GÖTEBORGS UNIVERSITET

Språk-BANKEN

Linguistic resources for the languages of the world

Lars Borin

Språkbanken, Dept. of Swedish Language, University of Gothenburg

GF summer school 26 August, 2009

GÖTEBORGS UNIVERSITET

Språk-BANKEN

purpose

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

GÖTEBORGS UNIVERSITET

Språk-BANKEN

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

GÖTEBORGS UNIVERSITET

Språk-BANKEN

linguistic demography and LT

GÖTEBORGS UNIVERSITET

Språk-BANKEN

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

GÖTEBORGS UNIVERSITET

Språk-BANKEN

language community statistics

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)

GÖTEBORGS UNIVERSITET

Språk-BANKEN

language death

- 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

GÖTEBORGS UNIVERSITET

Språk-BANKEN

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

GÖTEBORGS UNIVERSITET

Språk-BANKEN

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

GÖTEBORGS UNIVERSITET

Språk-BANKEN

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
-)

GÖTEBORGS UNIVERSITET

Språk-BANKEN

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

GÖTEBORGS UNIVERSITET

Språk-BANKEN

"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

GÖTEBORGS UNIVERSITET

Språk-BANKEN

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

GÖTEBORGS UNIVERSITET

Språk-BANKEN

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)

GÖTEBORGS UNIVERSITET

Språk-BANKEN

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

GÖTEBORGS UNIVERSITET

Språk-BANKEN

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

GÖTEBORGS UNIVERSITET

Språk-BANKEN

LDC LoDL criteria: writing

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

GÖTEBORGS UNIVERSITET

Språk-BANKEN

LDC LoDL criteria: non-digital resources

- Dictionary
- Newspaper
- Bible

GÖTEBORGS UNIVERSITET

Språk-BANKEN

LDC LoDL criteria: digital resources

- Standard digital encoding
- 100 kW news text
- 10 kW translation dictionary
- 100 kW parallel text
- Simple morphology
- Morphological analyzer

GÖTEBORGS UNIVERSITET

Språk-BANKEN

the LDC LoDL survey report

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Language	Country	# Speakers	Written	Sentence_Punctuation	Words separated	News_Text	Newspaper	Parallel_Text	Bible	Xltn_Dictionary	Dictionary	Morphology	Morph_Analyzer	Sort order		
91	FARSI_WESTERN	Iran	24 280 000	T	T	T	T	T	T	T	T	T	F	T	1161105000		
92	FINNISH	Finland	6 000 000	T	T	T	T	T	F	T	F	T	F	T	1160555000		
93	FON_GBE	Benin	1 436 000	T											1000000000		
94	FULFULDE_MAASINA	Mali	1 168 500	T	T	T	F	F	F	F	F	T		F	1110500000		
95	FULFULDE_NIGERIAN	Nigeria	7 611 000	F											0		
96	FUUTA_JALON	Guinea	2 900 000	T											1000000000		
97	GALICIAN	Spain	4 000 000	T				T							1000000000		
98	GAMO_GOFA_DAWRO	Ethiopia	1 236 637	T										F	1000000000		
99	GANDA	Uganda	3 025 000	T	T	T		T	F	T	F	T	F	F	1110550000		
100	GARHWALI	India	2 081 756	T											1000000000		
101	GEORGIAN	Georgia	4 103 000	T	T	T	T	T	F	Q	T	T	F	F	1161000000		
102	GIKUYU	Kenya	5 347 000	T											1000000000		
103	GILAKI	Iran	3 265 000	F											0		
104	GOGO	Tanzania	1 300 000	T	T	T	F	F	F	T	F	F		F	1110050000		
105	GREEK	Greece	12 000 000	T	T	T	T	T	T	T	T	T	F	T	1161105000		
106	GUARANI_PARAGUAYAN	Paraguay	4 848 000	T	T	T	Q	F	F	Q	F	F	Q	F	1110000000		
107	GUJARATI	India	44 000 000	T	T	T	T	T	T	T	F	T	Q	F	1160600000		
108	GUSII	Kenya	1 582 000	T						Q				F	1000000000		
109	HAITIAN_CREOLE_FRENCH	Haiti	7 372 000	T				T		T			T		1000060000		
110	HARYANVI	India	13 000 000	T			F	F	F	F	F	F		F	1000000000		
111	HASSANIYYA	Mauritania	2 511 000	F						F			F	F	0		
112	HAUSA	Nigeria	24 200 000	T	T	T	T	T	T	T	T	T	F	F	1161100000		
113	HAYA	Tanzania	1 200 000												0		
114	HAZARAGI	Afghanistan	1 756 000	T						F				F	1000000000		
115	HEBREW	Israel	4 612 000	T	T	T	T	T	F	T	T	T	F	T	1161055000		
116	HILIGAYNON	Philippines	7 000 000	T								T			1000500000		
117	HINDI	India	182 000 000	T	T	T	T	T	T	T	T	T	Q	T	1161105000		
118	HINDKO_NORTHERN	Pakistan	1 875 000	T											1000000000		
119	HMONG_NJUA	China	1 245 000	T								T			1000500000		
120	HO	India	1 026 000	T	Q	Q	F	F	F	Q	F	T		F	1000500000		
121	HUNGARIAN	Hungary	14 500 000	T	T	T	T	T	T	T	T	T	F	T	1161105000		
122	IBIBIO	Nigeria	1 500 000	T	T	T						T			1110500000		
123	IGBO	Nigeria	17 000 000	T	T	T	F	T	F	T	F	T	T	F	1110560000		
124	ILOCANO	Philippines	8 000 000	T	T	T	T	F	F	T	F	T	F	F	1160550000		
125	INDIAN_SIGN_LANGUAGE	India	1 500 000	F											0		
126	INDONESIAN	Indonesia	17 050 000	T	T	T	T	T	Q	T	T	T	T	T	1161060000		
127	ITALIAN	Italy	37 000 000	T	T	T		T					T	T	1110010000		

Helskärm

Helskärm

SurveyReport

GÖTEBORGS UNIVERSITET

Språk-BANKEN

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

GÖTEBORGS UNIVERSITET

Språk-BANKEN

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)

GÖTEBORGS UNIVERSITET

Språk-BANKEN

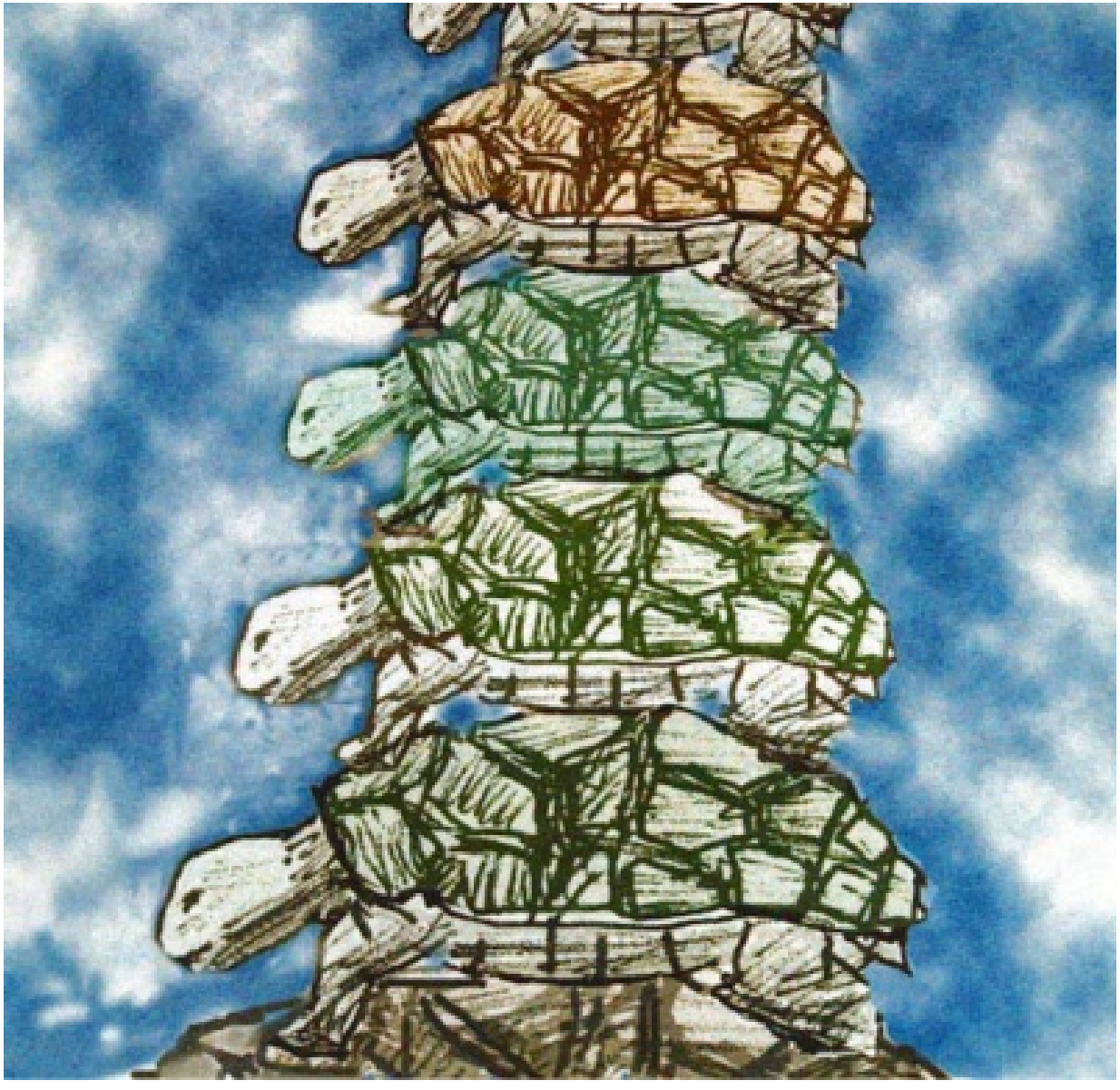
language status and LT

- 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

GÖTEBORGS UNIVERSITET

Språk-BANKEN

linguistic resources – turtles all the way up



GÖTEBORGS UNIVERSITET

Språk-BANKEN

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

GÖTEBORGS UNIVERSITET

Språk-BANKEN

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

GÖTEBORGS UNIVERSITET

Språk-BANKEN

strategic musings

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?

GÖTEBORGS UNIVERSITET

Språk-BANKEN

a foot in the door of the information society

- 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

GÖTEBORGS UNIVERSITET

Språk-BANKEN

from textual web . . .

- 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?

GÖTEBORGS UNIVERSITET

Språk-BANKEN

. . . 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

GÖTEBORGS UNIVERSITET

Språk-BANKEN

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

GÖTEBORGS UNIVERSITET

Språk-BANKEN

thank you for listening!