Selection Criteria for Low Resource Language Programs

Christopher Cieri[◦], Mike Maxwell[□], Stephanie Strassel[◦], Jennifer Tracey[◦]

Linguistic Data Consortium

UNIVERSITY OF MARYLAND

Background

- Despite increased work on resource-poor languages, differences in
  - terminology
  - available information
  - goals

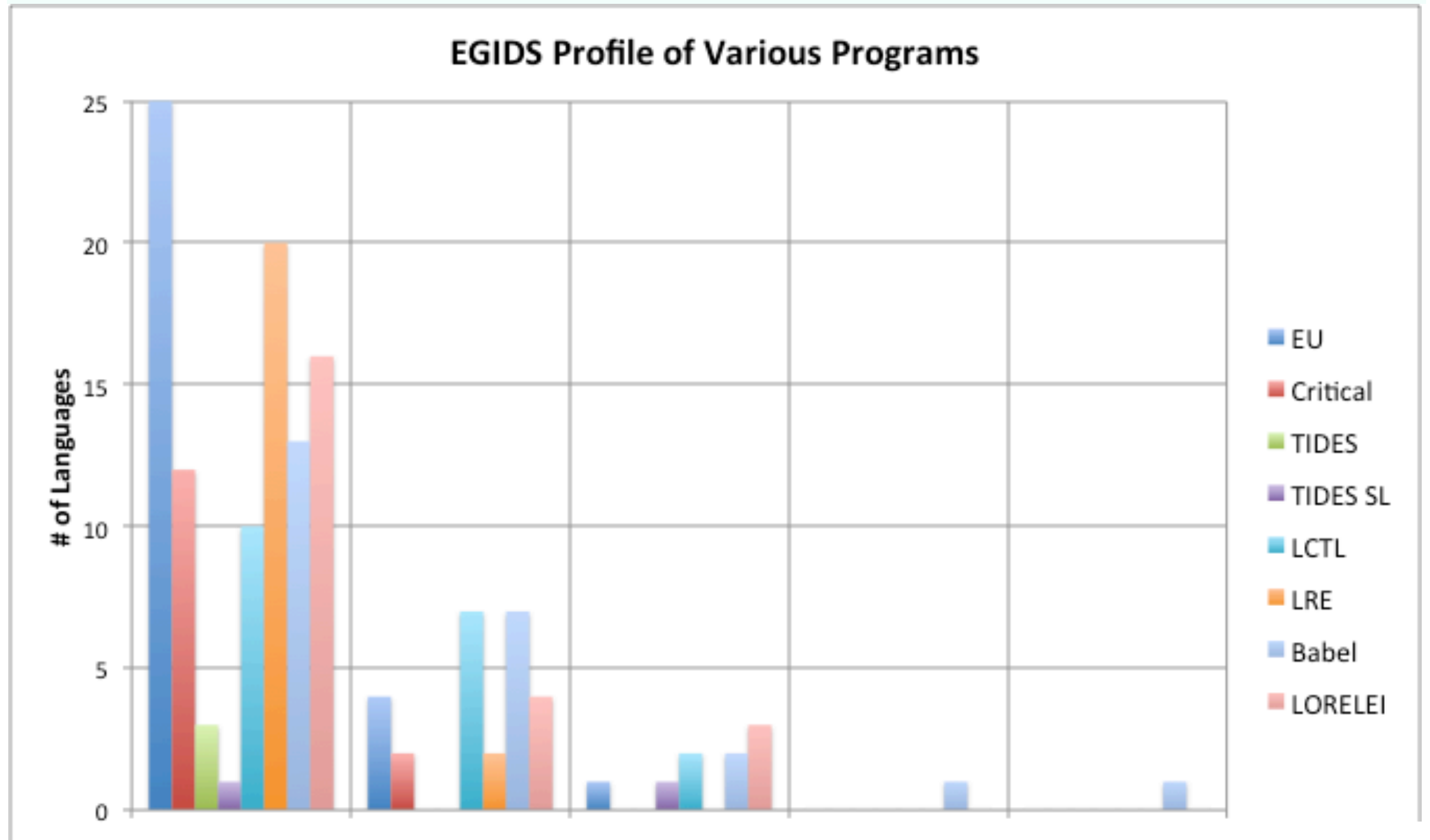
  yield obscurity in language selection criteria
- Program goals range from:
  - LORELEI: facilitate situational awareness in the event of a disaster
  - METANET: create missing technologies and transfer languages facing digital extinction
  - NSF DEL: "*document living endangered languages and their associated cultural and scientific information before they disappear*"
- Program Effects
  - time & funding commitments
  - create critical language resources
  - enable human language technologies
  - increase native speaker information access

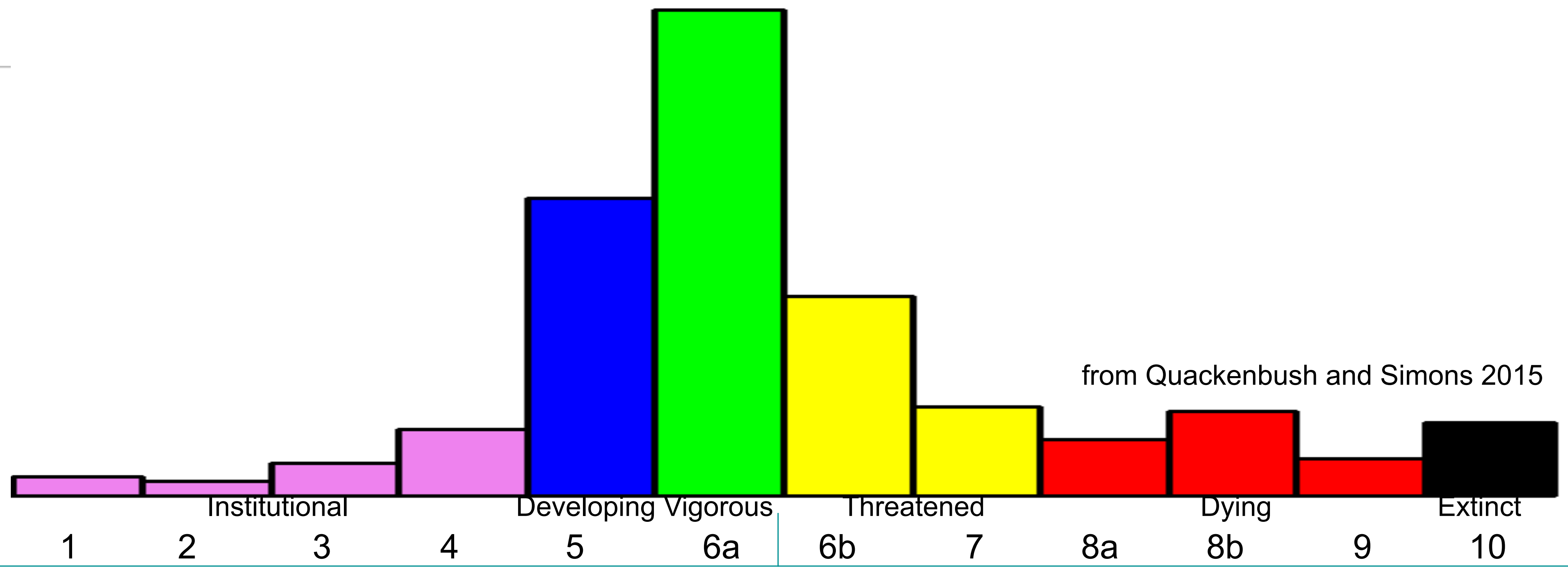
The potential impact – on research and daily life – of resource development efforts make language selection criteria a worthy topic.

Goal: begin dialog on how community decides which languages to study, survey selection criteria used by low resource language research and available

Programs

- DARPA TIDES: translingual IR/IE & summarization in 3 languages + 1 surprise
- REFLEX LCTL: translingual technologies, language packs in 20 languages
- NIST LRE: 1996-present evaluation campaign, language variety, confusability, not specifically low resource
- IARPA Babel: escape English bias in speech recognition
- DARPA LORELEI: information awareness for disaster events in low resource languages

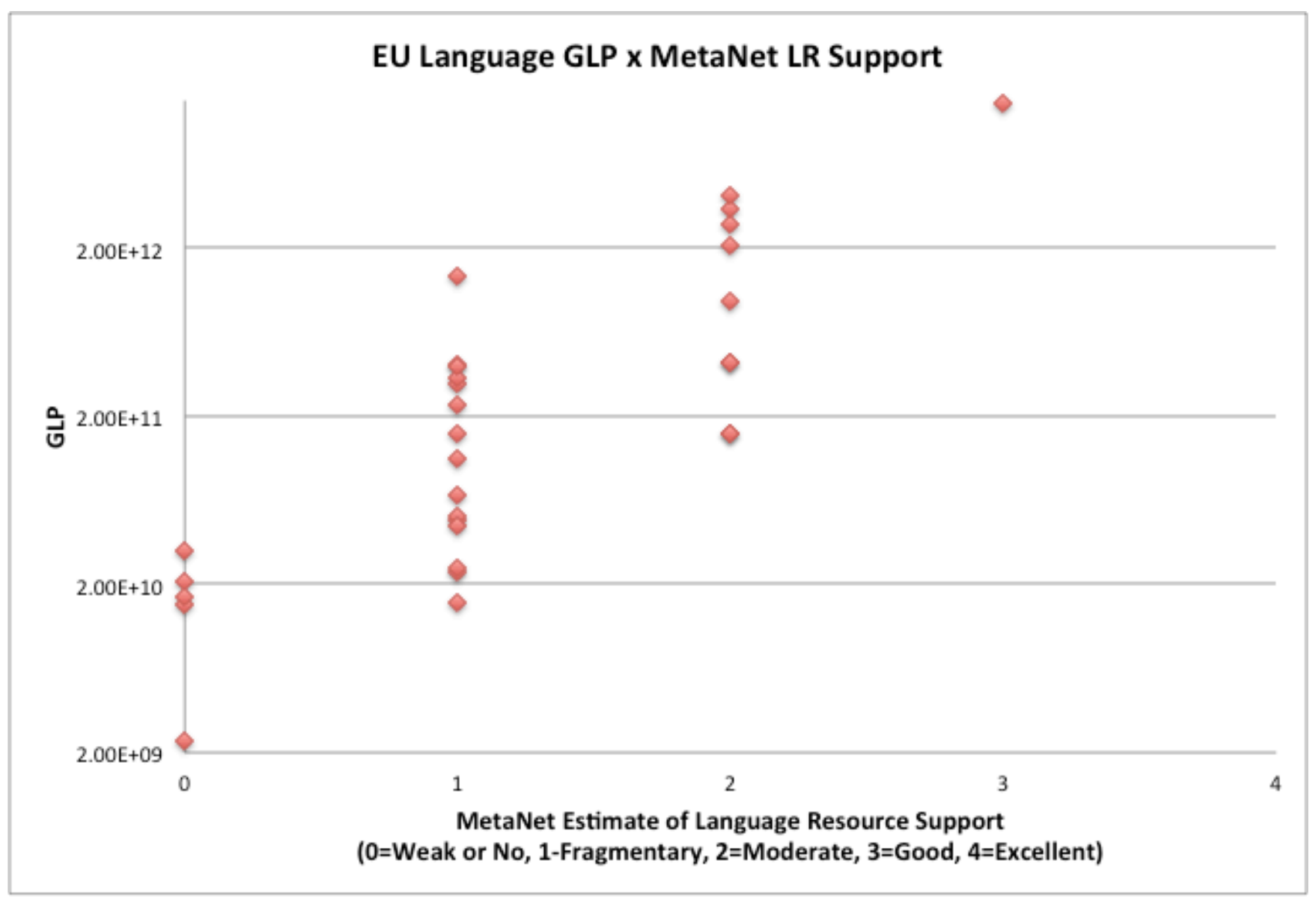




from Quackenbush and Simons 2015

Terms

- endangered: risk of losing native speakers
- critical: undesirable supply/demand ratio
- low density: few online resources (under-resourced, low resource)
- less commonly taught: e.g. specific market
- surprise: within common task program
- low-affluence: defined via GLP



Selection Criteria: Demographic

- importance, influence
  - population:
    - but see English vs Spanish, Mandarin
  - GLP= per capita GDP * native speakers per country
    - does not predict LR presence in EU
- # speakers of more 'important' language
  - e.g Italian versus 6 other languages of Italy among 60 most affluent
- total # speakers 1st or 2nd language
  - e.g. Swahili
- speakers involved in high profile event
  - e.g. Haitian Creole during earthquake

Selection Criteria: Linguistic

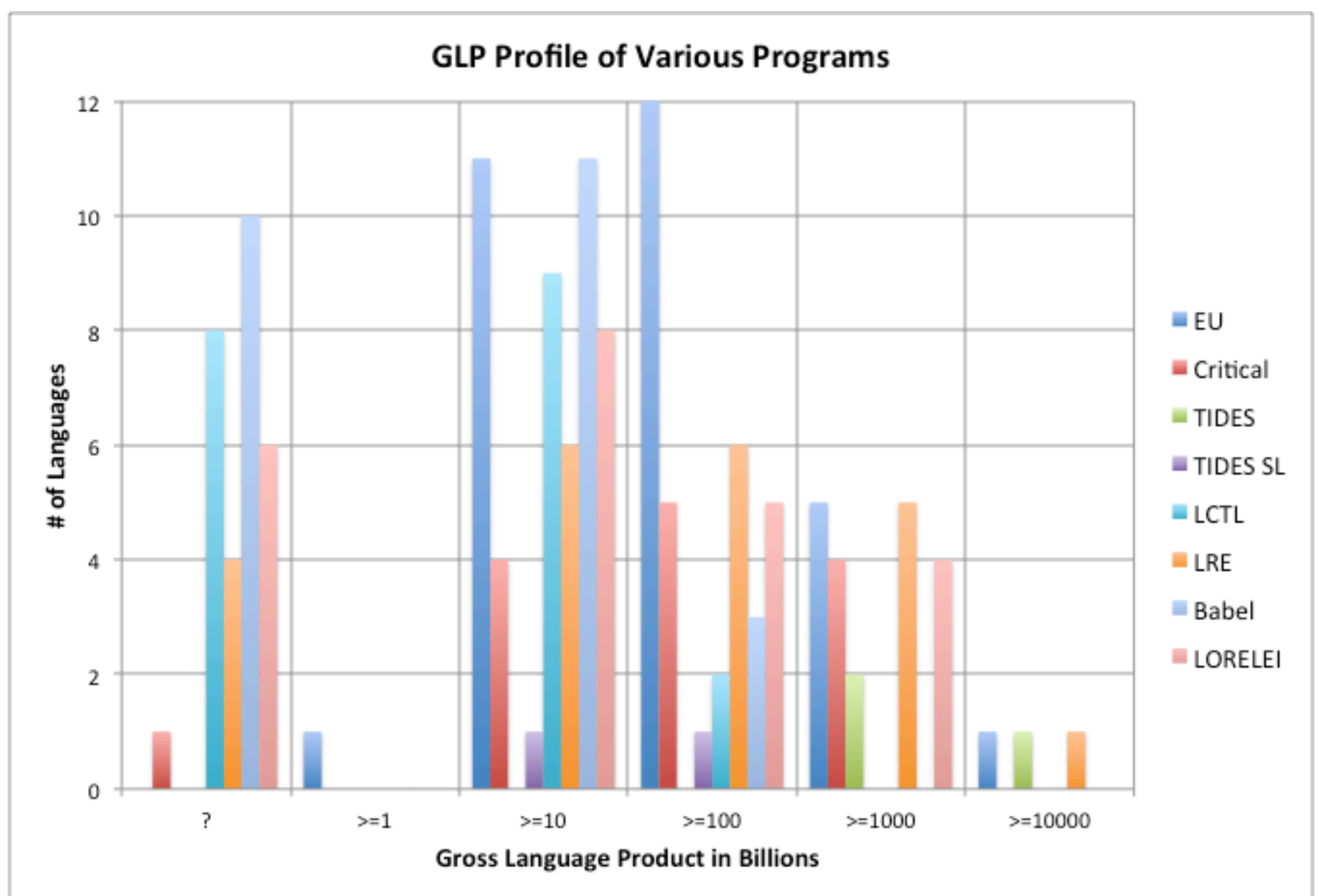
- similarity as measured via family tree
  - numerous resource porting experiments (Elmahdy et al. 2014, Vergyri 2005 et al., Beyerlein et al. 1999)
  - REFLEX LCT: Bengali, Punjabi, Urdu
  - LRE confusable clusters sometimes family tree related – but French-Haitian Creole
- written by native speakers
- orthography standardized
- words & sentences delimited in writing
- ease of letter to sound mapping
- nature of morphology
  - analytic or synthetic, number of morphological classes, degree of irregularity, syncretism
- typological diversity across program

Selection Criteria: Resource

- # resources
  - too few mires technology development
  - too many not representative
- specific resource types
  - monolingual & parallel text, speech
  - dictionaries, gazetteers, grammars
- human resources
  - previously, local speaker population
  - in-country partners
  - in-country infrastructure
- elaborated
  - standard digital encoding, news & parallel text, translation dictionaries, tokenizers, segmenters, taggers, morph analyzers
- different weightings of the above

Implementation Challenges

- different notions of 'language'
- different language names
- difficulty collecting data on
  - demographics
  - linguistic features
  - available resources
- demographics change over time
  - # Syrian Arabic speakers in Europe
- resource availability, change over time
  - Quechuan in the LCTL era versus today
- language attitudes
  - suppression, language death



Select References

- Beyerlein, P., W. Byrne, J. M. Huerta, S. Khudanpur, B. Marthi, J. Morgan, N. Peterek, J. Picone, W. Wang. 1999. Towards Language Independent Acoustic Modeling. IEEE Workshop on Automatic Speech Recognition and Understanding, December 12 - 15, Keystone, Colorado, U.S.A
- Elmahdy, Mohamed, Mark Hasegawa-Johnson, and Eiman Mustafawi, "Development of a tv broadcasts speech recognition system for Qatari Arabic," in The 9th edition of the Language Resources and Evaluation Conference (LREC 2014), Reykjavik, Iceland, 2014.
- Hammarström, H.. A survey of computational morphological resources for low-density languages. Journal of the NEALT, 2009.
- Lewis, M. Paul and Gary F. Simons. 2010. Assessing Endangerment: Expanding Fishman's GIDS. Revue Roumaine de Linguistique 55(2):103–120. http://www.lingv.ro/RRL 2 2010 art01Lewis.pdf. Accessed March 21, 2011.
- Lewis, M. Paul, Gary F. Simons, and Charles D. Fennig (eds.). 2015. Ethnologue: Languages of the World, Eighteenth edition. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com.
- Maxwell, Mike, Baden Hughes. 2006. Frontiers in Linguistic Annotation for Lower-Density Languages in Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora 2006, Sydney, Australia, Association for Computational Linguistics, pp. 29—37, URL: http://www.aclweb.org/anthology/W/W06/W06-0605
- National Science Foundation Documenting Endangered Languages Program. 2014. Press Release 14-098: Federal agencies provide new opportunities for dying languages, August 15, 2014, http://www.nsf.gov/news/news_summ.jsp?cntn_id=132370, accessed March 16, 2016.
- J. Stephen Quakenbush, Gary F. Simons. 2015. Looking at Austronesian language vitality and endangerment through EGIDS and the sustainable use model in WayanArka, I., Ni LuhNyoman Seri Malini, Ida Ayu Made Puspani (eds.) Language Documentation and Cultural Practices in the Austronesian World, Papers from 12-ICAL, Volume 4. Australian National University.
- Rehm, Georg, Hans Uszkoreit, eds. 2012. META-NET White Paper Series: Europe's Languages in the Digital Age, URL: www.meta-net.eu/whitepapers
- Simpson, Heather, Christopher Cieri, Kazuaki Maeda, Kathryn Baker, Boyan Onyshkevych. 2008. Human Language Technology Resources for Less Commonly Taught Languages: Lessons Learned Toward Creation of Basic Language Resources, paper presented at the SALTMIL Workshop: Free/Open-Source Language Resources for the Machine Translation of Less-Resourced Languages satellite to the 7th International Conference on Language Resources and Evaluation, Marrakesh, May 28-30
- Vergyri, D., K. Kirchhoff, R. Gadde, A. Stolcke, J. Zheng. 2005. Development Of A Conversational Telephone Speech Recognizer For Levantine Arabic. Proceedings of Interspeech, Lisboa, Portugal.