GF and the Languages of the World

Harald Hammarström
harald2@chalmers.se

August 27, 2009

Today's Talk

- Challenges to the GF-model posed by the diversity of the world's languages
- Outlook for extending the GF grammar library to further languages of the world

Languages of the World: Speakers

- Approximately (excluding ca 100 known uncontacted peoples)

6 783	living spoken
466	extinct (formerly) spoken
126	signed
7 357	total attested

- Most languages are tiny, e.g.

3 648	have less than 5 000 speakers
4 332	have less than 10 000 speakers
4 940	have less than 20 000 speakers

- A select few languages cover most of the planet's speakers

The 5 biggest languages have 1.9 billion (30.8%) speakers in total
The 10 biggest languages have 2.6 billion (42.0%) speakers in total
The 20 biggest languages have 3.3 billion (53.2%) speakers in total
The 100 biggest languages have 4.9 billion (80.7%) speakers in total

Languages of the World: Description

- Most languages are not written
  - 2882 are listed with an entry “Writing System” in Ethnologue 16th ed
  - 2523 are listed as having (a part of) the Bible translation in in Ethnologue 16th ed
  - The union of the is 3 233
  - Own estimate of languages for which the speakers tend to write at all, and if so, in their own language: 500
- Most languages are not described (yet) by linguists

Less than sketch (wordlist, phonology, . . .)	?4 729
Grammar Sketch (typically ca 50pp.)	3 337
Grammar (typically 100pp. or more)	2 215

The numbers refer to published materials or theses, unpublished manuscripts are **not** counted

Some Challenges to the GF-model

- Dividing up the **compositional** semantic space differently across languages

  Difficult to have a common abstract syntax

- Information outside the clause is needed to correctly generate the right one from an **open** set of clauses

  Departs from the (commonly assumed, but not inherent in GF) status of the clause as the basic unit

- Grammar-lexicon dichotomy is problematic even for the most **basic sentences**

  Difficult for many formalisms to deal with in fine granularity

Numeral Systems and Number Bases

Numeral expressions can be analysed as having (one or more) **bases**. English has 10-100-1000-1000000-10^9.

1	kéti
2	moru
3	súba
4	pā́da
5	bī́ya
6	bébènı
7	bḗmodu
8	bḗjiba
9	bḗfada
10	óraga
11	óraga buti kéti
…	…
20	óraga moru
30	óraga súba
…	…
100	áru

Sokoro (Central Chadic/Afroasiatic, Cameroon) has base 5-10-100

Number Bases Across the World

We count statistics as follows:

- For every language *family*, what is its numeral system?

  English and Swedish are counted as one

- If a family has more than one system with independent morphemes, count both separately

  Arabic has a 10-100 systems but the distantly related language Beja has a 5-10-100 system with completely different morphemes. They are counted separately.

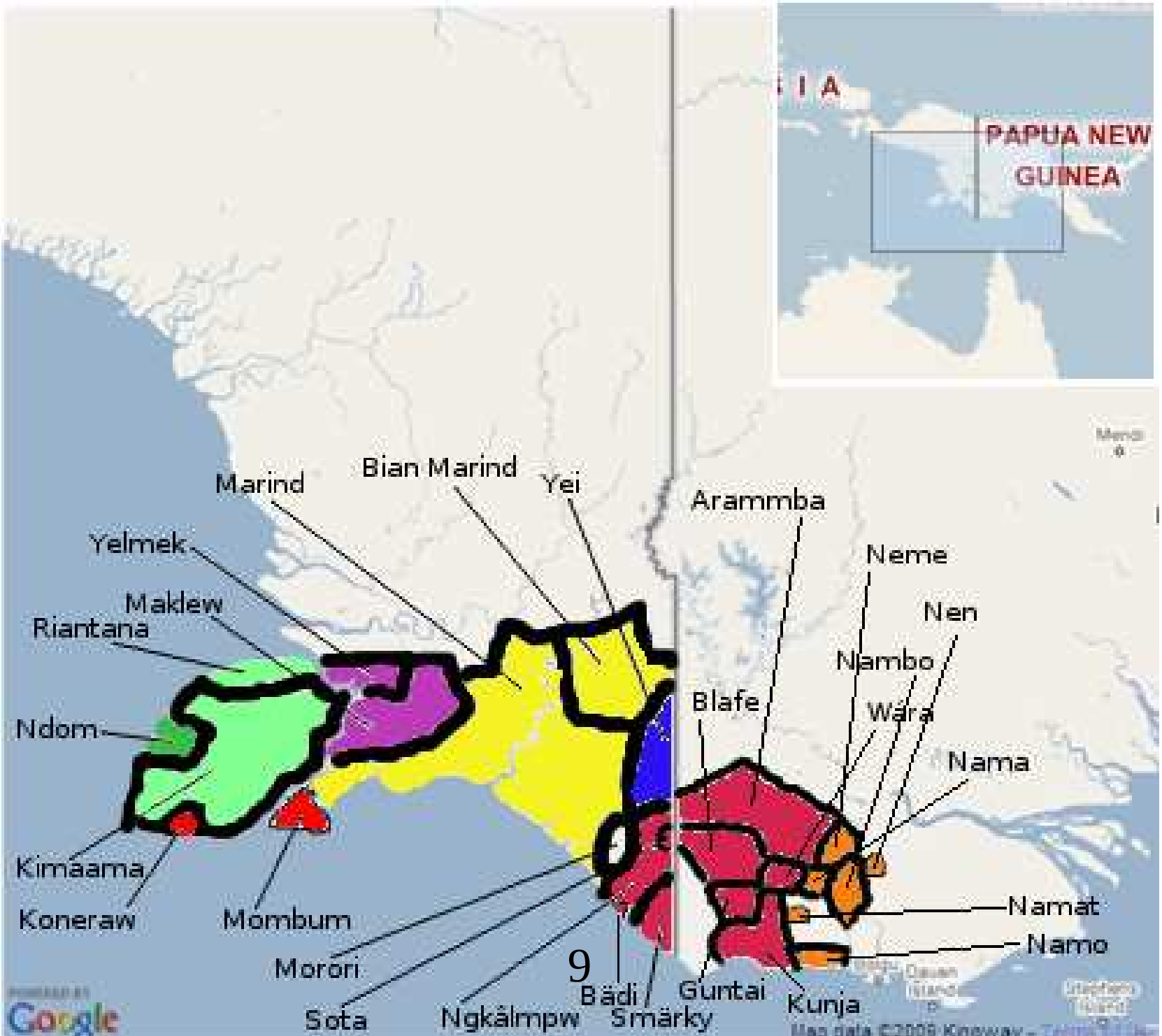
- Ca 400 families and ca 600 independent number systems in the world

Number Bases Across the World

Restricted	174	**29%**	Base-2	6	**1%**	Base-4	5	**0.8%**	Base-5	146	**24.3%**
			2-5-10-20	6	**1%**	4-32	1	**0.2%**	5-10-20	73	**12.2%**
						4-24	1	**0.2%**	5-10-100	61	**10.2%**
						4-20	1	**0.2%**	5-20	12	**2.0%**
						4-16	1	**0.2%**			
						4-10-100	1	**0.2%**			

Base-6	2	**0.3%**	Base-8	1	**0.2%**	Base-10	183	**30.5%**	Base-12	2	**0.3%**
6-36	2	**0.3%**	Base-8	1	**0.2%**	10-100	125	**20.8%**	12-144	2	**0.3%**
						10-20	54	**9.0%**			
						10-X	4	**0.7%**			

- NODATA 78 **13%**
- Total % where 5 is one of the bases 152 **25.3%**
- Total % where 10 is one of the bases 323 **53.8%**
- Total % where 20 is one of the bases 146 **24.3%**


PAPUA NEW GUINEA
Marind
Bian Marind
Yei
Arammba
Yelmek
Neme
Maklew
Nen
Riantana
Nambo
Blafe
Wára
Ndom
Nama
Kimaama
Koneraw
Mombum
Namat
Namo
Morori
Sota
Ngkâlmpw
Bädi
Smärky
Guntai
Kunja
Google

Base-6 forms in Ngkâlmpw Kanum

1	*aempy*	19	*aempy ntamnao*
2	*ynaoaempy*	20	*ynaoaemy ntamnao*
3	*ylla*	24	*wramaekr*
4	*eser*	25	*aempy wramaekr*
5	*tamp*	30	*ptae wramaekr*
6	*ptae*	31	*aempy ptae wramaekr*
7	*aempy ptae*	36	*(ntaop) ptae*
8	*ynaoaempy ptae*	37	*aempy (ntaop) ptae*
9	*ylla ptae*	50	*ynaoaempy tarwmpao (ntaop) ptae*
10	*eser ptae*	100	*eser wramaekr ptae ynaoaempy*
11	*tamp ptae*	216	*(ntaop) tarwmpao*
12	*tarwmpao*	1296	*(ntaop) ntamnao*
13	*aempy tarwmpao*	7776	*(ntaop) wramaekr*
14	*ynaoaempy tarwmpao*		
15	*ylla tarwmpao*		
16	*eser tarwmpao*		
17	*tamp tarwmpao*		
18	*ntamnao*		

Base-6 forms in Ndom

1	*sas*
2	*thef*
3	*ithin*
4	*thonìth*
5	*merègh*
6	*mer*
7	*(mer) abo sas*
8	*(mer) abo thef*
9	*(mer) abo ithin*
10	*(mer) abo thonìth*
11	*(mer) abo merègh*
12	*mer an thef*
13	*mer an thef abo sas*
14	*mer an thef abo thef*
15	*mer an thef abo ithin*
16	*mer an thef abo thonìth*
17	*mer an thef abo merègh*
18	*töndör*
19	*töndör abo sas*
20	*töndör abo thef*
21	*töndör abo ithin*
22	*töndör abo thonìth*
23	*töndör abo merègh*
24	*töndör abo mer*
36	*nif*
72	*nif thef*
108	*nif ithìn*
144	*nif thonìth*
180	*nif merègh*

Building blocks for Base-10 Languages

- A phrase structure grammar would want to look something like this:

```
N -> L10 | L100 | L1000 | L1000000
L10 -> one | two | ... | ten
L100 -> L10 | L10-ten | L10-ten L10
L1000 -> L100 | L10-hundred | L10-hundred L100
L1000000 -> L1000 | L100-thousand | L1000-thousand L1000
```

- That is, if X is a base, easy formation of non-atomic expressions should have the structure:

  some amount of X:s plus *some amount less than X*

Abstract Syntax in GF for Numerals

Ideally, all languages should share the same abstract syntax

- A base-10 backbone abstract syntax much like the CFG on previous slide
- Deviations from fully regular formations are handled by parameters and inherent features

  e.g., if 40 is irregular as compared to 20-30, 50-90, we keep an inherent feature for each expression less than 10, with the meaning "to be four or not" and use it to select the right formation of the tens
- In Hindi, where all of 11-99 are irregular, this same strategy is applied with 9 different values for the inherent feature and a 9x10 table
- For base-20 languages, the inherent-feature strategy is workable since 10 divides 20 and most base-20 languages have 100 as the next higher base
- For base-6 the inherent-feature/parameter boils down to a rote table …

Tense in Malay/Indonesian

- Malay: In Malaysia and adjacent countries, ca 40 million (mostly L2) speakers
- Indonesian: In the Indonesian archipelago, ca 240 million (mostly L2) speakers

  Essentially the same language with two slightly different prescriptive standards

- There is no morphological marking of tense

  Muhammed datang kemarin/hari ini/besok
  Muhammed come yesterday/today/tomorrow
  'Muhammed came/is coming/will come yesterday/today/tomorrow'

- If you want, you can mark tense with a variety of adverbs (see next slide)
- If tense is apparent from the context, it is not typically not marked with any of these adverbs

Marking Tense with an Adverb: #1

Muhammed sudah datang
Muhammed already come
'Muhammed came/has come'

Muhammed pernah datang
Muhammed once come
'Muhammed has come'

Muhammed baru datang
Muhammed new come
'Muhammed just came/has just come' (i.e, very recently)

Muhammed belum datang
Muhammed not.yet come
'Muhammed has not yet come'

Marking Tense with an Adverb: #2

Muhammed sekarang datang
Muhammed now come
'Muhammed is coming (now)'

Muhammed sedang datang
Muhammed now come
'Muhammed is coming (now)'

Muhammed biasa-nya datang
Muhammed habit-3PSG come
'Muhammed usually comes'

Muhammed akan datang
Muhammed will come
'Muhammed will come'

. . .

Use a Tense-marking Adverb?

- When there risk of confusion, use it
- If the time reference is clear from some adverbial phrase in the sentence, don't use it

  Muhammed ?sudah datang kemarin
  Muhammed ?already come yesterday
  'Muhammed came yesterday'

- If the time reference is irrelevant, don't use it

  Muhammed datang selalu
  Muhammed come always
  'Muhammed always comes'

- If the time reference is understood from the context, don't use it

  Apa yang buat Muhammed kemarin? *Dia *sudah tari*
  What REL do Muhammed yesterday He *already dance
  'What did Muhammed do yesterday?' 'He danced'

Translating Indonesian <> English in GF

- Ok: An Indonesian sentence with an adverb corresponds to exactly one English tense

  But every Indonesian adverb-phrase must carry a feature (past/present/future etc) to select the appropriate English one

- Ok: An Indonesian sentence without an adverb corresponds to a finite number of English tenses

  We can live with ambiguity on a finite level (generate all and let application decide)

- Problem: An English sentence corresponds to an unbounded number of Indonesian sentences (because of a variety of adverbial phrases) some of which should be taken out depending on the context

  We cannot generate all and the let application decide

Kalam and Grammar-Lexicon Dichotomy

- Spoken in Papua New Guinea highlands (Trans New Guinea family)
- ca 15 000 speakers
- first contact with the Western world in the 1950s
- Description by A. K. Pawley (ANU), 12 months fieldwork on site 1963-1964, 1965, 1969, 1972 and 1975 + Kalam speakers in Auckland

Kalam Verbs #1

- Closed set of about 130 members
- Corpus 14 000 tokens representing a variety of texts:
  - 10 verbs make up 78.5% of all verb tokens
  - 15 verbs make up 89.6% of all verb tokens
  - 35 verbs make up 97.6% of all verb tokens
- Serial verbs “usually spoken without perceptible internal pause” where e.g. English has lexical verbs
- Typical case

  b=ak am mon p=wk d=ap=ay-a-k
  man=that go wood hit=break get=come=put-3SG-PAST
  ’The man fetched some firewood’

Kalam Verbs #2

- A nine verb “conventional expression”:

  pk *wyk* *d* *ap* *tan* *d* *ap* *yap* *g-*
  strike rub hold come ascend hold come descend do
  ’to massage’

- Kalam is significantly more analytic in reporting everyday actions!

- Pawley wasn’t happy with the choice of composite items to put in the dictionary:

  > The most productive patterns of the language were not distinguished from those patterns that were grammatical but virtually never used.

Example dictionary entry

n- [n-], v. Generic for acts or processes of perception, sensing and cognition. The following is a list of English translation equivalents; it is not clear that all of these are separate senses for native speakers of Kalam. v. i. 1. Be conscious, aware. 2. Be awake. 3. Think, reason. v. tr. 1. Perceive, sense s.th. 2. Know s.th. 3. Understand, comprehend s.th. 4. Take notice of, pay attention to s.th. 5. Realise, become fully aware of a situation, see that s.th. is the case. 6. See (with the eyes). near syn. wdn n- (eye perceive). 7. Look, observe. 8. Hear. near syn. tmd n- (ear perceive). . Listen. 10. Feel. usu. d n- (touch perceive). 11. Smell (an odour). 12. Taste s.th. usu. ñb n- (consume perceive). 13. Think about s.th, have thoughts or opinions. near syn. gos n- (thought perceive). 14. Learn, acquire knowledge or understanding. 15. Discern, discriminate, work out a solution. 16. Be used to, familiar with, have experience of or in something. 17. Believe, be under the impression, think that (something is the case).

d n-	(touch perceive) ’feel by touching (deliberately)’
ñb n-	(consume perceive) ’taste s.th.’
pk n-	(hit perceive) ’feel by touching against, nudge’
pui n-	(pierce perceive) ’probe, test by probing’
wk n-	(burst perceive) ’test by cracking open’
ag n-	(say perceive) ’ask, enquire, ask for, request’
ap n-	(come perceive) ’visit s.o., come and see s.o.’
piow n-	(search perceive) ’find (what one is looking for, search and find’
tag n-	(travel perceive) ’sightsee, travel and see’
taw tag n-	(tread walk.about perceive) ’test (ground, etc.) by treading’
ñn ay n-	(hand put perceive) ’feel inside s.th., grope’
. . .	
gos n-	(thought perceive) ’think’
gos koay n-	(thought many perceive) ’be preoccupied’
gos mket n-	(thought heavy perceive) ’worry, be worried’
wsn n-	(sleep (N.) perceive) ’dream, have a dream’
wsn kab n-	(sleep untrue.dream perceive) ’have a dream that doesn’t come true’
gos tep n-	(thought good perceive) ’approve, like, admire s.th.’
gos tmey n-	(thought bad perceive) ’dislike, hate s.th.’
mapn n-	(liver perceive) ’feel sympathy’
mluk n-	(nose perceive) ’be resentful, feel angry’
wdn n-	(eye perceive) ’see (with one’s own eyes)’
tmd n-	(ear perceive) ’hear (with one’s own ears)’
. . .	

Kalam and GF

If:

- we put all productive rules in the grammar, and,
- the rest in lexicon

Then:

- we get all grammatical Kalam sentences
- some have one-lexeme English equivalents, some not
- but we have no means of distinguishing *idiomatic* from *non-idiomatic* Kalam
  - Idiomatic English: in two days
  - Non-Idiomatic English: the day after the day after tomorrow

  This problem occurs in all language, but in Kalam it occurs also in the most basic sentences!

Growth-Diversity Challenges and GF

- All kinds of things which are in the lexicon of most languages are grammaticalized in *some* language, e.g., frustrative verb ending, alcoholative noun class, nominal tense etc

  GF is not likely to face any of these

- Some things are common to grammaticalize, except in most European languages, e.g., evidentials, clusivity in pronouns, classifiers, obligatory focus marking

  Once GF expands the horizons, the *interlingua* or abstract syntax, will have to make very many distinctions

- Maybe the growth is side-stepped by domain specificity?

Challenges are not the end of GF

- Workaround
- Something I didn’t think of
- Inelegant GF-code solutions
- Live with mildly incorrect linguistic forms
- Nevermind, it’s so rare anyway
- …

GF outlook on the languages of the World

Which languages are strategic to include next?

- Languages that are likely to be used
- Languages with commercial potential
- Languages which are easy to add
  - Which are intrinsically lean to implement (orthography, regular morphology, etc.)
  - Which are similar to already GFed languages
  - Which have interested speakers/experts
- Languages which are not “covered” by someone else?
- Languages for which raw text data is not cheap and abundant?

A Survey of CMR for LDL

- CMR = Computational Morphological Resources
- LDL = Low-Density Languages

Aims to shed light on:

- **Languages that are likely to be used**
- **Languages with commercial potential**
- Languages which are easy to add
  - Which are intrinsically lean to implement (orthography, regular morphology, etc.)
  - Which are similar to already GFed languages
  - **Which have interested speakers/experts**
- **Languages which are not “covered” by someone else?**
- Languages for which raw text data is not cheap and abundant?

Motivation: Morphological Resources

- Morphological analysis is one of the bottom layers of the language resources pyramid
- Layer below, namely raw text data, is addressed already by others

  NOTE: Raw text data appears on the web for many more languages than those for which there is a published description of a morphological analyser.
- My own PhD work focusses on morphology

GLP: Measuring Language Momentum

- There is no point in surveying the richest languages, they inevitably have resources
- Practical definition: momentum of a language = economic power of its speakers
- Gross Language Product (GLP): total market value of all final goods and services produced by the speakers of the language within a calendar year
- Estimate by country averages:

$$GLP(L) = S(L) \cdot GDP\text{-per-capita}(Country(L))$$

- $S(L)$: Number of L1 speakers of L (from Ethnologue)
- $Country(L)$: Principal country of L (from Ethnologue)
- GDP-per-capita(C): GDP per capita for country C from CIA factbook

GLP Comments

- The GDP-figures used are not PPP-adjusted
- Ideally, one would like to count second language speakers along with first language speakers (but such data is not systematically available)
- Low-momentum language *defined* as low GLP
- Thus, whether low-momentum languages actually turn out to have low amounts of NLP infrastructure is an *empirical question*

Defining Low-Momentum Languages

Set threshold of low-momentum at 100 billion dollars of GLP

Chosen because:

- High even number with large number of zeroes
- Convenient number of non-low-momentum languages emerge

#	Language	iso-639-3	GLP	Pop.	GDP-per-capita
1	English	eng	14112019141500	309297750	45626
2	Spanish	spa	11466115307496	322299171	35576
3	Japanese	jpn	4210405702398	122388399	34402
4	German, Standard	deu	3845767908070	95392978	40315
5	Portuguese	por	3723229093580	177457180	20981
6	French	fra	2606363422200	64834911	40200
7	Italian	ita	2207898410784	60989984	36201
8	Chinese, Mandarin	cmn	2146742158782	873014298	2459
9	Russian	rus	2146466954800	145031551	14800
10	Korean	kor	1328565491640	66977490	19836
11	Dutch	nld	805812974253	17370777	46389
12	Turkish	tur	471145207462	50535794	9323
13	Polish	pol	465485547296	42658133	10912
14	Swedish	swe	443139531525	8789835	50415
15	Greek	ell	360217197900	12258540	29385
16	Bavarian	bar	349629329322	7667478	45599
17	Schwyzerdütsch	gsw	339134884000	6044000	56111
18	Lombard	lmo	330654684855	9133855	36201
19	Danish	dan	302298082240	5299756	57040
20	Napoletano-Calabrese	nap	255122891199	7047399	36201
21	Finnish	fin	244729455832	5232728	46769
22	Catalan-Valencian-Balear	cat	237196860928	6667328	35576
23	Czech	ces	197516975282	11525089	17138

#	Language	iso-639-3	GLP	Pop.	GDP-per-capita
24	Chinese, Wu	wuu	189773325000	77175000	2459
25	Hungarian	hun	189214851600	13611600	13901
26	Hindi	hin	175884141643	180764791	973
27	Sicilian	scn	174942056520	4832520	36201
28	Romanian	ron	173246830884	23248367	7452
29	Javanese	jav	139312813500	75508300	1845
30	Chinese, Yue	yue	134779260482	54810598	2459
31	Arabic, Najdi Spoken	ars	134439777600	9863520	13630
32	Malay	mly	132190335777	17604253	7509
33	Ukrainian	ukr	119705990470	39441842	3035
34	Hebrew	heb	117078855000	5055000	23161
35	Chinese, Min Nan	nan	113674565935	46227965	2459
36	Galician	glg	113430518400	3188400	35576
37	Piemontese	pms	112462750620	3106620	36201
38	Chinese, Jinyu	cjy	110655000000	45000000	2459
39	Azerbaijani, South	azb	109564908000	24364000	4497
40	Farsi, Western	pes	109237171137	24291121	4497
41	Tswana	tsn	91581075720	4407174	20780
42	Chinese, Xiang	hsn	88560885000	36015000	2459
43	Kurdish, Northern	kmr	84191398115	9030505	9323
44	Arabic, Algerian Spoken	arq	83227665000	21097000	3945
45	Bengali	ben	82284767162	171070202	481
46	Arabic, Hijazi Spoken	avx	81780000000	6000000	13630

#	Language	iso-639-3	GLP	Pop.	GDP-per-capita
47	Saxon, Upper	sxu	80630000000	2000000	40315
48	Venetian	vec	78932189787	2180387	36201
49	Thai	tha	76388430912	20229987	3776
50	Arabic, Egyptian Spoken	arz	73727112000	46311000	1592
51	Chinese, Hakka	hak	73617441181	29937959	2459
52	Emiliano-Romagnolo	eml	73130074512	2020112	36201
53	Yiddish, Eastern	ydd	72784832160	3142560	23161
54	Croatian	hrv	71033369490	6214643	11430
55	Limburgisch	lim	69583500000	1500000	46389
56	Ligurian	lij	69536654649	1920849	36201
57	Slovak	slk	68983077920	5011120	13766
58	Telugu	tel	67806694494	69688278	973
59	Marathi	mar	66212442751	68049787	973
60	Tamil	tam	64237654600	66020200	973
61	Thai, Northeastern	tts	56640000000	15000000	3776
62	Zulu	zul	55879074746	9563422	5843
63	Kazakh	kaz	55542767289	8178879	6791
64	Vietnamese	vie	55318273119	67379139	821
65	Sardinian, Logudorese	src	54301500000	1500000	36201
66	Auvergnat	auv	52863000000	1315000	40200
67	Vlaams	vls	52464896000	1202000	43648
68	Panjabi, Western	pnb	51629466957	60812093	849
69	Urdu	urd	51367538571	60503579	849

#	Language	iso-639-3	GLP	Pop.	GDP-per-capita
70	Chinese, Gan	gan	50606220000	20580000	2459
71	Sunda	sun	49815000000	27000000	1845
72	Walloon	wln	48885760000	1120000	43648
73	Bulgarian	bul	48436572699	8954811	5409
74	French, Cajun	frc	45959000000	1000000	45959
75	Serbian	srp	45739846348	11139758	4106
76	Slovenian	slv	45518829850	1984775	22934
77	Gujarati	guj	44861270328	46106136	973
78	Indonesian	ind	42699488130	23143354	1845
79	Arabic, Libyan Spoken	ayl	42581260000	4505000	9452
80	Arabic, Moroccan Spoken	ary	42292382600	19480600	2171
81	Xhosa	xho	42152091474	7214118	5843
82	Belarusan	bel	41809393608	9081102	4604
83	Luxembourgeois	ltz	40800831336	390618	104452
84	Afrikaans	afr	34858630997	5965879	5843
85	Malayalam	mal	34791658300	35757100	973
86	Kannada	kan	34391658000	35346000	973
87	Guadeloupean Creole French	gcf	34107850800	848454	40200
88	Okinawan, Central	ryu	33861372570	984285	34402
89	Turkmen	tuk	33810654240	6403533	5280
90	Lithuanian	lit	33521764006	3125281	10726
91	Arabic, Mesopotamian Spoken	acm	33008600000	15100000	2186
92	Frisian, Western	fri	32472300000	700000	46389

#	Language	iso-639-3	GLP	Pop.	GDP-per-capita
93	Arabic, Tunisian Spoken	aeb	31488759000	9247800	3405
94	Oriya	ori	30842673582	31698534	973
95	Cebuano	ceb	30666558060	20043502	1530
96	Arabic, Sa‘idi Spoken	aec	30088800000	18900000	1592
97	Friulian	fur	28743594000	794000	36201
98	Hausa	hau	28124568000	24162000	1164
99	Arabic, North Levantine Spoken	apc	27975144835	14309537	1955
100	Bashkir	bak	27696468400	1871383	14800
101	Hawai’i Creole English	hwc	27575400000	600000	45959
102	Gronings	gos	27462288000	592000	46389
103	Panjabi, Eastern	pan	27250522992	28006704	973
104	Azerbaijani, North	azj	27228603353	7059529	3857
105	Chuvash	chv	27149031200	1834394	14800
106	Bhojpuri	bho	25874537528	26592536	973
107	Chinese, Min Bei	mnp	25312946000	10294000	2459
108	Madura	mad	25267090500	13694900	1845
109	Zhuang, Northern	ccx	24590000000	10000000	2459
110	Welsh	cym	24467307508	536258	45626
111	Tagalog	tgl	24327149940	15900098	1530
112	Maithili	mai	24128047286	24797582	973
113	Réunion Creole French	rcf	24120000000	600000	40200
114	Tatar	tat	23828473600	1610032	14800
115	Thai, Northern	nod	22691479296	6009396	3776

#	Language	iso-639-3	GLP	Pop.	GDP-per-capita
116	Yoruba	yor	22496628000	19327000	1164
117	Chinese, Min Dong	cdo	22384663063	9103157	2459
118	Gaelic, Irish	gle	22341570000	355000	62934
119	Arabic, Sudanese Spoken	apd	22251592000	18986000	1172
120	Sotho, Northern	nso	21675876431	3709717	5843
121	Breton	bre	21415424400	532722	40200
122	Igbo	ibo	20952000000	18000000	1164
123	Basque	eus	20922530208	588108	35576
124	Umbundu	umb	20026408640	4002880	5003
125	Awadhi	awa	20005603912	20560744	973
126	Tsonga	tso	19136438515	3275105	5843
127	Sinhala	sin	18957847104	13220256	1434
128	Thai, Southern	sou	18880000000	5000000	3776
129	Uyghur	uig	18691918829	7601431	2459
130	Latvian	lav	18677424712	1543844	12098
131	Sindhi	snd	18136338000	21362000	849
132	Armenian	hye	18040062720	6723840	2683
133	Plautdietsch	pdt	17465069122	401699	43478
134	Estonian	est	17391861987	1075497	16171
135	Arabic, South Levantine Spoken	ajp	16253525000	6145000	2645
136	Corsican	cos	16160400000	402000	40200
137	Icelandic	isl	15882232320	239768	66240
138	Uzbek, Northern	uzn	15487566984	18795591	824

. . .

Results: Low/High Momentum Lgs

- There are currently 40 non-low-momentum languages
- The rest, beginning with Tswana at rank #41, are low-momentum languages

Non-low-momentum languages:

- All have fair amounts of NLP infrastructure except
  - South Azerbaijani
  - Languages with the following two properties:
    - They are not popularly written
    - In the country where they are spoken, there is a standardized close relative which is the preferred language for written communication

Survey

- Collection Methodology:
  - General knowledge
  - Browsing of the meta-literature
  - Corpora-list
  - Googling suitably for each of the 100 densest low-momentum languages
- Criteria for “morphology”:
  - Even if not the entire morphology for a language is covered, we count it anyway
  - If there is work on MT without a morphological component, we do not count it
  - For languages which have very little morphology we take some other NLP work at a comparable stage

Cont.	Langugae		$GLP(L)$	#	$S(N)$
Europe	Bulgarian	bul	48436572699	73	8954811
	Serbian	srp	45739846348	75	11139758
	Slovenian	slv	45518829850	76	1984775
	Lithuanian	lit	33521764006	90	3125281
	Welsh	cym	24467307508	110	536258
	Irish	gle	22341570000	118	355000
	Basque	eus	20922530208	123	588108
	Latvian	lav	18677424712	130	1543844
	Estonian	est	17391861987	134	1075497
	Icelandic	isl	15882232320	137	239768
	Udmurt	udm	8373632800	184	565786
	Mordvin	myv	7660110000	193	517575
	Komi	kpv	3880560000	257	262200
	Faroese	fao	2589616000	310	45400
	Sámi	sme	1776495000	369	21000
	Tundra Nenets	yrk	395604000	715	26730
	Khanty	kca	177600000	1055	12000
	Mansi	mns	47123200	2001	3184
	Nganasan	nio	7400000	3667	500
	Latin	lat	0	6786	0
	Ancient Greek	grc	0	7271	0

Cont.	Langugae		$GLP(L)$	#	$S(N)$
Asia	Bengali	ben	82284767162	45	171070202
	Thai	tha	76388430912	49	20229987
	Telugu	tel	67806694494	58	69688278
	Marathi	mar	66212442751	59	68049787
	Tamil	tam	64237654600	60	66020200
	Vietnamese	vie	55318273119	64	67379139
	Urdu	urd	51367538571	69	60503579
	Gujarati	guj	44861270328	77	46106136
	Malayalam	mal	34791658300	85	35757100
	Kannada	kan	34391658000	86	35346000
	Turkmen	tuk	33810654240	89	6403533
	Oriya	ori	30842673582	94	31698534
	Tagalog	tgl	24327149940	111	15900098
	Sinhala	sin	18957847104	127	13220256
	Uigur	uig	18691918829	129	7601431
	Assamese	asm	14958902000	141	15374000
	Burmese	mya	9238252166	176	32301581
	Pashto	pbu	8235300000	186	9700000
	Mongolian	khk	3091976685	288	2337095
	Lao	lao	1970540586	351	3188577
	Manipuri	mni	1226953000	434	1261000
	Sanskrit	san	5941138	3878	6106
	Great Andamanese	apq	23352	6516	24
	Syriac	syc	0	7068	0
	Akkadian		0	6945	0

Cont.	Langugae		$GLP(L)$	#	$S(N)$
Americas	Aymara	ayr	3221170332	279	2227642
	Greenlandic	kal	3125792000	283	54800
	Mapudungun	arn	3014100000	291	300000
	Plains Cree	crk	1482599800	402	34100
	Quiché	qut	661750000	573	250000
	Ralámuri	tar	452045000	679	55000
	Iñupiaq	esi	158558550	1121	3450
	Chuj	cnm	83724610	1555	31630
	Cayuga	cay	2173900	4713	50
Africa	Zulu	zul	55879074746	62	9563422
	Xhosa	xho	42152091474	81	7214118
	Afrikaans	afr	34858630997	84	5965879
	Sesotho	nso	21675876431	120	3709717
	Kikuyu	kik	4245518000	246	5347000
	Somali	som	3416439600	273	12653480
	Luo	luo	2751210000	307	3465000
	Kinyarwanda	kin	2437222820	316	7275292
	Malagasy	plt	2236711200	330	5948700
	Bambara	bam	1565948370	390	2786385
	Ekegusii	guz	1256108000	431	1582000
	Ha	haq	406890000	707	990000
	Swahili	swh	317555862	791	772642

Numbers Across Continents

Europe	22
Asia	25
Africa	13
Americas	9
	69

- GLP and popular writing account for the disproportionately high number of European languages
- High incidence of “dedicated individuals” in the Americas

Discussion

Which Languages get CMR?

- GLP-high languages
- GLP-low languages with *state support*
- A handful languages get CMR b/c dedicated individuals
- Not much transfer-effect so far observed

  Language A has CMR, Language B is closely related to A and has low GLP
- Caveat: Probably commercial actors have developed CMRs too, but this work is *not* published!

General Conclusions

- Languages with no popular writing system very rarely get CMR
- Languages with high GLP tend to get CMR
- Explicit state sponsorship is the mechanism by with GLP-high languages get CMR
- Private/commercial actors play no role or do not publish
- Unsupervised solutions for CMR should focus on GLP-low languages for which there is raw data

Conclusions as to GF-extension

- Which languages are likely to be used? & Which languages have commercial potential?

  GLP-high languages with a popular writing system

- Which have interested speakers/experts?
  - Dedicated individuals have so far played less prominent role than state-sponsored endeavors
  - Can GF be better at recruiting them?
- Languages which are not “covered” by someone else?
  - Languages with likely state support will presumably be covered
  - GF should pick others or expect competition

The End

Thank You for Listening!