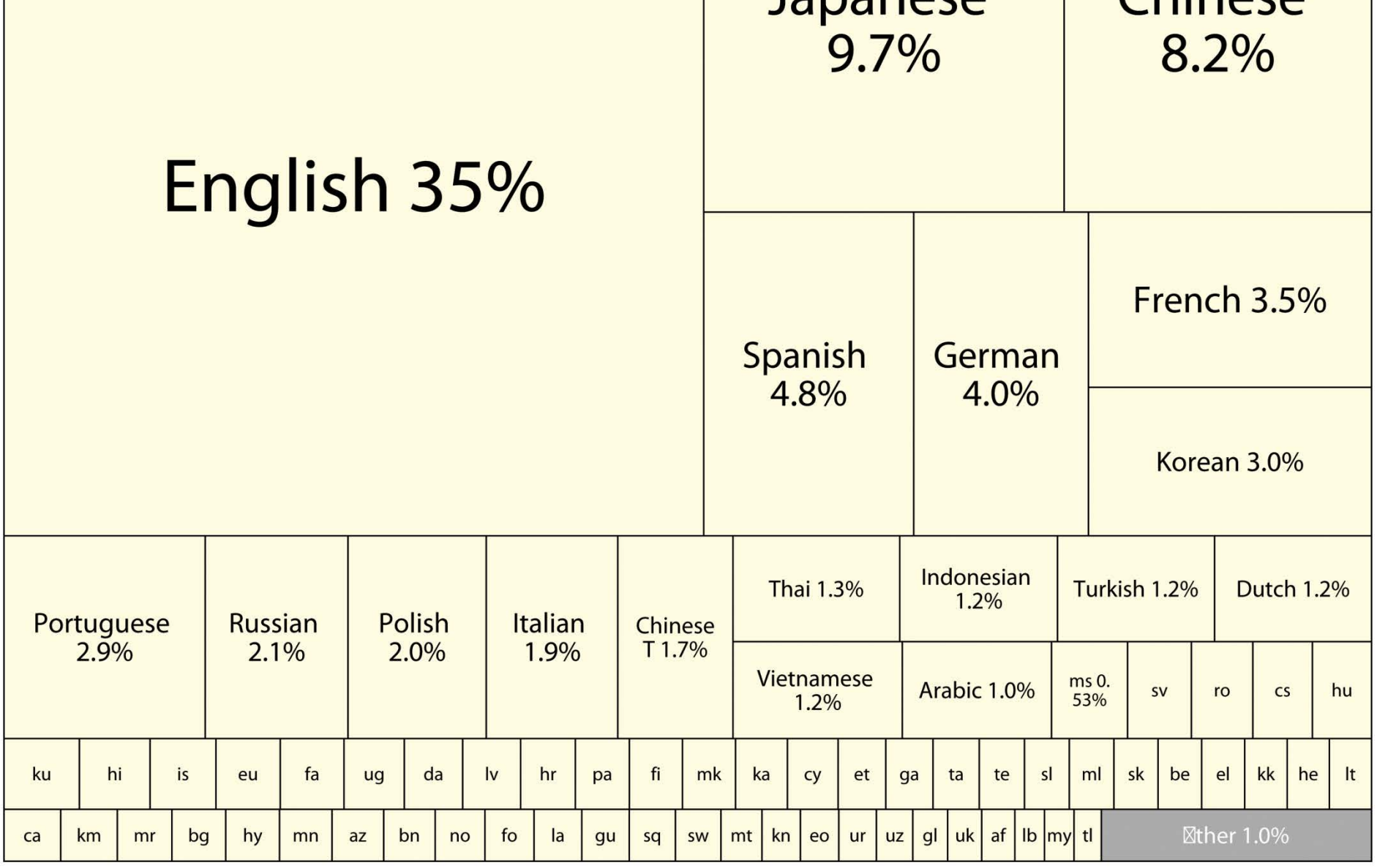
- There is a lot of garbage on the web

Webshare 2001 => 2011

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







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?