Linguistic Hacking

How to know what a text in an unknown language is about?

Martin.Haase@uni-bamberg.de
maha@jabber.ccc.de

Contents

- how to identify the language of a written text
  - in traditional ways,
    - with the help of computer technology.
  - how to get at least some information out of an unknown text.

Hacking

"the intellectual challenge of creatively overcoming or circumventing limitations."

Eric Raymond (1996): *The New Hacker's Dictionary*.

Spoken texts?

multi-language corpus of telephone calls

Writing Systems

- Roman (thousands of languages)
- Cyrillic (> 60 languages)
- Arabic (> 20 languages)
- Devanāgarī (> 10 languages, not counting derivative writing systems)
- Hebrew (~ 3 – 5 languages)

Devanāgarī

देवनागरी

DEVANAGARI.

(Ältere Druckschrift nach A. W. v. SCHLEGEL.)

Zeichen	Wert	Zeichen	Wert	Zeichen	Wert	Zeichen	Wert	Zeichen	Wert	Zeichen	Wert
अ	a	ऌ	ḷ	ग	a	ट	ṭa	ध	dha	र	ra
आ	ā	ॡ	ḹ	घ	ha	ठ	ṭha	न	na	ल	la
इ	i	ए	e	ङ	ṅa	ड	ḍa	प	pa	ळ	ḷa
ई	ī	ऐ	ai	च	tša	ढ	ḍha	फ	pha	व	va
उ	u	ओ	o	छ	tšha	ण	ṇa	ब	ba	स	sa
ऊ	ū	औ	au	ज	dža	त	ta	भ	bha	श ष	šа
ऋ	ṛ	क	ka	झ	džha	थ	tha	म	ma	ष	ša
ॠ	ṝ	ख	kha	ञ	ńa	द	da	य	ya	ह	ha

DEVANAGARI.

Neuere Druckschrift nach alten Handschriften.

Zeichen	Wert	Zeichen	Wert	Zeichen	Wert	Zeichen	Wert	Zeichen	Wert	Zeichen	Wert
अ	a	ऌ	ḷ	ग	ga	ट	ṭa	ध	dha	र	ra
आ	ā	ॡ	ḹ	घ	gha	ठ	ṭha	न	na	ल	la
इ	i	ए	e	ङ	ṅa	ड	ḍa	प	pa	ळ	ḷa
ई	ī	ऐ	ai	च	tša	ढ	ḍha	फ	pha	व	va
उ	u	ओ	o	छ	tša.	ण	ṇa	ब	ba	स	sa
ऊ	ū	औ	au	ज	dža	त	ta	भ	bha	श	ša
ऋ	ṛ	क	ka	झ	džha	थ	tha	म	ma	ष	ša
ॠ	ṝ	ख	kha	ञ	ña	द	da	य	ya	ह	ha

प	पा	पि	पी
pa	pā	pi	pī
पु	पू	पृ	पॄ
pu	pū	pṛ	pṝ
प्प	प्त	प्र	ङ्ख्य
ppa	pta	pra	ṅkhya

Sample of Armenian

Armenian:	Ի	սրտին	գործարանի	երեւութացեալ	հոգւոյն	
Transliteration:	I	srtin	gorcarani	erewut‘ac‘eal	hogwoyn	
Transcription:	i	səɹtin	goɹtsarani	eɹevut^{h}ats^{h}ial	hogwoin	
Gloss:	in	the.heart	of.the.organ	appearing	of.the.soul	

աչաց	թաթ	ձեռին	աջոյ	գրելով	ի վերայ	վիմի.
ač‘ac‘	t‘at‘	jeŕin	aǰoy	grelov	i veray	vimi.
atʃ^{h}ats^{h}	t^{h}at^{h}	dzeɹin	adʒo	gəɹelow	i veɹa	vimi
of.the.eyes	the.palm	of.the.hand	of.the.right	writing	on	the.rock

զի	որպէս	ի	ձեան	վերջք	գծին	կուսեալ ունէր	քարն:
zi	orpēs	i	jean	veŕǰk‘	gcin	kuteal unēr	k‘arn.
zi	voɹpɛs	i	dzian	veɹdʒk^{h}	gətsin	kutial unɛɹ	k^{h}aɹn
for	as	in	snow	traces	of.the.lines	had.retained	the.stone

եւ	յարուցեալ	յաղօթից	եստեղծ	զնշանագիրս	մեր,	հանդերձ
Ew	yaruc‘eal	yałot‘ic‘n	estełc	znšanagirs	mer,	handerj
jev	haɹuts^{h}ial	haɣot^{h}its^{h}ən	jesteɣts	əznəʃanagiɹəs	meɹ	handeɹʒ
and	arising	from.prayer	he.fashioned	the.letters	our	together

Հռոփիանոսիւ	կերպաձեւեալ	զգիրն	առ	ձեռն	պատրաստ	Մեսրոպայ,
Hrop‘ianosiw	kerpajeweal	zgirn	ar	jeŕn	patrast	Mesropay,
hrop^{h}ianosiv	keɹpadzevial	əzgiɹən	ar	dzerən	patɹast	mesɹopa
with.Rufinus	by.giving.shape	the.letters	by	the.hand	prepared	of.Mesrop

Hebrew

- Old and Modern Hebrew,
- Ladino (with different varieties),
- Judeo-Arabic,
- Yiddish.

א בְּרֵאשִׁית בָּרָא אֱלֹהִים אֵת הַשָּׁמַיִם וְאֵת הָאָרֶץ׃ ב וְהָאָרֶץ הָיְתָה
תֹהוּ וָבֹהוּ וְחֹשֶׁךְ עַל־פְּנֵי תְהוֹם וְרוּחַ אֱלֹהִים מְרַחֶפֶת עַל־פְּנֵי
הַמָּיִם׃ ג וַיֹּאמֶר אֱלֹהִים יְהִי אוֹר וַיְהִי־אוֹר׃ ד וַיַּרְא אֱלֹהִים אֶת־
הָאוֹר כִּי־טוֹב וַיַּבְדֵּל אֱלֹהִים בֵּין הָאוֹר וּבֵין הַחֹשֶׁךְ׃ ה וַיִּקְרָא
אֱלֹהִים ׀ לָאוֹר יוֹם וְלַחֹשֶׁךְ קָרָא לָיְלָה וַיְהִי־עֶרֶב וַיְהִי־בֹקֶר יוֹם אֶחָד׃

1 איִן אָנהייב האָט גאָט באַשאַפֿן דעם הימל און די ערד. 2 זן די ערד איז געווען וויסט
און ליידיק, און פֿינצטערניש איז געווען אויפֿן געזיכט פֿון טהאָם, און דער גייַסט פֿון גאָט
האָט געשוועבט אויפֿן געזיכט פֿון די וואַסערן. 3 חאָט גאָט געזאָגט: זאָל ווערן ליכט. זן עס
איז געוואָרן ליכט. 4 זן גאָט האָט געזען דאָס ליכט אַז עס איז גוט; און גאָט האָט
פֿאַנאַנדערגעשיידט צווישן דעם ליכט און צווישן דער פֿינצטערניש. 5 זן גאָט האָט גערופֿן
דאָס ליכט טאָג, און די פֿינצטערניש האָט ער גערופֿן נאַכט. זן עס איז געווען אָוונט, און
עס איז געווען פֿרימאָרגן, איין טאָג.

bu	3, 10, 11, 19, 23, 35, 40, 41, 42, 50
bú	27
bü	3
by	20, 22, 23, 24, 26, 28, 30, 32, 34, 35, 43
bý	27
bij	25
ċ-	56
ca	10, 14, 15, 17a, 17c, 17d, 22, 23, 24, 40, 41, 53
cá	12, 41
cà	17c, 18
că	14
cã	12
ça	11, 15, 16, 53
çà	11
ce	11, 13, 14, 17c, 41, 42
cé	17c, 41
će	31, 17d, 35
ĉe	2
če	33
ci	2, 11, 13, 14, 17c, 18, 30, 40, 41
ći	35
ĉi	2
či	32, 34
čí	32
co	16, 17, 17a, 17b, 17c, 17d, 19, 30, 32, 41
có	41, 42
cò	16, 17
ço	16, 53
çò	16
čo	32, 34
cu	10, 14, 17, 17a, 17c, 17d, 40, 41, 53, 54
cú	12, 41
cù	18, 42
ća	31
ću	31
ĉu	2

1 Latin
2 Esperanto
3 Volapük
10 Spanish
11 French
12 Portuguese
13 Italian
14 Romanian
15 Catalan
16 Occitan
17 Ladin (Engadine)
17a Sursilvan
17b Surmiran
17c Friulan
17d Ladin (Dolomite)
18 Corsican
19 Jersey language
20 English
21 German
22 Danish
23 Swedish
24 Norwegian
25 Dutch
26 Afrikaans
27 Icelandic
28 Frisian
30 Polish
31 Serbo-Croat
32 Czech
33 Slovenian
34 Slovak
35 Sorbian (Wend)
40 Welsh
41 Irish Gaelic
42 Scottish Gaelic
43 Manx
44 Breton
50 Turkish
51 Finnish
52 Hungarian
53 Albanian (Tosk)
54 = 53 Albanian (Tosk) + some Geg words
55 Basque
56 Maltese
57 Estonian
58 Lithuanian
59 Latvian

* See Page 7

Norman C. Ingle (1980): Language Identification Table. London: Technical Translation International.

bu 3, 10, 11, 19, 23, 35, 40, 41, 42, 50
bú 27
bü 3
by 20, 22, 23, 24, 26, 28, 30, 32, 34, 35, 43
bý 27
bij 25
ċ- 56
ca 10, 14, 15, 17a, 17c, 17d, 22, 23, 24, 40, 41, 53
cá 12, 41
cà 17c, 18
că 14
cã 12
ça 11, 15, 16, 53
çà 11
ce 11, 13, 14, 17c, 41, 42
cé 17c, 41
će 31, 17d, 35

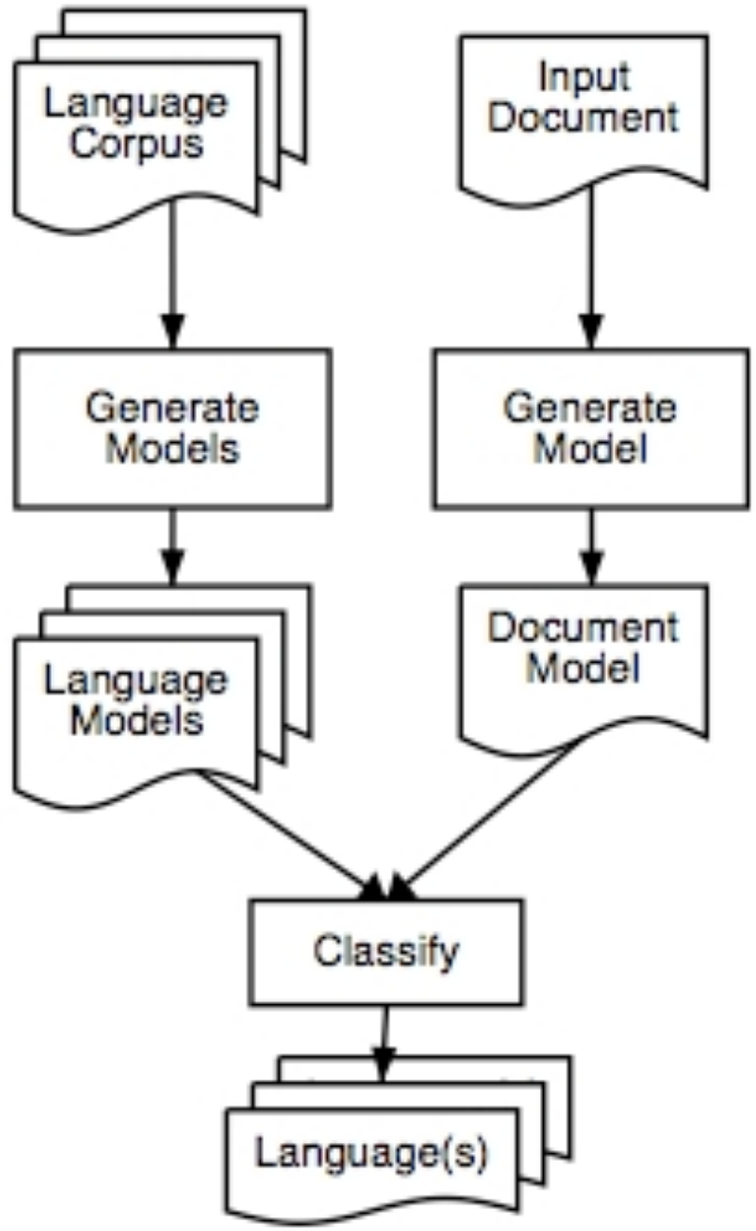
1 Latin
2 Esperanto
3 Volapük
10 Spanish
11 French
12 Portuguese
13 Italian
14 Romanian
15 Catalan
16 Occitan
17 Ladin (Engadine)
17a Sursilvan
17b Surmiran
17c Friulan
17d Ladin (Dolomite)
18 Corsican
19 Jersey language
20 English
21 German
22 Danish
23 Swedish
24 Norwegian

Computer-aided identification

- frequencies of unique characters and character strings
- common words recognition
- *n-gram* analysis

“Text”

(TE), (TEX), (EXT), (XT)


Language Corpus
Input Document
Generate Models
Generate Model
Language Models
Document Model
Classify
Language(s)

variant of the unique character string approach

compression efficiency

reference

model

reference model + text in language to be identified

gzip

reference model

+

text in language to be identified

gzip

reference model + text in language to be identified

compression efficiency

Interesting applications

- measuring linguistic difference
  - > language families
- determining types of text
  - spam detection?

- TextCat (http://odur.let.rug.nl/vannoord/TextCat/Demo/), *n-gram* based, 76 languages, usable as a web application,
- Languid (http://languid.cantbedone.org/), downloadable program, web application not running,
- Langid (http://complingone.georgetown.edu/~langid/), *n-gram* based, 65 languages, web application,
- LanguageGuesser (http://www.xrce.xerox.com/cgi-bin/mltt/LanguageGuesser), frequency tests on characters and character sequences, about 40 languages, web application,
- Polyglot 3000 (http://www.polyglot3000.com/), corpora, method unknown, currently 441 languages, closed-source Windows freeware. :-(

approaching “content analysis”

Hacker’s approach

- numbers, dates, words from another language
- typographic hints:
  - bold or italic print,
  - colored or underlined text chunks,
  - capital letters

Zipf’s law

Very frequent words are shorter and contain less lexical information, whereas infrequent words are longer and contain more lexical information.

less lexical information implies more grammatical information and vice versa

most interesting for us:

words with more specific lexical information

Ignore all short words!
(even if they reiterate throughout the text)

Ua salalau lenei gagana i le lalolagi atoa. ‘O lenei fo‘i gagana, ‘ua ‘avea ma gagana lona lua a le tele o tagata ‘o le vasa Pasefika, e pei ‘o Samoa. E iai le manatu, ‘o le gagana fa‘aperetania, ‘ua matuā talitonu i ai le tele o tagata Samoa e fa‘apea ‘o le gagana e maua ai le atamai ma le poto. ‘E talitonu fo‘i nisi o i latou, ‘e lē aoga la latou gagana. E lē sa‘o lea tāofi, ‘auā e ‘avatu le gagana fa‘aperetania i Samoa, ‘ua leva ona atamamai ma popoto tagata Samoa e fai lo latou soifua ma lo latou lalolagi.

Ua salalau lenei gagana i le lalolagi atoa. ‘O lenei fo‘i gagana, ‘ua ‘avea ma gagana lona lua a le tele o tagata ‘o le vasa Pasefika, e pei ‘o Samoa. E iai le manatu, ‘o le gagana fa‘aperetania, ‘ua matuā talitonu i ai le tele o tagata Samoa e fa‘apea ‘o le gagana e maua ai le atamai ma le poto. ‘E talitonu fo‘i nisi o i latou, ‘e lē aoga la latou gagana. E lē sa‘o lea tāofi, ‘auā e ‘avatu le gagana fa‘aperetania i Samoa, ‘ua leva ona atamamai ma popoto tagata Samoa e fai lo latou soifua ma lo latou lalolagi.

Ua salalau lenei gagana i le lalolagi atoa. ‘O lenei fo‘i gagana, ‘ua ‘avea ma gagana lona lua a le tele o tagata ‘o le vasa Pasefika, e pei ‘o Samoa. E iai le manatu, ‘o le gagana fa‘aperetania, ‘ua matuā talitonu i ai le tele o tagata Samoa e fa‘apea ‘o le gagana e maua ai le atamai ma le poto. ‘E talitonu fo‘i nisi o i latou, ‘e lē aoga la latou gagana. E lē sa‘o lea tāofi, ‘auā e ‘avatu le gagana fa‘aperetania i Samoa, ‘ua leva ona atamamai ma popoto tagata Samoa e fai lo latou soifua ma lo latou lalolagi.

Ua salalau lenei gagana i le lalolagi atoa. ‘O lenei fo‘i gagana, ‘ua ‘avea ma gagana lona lua a le tele o tagata ‘o le vasa Pasefika, e pei ‘o Samoa. E iai le manatu, ‘o le gagana fa‘aperetania, ‘ua matuā talitonu i ai le tele o tagata Samoa e fa‘apea ‘o le gagana e maua ai le atamai ma le poto. ‘E talitonu fo‘i nisi o i latou, ‘e lē aoga la latou gagana. E lē sa‘o lea tāofi, ‘auā e ‘avatu le gagana fa‘aperetania i Samoa, ‘ua leva ona atamamai ma popoto tagata Samoa e fai lo latou soifua ma lo latou lalolagi.