Linguistic resources for the languages of the world

Lars Borin

Språkbanken, Dept. of Swedish Language, University of Gothenburg

GF summer school 26 August, 2009
Provide some background information as well as a broader picture of some of the issues involved in developing language technology for the world’s languages, especially lower-density languages.
overview

- linguistic demography and language technology (LT):
  - spoken, signed and written languages
  - lower-density languages
- sociology of language and LT
- linguistic resource building for lower-density languages
- strategic considerations
- conclusion
Introduction

lg demography and LT

lg modalities LoDL

Sociology of lg and LT

LoDL LR creation

Strategic considerations

Conclusion

linguistic demography and LT
language statistics

- 5–7,000 living languages
- *Ethnologue* lists almost 7,000, but actual number unknowable
- (first-language speakers of) top 30 languages account for more than 60% of world population
There are close to 7,000 languages in the world, and half of them have fewer than 7,000 speakers each, less than a village. What is more, 80% of the world's languages have fewer than 100,000 speakers, the size of a small town.

(Nicholas Ostler)
the world’s linguistic diversity is under threat

according to an estimate by linguist Michael Krauss, half of the languages spoken today will be gone by the end of this century

on average, the last speaker of some language dies every two weeks
Globalization and ICT

- Some see globalization as the main threat to linguistic diversity,
- With modern ICT – information and communication technologies – one of its chief instruments (television is “cultural nerve gas”, according to Michael Krauss)
- Others see in ICT – especially the computer and internet – a potential means of reversing or at least slowing down language attrition and extinction
- Recall that language technology is a kind of ICT
spoken, signed and written languages

- the modalities of naturally occurring language are
  - speech
  - sign
  - writing
- the *Ethnologue* lists 6,909 living languages
- out of these, 126 are sign languages
- I will have nothing further to say about sign languages here, apart from noting that there is some LT work on them reported in the literature
how many written languages are there?

- difficult to find solid estimates of number of written languages
- the *Ethnologue* lists has a “Script” entry for 2844 languages ("Roman script", "Arabic script", etc.)
- an additional 372 languages have Bible/NT translations, but no script information
- (script and translations:
  - "(portions of the) Bible": 1090
  - "NT": 1080
- )
orthography vs. writing tradition

- half the world’s languages have an orthography/script
- but how many have a tradition of writing?
- linguists and missionaries (often the same people) have for centuries been devising orthographies for unwritten languages in order to translate the Bible and other religious works
- the existence of an orthography for a language does not automatically mean that the speakers use the orthography on a regular basis, or even that they are literate
- instead, writing may be used as a crutch for memory in oral presentation, rather than a means of communication
“fully developed” languages

- In the *Ethnologue* a “fully developed” language is one for which
- “extensive literature and media exist”
- Only 62 languages are “fully developed”
- E.g., Basque, Faroese, Macedonian and Welsh are European languages missing from this list
Languages with a writing tradition

▸ over 3,000 is too high a figure
▸ but 62 is far too low
▸ a generous ballpark estimate would be that no more than 15–20% of the world’s languages have a tradition of writing, i.e., on the order of a thousand languages, give or take a few hundred
why is this relevant?

- the most mature and sophisticated language technology is in effect written language technology
- most proposed applications presuppose a (standardized) written language
- we work with texts, rather than speech
- even much of the speech technology that is being developed in the field is geared toward the written language (speech-to-text and text-to-speech systems)
lower-density languages

- related, but separate issue:
- how much language resources and LT tools exist for a language?
- the expression “density” introduced to LT by LDC in connection with the DARPA TIDES surprise language exercise in 2003
- high-, medium- and low-density (or lower-density) languages
The LDC LoDL survey

- languages with at least a million native speakers (some 300 languages), excluding a few known high-density languages
- surveyed w.r.t. a list of criteria – prerequisites for resource creation and low-level resources
- notably, the density scale applicable only to written languages
- but in principle orthogonal to size of language
LDC LoDL criteria: writing

- Language written
- Words separated in writing
- Simple orthography
- Sentence punctuation
Introduction

LDC LoDL criteria: non-digital resources

- Dictionary
- Newspaper
- Bible
LDC LoDL criteria: digital resources

- Standard digital encoding
- 100 kW news text
- 10 kW translation dictionary
- 100 kW parallel text
- Simple morphology
- Morphological analyzer
the LDC LoDL survey report

<table>
<thead>
<tr>
<th>Language</th>
<th>Country</th>
<th>Speakers</th>
<th>Written</th>
<th>Sentence</th>
<th>Punctuation</th>
<th>Words separated</th>
<th>News</th>
<th>Text</th>
<th>Newspaper</th>
<th>Parallel</th>
<th>Text</th>
<th>Bifile</th>
<th>Xtm Dictionary</th>
<th>Morphology</th>
<th>Morph. Analyser</th>
<th>Sort order</th>
</tr>
</thead>
<tbody>
<tr>
<td>Farsi Western</td>
<td>Iran</td>
<td>24,930,000</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>1161100000</td>
<td></td>
</tr>
<tr>
<td>Finnish</td>
<td>Finland</td>
<td>6,000,000</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>1160550000</td>
<td></td>
</tr>
<tr>
<td>Fon GB</td>
<td>Benin</td>
<td>1,436,000</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>1100000000</td>
<td></td>
</tr>
<tr>
<td>Fulfulde_Maasina</td>
<td>Mali</td>
<td>1,168,000</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>1100000000</td>
<td></td>
</tr>
<tr>
<td>Fulfulde_Nigerian</td>
<td>Nigeria</td>
<td>7,611,000</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>1100000000</td>
<td></td>
</tr>
<tr>
<td>Gulua Jalon</td>
<td>Guinea</td>
<td>2,900,000</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>1000000000</td>
<td></td>
</tr>
<tr>
<td>Galiian</td>
<td>Spain</td>
<td>4,000,000</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>1000000000</td>
<td></td>
</tr>
<tr>
<td>Gamto Gofa Dawro</td>
<td>Ethiopia</td>
<td>1,236,637</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>1000000000</td>
<td></td>
</tr>
<tr>
<td>Ganda</td>
<td>Uganda</td>
<td>3,025,000</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>1100000000</td>
<td></td>
</tr>
<tr>
<td>Garhwali</td>
<td>India</td>
<td>2,081,768</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>1000000000</td>
<td></td>
</tr>
<tr>
<td>Georgian</td>
<td>Georgia</td>
<td>4,105,000</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>1100000000</td>
<td></td>
</tr>
<tr>
<td>Bikuyu</td>
<td>Kenya</td>
<td>5,347,000</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>1000000000</td>
<td></td>
</tr>
<tr>
<td>Guají</td>
<td>Colombia</td>
<td>3,325,000</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>1000000000</td>
<td></td>
</tr>
<tr>
<td>Socio</td>
<td>Tanzania</td>
<td>1,300,000</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>1100050000</td>
<td></td>
</tr>
<tr>
<td>Greek</td>
<td>Greece</td>
<td>12,000,000</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>1161100000</td>
<td></td>
</tr>
<tr>
<td>Guarani Paraguay</td>
<td>Paraguay</td>
<td>4,848,000</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>1100000000</td>
<td></td>
</tr>
<tr>
<td>Guarani</td>
<td>India</td>
<td>44,000,000</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>1160500000</td>
<td></td>
</tr>
<tr>
<td>Gusi</td>
<td>Kenya</td>
<td>1,682,000</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>1000000000</td>
<td></td>
</tr>
<tr>
<td>Haitian Creole French</td>
<td>Haiti</td>
<td>7,920,000</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>1000000000</td>
<td></td>
</tr>
<tr>
<td>Haryanní</td>
<td>India</td>
<td>13,000,000</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>1000000000</td>
<td></td>
</tr>
<tr>
<td>Hasseniyya</td>
<td>Mauritania</td>
<td>2,311,000</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>1000000000</td>
<td></td>
</tr>
<tr>
<td>Hausa</td>
<td>Nigeria</td>
<td>24,200,000</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>1161100000</td>
<td></td>
</tr>
<tr>
<td>Hayá</td>
<td>Tanzania</td>
<td>1,200,000</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>1000000000</td>
<td></td>
</tr>
<tr>
<td>Hazaragi</td>
<td>Afghanistan</td>
<td>1,756,000</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>1000000000</td>
<td></td>
</tr>
<tr>
<td>Hebrew</td>
<td>Israel</td>
<td>4,612,000</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>1161050000</td>
<td></td>
</tr>
<tr>
<td>Hiligaynon</td>
<td>Philippines</td>
<td>7,000,000</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>1000000000</td>
<td></td>
</tr>
<tr>
<td>Hindi</td>
<td>India</td>
<td>18,000,000</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>1161100000</td>
<td></td>
</tr>
<tr>
<td>Hindo Northern</td>
<td>Pakistan</td>
<td>1,875,000</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>1000000000</td>
<td></td>
</tr>
<tr>
<td>Hmong Njua</td>
<td>China</td>
<td>1,275,000</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>1000000000</td>
<td></td>
</tr>
<tr>
<td>Ho</td>
<td>India</td>
<td>1,025,000</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>1000000000</td>
<td></td>
</tr>
<tr>
<td>Hungarian</td>
<td>Hungary</td>
<td>14,500,000</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>1161100000</td>
<td></td>
</tr>
<tr>
<td>Ibibi</td>
<td>Nigeria</td>
<td>1,500,000</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>1161050000</td>
<td></td>
</tr>
<tr>
<td>Igbo</td>
<td>Nigeria</td>
<td>17,000,000</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>1161050000</td>
<td></td>
</tr>
<tr>
<td>Ilocano</td>
<td>Philippines</td>
<td>8,000,000</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>1165500000</td>
<td></td>
</tr>
<tr>
<td>Indian Sign Language</td>
<td>India</td>
<td>1,600,000</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Indonesian</td>
<td>Indonesia</td>
<td>17,050,000</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>1161060000</td>
<td></td>
</tr>
<tr>
<td>Italian</td>
<td>Italy</td>
<td>37,000,000</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>1110010000</td>
<td></td>
</tr>
</tbody>
</table>
sociology of language and LT

- language (non-)use mainly determined by attitudes
- languages are more or less prestigious, have higher or lower status
- linguistic inferiority complexes seem to be common in the world
speaking with forked tongue

In situations of language shift, we often observe a pattern of parents speaking a more prestigious language to their children at home, rather than their first, less prestigious, language, even while paying lip service to the need for preserving the lower-status language, because they are grappling with

(...) a conflict between wanting to do something for the language and wanting to improve the chances of the children to succeed in the macrosociety of which they are, and always will be, part. The linguist observing this state of affairs may feel regret at what is happening here; but if it is a fact that maintaining a small language at the expense of a major or national one means severely reducing prospects of an economically satisfactory life for one’s children, does one have a right to blame the parents? (Werner Winter)
language status and LT

- status is not an inherent and immutable characteristic of a language
- rather, it is something that lies in the eye of the beholder
- importantly for us, it has been suggested that the creation of linguistic resources and language technology for a language may serve to raise its status
linguistic resources – turtles all the way up
approaches to LR creation

- much recent interest in LR creation for lower-density languages
- substantial results mainly by grammar-based approaches, which are labor- and knowledge-intensive (case in point: North and Lule Sámi)
- pure data-driven approaches tend to be small-scale proof-of-concept experiments, but this is a fast-moving field
- surprise language exercise teams made good progress in short time (using very eclectic methodology), but still quite a bit short of state-of-the-art performance
- collaborative voluntary efforts à la Wikipedia are emerging
untouched by human hands?

- If we want guaranteed results, there is still no way of avoiding good old-fashioned linguistics entirely.
- In this case, it is important that tools for providing systems with linguistic knowledge use a conceptual apparatus and notation familiar to the linguists who are supposed to be working with them.
- In some cases one may get away with more naive approaches provided that the interaction with the user is arranged in a suitable way that compensates for the lack of linguistic knowledge in the system, the paradigm example being web search engines.
Given that we have limited resources – in terms of money, manpower and expertise – and that there is a choice of which resources we could realize within these limitations, how should we set our priorities?
Introduction

In order to survive in the modern world, it is claimed, low-density languages need to establish a presence in the information society.

Increasingly, people use the internet as their main or only source of information and means of communication.

This creates an opportunity for promoting LRs and LT for low-density languages, for concrete practical aims as well as a means of raising the status of these languages.
the vision of the emerging Semantic Web is a global information structure interlinked using logical representations and formal reasoning over these representations

today’s WWW is predominantly textual (and increasingly multilingual)

the question is: By which magical means will the WWW be turned into the Semantic Web?
... to semantic web

- the answer, according to, e.g., Yorick Wilks, is “language technology” – especially information extraction (IE) and related approaches

- thus, those languages for which IE technology will be available, will probably be more visible on the Semantic Web than those lacking such resources, and as a consequence, enjoying the associated status
summing up

- IE technologies will be important to lower-density language communities if they want to carve a niche for their languages and cultures in the information society of the future, ensuring that the world of the Semantic Web remains a linguistically and culturally rich and diverse place

- suitable IE applications will differ according to language, and depend on the kinds of textual material available and produced in a language

- for the near future at least, grammar-based approaches will be indispensable for realizing these applications
thank you for listening!