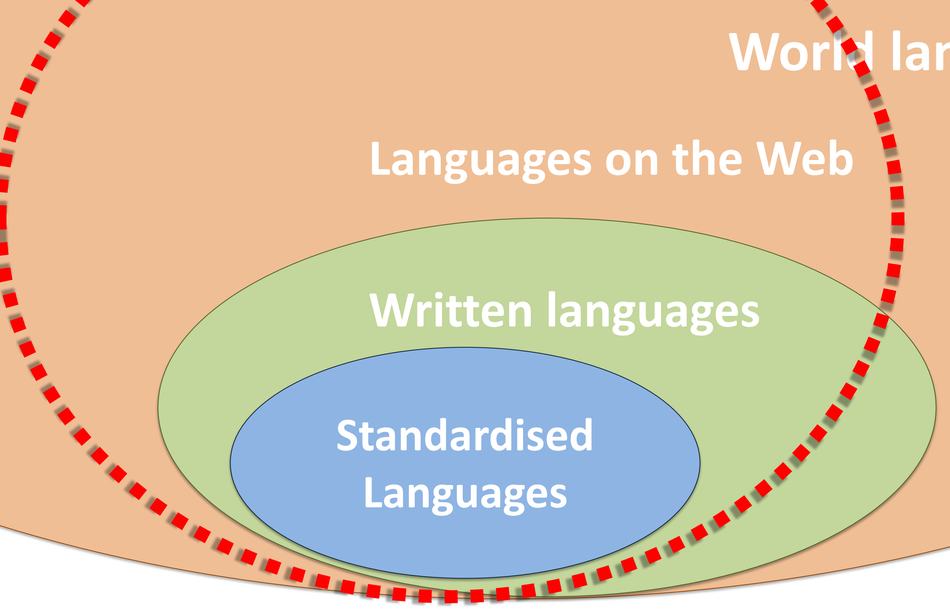


Ethnologue lists **7,413** languages

ISO 639-3:2007 Codes**: 7,589** entries

Books published in **480** languages (source: Google Books Project)

Official languages of country or region: 238 (source: Wikipedia)

Possibly **1,500** written languages on the **Web** (source: Scannel's study)

Web Language Identification Testing Tool - (IIT & ILC CNR)

matteo.abrate@iit.cnr.it clara.bacciu@iit.cnr.it francesca.frontini@ilc.cnr.it mariantonietta.lapolla@iit.cnr.it andrea.marchetti@iit.cnr.it monica.monachini@ilc.cnr.it

Language identification or detection:
given a text, determine the language it is written in

How does the best open source language identification library behave with respect to the languages on the Web?

Language identification techniques

- *Common words technique*: (Grefenstette)
- *Unique letter combinations*: (Churcher)
- *Statistical approach*: (Dunning)
- *N-grams approach*: (Canvar et al.)
- *Compression based approach*: (Teahan et al.)

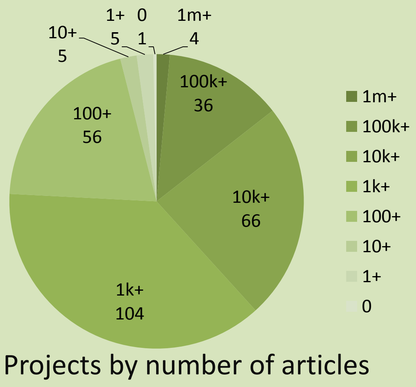
Some open-source language identification tools

- *Chromium Compact Language Detector (CLD):* declaring support for about **160 languages**. Method : n-grams. Language: C++ (we used an incomplete binding for Python)
- *Lingua::Identify:* Supports **33 languages**. Methods: small words, prefix, suffix, n-grams. Language : Perl.

Known problems:

Underrepresented languages Short sentences
Mix of languages in one page
Unusual script/orthography Non-standard register

Wikipedia:
a valuable source for underrepresented languages



Projects by number of articles

Test set: 1000 articles from 147 projects having more than 3000 articles

Testing Chromium CLD library

Mapping of each 283 **Wikipedia project prefix** (i.e. **fr**.wikipedia.org) on a valid **ISO639 language code** (i.e. **fr**)

↓

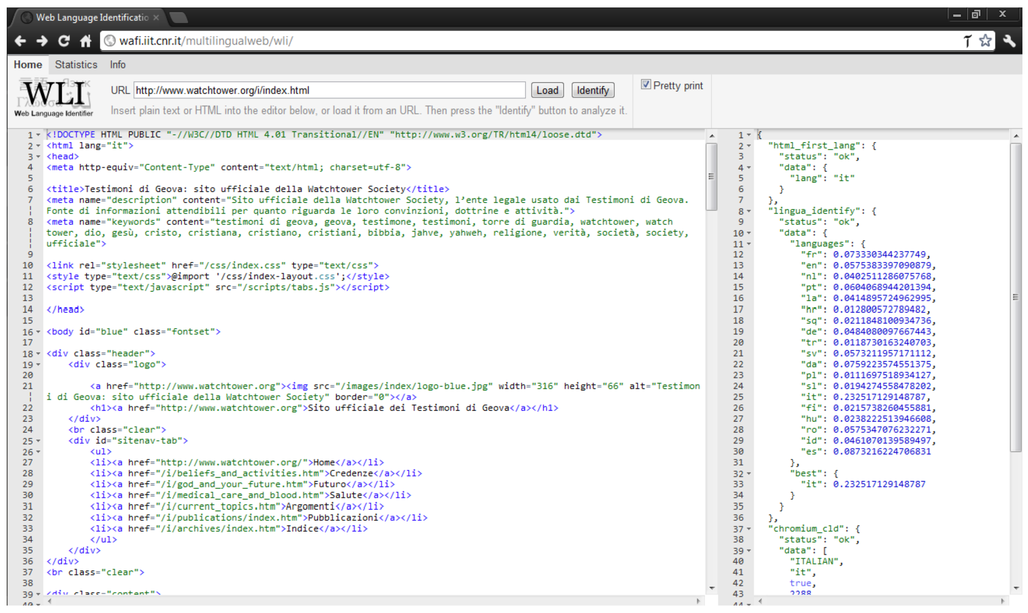
Acquisition of a random set of Wikipedia articles for 147 languages*

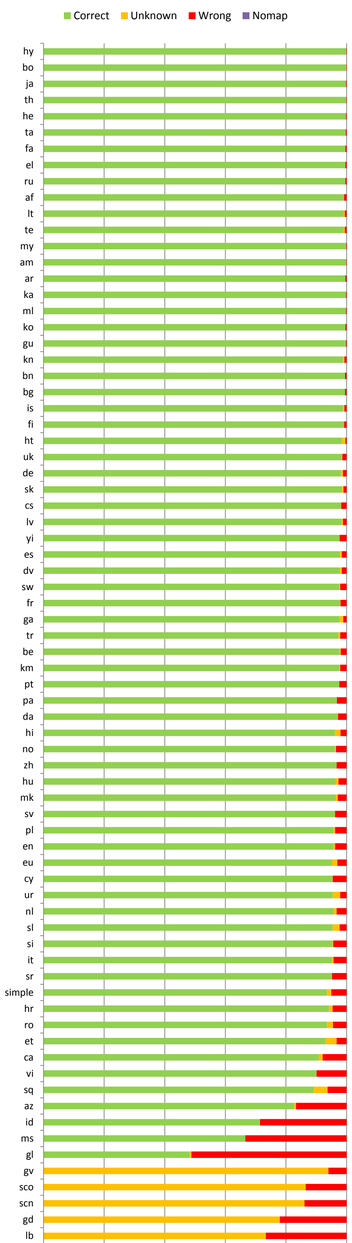
↓

*Leaving out closed or unmapped projects or projects with less than 3000 articles
Language identification using Chromium CLD

↓

Developement of a Web interface to test language detectors





Testing result:
a lot of data is still needed to build more language models

Testing Chromium Compact Language Detector library against 147.000 articles from 147 Wikipedia projects

Language identification for the Wikipedia home pages

Prefix	ISO 639 code	Chromium CLD	Lingua:: Identify	HTML lang
en	en	en	en	en
de	de	de	de	de
fr	fr	fr	fr	fr
nl	nl	nl	nl	nl
it	it	it	it	it
pl	pl	pl	pl	pl
es	es	es	es	es
ru	ru	ru	en	ru
ja	ja	ja	en	ja
pt	pt	pt	fr	pt
sv	sv	sv	sv	sv
zh	zh	zh	en	zh
ca	ca	ca	fr	ca
uk	uk	uk	en	uk
no	nb	nb	da	no
vi	vi	vi	en	vi
fi	fi	fi	fi	fi
cs	cs	cs	en	cs
hu	hu	hu	en	hu
ko	ko	ko	en	ko
id	id	id	id	id
tr	tr	tr	tr	tr
ro	ro	ro	fr	ro
fa	fa	fa	en	fa
ar	ar	ar	en	ar
da	da	da	da	da
eo	eo	sk	fr	eo
sr	sr	sr	en	sr
lt	lt	lt	en	lt
sl	sl	sl	sl	sl
sk	sk	sk	en	sk
ms	ms	ms	id	ms
he	he	he	en	he
bg	bg	bg	en	bg
kk	kk	ru	en	kk
eu	eu	eu	en	eu
vo	vo	ms	fr	vo
war	war	fil	id	war
hr	hr	hr	hr	hr
hi	hi	hi	en	hi
et	et	et	fi	et
az	az	az	en	az
gl	gl	gl	es	gl
nn	nn	nb	da	nn
simple	en	en	en	en

Prefix	ISO 639 code	Chromium CLD	Lingua:: Identify	HTML lang
th	th	th	en	th
la	la	xxx	en	la
new	new	hi	en	new
el	el	el	en	el
oc	oc	ca	fr	oc
tl	tl	fil	id	tl
ka	ka	ka	en	ka
mk	mk	mk	en	mk
sh	sh	hr	hr	sh
ht	ht	ht	fr	ht
pms	pms	xxx	id	pms
te	te	te	en	te
ta	ta	ta	en	ta
ceb	ceb	fil	en	ceb
br	br	nb	id	br
sq	sq	sq	sq	sq
lv	lv	lv	en	lv
be	be	be	en	be
jv	jv	id	en	jv
mg	mg	ms	en	mg
lb	lb	xxx	de	lb
cy	cy	cy	id	cy
mr	mr	xxx	en	mr
is	is	is	da	is
bs	bs	hr	hr	bs
yo	yo	xxx	en	yo
an	an	es	fr	an
bpy	bpy	bn	en	bpy
lmo	lmo	xxx	id	lmo
fy	fy	da	nl	fy
hy	hy	hy	en	hy
sw	sw	sw	id	sw
bn	bn	bn	en	bn
io	io	xxx	fr	io
ml	ml	ml	en	ml
gu	gu	gu	en	gu
af	af	af	nl	af
pnb	pnb	ur	en	pnb
ne	ne	hi	en	ne
nds	nds	de	nl	nds
scn	scn	it	it	scn
ku	ku	ms	fr	ku
ur	ur	ur	en	ur
zh-yue	yue	zh-TW	en	zh-yue
su	su	ms	en	su

Prefix	ISO 639 code	Chromium CLD	Lingua:: Identify	HTML lang
qu	qu	ms	en	qu
diq	diq	xxx	en	diq
ba	ba	ru	en	ba
ast	ast	es	fr	ast
tt	tt	bg	en	tt
ga	ga	ga	id	ga
nap	nap	it	it	nap
ia	ia	xxx	fr	ia
cv	cv	ru	en	cv
wa	wa	xxx	id	wa
am	am	am	en	am
kn	kn	kn	en	kn
bug	bug	xx-Bugi	id	bug
my	my	my	en	my
zh-min-nan	nan	zh-TW	fr	zh-min-nan
tg	tg	bg	en	tg
vec	vec	xxx	fr	vec
yi	yi	yi	id	yi
gd	gd	ga	id	gd
hif	hif	ms	en	hif
os	os	ru	en	os
sco	sco	en	en	sco
arz	arz	ar	en	arz
sah	sah	ru	en	sah
uz	uz	tr	en	uz
pam	pam	ms	id	pam
sa	sa	hi	en	sa
mn	mn	ru	en	mn
hsb	hsb	cs	en	hsb
mi	mi	ms	en	mi
li	li	nl	nl	li
si	si	si	en	si
co	co	it	it	co
gan	gan	zh-TW	en	gan
glk	glk	fa	en	glk
bcl	bcl	fil	en	bcl
fo	fo	is	en	fo
ckb	ckb	ur	en	ckb
bar	bar	de	de	bar
bo	bo	bo	en	bo
mrj	mrj	ru	en	mrj
ky	ky	ru	en	ky
ilo	ilo	en	hu	null
tk	tk	id	en	tk
vls	vls	nl	nl	vls

Prefix	ISO 639 code	Chromium CLD	Lingua:: Identify	HTML lang
se	se	xxx	en	se
gv	gv	fr	id	gv
dv	dv	dv	en	dv
rue	rue	uk	en	rue
pag	pag	uk	en	pag
pa	pa	pa	en	pa
ps	ps	ar	en	ps
rm	rm	xxx	id	rm
mzn	mzn	fa	en	mzn
km	km	km	en	km
udm	udm	ru	en	udm
fur	fur	it	fr	fur
csb	csb	sk	en	csb
mt	mt	mt	hr	mt
wuu	wuu	zh	id	wuu
ug	ug	ar	en	ug
lij	lij	xxx	fr	lij
koi	koi	ru	en	koi
kv	kv	ru	en	kv
mhr	mhr	ru	en	mhr
ang	ang	xxx	en	ang
pi	pi	hi	en	pi
lad	lad	es	fr	lad
sc	sc	sk	fr	sc
zea	zea	nl	nl	zea
zh-classical	lzh	zh-TW	en	zh-classical
nov	nov	en	en	nov
ksh	ksh	ru	fr	ksh
so	so	en	id	so
nv	nv	cs	en	nv
hak	hak	zh-TW	en	hak
or	or	or	en	or
kw	kw	da	en	kw
frp	frp	id	fr	frp
stq	stq	xxx	en	stq
szl	szl	sk	id	szl
ext	ext	es	fr	ext
xmf	xmf	ka	en	xmf
pcd	pcd	xxx	fr	pcd
ie	ie	xxx	fr	ie
frr	frr	en	en	frr
xal	xal	ru	en	xal
haw	haw	id	id	haw
rw	rw	xxx	en	rw
ln	ln	fr	fr	ln

Prefix	ISO 639 code	Chromium CLD	Lingua:: Identify	HTML lang
pdc	pdc	de	de	pdc
ay	ay	es	es	ay
pfl	pfl	de	de	pfl
krc	krc	ru	en	krc
ace	ace	ms	en	ace
to	to	en	id	to
crh	crh	tr	en	crh-latn
ce	ce	bg	en	ce
gn	gn	es	fr	gn
myv	myv	ru	en	myv
kl	kl	ms	en	kl
arc	aii	syr	en	arc
pap	pap	xxx	id	pap
gag	gag	tr	en	gag
bjn	bjn	ms	id	bjn
lbe	lbe	ru	en	lbe
tpi	tpi	xxx	id	tpi
as	as	bn	en	as
wo	wo	ms	id	wo
jbo	jbo	xxx	en	jbo
dsb	dsb	sk	en	dsb
mdf	mdf	ru	en	mdf
av	av	ru	en	av
kab	kab	id	en	kab
sn	sn	uk	en	sn
ty	ty	fr	id	ty
srn	srn	nl	en	srn
lo	lo	lo	en	lo
ab	ab	ru	en	ab
ig	ig	xxx	id	ig
na	na	en	en	na
kg	kg	en	en	kg
tet	tet	ms	fr	tet
mwl	mwl	pt	fr	mwl
ltg	ltg	xxx	en	ltg
kaa	kaa	tr	en	kaa
nso	nso	xxx	fr	nso
rmy	rmy	ro	fr	rmy
cu	cu	bg	en	cu
tn	tn	en	en	tn
kbd	kbd	ru	en	kbd
sm	sm	ms	en	sm
bi	bi	ru	en	bi
sd	sd	ur	en	sd
bm	bm	fr	fr	bm

Prefix	ISO 639 code	Chromium CLD	Lingua:: Identify	HTML lang
ik	ik	bn	en	ik
iu	iu	iu	en	iu
ss	ss	en	en	ss
pih	pih	en	en	pih
za	za	en	en	za
chr	chr	chr	en	chr
pnt	pnt	el	en	pnt
got	got	xx-Goth	en	got
cdo	cdo	zh-TW	en	cdo
ee	ee	da	id	ee
ha	ha	en	id	ha
zu	zu	xxx	en	zu
ti	ti	am	en	ti
ks	ks	fa	en	ks
om	om	mk	en	om
bxr	bxr	ru	en	bxr
ve	ve	en	id	ve
ts	ts	xxx	en	ts
sg	sg	xxx	en	sg
rn	rn	en	en	rn
cr	cr	en	id	cr
dz	dz	bo	en	dz
lg	lg	xxx	id	lg
ak	ak	en	en	ak
fj	fj	bn	en	fj
tum	tum	en	en	tum
chy	chy	en	en	chy
ff	ff	fr	fr	ff
ch	ch	xxx	id	ch
st	st	en	id	st
ny	ny	en	en	ny
tw	tw	en	id	tw
xh	xh	en	en	xh
ki	ki	en	en	ki
vep	vep	en	en	vep
ng	ng	en	en	ng
cho	cho	en	en	cho
mh	mh	en	en	mh
aa	aa	en	en	aa
kj	kj	en	en	kj
ho	ho	en	en	ho
mus	mus	en	en	mus
kr	kr	en	en	kr
hz	hz	en	en	hz

SEVENTH FRAMEWORK PROGRAMME
MultilingualWeb