Assessing digital language support as a factor in language vitality

Gary F. Simons and Abbey Thomas

SIL International

6th Int’l Conf. on Language Documentation and Conservation

—*Connecting Communities, Languages, and Technology*

University of Hawaii at Manoa, 28 February to 3 March 2019

Background

- *Ethnologue* has been inspired by Kornai's work to incorporate digital language vitality into our reporting on the world's languages
  - Kornai, András. (2013). Digital language death. *PLoS ONE* 8(10), e77056. doi:10.1371/journal.pone.0077056 http://www.plosone.org/article/info:doi/10.1371/journal.pone.0077056
- In the 21st century context of accelerating globalization and technology on the one hand and the endangerment of non-dominant languages on the other, crossing the Digital Divide may be essential to the long-term survival of a language
  - We therefore want to be able to monitor and report what is happening in that regard

Our model of digital language vitality

- Studies of digital adoption typically look at two major components:
  - The access to digital technology that is achieved
  - The usage that is actually made
- For a language this corresponds to:
  - The degree to which the language is supported in various digital technologies
  - The degree to which users of the language actually use the language in digital communication

Toward a Digital Language Vitality Index

- Our ultimate aim is to develop a Digital Language Vitality Index that will combine the place of a language on two scales
  - a Digital Language Support Scale
  - a Digital Language Use Scale
- Digital language support is the easier one
  - The raw data are available on web pages that list the languages supported by various digital tools
  - Data on digital language use are not similarly open
- Thus we have begun with digital language support

Overview

1. Describe the methodology we used to develop a scale for measuring digital language support
2. Show the results of calculating a score on the Digital Language Support Scale for every known language
3. Examine the relationship between digital language support (as measured by DLSS) and overall language vitality (as measured by EGIDS)

System requirements

- In order to periodically monitor and report, we require a system that is fully automated, i.e.
    - Scrape names of supported languages from web pages
    - Map the language names to ISO 639-3 codes
    - Calculate a support score for each ISO 639-3 language
- The main research challenges
    - How do we construct a representative sample of digital tools that support multiple languages?
    - How do we transform the harvested data on languages supported into a Digital Language Support score for each language?

Categories of digital language support

- We began by identifying six categories of digital language support:
  - Encoding support (*e.g.*, keyboards, fonts)
  - Localized user interface (*e.g.*, OS, browsers, messaging)
  - Surface-level processing (*e.g.*, spell-checking, stemming)
  - Meaning-level processing (*e.g.*, machine translation)
  - Speech processing (*e.g.*, speech-to-text, text-to-speech)
  - Virtual assistance (*e.g.*, Siri, Alexa)

Finding the most widely used exemplars in each category

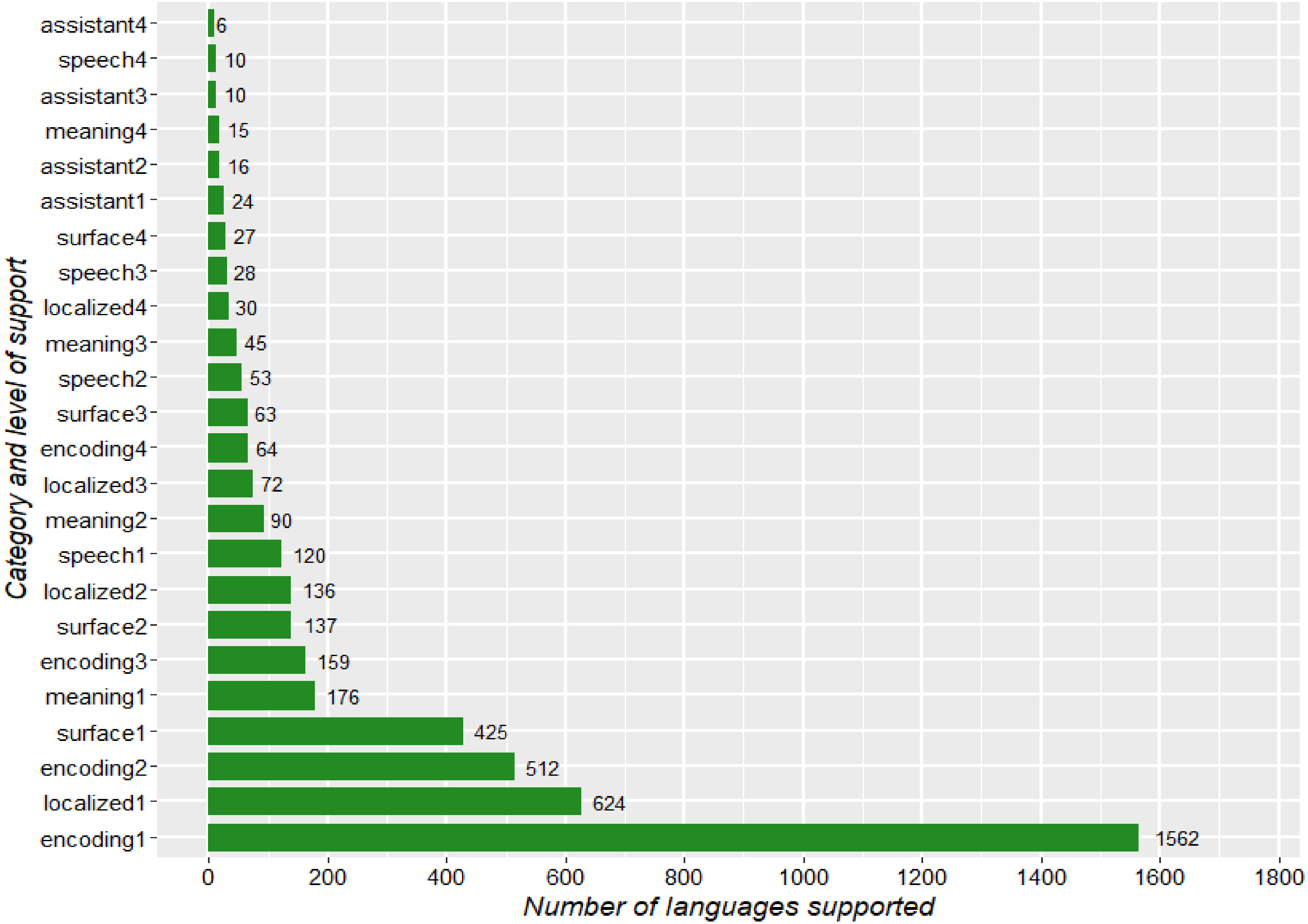
- For each category, we identified:
  - Top 10 tools of its kind globally
  - Top 5 tools in each of the 10 most populous countries of the world (to ensure we included the major tools in use outside the English-speaking world)
  - The reference authority for these rankings was similarweb.com
  - Then we added any other tools found to support more than 10% of the median number of languages supported by the top tools in the category
- The full sample comprised 126 digital tools

Scoring support by category

- Each language is scored 0 to 4 for each category based on the number of tools that support it — Jenks optimization was used to find the natural breaks in each distribution

Category	Tools in category	Score				
		0	1	2	3	4
Encoding	9	0	1-2	3-5	6-8	9
Localized	51	0	1-15	16-30	31-48	49-51
Surface	16	0	1-5	6-9	10-15	16
Meaning	16	0	1-6	7-11	12-15	16
Speech	22	0	1-8	9-15	16-20	21-22
Assistant	12	0	1	2-3	4-6	7-12

Number of languages supported at each level of the six categories



Does this form a hierarchical scale?

- Hypothesis: “This looks a lot like a Guttman scale”
  - A set of items measuring the same underlying trait that are hierarchically ordered by degree of “difficulty” — correctly answering or agreeing with a more difficult item on the scale implies you also have all the easier items
  - The preceding plot is not a perfect Guttman scale since there are many exceptions (i.e., having a higher item and missing some of the lower ones), but it feels a lot like one
- Is there a statistical method for testing how well a set of items fits this notion of being a hierarchical scale?
  - Yes, Mokken Scale Analysis

Measuring the fit as a scale

- The coefficient of scalability, *H*, measures how well the items fit into a hierarchical scale — the more exceptions there are to not having all the easier items, the lower the value
    - 0.3 to 0.4 = weak
    - 0.4 to 0.5 = medium
    - > 0.5 = strong
- These digital language support items form a *very* strong scale at *H* = 0.898
    - We can thus use the raw score (number of items) as an estimate of the trait (level of digital language support)

From raw score to adjusted

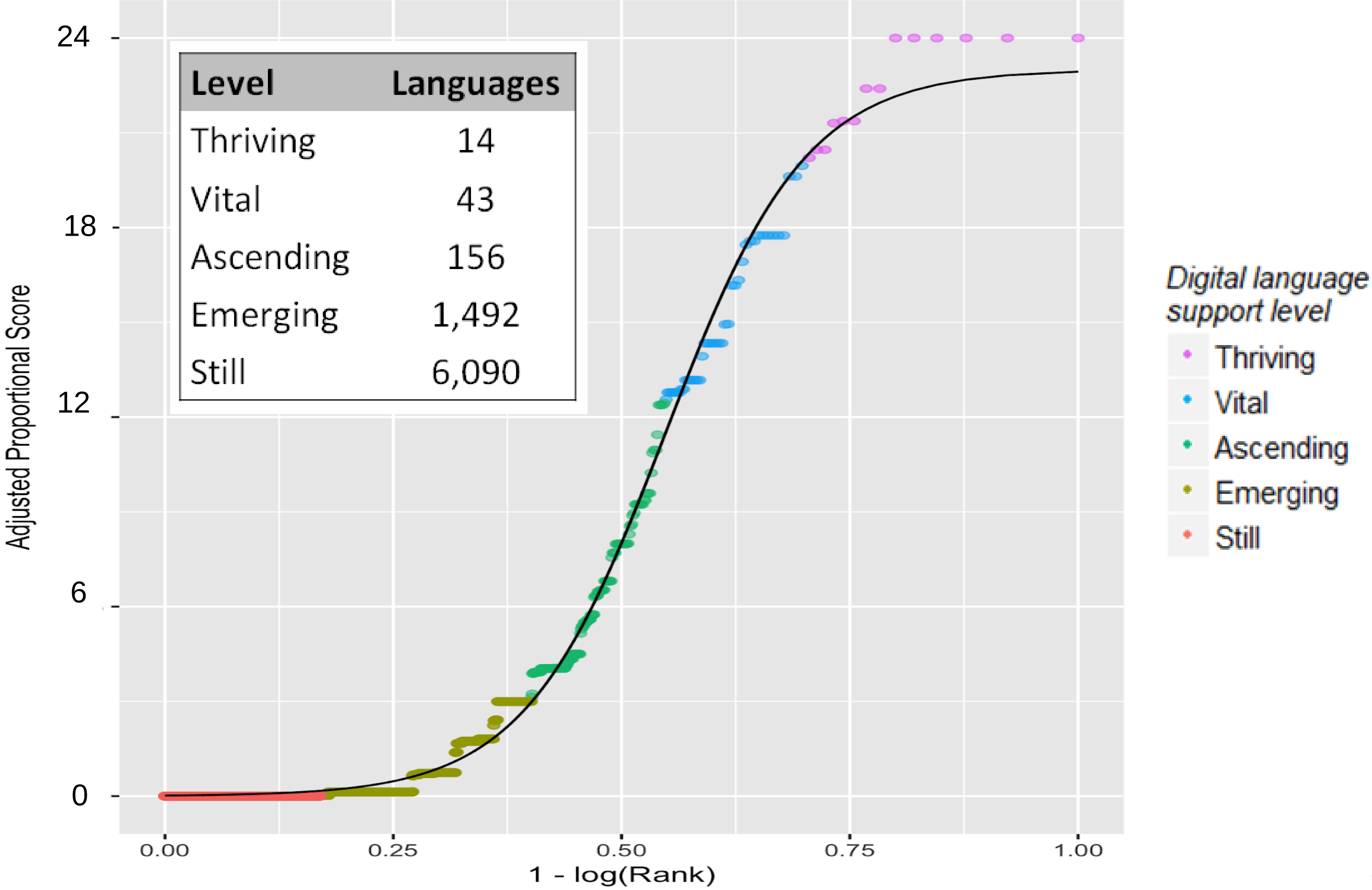
- But we can use the fact that it is such a strong scale to create an even more accurate score
- Mokken analysis is based on Item Response Theory which is a methodology developed for educational testing
  - The scale measures a single latent trait (θ): Any subject (*s*) has a scale value for the latent trait (θ_s), and any test item (*i*) has a value on the same scale, known as its difficulty level (δ_i).
  - The adjusted score counts each correct item not as 1, but as the probability that a subject with the same raw score on the rest of the items on the test would get a correct response on that item

Plotting the distribution of DLS scores for all known languages

- On y-axis we plot the adjusted Digital Language Support score
- On x-axis we plot the rank of the language (1 high to 7795 low)
  - But converted to a log scale and flipped so that lowest rank is on the left and highest is on the right
- What emerges is a classic "diffusion of innovation" S-curve
- Digital Language Support levels defined by geometry of the curve
  - Still = 0
  - Ascending/Vital divide is midpoint
  - The two other divides are where the slope changes from mostly horizontal to mostly vertical

Level	Languages
Thriving	14
Vital	43
Ascending	156
Emerging	1,492
Still	6,090

Growing digital language support as a diffusion of innovation curve

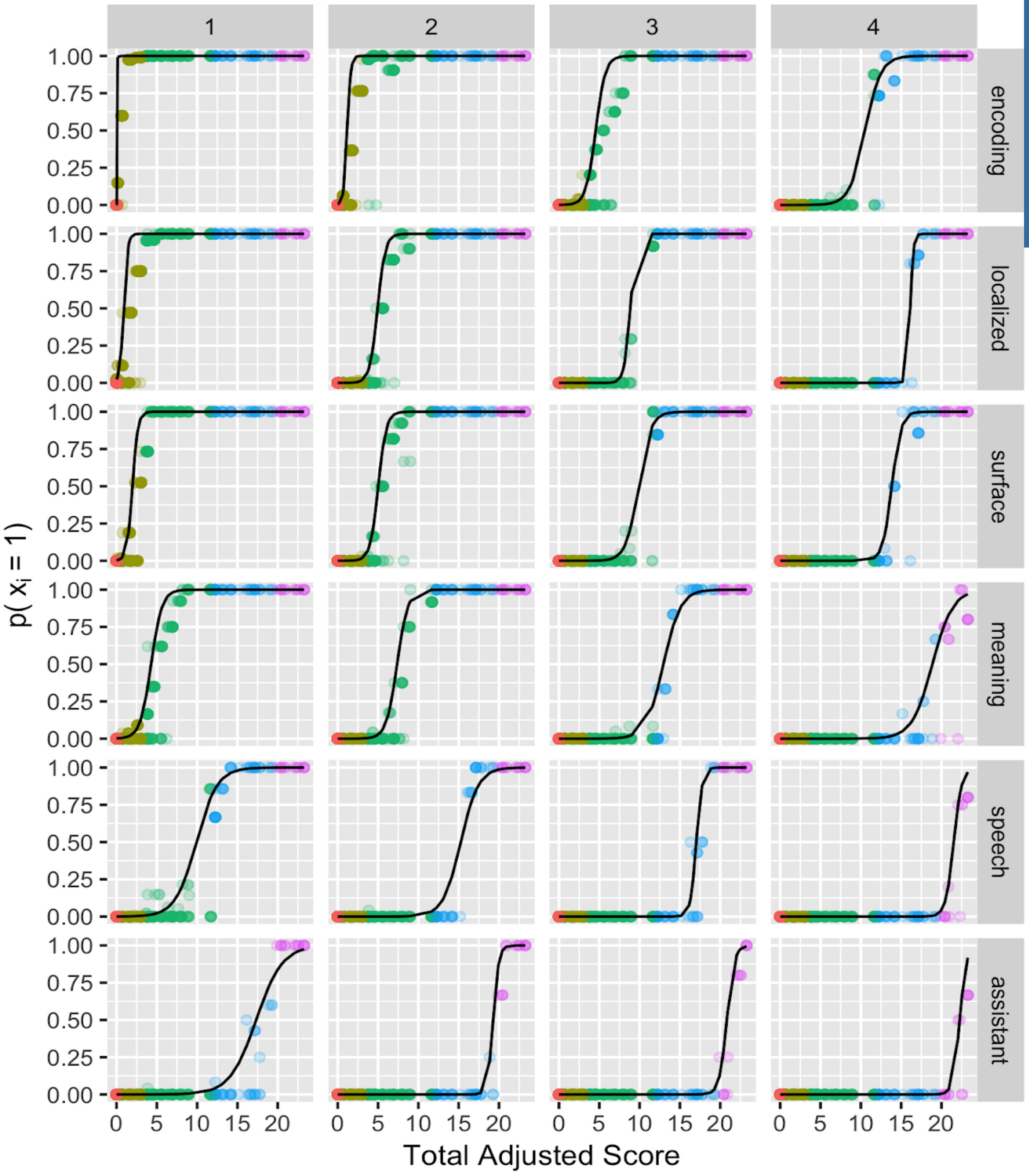


Level	Languages
Thriving	14
Vital	43
Ascending	156
Emerging	1,492
Still	6,090

Relating difficulty of items to DLS levels

- Item Response Theory uses an Item Response Function for each item to plot the probability (*p*) that a subject with a given score on the scale will get a “correct” response on the item
  - The “difficulty” of an item is the score where *p* = 0.5
- The next two slides show
  - The Item Response Functions for the 24 items on the Digital Language Support Scale
  - A plot of how the items correspond to the four Digital Language Support levels

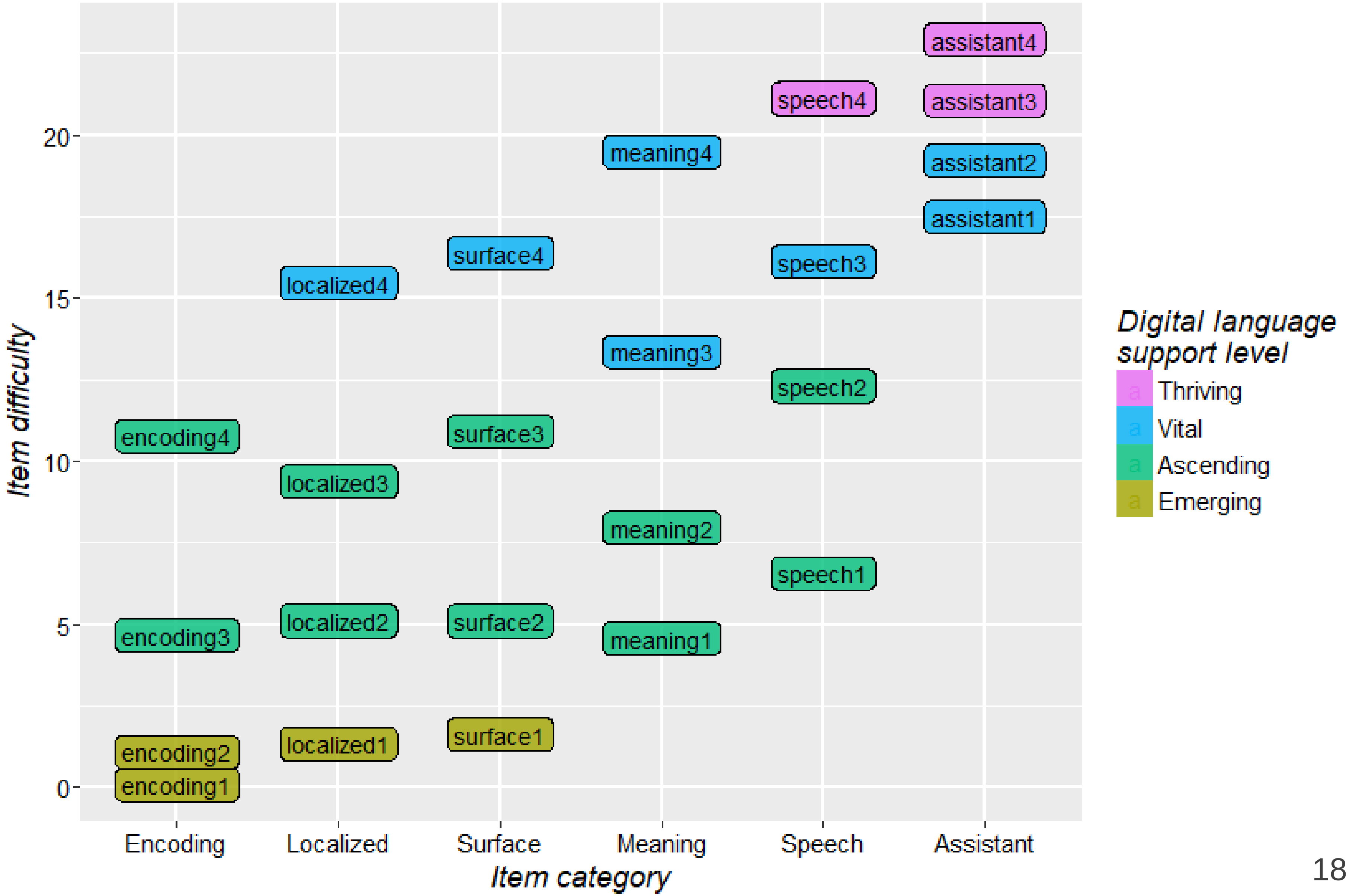
Item Response Functions



Plots of difficulty of each item

- X-axis is the total score for the language (e.g., 0–24)
- Y-axis is the probability (0–1) that a language has the item

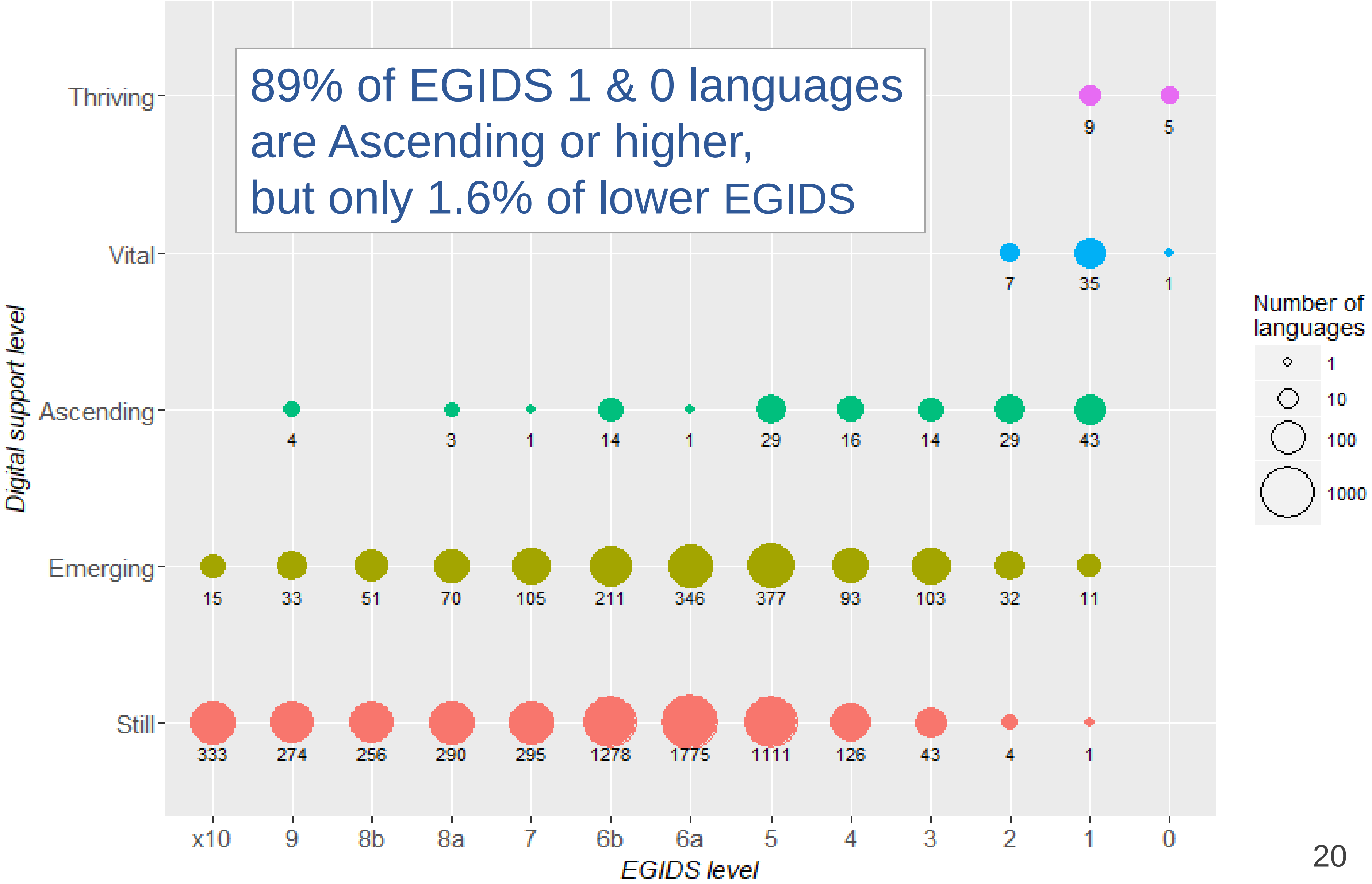
Correspondence of items to levels



Digital language support and language vitality

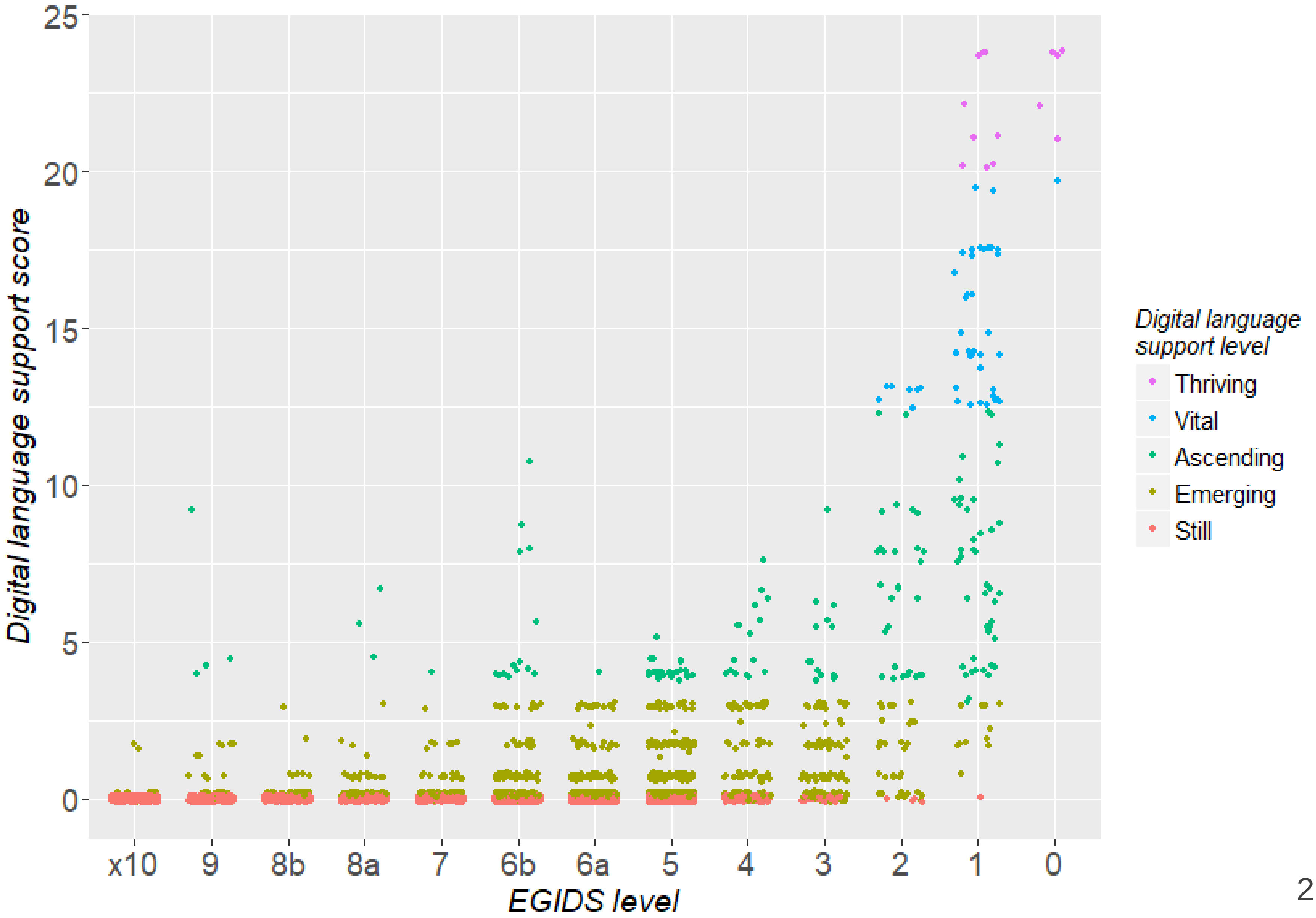
- The next two slides use the EGIDS ratings in *Ethnologue* to plot the relationship between digital language support overall language vitality
  - A bubble plot shows the number of languages that are at each combination of DLS level and EGIDS level
  - A “jittered” scatter graph plots each language by its DLS score and EGIDS level. (At the Still and Emerging levels there are hundreds of dots on top of each other as documented by the bubble plot.)
- There is a clear trend: the stronger the vitality level, the greater the digital language support

DLS level by EGIDS level



89% of EGIDS 1 & 0 languages are Ascending or higher, but only 1.6% of lower EGIDS

DLS score by EGIDS level



Digital support and long-term vitality

- There has been a general consensus that 10% of languages are safe for the long term, 50% are likely to die, and it remains to be seen for the rest
- The DLS scale could help us to monitor the progress that languages are making toward becoming safe
  - If the information industry is significantly investing in a language, it is probably a good sign that it is safe
  - Only 53 languages are now digitally Vital or Thriving
  - Note that 659 languages (= ~10%) have now reached the second easiest item on the scale: Localized 1

Digital language support may pattern differently for non-dominant languages

- Development at the leading edge of the innovation curve is being driven by dominant languages
  - The monolingual members of these societies need every-thing they do on their devices to be in their language
- The story is different for non-dominant languages
  - Their speakers are typically multilingual and have learned how to use their devices with user interfaces that are in the dominant language
  - They may not even want localized user interfaces
  - The Digital Language Support Scale for those languages could conceivably skip over the Localized category

Conclusion

- Growth of digital language support has been rapid
    - E.g., Google Translate: 2009, 41 lgs; 2019, 103 lgs
- The Digital Language Support Scale holds promise for helping us to monitor the progress languages make toward greater relevance in a digital world
- But the current scale is most well elaborated at the higher end; to better monitor threatened languages
    - We may require more elaboration at the lower end
    - We may discover that a slightly different scale is in play for non-dominant languages whose speakers are accustomed to using interfaces in dominant languages