THE LOW-COMPLEXITY-BELT: EVIDENCE FOR LARGE-SCALE LANGUAGE CONTACT IN HUMAN PREHISTORY?

CHRISTIAN BENTZ

Department of Linguistics, University of Tübingen
Tübingen, Germany

Department of Theoretical and Applied Linguistics, University of Cambridge
Cambridge, United Kingdom
chris@christianbentz.de

The quantitative measurement of language complexity has witnessed a recent rise of interest, not least because language complexities reflect the learning constraints and pressures that shape languages over historical and evolutionary time. Here, an information-theoretic account of measuring language complexity is presented. Based on the entropy of word frequency distributions in parallel text samples, the complexities of overall 646 languages are estimated. A large-scale finding of this analysis is that languages just above the equator exhibit lower complexity than languages further away from the equator. This geo-spatial pattern is here referred to as the *Low-Complexity-Belt (LCB)*. The statistical significance of the positive latitude/complexity relationship is assessed in a linear regression and a linear mixed-effects regression, suggesting that the pattern holds *between* different families and areas, but not *within* different families and areas. The lack of systematic within-family effects is taken as potential evidence for a phylogenetically "deep" explanation. The pressures shaping language complexities probably pre-date the expansion of language families from their proto-languages. Large-scale prehistoric contact around the equator is tentatively given as a possible factor involved in the evolution of the LCB.

1. Introduction

Languages are cultural "tools" shaped to successfully transmit information. Due to different pathways and pressures of cultural evolution, they can differ widely with regards to their exact structural characteristics. In this context, there has been a rise of interest in the description, measurement and modelling of language complexity (Sampson, Gil, & Trudgill, 2009; Dahl, 2004; Newmeyer & Preston, 2014; Trudgill, 2011; Baerman, Brown, Corbett, et al., 2015).

This contribution focuses on information-theoretic complexities (Ehret & Szmrecsanyi, in press; Juola, 2008, 1998), and their implications for the evolutionary pressures that have shaped languages. *Information-theoretic complexity* is here defined with reference to the distribution of word tokens over word types - often called *lexical diversity*. It is measured across 1155 parallel texts - i.e. translations of the same content - into 885 different languages (see Section 2).

Imagine a language that uses a single word type over and over again, thus having *minimum* information-theoretic complexity. The word type effectively tells us nothing about the meaning encoded. In contrast, a language using a new word type for any conceivable meaning has *maximal* information-theoretic complexity. That is, every word type is exactly paired with one meaning, and is hence maximally informative.

Note that "complexity" can here be interpreted in two different senses: namely as *learning difficulty* and as *information encoding potential*. A minimum complexity language is extremely easy to learn, but meaningless. A maximum complexity language is hard (or impossible) to learn, but meaningful. The evolutionary trade-off between these two aspects of information encoding has been modelled computationally, and tested experimentally (Kirby, Cornish, & Smith, 2008; Kirby, Tamariz, Cornish, & Smith, 2015; Berdicevskis, 2012; Berdicevskis & Semenuks, in press). Human languages range in between these extremes (Bentz, Verkerk, Kiela, Hill, & Buttery, 2015), falling on a limited spectrum between minimum and maximum complexity. This has far-reaching implications. Minimum and maximum complexities of languages reflect the limits of human learning capacities, and the distribution of complexities across languages gives us a window into the interplay of language learning, usage and linguistic structure on historical and evolutionary timescales.

This study illustrates a systematic geo-spatial pattern relating to information-theoretic complexities across languages of the world. Namely, languages close to the equator have systematically lower information theoretic complexity than languages further away from the equator - given constant content of texts. This phenomenon is called the *Low-Complexity-Belt* (LCB), and is illustrated in Section 3.1. Its statistical significance is tested in Section 3.2. Moreover, it is shown that though the pattern holds *between* language families and areas, there are differences *within* families and areas (Section 3.3).

Finally, it is argued that the presence of *between-family* correlations - and the absence of reliable *within-family* correlations - suggest that the LCB is a phenomenon with a "deep" phylogenetic explanation. Prehistoric *language contact* is given as a promising candidate for explaining the evolution of the LCB (Section 4).

2. Materials and methods

2.1. *Parallel corpora*

The parallel corpora used here come from the *Universal Declaration of Human Rights* (UDHR) in unicode,[a] the *Parallel Bible Corpus* (PBC),[b] and the *European*

[a] hp://www.unicode.org/udhr/

[b] (Mayer & Cysouw, 2014), http://paralleltext.info/data/

Parliament Corpus (EPC).[c] These add up to an overall sample of around 200 million words, 1529 texts, and 1050 languages (i.e. unique ISO-639-3 codes).

Each text is tokenized by using an algorithm that splits strings of unicode characters on non-alphanumeric characters (i.e. white spaces, punctuation, special characters, etc.).[d] The resulting tokens are then added up to the frequency per unique type. For example, the word type *right* occurs 33 times in the English UDHR. Note that this process does not involve lemmatization or stemming, i.e. *right* and *rights* are counted as two separate types here.

2.2. *Estimating entropies*

For each text the information-theoretic complexity is then calculated as the entropy of the distribution of word tokens over word types. The classic Shannon entropy (Shannon & Weaver, 1949) is defined as

$$H = -K \sum_{i=1}^{r} p_i \, log_2(p_i). \tag{1}$$

Where K is a positive constant determining the unit of measurement (which is bits for K=1 and log to the base 2), r is the number of ranks (or different word types) in a word frequency distribution, and p_i is the probability of occurrence of a word of i^{th} rank. According to the maximum likelihood account, the probability p_i is simply the frequency of a type divided by the overall number of tokens in a text. However, it has been shown that the maximum likelihood method is somewhat unreliable, especially for small texts (Hausser & Strimmer, 2009; Nemenman, Shafee, & Bialek, 2001). To estimate entropies reliably, the *James-Stein shrinkage* estimator (Hausser & Strimmer, 2009) is used here instead.

Moreover, texts are taken from three different corpora (UDHR, PBC, EPC) with vastly differing average numbers of tokens (ca. 2K, ca. 10K, ca. 7M), which can additionally bias the estimation of entropy values. To reduce this bias, entropy values are centered and scaled per corpus.

Finally, information on latitudes and longitudes per language, as well as information on language stocks (i.e. language families) and language areas, is taken from the AUTOTYP database (Bickel & Nichols, 1999). Merging the scaled entropy values per language (i.e. ISO code) with AUTOTYP information reduces the sample to 1422 texts of 646 languages.

---

[c] (Koehn, 2005), http://www.statmt.org/europarl/

[d] Note that in the PBC - due to careful automated processing and manual double-checking - word types are reliably delimited by white spaces. This makes tokenization fairly robust across many different scripts. The UDHR and the EPC texts have not yet been pre-processed this way. This means they are more prone to errors when splitting strings of characters into word types, especially when problematic characters such as the apostrophe or tone numbers are involved.

3. Results

3.1. *The Low-Complexity-Belt*

In Figure 1 entropy values are plotted on a world map using the latitudes and longitudes from the AUTOTYP database. In a) the size of dots reflects entropy values, and their colour reflects family membership. Visual inspection reveals that texts of languages located just above the equator ($0\,^\circ$ to ca. $30\,^\circ\,north$) are systematically represented by smaller dots, i.e. lower entropy. This is even more apparent in b), where the longitude is replaced by scaled entropy (and all dots are of the same size now). A loess smoother (black line) again indicates that texts of languages falling on the "belt" between the equator and a latitude of $30\,^\circ\,north$ have systematically lower entropies, with the lowest point at $15\,^\circ\,north$. The statistical significance of this pattern is assessed in the following subsections.

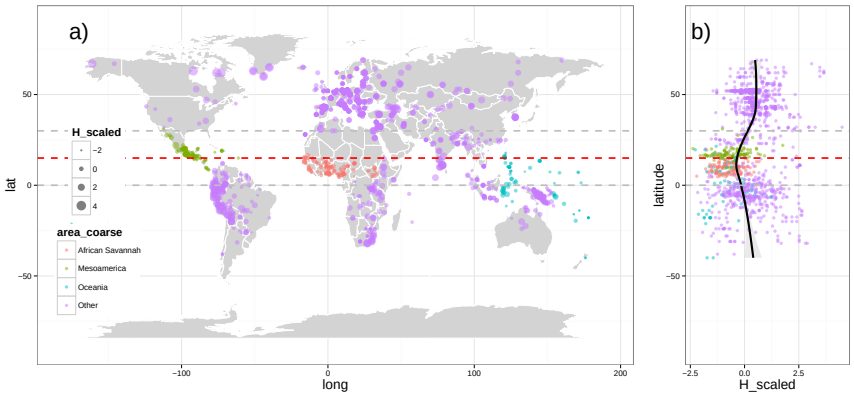

Figure 1. The LCB on a world map. a) World map with scaled entropy values (dots) for 1422 texts (646 languages). The size of dots reflects the scaled entropy. The colour of dots reflects selected language areas: African-Savannah (red), Mesoamerica (green), Oceania (blue), and all others (purple). The LCB is indicated by grey dashed lines at a latitude of $30\,^\circ\,north$ and $0\,^\circ$ (i.e. the equator). The core belt is at around $15\,^\circ\,north$ (red dashed line). b) Cross-section with the x-axis reflecting scaled entropy values, and the y-axis representing latitude. A loess smoother with confidence intervals (black line with grey areas) is overlaid to illustrate the systematically lower entropy values around $15\,^\circ\,north$.

3.2. *Simple regressions*

3.2.1. *Individual languages*

If the LCB is an empirical phenomenon that does not derive from random fluctuations in entropies, then the distance from the core of the belt should be a significant

predictor of entropy values: bigger distance from latitude $15\,^{\circ}\,north$ should predict higher entropies.

This is tested in a simple linear regression model run in R (R Core Team, 2013), with scaled entropies per text as dependent variable, and distance from $15\,^{\circ}\,north$ as predictor variable. Homoscedasticity, linearity and normality of residuals are checked visually. In this model the positive association is highly significant ($\beta = 0.023, p < 2.2e^{-16}, R^2 = 0.10$),[e] with distance from $15\,^{\circ}\,north$ explaining 10% of the variance in scaled entropies (see also Figure 2).

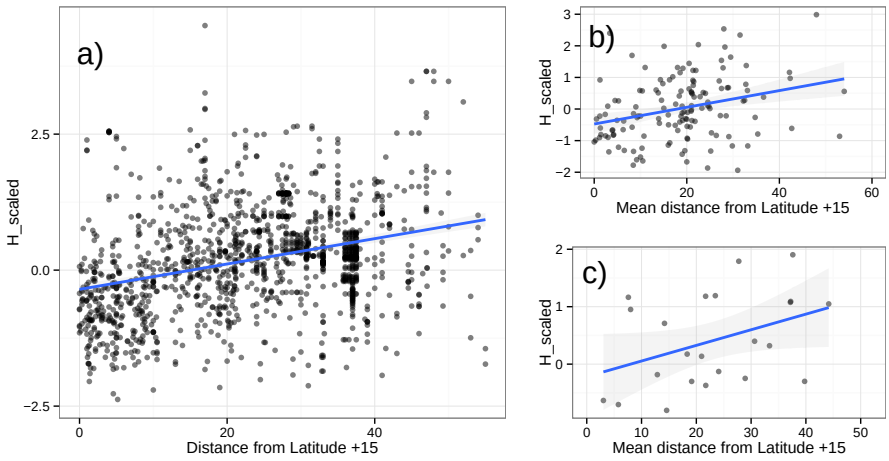

Figure 2. Entropy and distance from the LCB. a) Scaled entropy values (y-axis) for 1422 texts (646 languages) as a function of the distance from $15\,^{\circ}\,north$ (x-axis). Note that 0 does not indicate the equator here, but distance from $15\,^{\circ}\,north$. The positive trend is indicated by a linear regression line with 95% confidence intervals (blue line with grey areas). b) Mean scaled entropy values and mean distances for 140 families. c) Mean scaled entropy values and mean distances for 23 areas.

3.2.2. *Mean entropies per family and area*

The significant positive association might be driven by specific language families and areas, rather than being a pattern holding across different families and areas. A way to test this is to use mean entropy and mean distance values per family and area, rather than individual languages. This method is illustrated in Figure 2, panels b) and c).

[e] $p < 2.2e^{-16}$ is the smallest p-value that R reports, i.e. effectively 0.

This time, two simple linear regression models are fitted, with mean entropies per family and area as dependent variables, as well as mean distance from $15^\circ\, north$ per family and area as predictors. The β-coefficients for both regressions per families ($\beta = 0.026$) and areas ($\beta = 0.027$) are very similar to the original one (0.023). Note that only for families the positive coefficient is significant ($p = 0.0004, R^2 = 0.08$), for areas it is not ($p = 0.06, R^2 = 0.12$). The non-significance of this p-value is certainly related to the drastic reduction of sample size from originally 1422 texts to just 23 areas. However, the positive β-coefficients still indicate that the pattern holds both across different families and across different areas.

3.3. *Mixed-effects regression*

If the positive association between distance from the core of the LCB and scaled entropies holds *between* different families and areas, does it also hold *within* different families and areas? To further assess this, we can fit linear mixed-effects models (Baayen, Davidson, & Bates, 2008; Jaeger, Graff, Croft, & Pontillo, 2011; Winter, 2013) with distance from $15^\circ\, north$ as fixed effect, and family, area, text type and ISO code[f] as random effects.

A "maximal" model according to Barr, Levy, Scheepers, and Tily (2013) is fitted with package *lme4* (Bates, Maechler, & Bolker, 2012) in *R*. This is a model with random slopes and intercepts per family, area and text type, and random intercepts for ISO codes.[g] Again, linearity, homoscedasticity and normality of residuals are checked visually.

It turns out in a likelihood ratio test that this model is not significantly better than a null model without the fixed effect (distance from the LCB) ($\chi^2(13) = 2.45, p = 0.12$). This means that when adjusting for idiosyncratic variation within families, areas, text types and languages, the positive association between distance from $15^\circ\, north$ and scaled entropy vanishes. In other words, though this association holds *between* families and areas, it does not hold *within* families and areas.[h]

A visual way of illustrating this is to plot data points for families and areas separately, as seen in Figure 3. Here, it is apparent that though the positive relationship holds for Benue-Congo languages, it does not hold - and is even inverted - for Austronesian and Indo-European languages.

---

[f] "Text type" here refers to whether the text is from the PBC, UDHR or EPC. ISO codes have to be included as random effects since there are sometimes multiple texts per ISO code and it is conceivable that there is within-language variation.

[g] For ISO codes only random intercepts make sense, since entropy values can only vary for constant distances, and distances can only vary for constant entropies

[h] Note that this is not due to different slopes (i.e. coefficients) per text type, since they are all positive if we look at the PBC, UDHR and EPC separately.

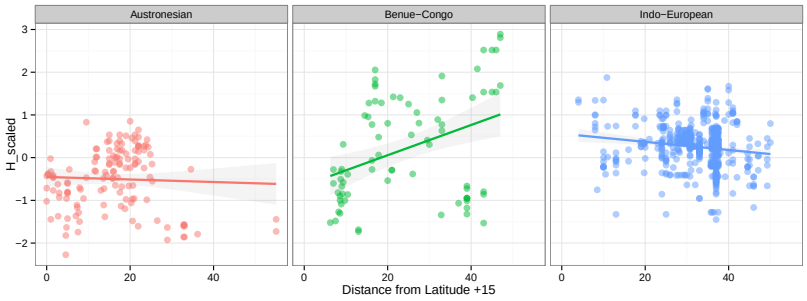

Figure 3. The entropy/latitude relationship for different language families. Scatterplots facetted by the biggest language families (with more than 50 members): Austronesian (red), Benue-Congo (green), and Indo-European (blue). Linear regression lines with 95% confidence intervals (lines with transparent areas) are overlaid.

4. Discussion

Languages falling inside the *Low-Complexity-Belt*, spanning an area from the equator to ca. $30^\circ\, north$, have significantly lower entropies than languages north and south of them. This pattern is strongly significant in a simple regression model across 1422 texts and 646 languages, and it also holds (with reduced significance) for average values of 140 language families and 23 areas. However, it does not hold if slopes and intercepts per families, areas, texts and languages (random effects) are adjusted. Hence, whatever *causally* explains the positive relationship between latitude and information-theoretic complexity, it is an effect that seems to work at the between-family and between-area level, but is strongly weakened at the within-family and within-area level.

A possible explanation for this could be that the effect had an impact in prehistory on proto-languages of modern day language families, before they started to fan out into different branches and wider areas, explaining the between-family and between-area variation. As the effect started to cease or change in recent history, it left no systematic traces at the within-family and within-area level.

4.1. *Language contact*

A potential effect on complexity that has been proposed in the literature is the proportion of non-native adults (L2 speakers) learning a language, i.e. *language contact*. Lupyan and Dale (2010) illustrated that morphological complexity is linked with population sizes in a sample of more than 2000 languages. Population size was here taken as an approximation for language contact. Bentz and Winter (2013) tested this hypothesis more explicitely with regards to nominal case morphology and L2 speaker ratios. Furthermore, a direct link between L2 speaker proportions and entropy (as lexical diversity measure) was established recently

(Bentz et al., 2015). Namely, languages with higher proportions of L2 speakers tend to be those with lower entropies. Potential mechanisms of entropy reduction by means of imperfect learning were elicited in a series of iterated learning experiments (Kirby et al., 2008, 2015; Berdicevskis, 2012; Berdicevskis & Semenuks, in press).

4.2. *Deep phylogenetic signals of complexity*

Based on these findings, it is conceivable that areas and families that contribute most to the LCB are those that had the biggest potential in terms of language contact in human prehistory. This makes sense, for instance, for the Benue-Congo family in the African Savannah (and South Africa). It is known as a "deep" family, with migrations and language contact in its early history, such as the Bantu expansion 3000 BC (Pereltsvaig, 2012, p.118). It might be worth considering similar scenarios for languages in Mesoamerica and Oceania. Interestingly, it was shown that entropies in Bantu languages, as well as Austronesian and Indo-European languages, have relatively strong "phylogenetic signals", meaning that they follow closely the evolution reconstructed from cognate data (Bentz et al., 2015). In other words, entropies of extant languages are "conservative". They reflect the situation of the past, going back to the roots of the language families several thousand years ago. This suggests that the pressures of the deep phylogenetic past - such as early language contact - might still be reflected in language complexities of today, even if the pressure has ceased to be relevant in recent history.

5. Conclusions

Languages tend to have lower information-theoretic complexity closer to the equator ($15\,^{\circ}\,north$ more precisely). This pattern is statistically strongly significant, and requires explanation. The differences in significance, relating to variation within and between families and areas, suggest that the effect causing this pattern might work at deep timescales. Language contact was here proposed as a possible explanation. It is attested as a factor driving lower morphological complexity and lower information-theoretic complexity in large-scale statistical studies, and tested in the lab via iterated learning experiments. However, further studies are necessary to illustrate comprehensively the link between latitude and higher language contact at different time depths. If this link is confirmed, it would constitute important evidence for explaining the evolution and diversity of human languages.

Acknowledgements

This project was funded by an Arts and Humanities Research Council (UK) doctoral grant and Cambridge Assessment (reference number: RG 69405), as well as a grant from the Cambridge Home and European Scholarship Scheme. At a later stage, this work was also supported by the EVOLAEMP project and the *Words, Bones, Genes, Tools* project at the University of Tübingen.

References

Baayen, H. R., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. doi: 10.1016/j.jml.2007.12.005

Baerman, M., Brown, D., Corbett, G. G., et al. (2015). *Understanding and measuring morphological complexity*. Oxford University Press.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. doi: 10.1016/j.jml.2012.11.001

Bates, D., Maechler, M., & Bolker, B. (2012). *lme4: Linear mixed-effects models using S4 classes.* Retrieved from `http://cran.r-project.org/package=lme4`

Bentz, C., Verkerk, A., Kiela, D., Hill, F., & Buttery, P. (2015). Adaptive communication: Languages with more non-native speakers tend to have fewer word forms. *PLoS ONE*, *10*(6), e0128254. doi: 10.1371/journal.pone.0128254

Bentz, C., & Winter, B. (2013). Languages with more second language speakers tend to lose nominal case. *Language Dynamics and Change*, *3*, 1–27. doi: 10.1163/22105832-13030105

Berdicevskis, A. (2012). Introducing pressure for expressivity into language evolution experiments. In T. C. Scott-Phillips, M. Tamariz, E. A. Cartmill, & J. R. Hurford (Eds.), *The Evolution of Language. Proceedings of the 9th International Conference (EVOLANG9).* Singapore: World Scientific.

Berdicevskis, A., & Semenuks, A. (in press). Non-native speakers wreak havoc, native speakers clean it up: experimental modeling of contact-induced language simplification.

Bickel, B., & Nichols, J. (1999). *The AUTOTYP database.* Retrieved from `http://www.autotyp.uzh.ch/`

Dahl, Ö. (2004). *The growth and maintenance of linguistic complexity*. John Benjamins Publishing.

Ehret, K., & Szmrecsanyi, B. (in press). An information-theoretic approach to assess linguistic complexity. In R. Baechler & G. Seiler (Eds.), *Complexity and isolation.* Berlin: de Gruyter.

Hausser, J., & Strimmer, K. (2009). Entropy inference and the james-stein estimator, with application to nonlinear gene association networks. *The Journal of Machine Learning Research*, *10*, 1469–1484.

Jaeger, T. F., Graff, P., Croft, W., & Pontillo, D. (2011). Mixed effect models for genetic and areal dependencies in linguistic typology. *Linguistic Typology*, *15*, 281–320. doi: 10.1515/LITY.2011.021

Juola, P. (1998). Measuring linguistic complexity: The morphological tier. *Journal of Quantitative Linguistics*, *5*(3), 206–213.

Juola, P. (2008). Assessing linguistic complexity. In M. Miestamo, K. Sinnemäki,

& F. Karlsson (Eds.), *Language complexity: typology, contact, change* (pp. 89–108). Amsterdam: John Benjamins.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, *105*(31), 10681–10686.

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, *141*, 87–102.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Mt summit* (Vol. 5, pp. 79–86).

Lupyan, G., & Dale, R. (2010, January). Language structure is partly determined by social structure. *PloS ONE*, *5*(1), e8559.

Mayer, T., & Cysouw, M. (2014). Creating a massively parallel bible corpus. In N. Calzolari et al. (Eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014* (pp. 3158–3163). European Language Resources Association (ELRA).

Nemenman, I., Shafee, F., & Bialek, W. (2001). Entropy and inference, revisited. *arXiv preprint*, physics/0108025.

Newmeyer, F. J., & Preston, L. B. (2014). *Measuring grammatical complexity*. Oxford University Press.

Pereltsvaig, A. (2012). *Languages of the world: an introduction*. Cambridge University Press.

R Core Team. (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from `http://www.r-project.org/` doi: ISBN3-900051-07-0

Sampson, G., Gil, D., & Trudgill, P. (2009). *Language complexity as an evolving variable*. Oxford University Press.

Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication*. Urbana: The University of Illinois Press.

Trudgill, P. (2011). *Sociolinguistic typology: social determinants of linguistic complexity*. Oxford: Oxford University Press.

Winter, B. (2013). Linear models and linear mixed effects models in R with linguistic applications. *arXiv preprint*, 1308.5499.