A restricted freedom of choice: Linguistic diversity in the digital landscape

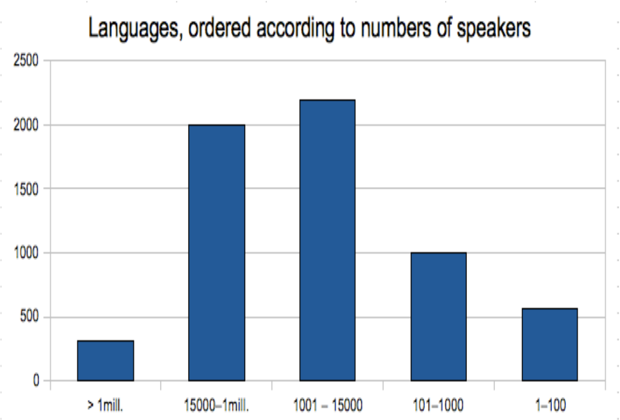
Trond Trosterud

30 мая 2011 г.

Today's topic:

Philosophy, the freedom of will
Bottom line:
Our freedom is restricted in many ways.
If we overlook these restrictions, we will not be able to understand linguistic behaviour

Today's topic: The upper third

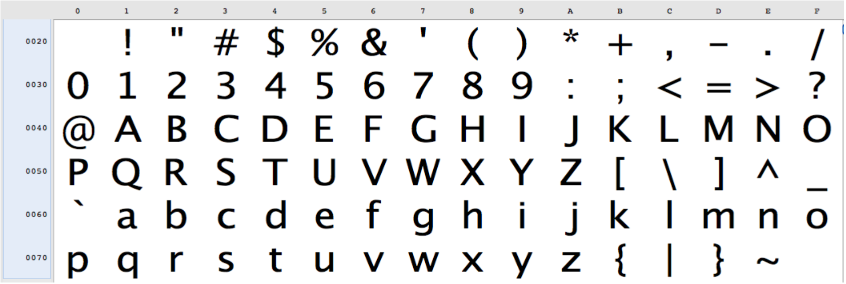


Read

To read you need the letters of the language

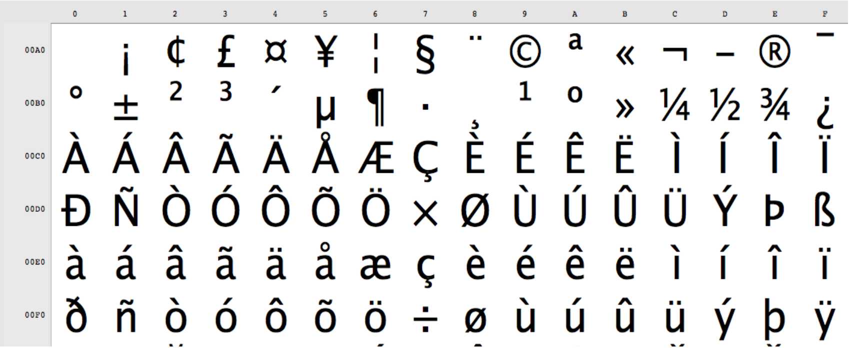
(no arms, no cake)

ASCII



	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0020		!	"	#	$	%	&	'	(	)	*	+	,	−	.	/
0030	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
0040	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
0050	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
0060	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
0070	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	

Latin 1

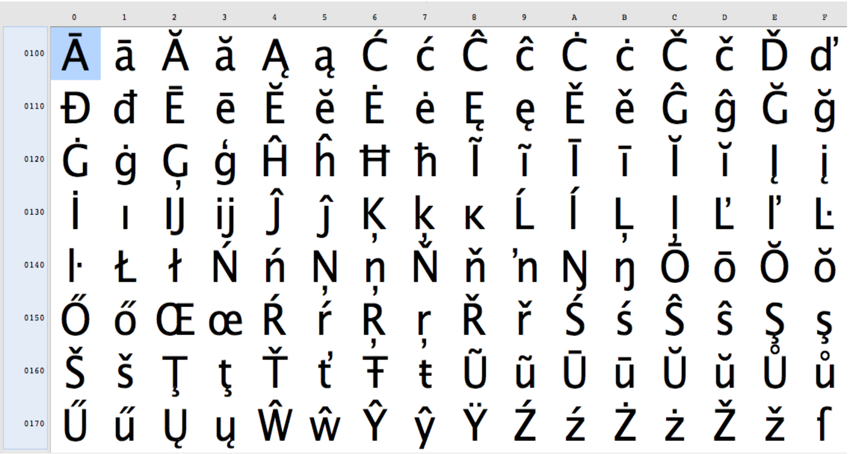


	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
00A0		¡	¢	£	¤	¥	¦	§	¨	©	ª	«	¬	-	®	¯
00B0	°	±	²	³	´	µ	¶	·	¸	¹	º	»	¼	½	¾	¿
00C0	À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï
00D0	Ð	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û	Ü	Ý	Þ	ß
00E0	à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï
00F0	ð	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ	ÿ

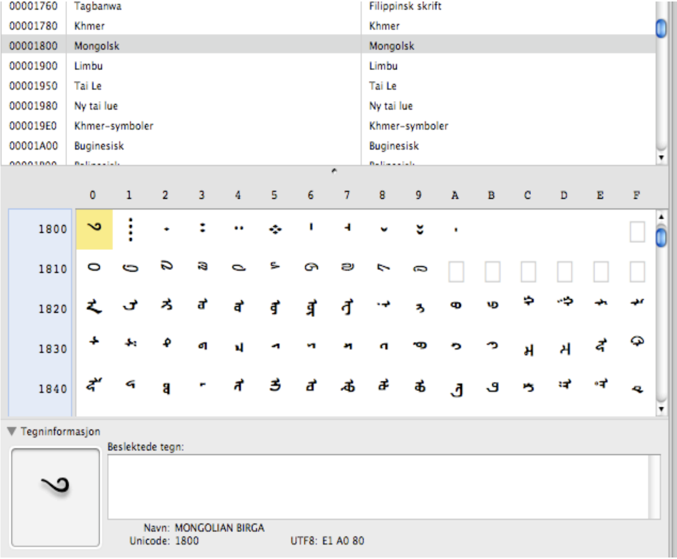
Unicode – the first milestone after Gutenberg

- Contains
  - all writing symbols of all living and most dead languages
  - auxiliary symbols for most linguistic and non-linguistic processing
    - Translitteration alphabets, Braille, ...
    - Mathematical symbols, chess symbols, ...
- Unicode is what makes it possible to publish text in all languages

Latin A



	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0100	Ā	ā	Ă	ă	Ą	ą	Ć	ć	Ĉ	ĉ	Ċ	ċ	Č	č	Ď	ď
0110	Đ	đ	Ē	ē	Ĕ	ĕ	Ė	ė	Ę	ę	Ě	ě	Ĝ	ĝ	Ğ	ğ
0120	Ġ	ġ	Ģ	ģ	Ĥ	ĥ	Ħ	ħ	Ĩ	ĩ	Ī	ī	Ĭ	ĭ	Į	į
0130	İ	ı	Ĳ	ĳ	Ĵ	ĵ	Ķ	ķ	ĸ	Ĺ	ĺ	Ļ	ļ	Ľ	ľ	Ŀ
0140	ŀ	Ł	ł	Ń	ń	Ņ	ņ	Ň	ň	ŉ	Ŋ	ŋ	Ō	ō	Ŏ	ŏ
0150	Ő	ő	Œ	œ	Ŕ	ŕ	Ŗ	ŗ	Ř	ř	Ś	ś	Ŝ	ŝ	Ş	ş
0160	Š	š	Ţ	ţ	Ť	ť	Ŧ	ŧ	Ũ	ũ	Ū	ū	Ŭ	ŭ	Ů	ů
0170	Ű	ű	Ų	ų	Ŵ	ŵ	Ŷ	ŷ	Ÿ	Ź	ź	Ż	ż	Ž	ž	ſ



00001760	Tagbanwa	Filippinsk skrift
00001780	Khmer	Khmer
00001800	Mongolsk	Mongolsk
00001900	Limbu	Limbu
00001950	Tai Le	Tai Le
00001980	Ny tai lue	Ny tai lue
000019E0	Khmer-symboler	Khmer-symboler
00001A00	Buginesisk	Buginesisk
00001B00	Balinesisk	Balinesisk

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
1800	᠀	᠁	᠂	᠃	᠄	᠅	᠆	᠇	᠈	᠉	᠊					□
1810	᠐	᠑	᠒	᠓	᠔	᠕	᠖	᠗	᠘	᠙	□	□	□	□	□	□
1820	ᠠ	ᠡ	ᠢ	ᠣ	ᠤ	ᠥ	ᠦ	ᠧ	ᠨ	ᠩ	ᠪ	ᠫ	ᠬ	ᠭ	ᠮ	ᠯ
1830	ᠰ	ᠱ	ᠲ	ᠳ	ᠴ	ᠵ	ᠶ	ᠷ	ᠸ	ᠹ	ᠺ	ᠻ	ᠼ	ᠽ	ᠾ	ᠿ
1840	ᡀ	ᡁ	ᡂ	ᡃ	ᡄ	ᡅ	ᡆ	ᡇ	ᡈ	ᡉ	ᡊ	ᡋ	ᡌ	ᡍ	ᡎ	ᡏ

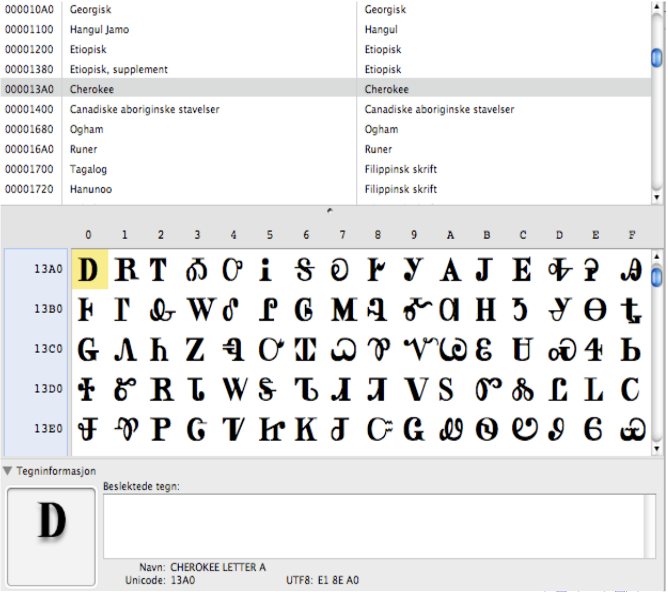
▼ Tegninformasjon

᠀

Beslektede tegn:

Navn: MONGOLIAN BIRGA
Unicode: 1800 UTF8: E1 A0 80

Freedom of choice?
└Written language
 └Read



000010A0	Georgisk	Georgisk
00001100	Hangul Jamo	Hangul
00001200	Etiopisk	Etiopisk
00001380	Etiopisk, supplement	Etiopisk
000013A0	Cherokee	Cherokee
00001400	Canadiske aboriginske stavelser	Canadiske aboriginske stavelser
00001680	Ogham	Ogham
000016A0	Runer	Runer
00001700	Tagalog	Filippinsk skrift
00001720	Hanunoo	Filippinsk skrift

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
13A0	Ꭰ	Ꭱ	Ꭲ	Ꭳ	Ꭴ	Ꭵ	Ꭶ	Ꭷ	Ꭸ	Ꭹ	Ꭺ	Ꭻ	Ꭼ	Ꭽ	Ꭾ	Ꭿ
13B0	Ꮀ	Ꮁ	Ꮂ	Ꮃ	Ꮄ	Ꮅ	Ꮆ	Ꮇ	Ꮈ	Ꮉ	Ꮊ	Ꮋ	Ꮌ	Ꮍ	Ꮎ	Ꮏ
13C0	Ꮐ	Ꮑ	Ꮒ	Ꮓ	Ꮔ	Ꮕ	Ꮖ	Ꮗ	Ꮘ	Ꮙ	Ꮚ	Ꮛ	Ꮜ	Ꮝ	Ꮞ	Ꮟ
13D0	Ꮠ	Ꮡ	Ꮢ	Ꮣ	Ꮤ	Ꮥ	Ꮦ	Ꮧ	Ꮨ	Ꮩ	Ꮪ	Ꮫ	Ꮬ	Ꮭ	Ꮮ	Ꮯ
13E0	Ꮰ	Ꮱ	Ꮲ	Ꮳ	Ꮴ	Ꮵ	Ꮶ	Ꮷ	Ꮸ	Ꮹ	Ꮺ	Ꮻ	Ꮼ	Ꮽ	Ꮾ	Ꮿ

Navn: CHEROKEE LETTER A
Unicode: 13A0 UTF8: E1 8E A0

Why such a generous policy?

- because of Chinese
- But also linguists devoted to making language representation possible

In practice, there still are obstacles

Obstacle: The Norwegian census registry (Folkeregisteret)

- Allows a-z and æøå and äéèôöü
- but not the Sámi letters á or čđŋšž
- Transition period 1.1.2011 - 1.1.2020 (!!!)
    - Ánde, Behkká, Iŋgá, Máret are thus illegal names
    - 63% of the Sámi first names in our base contain á (9.2% contain other Sámi letters)
- The lesson learned: Sámi has no status

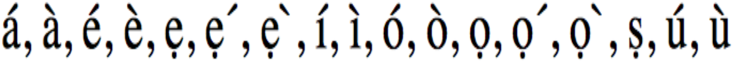
Obstacle: The Norwegian company registry (Brønnøysundregistra):

- Allows Latin 1 (also á), but not the other Sámi letters
  - The newspaper Ávvir can be registered, but Šillju Gatekjøkken & Café Karasjok cannot

There are also languages with real difficulties: Yoruba

- 20 million speakers
- Official status
- No official support on any OS, but Linux work in progress
- Now, there is a discussion on skipping the diacritics

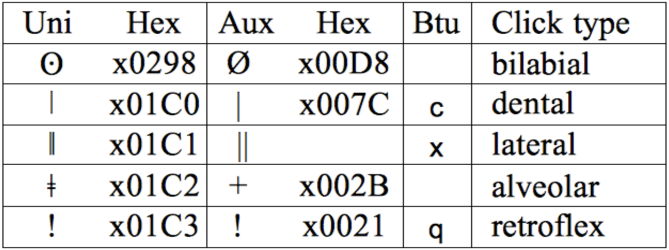
á, à, é, è, ẹ, ẹ́, ẹ̀, í, ì, ó, ò, ọ, ọ́, ọ̀, ṣ, ú, ù



á, à, é, è, ẹ, ẹ´, ẹ`, í, ì, ó, ò, ọ, ọ´, ọ`, ṣ, ú, ù

img/logoWeb070sh

South-West Africa: Click letters



Uni	Hex	Aux	Hex	Btu	Click type
ʘ	x0298	Ø	x00D8		bilabial
ǀ	x01C0	|	x007C	c	dental
ǁ	x01C1	||		x	lateral
ǂ	x01C2	+	x002B		alveolar
ǃ	x01C3	!	x0021	q	retroflex

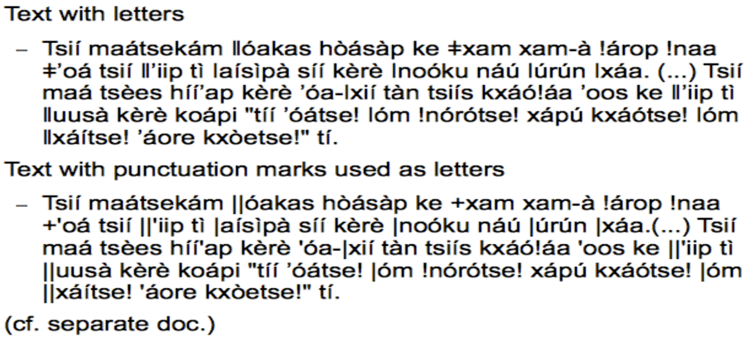
Should one use the click letters?

- Arguments for these click letters
  - They are already in use (conservativity)
  - They differ from other letters (as the clicks differ from other sounds)
- Arguments against these click letters
  - Problem: They look like punctuation marks
  - → confusion, people use the punctuation marks instead

Letters vs. punctuation marks

- These are English WORDS with letters, not numbers like 1980
- These are English WORDS with letters, not numbers like 1980
- These are Eng1ish W0RDS with 1etters and numbers 1ike l98O
- These are Eng1ish WORDS with 1etters and numbers 1ike l98O

Why letters and not punctuation marks?



Text with letters

- Tsií maátsekám ǁóakas hòásàp ke ǂxam xam-à !árop !naa ǂ’oá tsií ǁ’iip tì ǀaísìpà síí kèrè ǀnoóku náú ǀúrún ǀxáa. (...) Tsií maá tsèes híí’ap kèrè ’óa-ǀxií tàn tsiís kxáó!áa ’oos ke ǁ’iip tì ǁuusà kèrè koápi "tíí ’óátse! ǀóm !nórótse! xápú kxáótse! ǀóm ǁxáítse! ’áore kxòetse!" tí.

Text with punctuation marks used as letters

- Tsií maátsekám ||óakas hòásàp ke +xam xam-à !árop !naa +'oá tsií ||'iip tì |aísìpà síí kèrè |noóku náú |úrún |xáa.(...) Tsií maá tsèes híí'ap kèrè 'óa-|xií tàn tsiís kxáó!áa 'oos ke ||'iip tì ||uusà kèrè koápi "tíí ’óátse! |óm !nórótse! xápú kxáótse! |óm ||xáítse! 'áore kxòetse!" tí.

(cf. separate doc.)

Write

Keyboards

For Nama, we need a keyboard to write ǂ, ǃ, not +, !

Keyboards

Out-of-the-box on 3 different platforms (2004 (2011*))

OS	keyboard	GUI
Windows XP	51	33
Mac OS X	78*	-
Linux KDE	-	88

Language keyboards out-of-the-box

12 largest lgs not support out-of-the-box				12 smallest lgs with basic support or more			
Rank	Speakers	Name	Country	Rank	Speakers	Name	Country
26	41.0	Bhojpuri	India	2108	0.014	Inuktitut	Canada
33	30.0	Siraki	Pakistan	1971	0.017	North Sámi	Nordic
35	24.0	Maithili	India	1752	0.022	Cherokee	USA
37	23.0	Oriya	India	1344	0.047	Greenlandic	Greenland
39	22.0	Burmese	Myanmar	1343	0.047	Faroese	Denmark
40	22.0	Hausa	Nigeria	1304	0.050	Maori	NZ
44	20.3	Awadhi	India	991	0.940	Gaelic	Scotland
47	20.0	Yoruba	Nigeria	601	0.250	Icelandic	Iceland
51	17.0	Sindhi	Pakistan	517	0.330	Maltese	Malta
53	16.0	Nepali	Nepal	407	0.500	Breton	France
55	15.0	Amharic	Ethiopia	370	0.580	Welsh	UK
59	13.7	Assamese	India	292	0.910	Basque	Spain
60	13.0	Haryanvi	India	130	4.000	Georgian	Georgia

The haves and the havenots of the linguistic scene

The haves

- Languages with IT support (the top 100 lgs)
  1. Languages with official status in an independent country, and rich and monolingual speakers
  2. (Most) official state languages of India
  3. Minority languages with a strong government backing them up (W Europe, Canada, NZ)

The havenots

- Languages with marginal or no IT support (the remaining 6400 lgs)
  1. African languages
  2. Indian languages other than the official state lgs
  3. Languages without official status in an independent country, especially in former British and French colonies

How to get what you do not have – Sámi Localisation

Sámi Localisation – a success story

- North Sámi keyboard layout is now included, out of the box, no matter where you buy your computer,
- from Linux KDE 3.0, Mac OS 10.3, Win XP SP2 onwards

Sámi Localisation – a success story

- a decade of hard work, involving experts and language users
- consensus-seeking conferences among users
- standardisation (ISO, CEN, national standards)
- pressure from our state administrations upon the OS vendors
- the open source movement

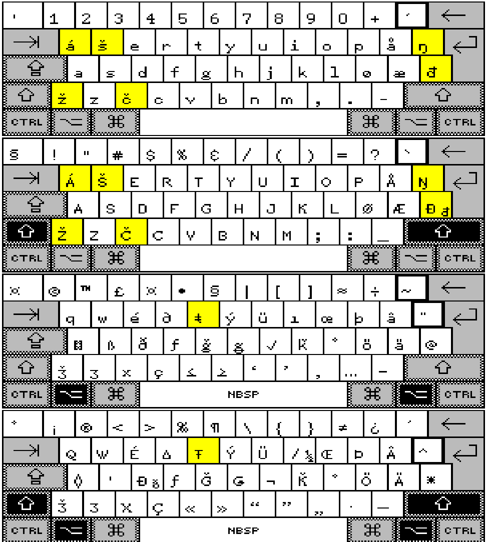
What we did for Sámi

- Already many keyboard layouts available
  1. We compared them to each other
  2. Letters that had the same positions in all former keyboards kept their positions
- The layouts in different countries were based on different keyboards
  1. We made one Sámi keyboard for each country
  2. @, §, ’, etc. were placed as in the national keyboards
  3. ... but the letters were kept in the same positions

Placement of Sámi letters varied on existing keyboards

- which keys to use for Sámi letters?
  - Strategy: Keep both maj lg letters and Sámi, and sacrifice non-Nordic q, w, x
- How to place the Sámi letters?
  - According to text frequency
  - The most common letters were given more prominent positions.
- Where to place the replaced letters?
  - As a rule, we put the replaced letters one level up.
  - So, when the key `W` gives š, then, in order to get w, you press option-w, etc..

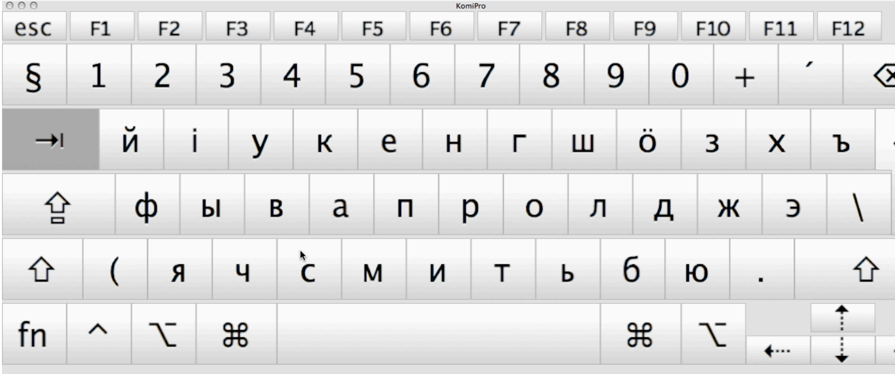
Northern Sámi keyboard for Macintosh



Survey of non-Russian letters in some Uralic languages of Russia

	Ltr	Rank	%	Ltr	Rank	%
Komi Permyak	ӧ	3	6,92	ъ	34	0,03
	і	22	1,32	щ	35	0,02
Komi	ӧ	4	6,38	щ	34	0,02
	і	27	0,81	ц	33	0,16
Meadow Mari	ӱ	22	1,12	щ	35	0,02
	ӧ	24	0,61	ъ	34	0,02
	ҥ	26	0,52	ф	33	0,09
Hill Mari	ӹ	2	6,84	щ	35	0,02
	ӓ	11	4,37	ю	34	0,07
	ӱ	31	0,29	ф	32	0,27
	ӧ	33	0,16	б	30	0,35

Keyboard for Komi


KomiPro
esc F1 F2 F3 F4 F5 F6 F7 F8 F9 F10 F11 F12
§ 1 2 3 4 5 6 7 8 9 0 + ´ ⌫
⇥ й і у к е н г ш ӧ з х ъ
⇪ ф ы в а п р о л д ж э \
⇧ (я ч с м и т ь б ю . ⇧
fn ^ ⌥ ⌘ ⌘ ⌥

The computer should adjust to humans, and not vice versa

- Orthographies and keyboard layouts should be designed according to linguistic and ergonomic principles

The computer should adjust to humans, and not vice versa

- Orthographies and keyboard layouts should be designed according to linguistic and ergonomic principles
- We linguists invented these diacritic signs – we should help the speakers out

The computer should adjust to humans, and not vice versa

- Orthographies and keyboard layouts should be designed according to linguistic and ergonomic principles
- We linguists invented these diacritic signs – we should help the speakers out
- Do not change the orthography – change the computer

And status quo?

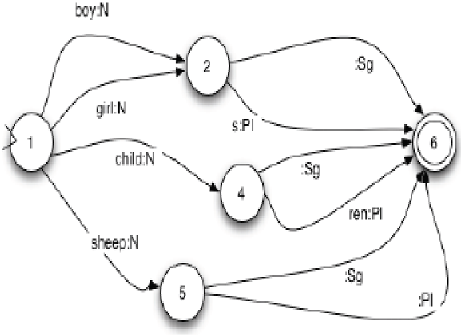
- Komi
  - Search-and-replace in MS Word
  - ... but often with Latin i, ö, not with Cyrillic і, ӧ
- Udmurt
  - 25 ӥ, 30 ӧ, 34 ӟ, 36 ӵ, 37 ӝ (low rank)
  - No Latin letters available: Udmurt letters on the number row

Basic language technology

Grammatical analysers

- ... take text or words as input and deliver a grammatical analysis, or vice versa.
  - Analysis – based upon morphological transducers
  - Disambiguation of grammatical homonymy

Morphological transducers: The language machine



Language with more than a rudimentary morphology need morphological transducers

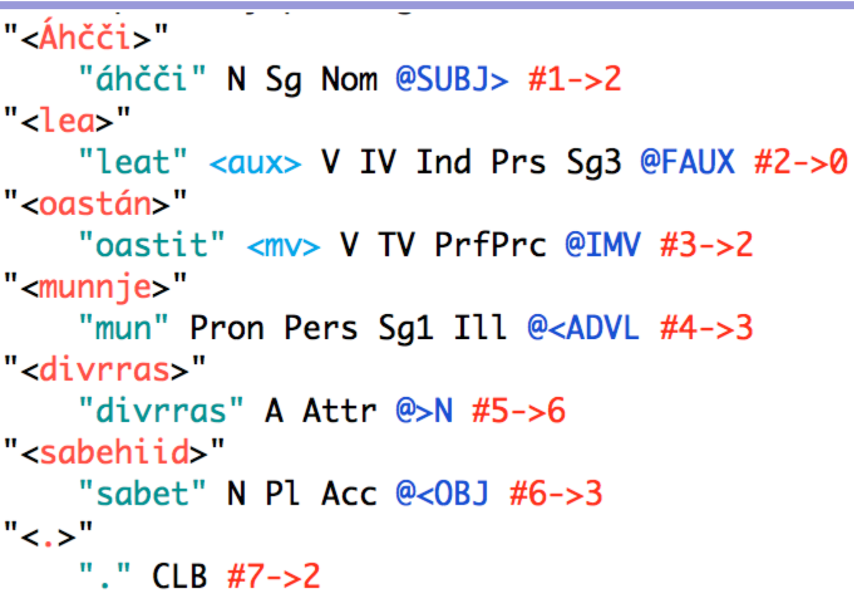
Such transducers can be written within a year or so

Syntactic analysis

- Áhčči lea oastán munnje divrras sabehiid
- ’Father has bought me an expensive pair of skis’
  - Morphological analysis
  - Disambiguation
  - Dependency analysis

```
"<Áhčči>"
          "áhčči" N Sg Nom
"<lea>"
          "leat" V Ind Prs Sg3
"<oastán>"
          "oastit" V PrfPrc
          "oastit" V* N Actor Sg Nom PxSg1
          "oastit" V* N Actor Sg Gen PxSg1
          "oastit" V* N Actor Sg Acc PxSg1
          "oasti" N Sg Nom PxSg1
          "oasti" N Sg Gen PxSg1
          "oasti" N Sg Acc PxSg1
"<munnje>"
          "mun" Pron Pers Sg1 Ill
"<divrras>"
          "divrras" A Attr
          "divrras" A Sg Nom
"<sabehiid>"
          "sabet" N Pl Gen
          "sabet" N Pl Acc
"<.>"
```

```
"<Áhčči>"
    "áhčči" N Sg Nom
"<lea>"
    "leat" V IV Ind Prs Sg3
"<oastán>"
    "oastit" V TV PrfPrc
"<munnje>"
    "mun" Pron Pers Sg1 Ill
"<divrras>"
    "divrras" A Attr
"<sabehiid>"
    "sabet" N Pl Acc
"<.>"
    "." CLB
```



```
"<Áhčči>"
      "áhčči" N Sg Nom @SUBJ> #1->2
"<lea>"
      "leat" <aux> V IV Ind Prs Sg3 @FAUX #2->0
"<oastán>"
      "oastit" <mv> V TV PrfPrc @IMV #3->2
"<munnje>"
      "mun" Pron Pers Sg1 Ill @<ADVL #4->3
"<divrras>"
      "divrras" A Attr @>N #5->6
"<sabehiid>"
      "sabet" N Pl Acc @<OBJ #6->3
"<.>"
      "." CLB #7->2
```

Summing up with an overview over language technology activity

The languages found on aclWiki

- Afrikaans
- Albanian
- Amharic
- Arabic
- Basque
- Breton
- Bulgarian
- Catalan
- Chinese
- Croatian
- Czech
- Danish
- Dutch
- English
- Estonian
- Faroese
- Finnish
- French
- Galician
- Georgian
- German
- Greek
- Greenlandic
- Haitian
- Hebrew
- Hindi
- Icelandic
- Iranian
- Irish
- Italian
- Iñupiaq
- Japanese
- Komi
- Korean
- Kurdish
- Lithuanian
- Macedonian
- Malay
- Montenegrin
- Norwegian
- Navajo
- Occitan
- Persian
- Polish
- Portugese
- Punjabi
- Quechua
- Romanian
- Russian
- Sanskrit
- Serbian
- Slovak
- Slovenian
- Sorbian
- Spanish
- Swahili
- Swedish
- Sámi
- Tajik
- Turkish
- Tigrinya
- Ukrainian
- Urdu
- Vietnbamese
- Welsh

The languages found on aclWiki

State-level official lgs	46
Regional-level official lgs	18
Languages with no official status	0

Native American languages Iñupiaq

African languages Amharic, Swahili

Asian languages Chinese, Hindi, Japanese, Korean, Malay, Pashto, Persian, Sanskrit

Largest languages not found on aclWiki

Rank	Speakers	Name	Rank	Speakers	Name	Rank	Speakers	Name
6	168.0	Bengali	44	20.3	Awadhi	84	8.0	Bundeli
12	75.5	Javanese	47	20.0	Yoruba	85	8.0	Ilocano
14	69.0	Telugu	50	17.0	Indonesian	86	8.0	Kazakh
16	61.0	Marathi	51	17.0	Sindhi	87	8.0	Rwanda
17	59.0	Tamil	53	16.0	Nepali	88	7.5	Uyghur
18	59.0	Vietnamese	54	15.0	Uzbek	90	7.1	Marwari
20	51.3	Urdu	56	15.0	Tai	91	7.1	Khmer
23	46.0	Ukrainian	57	14.5	Hungarian	92	7.0	Neapolitan
25	41.5	Gujarati	60	13.8	Azerbaijani	93	7.0	Akan
26	41.0	Bhojpuri	60	13.7	Assamese	94	7.0	Farsi
30	33.0	Kannada	60	13.0	Haryanvi	95	7.0	Kurmanji
32	30.0	Panjabi	61	13.0	Sinhala	96	7.0	Shona
33	30.0	Siraiki	62	12.2	Igbo	97	7.0	Somali
35	24.0	Maithili	63	12.0	Cebuano	98	7.0	Tatar
37	23.0	Oriya	70	10.7	Deccan	99	6.8	Azerbaijani
38	23.0	Panjabi	70	10.5	Tagalog	100	6.5	Xhosa
39	22.0	Burmese	72	10.0	Magahi	102	6.3	Luba-Kasai
40	22.0	Hausa	73	10.0	Zhuang	103	6.1	Haitian
41	21.0	Thai	76	9.1	Lombard	104	6.0	Kurdi
43	20.5	Farsi	80	8.2	Chattisgarhi	105	6.0	Tai

So far, no big surprises

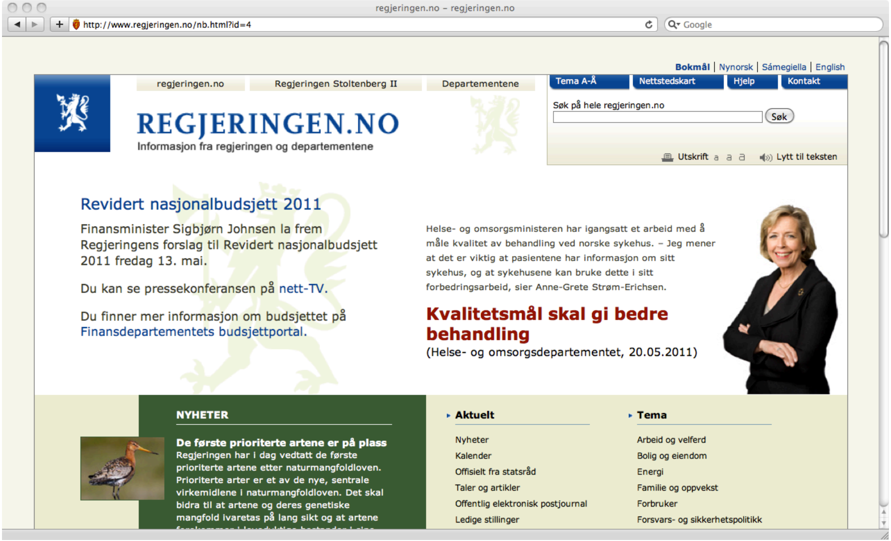
- The rich ones get richer...
  - the number of languages with lg tech resources is small
  - having such resources become more and more important

Demo case: The ultimate LT challenge – Machine translaton

- The first language technology project: cold war MT
- Hard task, impossible to cheat (unlike (socio)linguists, you cannot select data, your one pet sentence, or social variable)
- We linguists lost the cold war

The challenge today

In the future there will be no bilingual administration without a MT system facilitating text production


regjeringen.no – regjeringen.no
http://www.regjeringen.no/nb.html?id=4
Google
Bokmål | Nynorsk | Sámegiella | English
regjeringen.no
Regjeringen Stoltenberg II
Departementene
Tema A-Å
Nettstedskart
Hjelp
Kontakt
REGJERINGEN.NO
Informasjon fra regjeringen og departementene
Søk på hele regjeringen.no
Søk
Utskrift a a a
Lytt til teksten

Revidert nasjonalbudsjett 2011

Finansminister Sigbjørn Johnsen la frem Regjeringens forslag til Revidert nasjonalbudsjett 2011 fredag 13. mai.

Du kan se pressekonferansen på nett-TV.

Du finner mer informasjon om budsjettet på Finansdepartementets budsjettportal.

Helse- og omsorgsministeren har igangsatt et arbeid med å måle kvalitet av behandling ved norske sykehus. – Jeg mener at det er viktig at pasientene har informasjon om sitt sykehus, og at sykehusene kan bruke dette i sitt forbedringsarbeid, sier Anne-Grete Strøm-Erichsen.

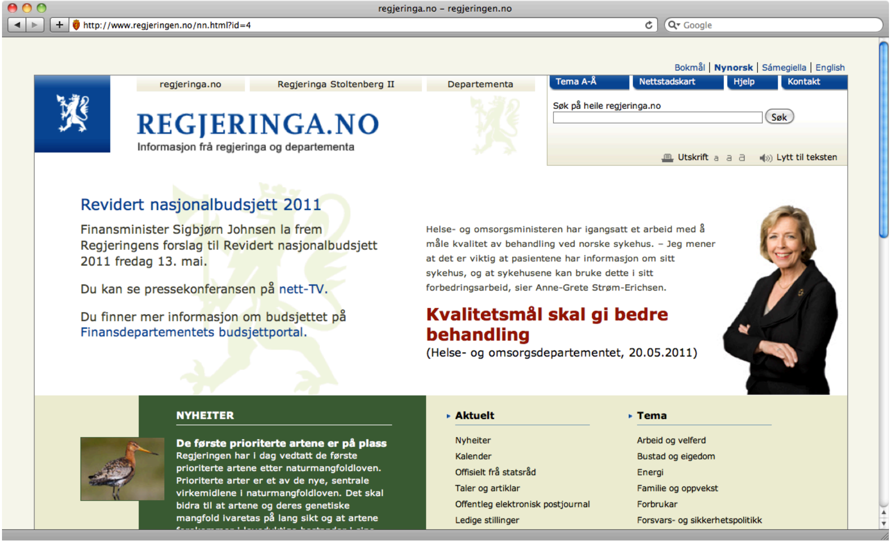
Kvalitetsmål skal gi bedre behandling

(Helse- og omsorgsdepartementet, 20.05.2011)

NYHETER

De første prioriterte artene er på plass
Regjeringen har i dag vedtatt de første prioriterte artene etter naturmangfoldloven. Prioriterte arter er et av de nye, sentrale virkemidlene i naturmangfoldloven. Det skal bidra til at artene og deres genetiske mangfold ivaretas på lang sikt og at artene

../../img/logoWeb070sh



regjeringa.no – regjeringen.no
http://www.regjeringen.no/nn.html?id=4
Google

Bokmål | **Nynorsk** | Sámegiella | English

REGJERINGA.NO

Informasjon frå regjeringa og departementa

Søk på heile regjeringa.no

Revidert nasjonalbudsjett 2011

Finansminister Sigbjørn Johnsen la frem Regjeringens forslag til Revidert nasjonalbudsjett 2011 fredag 13. mai.

Du kan se pressekonferansen på nett-TV.

Du finner mer informasjon om budsjettet på Finansdepartementets budsjettportal.

Helse- og omsorgsministeren har igangsatt et arbeid med å måle kvalitet av behandling ved norske sykehus. – Jeg mener at det er viktig at pasientene har informasjon om sitt sykehus, og at sykehusene kan bruke dette i sitt forbedringsarbeid, sier Anne-Grete Strøm-Erichsen.

Kvalitetsmål skal gi bedre behandling

(Helse- og omsorgsdepartementet, 20.05.2011)

NYHEITER

De første prioriterte artene er på plass

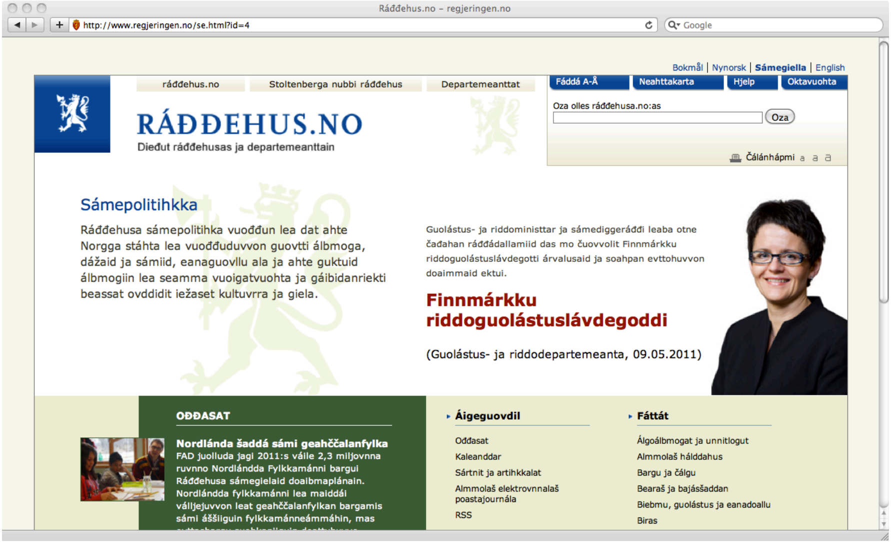
Regjeringen har i dag vedtatt de første prioriterte artene etter naturmangfoldloven. Prioriterte arter er et av de nye, sentrale virkemidlene i naturmangfoldloven. Det skal bidra til at artene og deres genetiske mangfold ivaretas på lang sikt og at artene

Aktuelt

- Nyheiter
- Kalender
- Offisielt frå statsråd
- Taler og artiklar
- Offentleg elektronisk postjournal
- Ledige stillinger

Tema

- Arbeid og velferd
- Bustad og eigedom
- Energi
- Familie og oppvekst
- Forbrukar
- Forsvars- og sikkerhetspolitikk


Ráđđehus.no – regjeringen.no
http://www.regjeringen.no/se.html?id=4
Google
Bokmål | Nynorsk | Sámegiella | English
ráđđehus.no
Stoltenberga nubbi ráđđehus
Departemeanttat
Fáddá A-Å
Neahttakarta
Hjelp
Oktavuohta
RÁĐĐEHUS.NO
Dieđut ráđđehusas ja departemeanttain
Oza olles ráđđehusa.no:as
Oza
Čálánhápmi a a a

Sámepolitihkka

Ráđđehusa sámepolitihka vuođđun lea dat ahte Norgga stáhta lea vuođđuduvvon guovtti álbmoga, dážaid ja sámiid, eanaguovllu ala ja ahte guktuid álbmogiin lea seamma vuoigatvuohta ja gáibidanriekti beassat ovddidit iežaset kultuvrra ja giela.

Guolástus- ja riddoministtar ja sámediggeráđđi leaba otne čađahan ráđđádallamiid das mo čuovvolit Finnmárkku riddoguolástuslávdegotti árvalusaid ja soahpan evttohuvvon doaimmaid ektui.

Finnmárkku riddoguolástuslávdegoddi

(Guolástus- ja riddodepartemeanta, 09.05.2011)

ODĐASAT

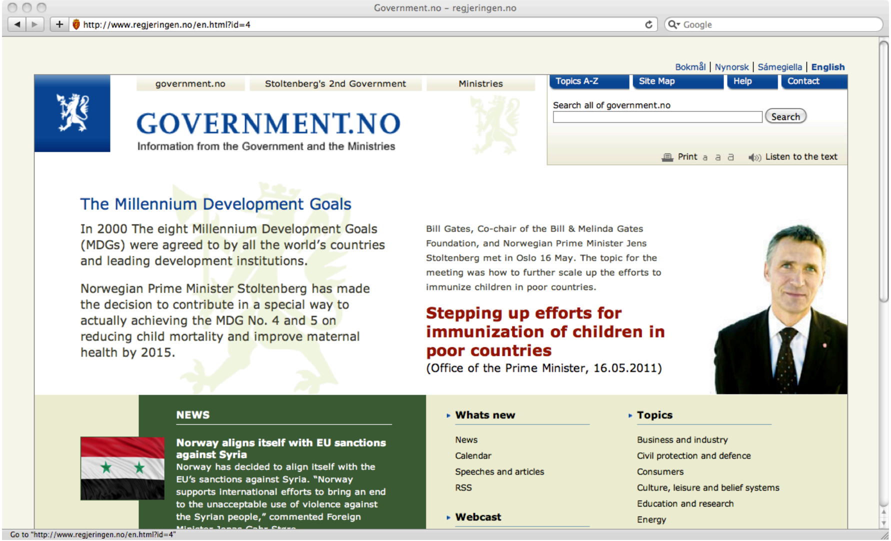
Nordlánda šaddá sámi geahččalanfylka
FAD juolluda jagi 2011:s váile 2,3 miljovnna ruvnno Nordlándda Fylkkamánni bargui Ráđđehusa sámegielaid doaibmaplánain. Nordlándda fylkkamánni lea maiddái válljejuvvon leat geahččalanfylkan bargamis sámi áššiiguin fylkkamánneámmáhin, mas

Áigeguovdil

- Ođđasat
- Kaleanddar
- Sártnit ja artihkkalat
- Almmolaš elektrovnnalaš poastajournála
- RSS

Fáttát

- Álgoálbmogat ja unnitlogut
- Almmolaš hálddahus
- Bargu ja čálgu
- Bearaš ja bajásšaddan
- Biebmu, guolástus ja eanadoallu
- Biras



Government.no – regjeringen.no
http://www.regjeringen.no/en.html?id=4
Google
Bokmål | Nynorsk | Sámegiella | English
government.no
Stoltenberg's 2nd Government
Ministries
Topics A-Z
Site Map
Help
Contact

GOVERNMENT.NO

Information from the Government and the Ministries

Search all of government.no

Search

Print a a a Listen to the text

The Millennium Development Goals

In 2000 The eight Millennium Development Goals (MDGs) were agreed to by all the world's countries and leading development institutions.

Norwegian Prime Minister Stoltenberg has made the decision to contribute in a special way to actually achieving the MDG No. 4 and 5 on reducing child mortality and improve maternal health by 2015.

Bill Gates, Co-chair of the Bill & Melinda Gates Foundation, and Norwegian Prime Minister Jens Stoltenberg met in Oslo 16 May. The topic for the meeting was how to further scale up the efforts to immunize children in poor countries.

Stepping up efforts for immunization of children in poor countries

(Office of the Prime Minister, 16.05.2011)

NEWS

Norway aligns itself with EU sanctions against Syria

Norway has decided to align itself with the EU's sanctions against Syria. "Norway supports international efforts to bring an end to the unacceptable use of violence against the Syrian people," commented Foreign Minister Jonas Gahr Støre

Whats new

News
Calendar
Speeches and articles
RSS

Webcast

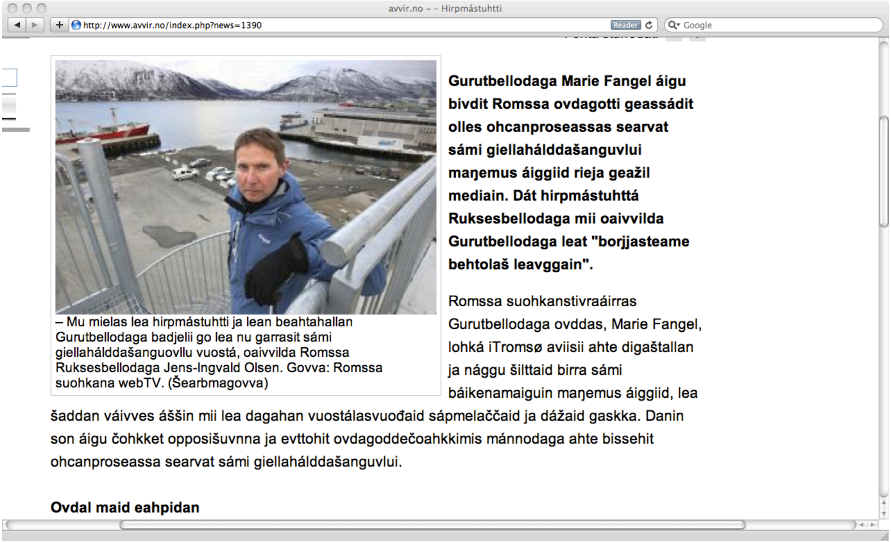
Topics

Business and industry
Civil protection and defence
Consumers
Culture, leisure and belief systems
Education and research
Energy

Go to "http://www.regjeringen.no/en.html?id=4"

MT in the north

- Minority to majority language
- – what do they write about me?

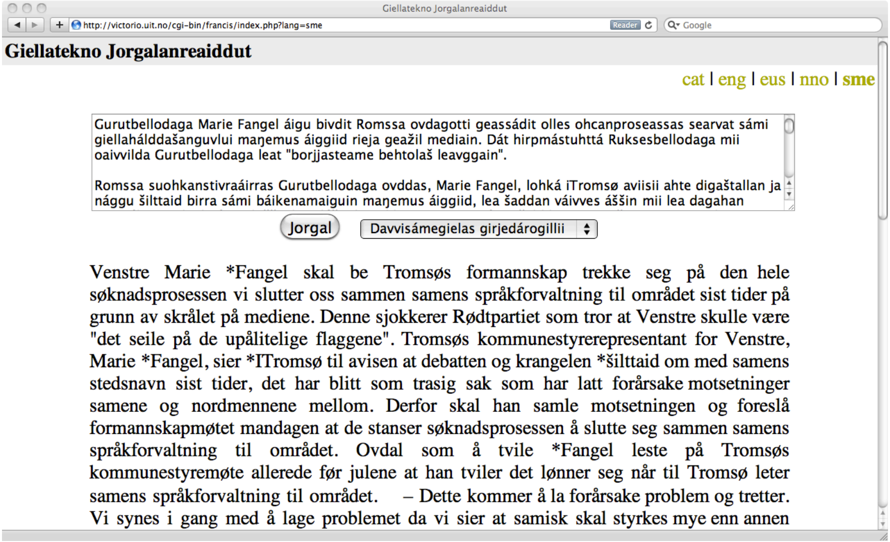

avvir.no – – Hirpmástuhtti
http://www.avvir.no/index.php?news=1390
Reader
Google

– Mu mielas lea hirpmástuhtti ja lean beahtahallan Gurutbellodaga badjelii go lea nu garrasit sámi giellahálddašanguovllu vuostá, oaivvilda Romssa Ruksesbellodaga Jens-Ingvald Olsen. Govva: Romssa suohkana webTV. (Šearbmagovva)

Gurutbellodaga Marie Fangel áigu bivdit Romssa ovdagotti geassádit olles ohcanproseassas searvat sámi giellahálddašanguvlui maŋemus áiggiid rieja geažil mediain. Dát hirpmástuhttá Ruksesbellodaga mii oaivvilda Gurutbellodaga leat "borjjasteame behtolaš leavggain".

Romssa suohkanstivraáirras Gurutbellodaga ovddas, Marie Fangel, lohká iTromsø aviisii ahte digaštallan ja nággu šilttaid birra sámi báikenamaiguin maŋemus áiggiid, lea šaddan váivves áššin mii lea dagahan vuostálasvuođaid sápmelaččaid ja dážaid gaskka. Danin son áigu čohkket opposišuvnna ja evttohit ovdagoddečoahkkimis mánnodaga ahte bissehit ohcanproseassa searvat sámi giellahálddašanguvlui.

Ovdal maid eahpidan


Giellatekno Jorgalanreaiddut
http://victorio.uit.no/cgi-bin/francis/index.php?lang=sme
Reader
Google

Giellatekno Jorgalanreaiddut

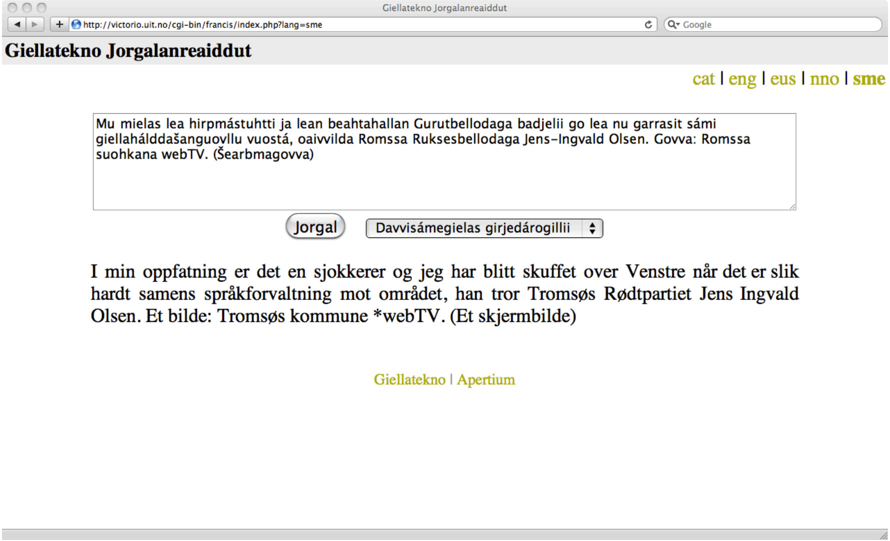
cat | eng | eus | nno | sme

Gurutbellodaga Marie Fangel áigu bivdit Romssa ovdagotti geassádit olles ohcanproseassas searvat sámi giellahálddašanguvlui maŋemus áiggiid rieja geažil mediain. Dát hirpmástuhttá Ruksesbellodaga mii oaivvilda Gurutbellodaga leat "borjjasteame behtolaš leavggain".

Romssa suohkanstivraáirras Gurutbellodaga ovddas, Marie Fangel, lohká iTromsø aviisii ahte digaštallan ja nággu šilttaid birra sámi báikenamaiguin maŋemus áiggiid, lea šaddan váivves áššin mii lea dagahan

Venstre Marie *Fangel skal be Tromsøs formannskap trekke seg på den hele søknadsprosessen vi slutter oss sammen samens språkforvaltning til området sist tider på grunn av skrålet på mediene. Denne sjokkerer Rødtpartiet som tror at Venstre skulle være "det seile på de upålitelige flaggene". Tromsøs kommunestyrerepresentant for Venstre, Marie *Fangel, sier *ITromsø til avisen at debatten og krangelen *šilttaid om med samens stedsnavn sist tider, det har blitt som trasig sak som har latt forårsake motsetninger samene og nordmennene mellom. Derfor skal han samle motsetningen og foreslå formannskapmøtet mandagen at de stanser søknadsprosessen å slutte seg sammen samens språkforvaltning til området. Ovdal som å tvile *Fangel leste på Tromsøs kommunestyremøte allerede før julene at han tviler det lønner seg når til Tromsø leter samens språkforvaltning til området. – Dette kommer å la forårsake problem og tretter. Vi synes i gang med å lage problemet da vi sier at samisk skal styrkes mye enn annen

../../img/logoWeb070sh


Giellatekno Jorgalanreaiddut
http://victorio.uit.no/cgi-bin/francis/index.php?lang=sme
Google

Giellatekno Jorgalanreaiddut

cat | eng | eus | nno | sme

Mu mielas lea hirpmástuhtti ja lean beahtahallan Gurutbellodaga badjelii go lea nu garrasit sámi giellahálddašanguovllu vuostá, oaivvilda Romssa Ruksesbellodaga Jens-Ingvald Olsen. Govva: Romssa suohkana webTV. (Šearbmagovva)

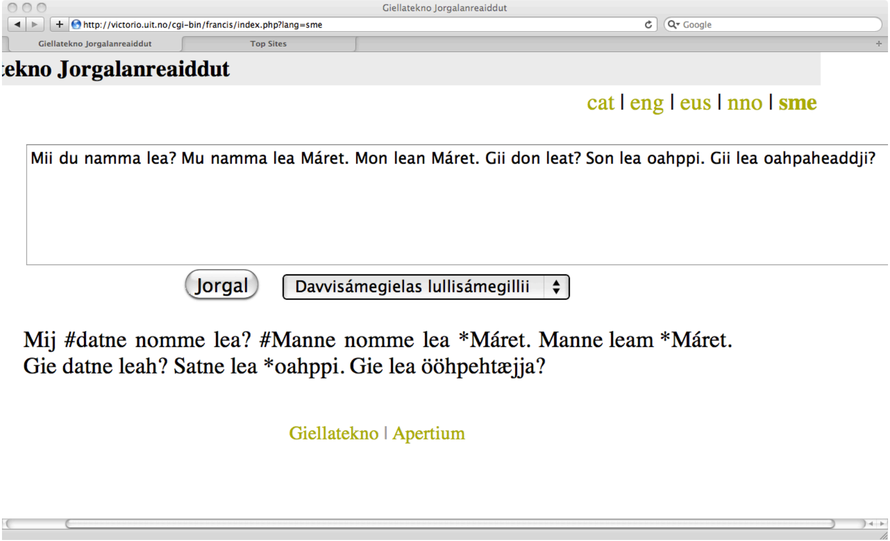
Jorgal
Davvisámegielas girjedárogillii

I min oppfatning er det en sjokkerer og jeg har blitt skuffet over Venstre når det er slik hardt samens språkforvaltning mot området, han tror Tromsøs Rødtpartiet Jens Ingvald Olsen. Et bilde: Tromsøs kommune *webTV. (Et skjermbilde)

Giellatekno | Apertium

MT in the north

- Translating between closely related languages
  - Greenlandic to Inuktitut
  - North Sámi to South Sámi
- Goal: Text production


Giellatekno Jorgalanreaiddut
http://victorio.uit.no/cgi-bin/francis/index.php?lang=sme
Google
Giellatekno Jorgalanreaiddut
Top Sites

ekno Jorgalanreaiddut

cat | eng | eus | nno | **sme**

Mii du namma lea? Mu namma lea Máret. Mon lean Máret. Gii don leat? Son lea oahppi. Gii lea oahpaheaddji?

Jorgal

Davvisámegielas lullisámegillii

Mij #datne nomme lea? #Manne nomme lea *Máret. Manne leam *Máret. Gie datne leah? Satne lea *oahppi. Gie lea ööhpehtæjja?

Giellatekno | Apertium

Note: Grammar-based lg tech is not the dominating approach

- Two flavours of language technology
- Grammatical (*symbolic*) approach
    - Good results for some grammatical frameworks
    - ... not so good for others
    - Much lg-specific work
    - Demands a good grammar, dictionary, and a modest text corpus (1+ mill)
    - Statistical (*stochastic*) approach
        - Good results for lgs with small morphologies
        - Marginal lg-specific work
        - Demands huge text corpora (100+ mill)
- The last decade and a half, the latter has dominated

The languages of Google Translate

- Western European
  - Basque, Catalan, Danish, Dutch, English, Estonian, Finnish, French, Galician, German, Icelandic, Irish, Italian, Latin, Maltese, Norwegian, Portuguese, Spanish, Swedish, Welsh, Yiddish
- Eastern European
  - Albanian, Belarussian, Bulgarian, Croatian(-), Czech, Georgian, Greek, Hungarian, Latvian, Lithuanian, Macedonian, Polish, Romanian, Russian, Serbian, Slovak, Slovene, Ukranian

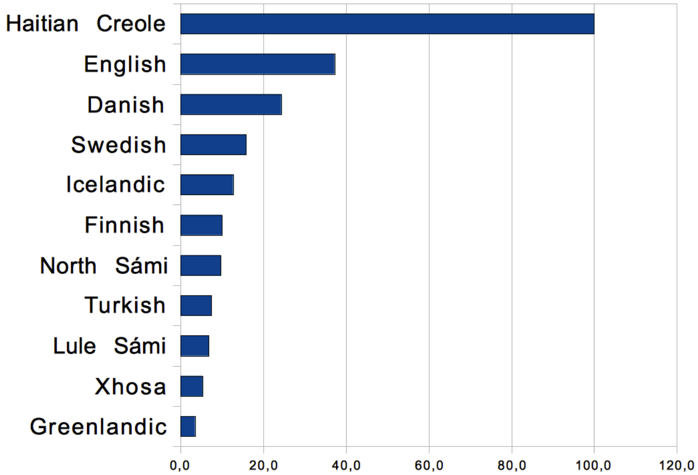
The languages of Google Translate

- Asian
  - Arabic, Armenian, Azerbaijani, Chinese, Filipino, Hebrew, Hindi, Indonesian, Japanese, Korean, Persian, Thai, Turkish, Urdu, Vietnamese
- African
  - Afrikaans, Swahili
- American
  - Haitian Creole

The languages excluded from Google Translate

- Languages with less than 100 million words of text
  - No text, no translate
- Languages with much morphology
  - Token-type ratio for some languages

Token-type ratio



MT as a challenge to linguistics

- MT requires full control over all aspects of language
- MT has a direct impact upon the language community

Bridging the gap: Language technology for minority langagues

- More basic tools become open source

Bridging the gap: Language technology for minority langagues

- More basic tools become open source
  - Helsinki finite transducer, Odense disambiguator, ...

Bridging the gap: Language technology for minority langagues

- More basic tools become open source
  - Helsinki finite transducer, Odense disambiguator, ...
  - Publically funded lexica become accessible (Finland, Norway, ...)

Bridging the gap: Language technology for minority langagues

- More basic tools become open source
  - Helsinki finite transducer, Odense disambiguator, ...
  - Publically funded lexica become accessible (Finland, Norway, ...)
- Lg tech might be run as academic projects for vanishing languages

Bridging the gap: Language technology for minority langagues

- More basic tools become open source
  - Helsinki finite transducer, Odense disambiguator, ...
  - Publically funded lexica become accessible (Finland, Norway, ...)
- Lg tech might be run as academic projects for vanishing languages
  - For lingustics, languages with few speakers are as interesting as languages with many speakers

Bridging the gap: Language technology for minority langagues

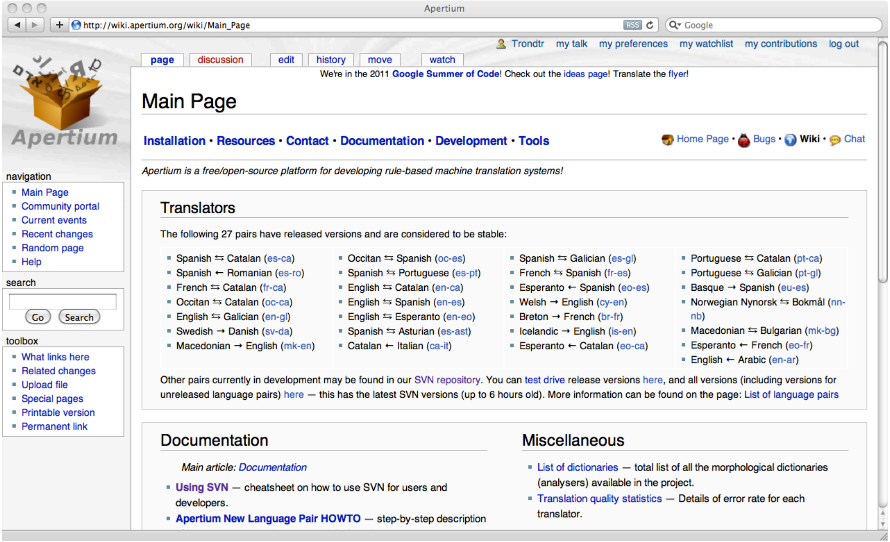
- More basic tools become open source
  - Helsinki finite transducer, Odense disambiguator, ...
  - Publically funded lexica become accessible (Finland, Norway, ...)
- Lg tech might be run as academic projects for vanishing languages
  - For lingustics, languages with few speakers are as interesting as languages with many speakers
  - Even more so: Languages where you may be a pioneer may be more attractive

Language technology as language documentation

- When languages are about to vanish, we want documentation
- It is not obvious that we should make transducers etc.
- but...
  - lexicographical work should be conducted in a structured way
  - if corpora are available, they could be annotated by a parser
  - a transducer may check the validity of the rules of the reference grammar
- ... so the researcher and the language community have common interests

A new paradigm for linguistic work

- New way of doing linguistics within Academia
  - Projects share sources openly:
    - lexica, grammatical rules, infrastructure
  - File sharing via version control systems
  - Open documentation pages
    - documentation via wikis, (you may contribute)
  - Academic computational linguists actually want their stuff to work on a realistic scale



We're in the 2011 **Google Summer of Code**! Check out the ideas page! Translate the flyer!

Main Page

Installation • **Resources** • **Contact** • **Documentation** • **Development** • **Tools**

Home Page • Bugs • **Wiki** • Chat

Apertium is a free/open-source platform for developing rule-based machine translation systems!

Translators

The following 27 pairs have released versions and are considered to be stable:

- Spanish ⇆ Catalan (es-ca)
- Spanish ← Romanian (es-ro)
- French ⇆ Catalan (fr-ca)
- Occitan ⇆ Catalan (oc-ca)
- English ⇆ Galician (en-gl)
- Swedish → Danish (sv-da)
- Macedonian → English (mk-en)
- Occitan ⇆ Spanish (oc-es)
- Spanish ⇆ Portuguese (es-pt)
- English ⇆ Catalan (en-ca)
- English ⇆ Spanish (en-es)
- English ⇆ Esperanto (en-eo)
- Spanish ⇆ Asturian (es-ast)
- Catalan ← Italian (ca-it)
- Spanish ⇆ Galician (es-gl)
- French ⇆ Spanish (fr-es)
- Esperanto ← Spanish (eo-es)
- Welsh → English (cy-en)
- Breton → French (br-fr)
- Icelandic → English (is-en)
- Esperanto ← Catalan (eo-ca)
- Portuguese ⇆ Catalan (pt-ca)
- Portuguese ⇆ Galician (pt-gl)
- Basque → Spanish (eu-es)
- Norwegian Nynorsk ⇆ Bokmål (nn-nb)
- Macedonian ⇆ Bulgarian (mk-bg)
- Esperanto ← French (eo-fr)
- English ← Arabic (en-ar)

Other pairs currently in development may be found in our SVN repository. You can test drive release versions here, and all versions (including versions for unreleased language pairs) here — this has the latest SVN versions (up to 6 hours old). More information can be found on the page: List of language pairs

Documentation

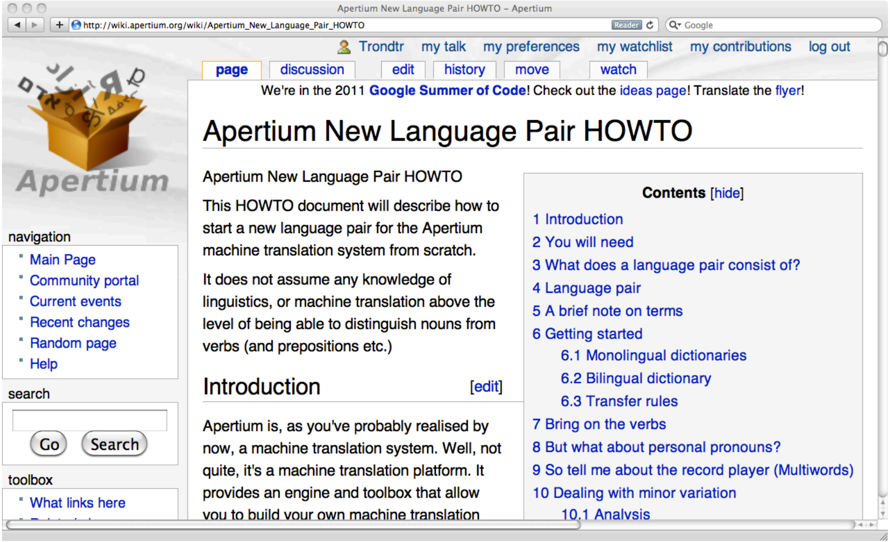
Main article: Documentation

- **Using SVN** — cheatsheet on how to use SVN for users and developers.
- **Apertium New Language Pair HOWTO** — step-by-step description

Miscellaneous

- List of dictionaries — total list of all the morphological dictionaries (analysers) available in the project.
- Translation quality statistics — Details of error rate for each translator.

../../img/logoWeb070sh



We're in the 2011 **Google Summer of Code**! Check out the ideas page! Translate the flyer!

Apertium New Language Pair HOWTO

Apertium New Language Pair HOWTO

This HOWTO document will describe how to start a new language pair for the Apertium machine translation system from scratch.

It does not assume any knowledge of linguistics, or machine translation above the level of being able to distinguish nouns from verbs (and prepositions etc.)

Contents [hide]

Introduction [edit]

Apertium is, as you've probably realised by now, a machine translation system. Well, not quite, it's a machine translation platform. It provides an engine and toolbox that allow you to build your own machine translation

../../img/logoWeb070sh

Status quo for Apertium

- 147 language pairs
- 93 languages
- 27 stable language pairs

Challenges for Academia

- Open source: Share what you do (not only the article, but the ground material)
- Cooperation: Work in teams
  - Learn from programmers: track your work, document what you do
  - Compose teams with mixed backgrounds

Conclusion

- There is now a will, and a way, to provide languages with necessary infrastructure

Conclusion

- There is now a will, and a way, to provide languages with necessary infrastructure
- Better grammatical methods make our analysers robust, and interesting both for linguists and the language communities

Conclusion

- There is now a will, and a way, to provide languages with necessary infrastructure
- Better grammatical methods make our analysers robust, and interesting both for linguists and the language communities
- Without these resources in place, the freedom of choosing the language of your desire remains an illusion

Conclusion

- There is now a will, and a way, to provide languages with necessary infrastructure
- Better grammatical methods make our analysers robust, and interesting both for linguists and the language communities
- Without these resources in place, the freedom of choosing the language of your desire remains an illusion
- The message to sociolinguistics: Remember the material base for linguistic practice

thank you!