GÖTEBORGS UNIVERSITET

Språkbanken

CLT

Taraka Rama

N-gram approaches to the historical dynamics of basic vocabulary

Taraka Rama

Språkbanken

University of Gothenburg

ESSLLI 2012

GÖTEBORGS UNIVERSITET

Språkbanken

CLT

Taraka Rama

Outline

Introduction I

GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

Taraka Rama

- Recent years has seen a surge in the number of papers in computational historical linguistics (CHL).
- Availability of huge datasets has attracted researchers from diverse fields such as particle physics, biology.
- Physicists invaded the field of historical linguistics *en masse* (Schulze et al. 2008)

Introduction II

Based on the type of datasets and methods, recent work in CHL could be classified into (Nichols & Warnow 2008):

- Based on typological data.
- Based on lexical data.
- Distances computed using some form of lexical similarity or vector similarity.
- Trees inferred using Parametric methods such as Maximum Likelihood, Bayesian Inference.
- Latest methods are based on Networks than trees.

Introduction III

- Item stability is defined as the degree of resistance of an item to lexical replacement over time.
- Holman et al. (2008) note that words for stable items yield higher number of cognates than the words for less stable items in closely related languages.

GÖTEBORGS UNIVERSITET

Språkbanken

CLT

Taraka Rama

Introduction IV

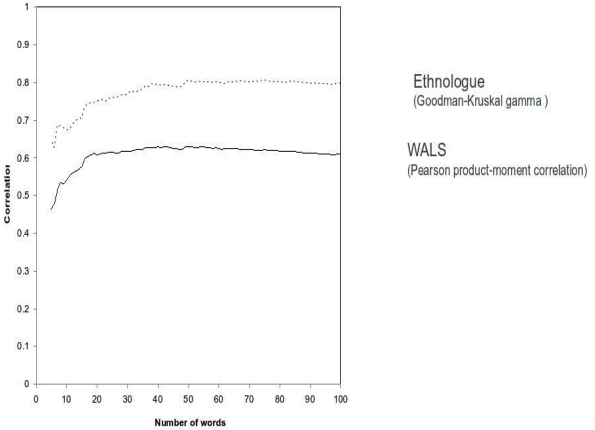

Figure: Correlations with WALS and Ethnologue

GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

Taraka Rama

Outline

- Dunning (1994) motivate the use of character *n*-grams for automatic language identification as well as computation of inter-language distances.
- Huffman & Mentor-Loritz (1998) use vector similarity measures for computing the inter-language distances for Mayan family.

- Singh & Surana (2007) use character *n*-grams extracted from raw corpora of ten languages from the Indian subcontinent for computing the pair-wise language distances among languages from two different language families (Indo-Aryan and Dravidian).
- Holman et al. (2008) defined a measure based on phonological matches to rank the items in a 100–item Swadesh list.

GÖTEBORGS UNIVERSITET

Språkbanken

CLT

Taraka Rama

Outline

GÖTEBORGS UNIVERSITET
Språkbanken
CLT

Taraka Rama

ASJP database

- A much larger sample of languages, 3000+ languages
- Around half of the world's languages
- 109 out of the world's 121 linguistic families
- 47 out of 123 isolates
- 40 out of 122 creoles, mixed languages, and pidgins

All the above language classifications are based on *Ethnologue*

- Word list admitted if and only if it has 70% of the entries

ASJP code I

GÖTEBORGS UNIVERSITET
Språk-BANKEN
CLT

Taraka Rama

- ASJP code is a simple code using QWERTY keyboard
  1. 34 symbols for consonants
  2. 7 symbols for vowels
  3. Two modifiers $\sim$ and \$for combining the previous segments
  4. ” indicates glottalization
- For instance, “kwy ” is a labialized velar with a palatal offglide

ASJP code II

GÖTEBORGS UNIVERSITET

Språkbanken

CLT

Taraka Rama

BLOOD, BONE, BREAST, COME, DIE, DOG, DRINK, EAR, EYE, FIRE, FISH, FULL, HAND, HEAR, HORN, I, KNEE, LEAF, LIVER, LOUSE, MOUNTAIN, NAME, NEW, NIGHT, NOSE, ONE, PATH, PERSON, SEE, SKIN, STAR, STONE, SUN, TONGUE, TOOTH, TREE, TWO, WATER, WE, YOU (SG).

GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

Taraka Rama

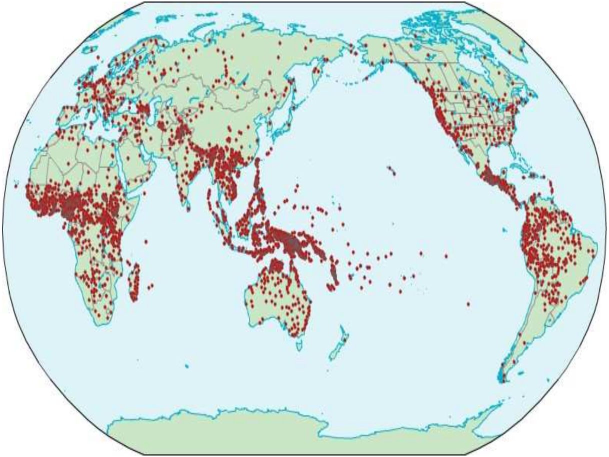

Figure: Language distribution across world

GÖTEBORGS UNIVERSITET
Språkbanken
CLT
Taraka Rama

Outline

SR as proxy I

- Confirm the validity of using segments extracted from the word list (SR)
- Match the UPSID (Maddieson & Precoda 1990) segment inventory sizes for 392 (out of 451) languages against SR
- The mean of UPSID/SR is .818 with s.d = .188

SR as proxy II

- Each UPSID language is matched to ASJP language(s) list based on the following criterion:
  1. Both should pertain to the same geographical dialect
  2. Have similar names
  3. If UPSID covers several word lists in ASJP list, then the ASJP SR is represented by the mean SR of the several ASJP lists

SR as proxy III

GÖTEBORGS UNIVERSITET

Språkbanken

CLT

Taraka Rama

- One might assume that a larger list allows us to represent better all the phonological segments
- The average length of word list is 35.7 for 3168 languages
- Very small correlation, $r = .17$ between the number of words attested and SR
- Very small correlation, $r = -.05$ between word list size and UPSID/SR
- Further, loanwords are excluded for exclusing the rare phonemes

GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

Taraka Rama

SR as proxy IV

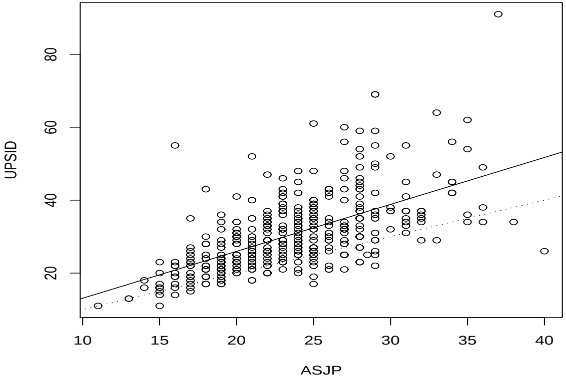

Figure: Pearson's $r = .61$

GÖTEBORGS UNIVERSITET

Språkbanken

CLT

Taraka Rama

Outline

Phoneme N-grams

- *n*-grams defined over the Swadesh list of a language *L*.
- Sample Space $\Omega = \{\phi | \phi \text{ is a phoneme}\}$
- Phoneme *n*-gram $P \in \Omega_n = \overbrace{\Omega \times \Omega \times \ldots \times \Omega}^{n}$
- Phoneme *N*-gram model for a language *L*, $M_p^l : \Omega_N \rightarrow \mathbb{R}$
- $\Omega_N = \bigcup_{i=1}^{N} \Omega_i$
- Relative frequency or an exponential estimator could be used for computing the above model.
- Size of a *N*-gram model is defined as $|\Omega_N|$.

GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

Taraka Rama

Outline

Idea I

- Related to the idea that words for highly stable items yield phonologically similar cognates.
- Cognate words for an item tend to be phonologically more similar.

GÖTEBORGS UNIVERSITET
Språk-BANKEN
CLT
Taraka Rama

Idea II

GÖTEBORGS UNIVERSITET

Språkbanken

CLT

Taraka Rama

- When words for a item are distributed across multiple unrelated cognate classes.
- The cognate classes for such an item would naturally share lesser number of phoneme *n*-grams than an item with fewer number of cognate classes.

Idea III

- A simple information theoretic measure such as self-entropy can be used to measure the amount of phonological divergence in a phoneme *n*-gram profile for a item in a language family.

Computing N-gram frequency

GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

Taraka Rama

$$rf^{i}_{ngram} = \frac{f^{i}_{ngram}}{\sum_{i=1}^{S} f^{i}_{ngram}} \tag{1}$$

$$H^{k}_{item} = -\sum_{i=1}^{S} rf^{i}_{ngram} \cdot log(rf^{i}_{ngram}) \tag{2}$$

GÖTEBORGS UNIVERSITET
Språkbanken
CLT
Taraka Rama

Outline

Properties of Phoneme Models I

- Rank of the *N*-grams follow a Zipfian distribution.
- Each profile is a signature of the family/language.
- The size of the N-gram model vs the rank of family follows a Zipfian distribution.

Properties of Phoneme Models II

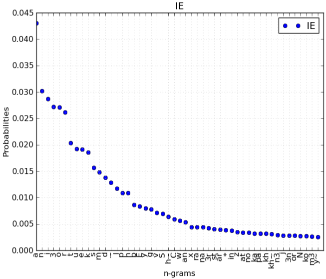

Figure: Indo-European

Properties of Phoneme Models III

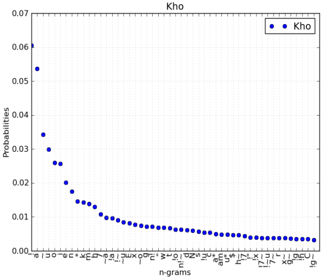

Figure: Khoisan

GÖTEBORGS UNIVERSITET
Språkbanken
CLT
Taraka Rama

Properties of Phoneme Models IV

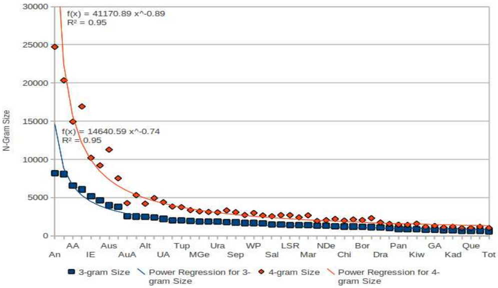

Figure: Power Law for the size vs rank for WALS families.

Outline

GÖTEBORGS UNIVERSITET

Språkbanken

CLT

Taraka Rama

Introduction
Related Work
ASJP
SR as proxy
Definitions
Method
Properties
Results
Acknowledgements
References

GÖTEBORGS UNIVERSITET

Språkbanken

CLT

Taraka Rama

Results I

Meaning	# in ASJP list	Stability $exp(H(\cdot))$
I	1	1717.3609
you	2	2134.5054
water	75	2150.416
horn	34	2323.5502
louse	22	2735.9681
hand	48	2837.8896
tree	23	2868.8678
we	3	2927.731
name	100	2940.973
drink	54	2998.3115
bone	31	3066.0844
fire	82	3084.6197
liver	53	3098.0558
person	18	3128.8495
tooth	43	3189.1238
eye	40	3202.9192
die	61	3267.3181
path	85	3371.6788
come	66	3429.0297
two	12	3431.9033

GÖTEBORGS UNIVERSITET

Språkbanken

CLT

Taraka Rama

Results II

Meaning	# in ASJP list	Stability $exp(H(\cdot))$
new	96	3435.0336
nose	41	3446.6322
breast	51	3458.9689
tongue	44	3500.0106
blood	30	3505.9971
stone	77	3567.2699
sun	72	3683.9486
dog	21	3693.7477
fish	19	3700.0209
one	11	3820.584
leaf	25	3834.6073
full	95	3857.6387
ear	39	3884.9767
skin	28	3887.211
mountain	86	4298.8018
hear	58	4429.0253
see	57	4449.0301
night	92	4549.2087
star	74	4754.1568
knee	47	4967.5705

Results III

GÖTEBORGS UNIVERSITET

Språkbanken

CLT

Taraka Rama

- ρ between the ranks given in Table 1 and the ranks given in Holman et al. (2008) is 0.35 ($p = 0.028$).
- The inter-hemisphere correlation ρ is 0.41, which is in the range of 0.37 reported by Holman et al. (2008).
- ρ between the item stability rank of Holman et al. (2008) and that of self-entropy, for 100-items list is 0.61 and is significant at the level of 0.01.
- The 40-item list given by the self-entropy method and that of Holman et al. (2008) has 28 items in common.

Outline

Acknowledgements

The author would like to thank:

- Søren Wichmann for stimulating discussions and comments on the paper.
- Swedish Graduate School in Language Technology (GSLT) for the financial assistance.

Outline

Taraka Rama

Introduction
Related Work
ASJP
SR as proxy
Definitions
Method
Properties
Results
Acknowledgements
References

GÖTEBORGS UNIVERSITET

Språkbanken

CLT

Taraka Rama

References

Dunning, T. (1994), Statistical identification of language, Technical Report CRL MCCS-94-273, Computing Research Lab, New Mexico State University.

Holman, E. W., Wichmann, S., Brown, C. H., Velupillai, V., Müller, A. & Bakker, D. (2008), 'Explorations in automated language classification', *Folia Linguistica* **42**(3-4), 331–354.

Huffman, S. & Mentor-Loritz, D. (1998), *The Genetic Classification of Languages by n-Gram Analysis: A Computational Technique*, Georgetown University.

Maddieson, I. & Precoda, K. (1990), 'UPSID-PC', *The UCLA Phonological Segment Inventory Database* .

Nichols, J. & Warnow, T. (2008), 'Tutorial on computational linguistic phylogeny', *Language and Linguistics Compass* **2**(5), 760–820.

Schulze, C., Stauffer, D. & Wichmann, S. (2008), 'Birth, survival and death of languages by Monte Carlo simulation', *Communications in Computational Physics* **3**(2).

Singh, A. & Surana, H. (2007), 'Can corpus based measures be used for comparative study of languages?', *Proceedings of Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology* pp. 40–47.