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
Freedom of choice?

Written language

Read
To read you need the letters of the language

(no arms, no cake)
Freedom of choice?

Written language

Read

ASCII
Freedom of choice?

Written language

Read

Latin 1
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
    - Transliteration alphabets, Braille, ...
    - Mathematical symbols, chess symbols, ...

- Unicode is what makes it possible to publish text in all languages
Freedom of choice?

Written language

Read

Latin A
Freedom of choice?
- Written language
- Read

<table>
<thead>
<tr>
<th>Code</th>
<th>Language</th>
<th>Code</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>000001760</td>
<td>Tagbanwa</td>
<td>000001780</td>
<td>Khmer</td>
</tr>
<tr>
<td>000001800</td>
<td>Mongolsk</td>
<td>000001800</td>
<td>Mongolsk</td>
</tr>
<tr>
<td>00001900</td>
<td>Limbu</td>
<td>00001910</td>
<td>Tai Le</td>
</tr>
<tr>
<td>00001980</td>
<td>Ny tai lue</td>
<td>00001980</td>
<td>Khmer-symbolet</td>
</tr>
<tr>
<td>00001980</td>
<td>Khmer-symbolet</td>
<td>00001980</td>
<td>Khmer-symbolet</td>
</tr>
<tr>
<td>00001980</td>
<td>Buginesisk</td>
<td>00001980</td>
<td>Buginesisk</td>
</tr>
<tr>
<td>00001980</td>
<td>Mongolsk</td>
<td>00001980</td>
<td>Mongolsk</td>
</tr>
</tbody>
</table>

| 1800   |                |       |                |
| 1810   |                |       |                |
| 1820   |                |       |                |
| 1830   |                |       |                |
| 1840   |                |       |                |

Tegninformasjon
Beslekte tegn:

Navn: MONGOLIAN BIRGA
Unicode: 1800
UTF8: E1 A0 80
Freedom of choice?

Written language

Read
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 čďŋšž
- 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, é, é, í, í, ó, ò, ò, ó, ò, ò, ş, ú, û, ä, à, é, è, é, é, é, í, í, ó, ò, ò, ò, ó, ò, ò, ò, ş, ú, û
South-West Africa: Click letters

<table>
<thead>
<tr>
<th>Uni</th>
<th>Hex</th>
<th>Aux</th>
<th>Hex</th>
<th>Btu</th>
<th>Click type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ø</td>
<td>x0298</td>
<td>Ø</td>
<td>x00D8</td>
<td></td>
<td>bilabial</td>
</tr>
<tr>
<td></td>
<td>x01C0</td>
<td></td>
<td>x007C</td>
<td>c</td>
<td>dental</td>
</tr>
<tr>
<td></td>
<td>x01C1</td>
<td></td>
<td>x007C</td>
<td>x</td>
<td>lateral</td>
</tr>
<tr>
<td></td>
<td>x01C2</td>
<td>+</td>
<td>x002B</td>
<td></td>
<td>alveolar</td>
</tr>
<tr>
<td></td>
<td>x01C3</td>
<td>!</td>
<td>x0021</td>
<td>q</td>
<td>retroflex</td>
</tr>
</tbody>
</table>
Should one use the click letters?

- Arguments for these click letters
  - They are already in use (conservativeness)
  - 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 English WORDS with letters and numbers like 1980
- These are English WORDS with letters and numbers like 1980
Why letters and not punctuation marks?

Text with letters
- Tsií maátsekám llóakas hòásàp ke $\approx$xam xam-à !árop !naa +'oá tsií l''iip tì laísipà síí kèrè Inoóku náú lúrún l'xáa. (...) Tsií maá tsëes híf'ap kèrè 'óa-lxíí tān tsiís kxáó!áa 'oos ke ll''iip tì lluusà kèrè koápi "tíí 'óátse! lóm !nórótse! xápú kxáótse! lóm llxáítse! 'áore kxòetse!" tí.

Text with punctuation marks used as letters
- Tsií maátsekám ||óakas hòásàp ke $\approx$xam xam-à !árop !naa +'oá tsií ||'iip tì ||aisipà síí kèrè ||noóku náú ||úrún ||xáa.(...) Tsií maá tsëes híf'ap kèrè 'óa-|lxíí tān tsiís kxáó!áa 'oos ke ||'iip tì ||luusà kèrè koápi "tíí 'óátse! |óm !nórótse! xápú kxáótse! |óm ||xáítse! 'áore kxòetse!" tí.

(cf. separate doc.)
Freedom of choice?

Read

Write
Freedom of choice?

Written language

Write

Keyboards

For Nama, we need a keyboard to write \( \pm, !, \) not \(+, !\)
Keyboards

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

<table>
<thead>
<tr>
<th>OS</th>
<th>keyboard</th>
<th>GUI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Windows XP</td>
<td>51</td>
<td>33</td>
</tr>
<tr>
<td>Mac OS X</td>
<td>78*</td>
<td>-</td>
</tr>
<tr>
<td>Linux KDE</td>
<td>-</td>
<td>88</td>
</tr>
</tbody>
</table>
Freedom of choice?

- Written language
- Write

Language keyboards out-of-the-box

<table>
<thead>
<tr>
<th>Rank</th>
<th>Speakers</th>
<th>Name</th>
<th>Country</th>
<th>Rank</th>
<th>Speakers</th>
<th>Name</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>26</td>
<td>41.0</td>
<td>Bhojpuri</td>
<td>India</td>
<td>2108</td>
<td>0.014</td>
<td>Inuktitut</td>
<td>Canada</td>
</tr>
<tr>
<td>33</td>
<td>30.0</td>
<td>Siraki</td>
<td>Pakistan</td>
<td>1971</td>
<td>0.017</td>
<td>North Sámi</td>
<td>Nordic</td>
</tr>
<tr>
<td>35</td>
<td>24.0</td>
<td>Maithili</td>
<td>India</td>
<td>1752</td>
<td>0.022</td>
<td>Cherokee</td>
<td>USA</td>
</tr>
<tr>
<td>37</td>
<td>23.0</td>
<td>Oriya</td>
<td>India</td>
<td>1344</td>
<td>0.047</td>
<td>Greenlandic</td>
<td>Greenland</td>
</tr>
<tr>
<td>39</td>
<td>22.0</td>
<td>Burmese</td>
<td>Myanmar</td>
<td>1343</td>
<td>0.047</td>
<td>Faroese</td>
<td>Denmark</td>
</tr>
<tr>
<td>40</td>
<td>22.0</td>
<td>Hausa</td>
<td>Nigeria</td>
<td>1304</td>
<td>0.050</td>
<td>Maori</td>
<td>NZ</td>
</tr>
<tr>
<td>44</td>
<td>20.3</td>
<td>Awadhi</td>
<td>India</td>
<td>991</td>
<td>0.940</td>
<td>Gaelic</td>
<td>Scotland</td>
</tr>
<tr>
<td>47</td>
<td>20.0</td>
<td>Yoruba</td>
<td>Nigeria</td>
<td>601</td>
<td>0.250</td>
<td>Icelandic</td>
<td>Iceland</td>
</tr>
<tr>
<td>51</td>
<td>17.0</td>
<td>Sindhi</td>
<td>Pakistan</td>
<td>517</td>
<td>0.330</td>
<td>Maltese</td>
<td>Malta</td>
</tr>
<tr>
<td>53</td>
<td>16.0</td>
<td>Nepali</td>
<td>Nepal</td>
<td>407</td>
<td>0.500</td>
<td>Breton</td>
<td>France</td>
</tr>
<tr>
<td>55</td>
<td>15.0</td>
<td>Amharic</td>
<td>Ethiopia</td>
<td>370</td>
<td>0.580</td>
<td>Welsh</td>
<td>UK</td>
</tr>
<tr>
<td>59</td>
<td>13.7</td>
<td>Assamese</td>
<td>India</td>
<td>292</td>
<td>0.910</td>
<td>Basque</td>
<td>Spain</td>
</tr>
<tr>
<td>60</td>
<td>13.0</td>
<td>Haryanvi</td>
<td>India</td>
<td>130</td>
<td>4.000</td>
<td>Georgian</td>
<td>Georgia</td>
</tr>
</tbody>
</table>
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
Freedom of choice?
- Written language
- Write

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
Freedom of choice?

- Written language
- Write

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..
Freedom of choice?

- Written language
- Write

Northern Sámi keyboard for Macintosh
Survey of non-Russian letters in some Uralic languages of Russia

<table>
<thead>
<tr>
<th>Language</th>
<th>Ltr</th>
<th>Rank</th>
<th>%</th>
<th>Ltr</th>
<th>Rank</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Komi Permyak</td>
<td>ö</td>
<td>3</td>
<td>6.92</td>
<td>ъ</td>
<td>34</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>i</td>
<td>22</td>
<td>1.32</td>
<td>щ</td>
<td>35</td>
<td>0.02</td>
</tr>
<tr>
<td>Komi</td>
<td>ö</td>
<td>4</td>
<td>6.38</td>
<td>щ</td>
<td>34</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>i</td>
<td>27</td>
<td>0.81</td>
<td>ц</td>
<td>33</td>
<td>0.16</td>
</tr>
<tr>
<td>Meadow Mari</td>
<td>ў</td>
<td>22</td>
<td>1.12</td>
<td>щ</td>
<td>35</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>ö</td>
<td>24</td>
<td>0.61</td>
<td>ъ</td>
<td>34</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>н</td>
<td>26</td>
<td>0.52</td>
<td>ф</td>
<td>33</td>
<td>0.09</td>
</tr>
<tr>
<td>Hill Mari</td>
<td>ў</td>
<td>2</td>
<td>6.84</td>
<td>щ</td>
<td>35</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>ä</td>
<td>11</td>
<td>4.37</td>
<td>ю</td>
<td>34</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>ў</td>
<td>31</td>
<td>0.29</td>
<td>ф</td>
<td>32</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>ö</td>
<td>33</td>
<td>0.16</td>
<td>б</td>
<td>30</td>
<td>0.35</td>
</tr>
</tbody>
</table>
Keyboard for Komi
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
Freedom of choice?
---
Written language
  Write

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
Freedom of choice?

Written language

Write

And status quo?

- Komi
  - Search-and-replace in MS Word
  - ... but often with Latin i, ö, not with Cyrillic i, ö

- Udmurt
  - 25 ё, 30 ö, 34 ъ, 36 ё, 37 ž (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
Freedom of choice?
Language technology
Overview

"<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 Pl1 Gen
  "sabet" N Pl1 Acc
"< >"
Freedom of choice?

Language technology

Overview

"<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
Freedom of choice?
Language technology
Overview

"<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
"<sabehiiday>"
  "sabet" N Pl Acc @<OBJ #6->3
"<.>"
  "." CLB #7->2
Summing up with an overview over language technology activity
The languages found on aclWiki

<table>
<thead>
<tr>
<th>Afrikaans</th>
<th>English</th>
<th>Icelandic</th>
<th>Norwegian</th>
<th>Slovenian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Albanian</td>
<td>Estonian</td>
<td>Iranian</td>
<td>Navajo</td>
<td>Sorbian</td>
</tr>
<tr>
<td>Amharic</td>
<td>Faroese</td>
<td>Irish</td>
<td>Occitan</td>
<td>Spanish</td>
</tr>
<tr>
<td>Arabic</td>
<td>Finnish</td>
<td>Italian</td>
<td>Persian</td>
<td>Swahili</td>
</tr>
<tr>
<td>Basque</td>
<td>French</td>
<td>Iñupiaq</td>
<td>Polish</td>
<td>Swedish</td>
</tr>
<tr>
<td>Breton</td>
<td>Galician</td>
<td>Japanese</td>
<td>Portugese</td>
<td>Sámi</td>
</tr>
<tr>
<td>Bulgarian</td>
<td>German</td>
<td>Komi</td>
<td>Punjabi</td>
<td>Tajik</td>
</tr>
<tr>
<td>Catalan</td>
<td>Greek</td>
<td>Korean</td>
<td>Quechua</td>
<td>Turkish</td>
</tr>
<tr>
<td>Chinese</td>
<td>Greenlandic</td>
<td>Lithuanian</td>
<td>Russian</td>
<td>Tigrinya</td>
</tr>
<tr>
<td>Croatian</td>
<td>Haitian</td>
<td>Macedonian</td>
<td>Sanskrit</td>
<td>Ukrainian</td>
</tr>
<tr>
<td>Czech</td>
<td>Hebrew</td>
<td>Malay</td>
<td>Serbian</td>
<td>Urdu</td>
</tr>
<tr>
<td>Danish</td>
<td>Hindi</td>
<td>Montenegrin</td>
<td>Slovak</td>
<td>Vietnamese</td>
</tr>
<tr>
<td>Dutch</td>
<td></td>
<td></td>
<td></td>
<td>Welsh</td>
</tr>
</tbody>
</table>
The languages found on aclWiki

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>State-level official lgs</td>
<td>46</td>
</tr>
<tr>
<td>Regional-level official</td>
<td>18</td>
</tr>
<tr>
<td>lgs</td>
<td></td>
</tr>
<tr>
<td>Languages with no official status</td>
<td>0</td>
</tr>
</tbody>
</table>

Native American languages  Iñupiaq
African languages  Amharic, Swahili
Asian languages  Chinese, Hindi, Japanese, Korean, Malay, Pashto, Persian, Sanskrit
### Language technology

#### Overview

**Largest languages not found on aclWiki**

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

- 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
Freedom of choice?

- Language technology
- Machine translation and multilingualism
Freedom of choice?

Language technology

Machine translation and multilingualism
Freedom of choice?

Language technology

Machine translation and multilingualism
Freedom of choice?

Language technology

Machine translation and multilingualism
MT in the north

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

Language technology

Machine translation and multilingualism

Gurutbellodaga Marie Fangel áigu bivdit Romssa ovdagott geassádit olles ohcanproseassas searvat sámi giellahálldašanguvluí mañemus áiggiid rieja geažil medain. Dát hirpmástuhttá Ruksesbellodaga mii oaiivilda Gurutbellodaga leat "borjasteame behtolaš leavggain".

Romssa suohkanstivraáiíres Gurutbellodaga ovdas, Marie Fangel, lohká iTromsø aviisii ahde digaštallan ja nággu šilttait birra sámi báikenamaigui mañemus áiggiid, lea šaddan váívves áššin mii lea dagahan vuostálasvuodaid sápmelaččait ja dážaid gaskka. Danin son áigu čohkket opposišuvnna ja evttōhit ovdagodečoahkkimis mánnodaga ahde bissēhit ohcanproseassas searvat sámi giellahálldašanguvluí.

Ovdal maid eahpidan
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 upålidelige flaggene". Tromsøs kommunestyrerepresentant for Venstre, Marie *Fangel, sier *ITromsø til avisen at debatten og krangelen *siltaid 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
Freedom of choice?

Language technology

Machine translation and multilingualism

Giellatekno Jorgalanreaiddut


Jorgal Davvisámegielas girjedárogíllii

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 Rødpartiet 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
Freedom of choice?

Language technology

Machine translation and multilingualism


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
Freedom of choice?
- Language technology
- Machine translation and multilingualism

Token-type ratio

- Haitian Creole
- English
- Danish
- Swedish
- Icelandic
- Finnish
- North Sámi
- Turkish
- Lule Sámi
- Xhosa
- Greenlandic
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 languages

- More basic tools become open source
Bridging the gap: Language technology for minority languages

- More basic tools become open source
  - Helsinki finite transducer, Odense disambiguator, ...
Bridging the gap: Language technology for minority languages

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

For linguistics, 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.
Bridging the gap: Language technology for minority languages

- 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 languages

- 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 linguistics, languages with few speakers are as interesting as languages with many speakers
Bridging the gap: Language technology for minority languages

- 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 linguistics, 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
Freedom of choice?

- Language technology
- A new paradigm for linguistic work
Freedom of choice?
Language technology
A new paradigm for linguistic work

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
Freedom of choice?

Conclusion

thank you!