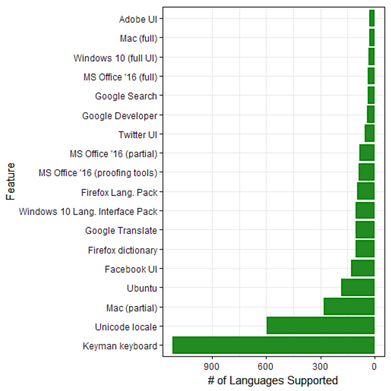
Measuring Digital Language Support

Abbey Thomas & Gary F. Simons
abbey.thomas@mavs.uta.edu, gary_simons@sil.org

Background

As digital modes of communicating and disseminating information become increasingly prevalent throughout the world, they also become increasingly relevant for endangered language communities. Opportunities afforded by ICT differ drastically by the language one uses; this has been dubbed the “digital language divide” (Mikami 2008, Young 2015, Soria 2016). As we explored the distribution of 18 digital support features across 7803 languages, we began to see the problem of a linguistic community’s access to digital resources not as a single divide, but instead as an ascent towards what Kornai (2013, 2015) and Gibson (2016) have termed “digital vitality.”



Data Collection

- Lists of supported languages harvested from settings pages for 18 digital support tools (“features”)
- Data collected for 7803 languages
- Figure 1 (Left) shows number of languages supported by each feature.

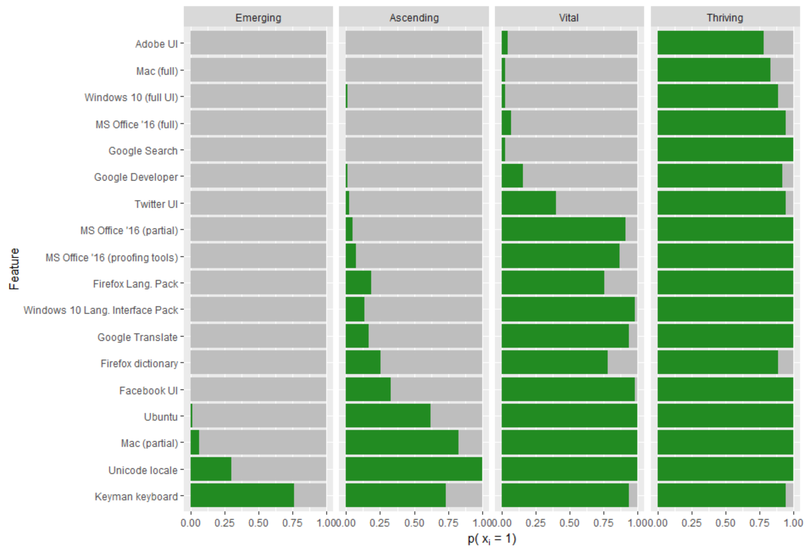
Analysis, Part 1: Item Response Theory

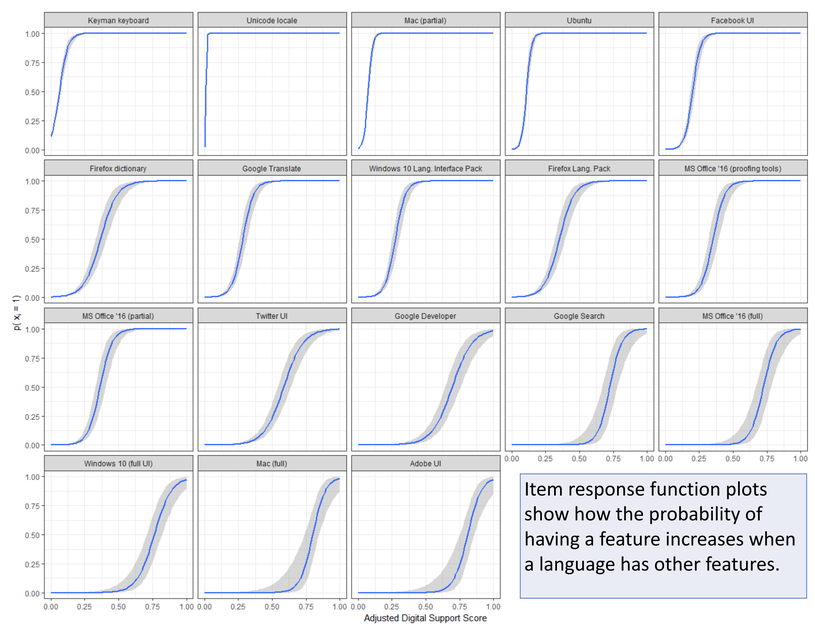
Researchers use Item Response Theory (IRT) to measure a test-taker's ability or attitude on an underlying trait with a set of test items that expect dichotomous responses. IRT predicts the probability of a response to a given item based on the test-taker's ability or attitude.

Element	Example from Other IRT Studies	Our Equivalent
Latent Trait	Knowledge of a particular subject	Digital support
Item	Question on a test	Digital support feature
Subject	Student	Language
Difficulty level	Difficulty of test question based on frequency of correct responses	Difficulty or cost of acquiring a feature for a given language
Item Response Function	How the probability of a correct response changes based on the student’s overall score on the test	How the probability of having a feature changes based on how many features the language has
Subject’s Scale Value	Student’s true score on the test, based on which questions were answered correctly	Language’s digital support score, adjusted to reflect which features the language has

Analysis, Part 2: Clustering

We utilized the Partitioning Around Mediods (PAM) clustering method to find groups of languages that had similar digital support profiles. PAM finds data points separated by minimal (Euclidean) distances.

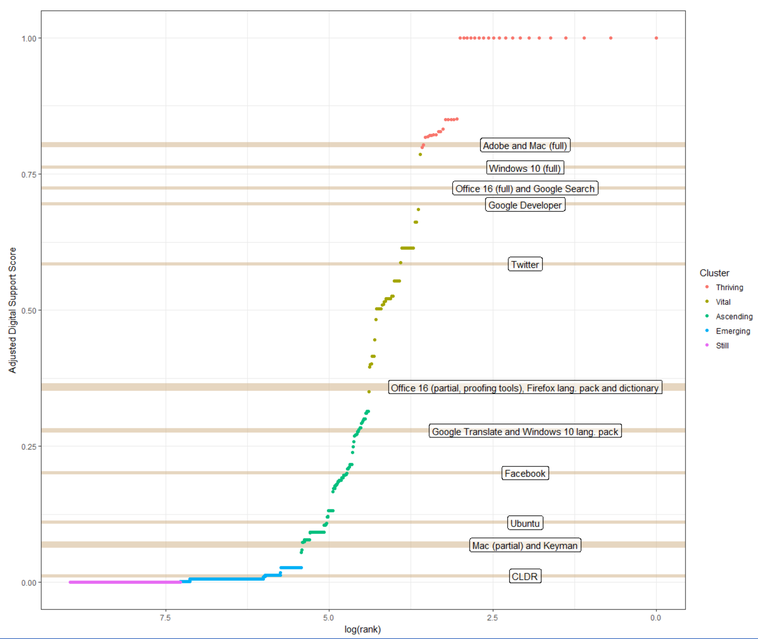




Item response function plots show how the probability of having a feature increases when a language has other features.

Results & Conclusions

A strongly homogenous scale	•Having one feature increases likelihood that a language will have another •Tools created by major corporations pattern together more strongly than those by non-profit organizations •Coefficient of Homogeneity = 0.88 (≥ 0.5 indicates a strong scale, according to Molenaar, 2002)
A definitive index	•Certain features are characteristic of clusters •Average silhouette width = 0.85 (> 0.7 indicates a strong structure has been found within the data, according to Struyf, et al, 1997)
A skewed distribution	•Languages per cluster: •Thriving=36 •Vital=45 •Ascending=146 •Emerging=1223 •Still=6353 •Urgent need for development of digital tools for under-supported and digitally endangered) languages
Data-driven methodology	•Easily allows for incorporation of new features •Allows for repeated testing to track digital language support diachronically



Works Cited

Gibson, M. L. (2015). A framework for measuring the presence of minority languages in cyberspace. In Linguistic and Cultural Diversity in Cyberspace. Proceedings of the 3rd International Conference (Yakutsk, Russian Federation, 30 June–3 July, 2014). Moscow: Interregional Library Cooperation Centre. 61–70.
Hilbert, M. (2011). The end justifies the definition: The manifold outlooks on the digital divide and their practical usefulness for policy-making. Telecommunications Policy 35(8), 715–736.
Kornai, A. (2013). Digital language death. PLoS ONE 8(10), e77056.
Kornai, A. (2015). A new method of language vitality assessment. In Linguistic and Cultural Diversity in Cyberspace. Proceedings of the 3rd International Conference (Yakutsk, Russian Federation, 30 June – 3 July, 2014). Moscow: Interregional Library Cooperation Centre. 132-138.
Mikami, Y. (2008). Digital language divide: Measuring linguistic diversity on the Internet. Presentation at the UNESCO/UNU Conference on Globalization and Languages: Building on our Rich Heritage, Tokyo, Japan, 27 – 28 August 2008.
Sijtsma, K. and I. W. Molenaar. (2002). Introduction to nonparametric item response theory. Thousand Oaks, CA: Sage.
Soria, C. (2016). What is digital language diversity and why should we care? In J. Cru (ed.), Digital media and language revitalisation. Linguapax Review 2016, 13–28.
Struyf, A., M. Hubert, and P. Rousseeuw. (1997). Clustering in an object-oriented environment. Journal of Statistical Software, 1(4), 1–30.