N-gram approaches to the historical dynamics of basic vocabulary

Taraka Rama

Språkbanken
University of Gothenburg

ESSLLI 2012
Outline

Introduction

Related Work

ASJP

SR as proxy

Definitions

Method

Properties

Results

Acknowledgements

References
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)
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.
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.
Figure: Correlations with WALS and Ethnologue
Outline

Introduction

Related Work

ASJP

SR as proxy

Definitions

Method

Properties

Results

Acknowledgements

References
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.
Outline

Introduction

Related Work

ASJP

SR as proxy

Definitions

Method

Properties

Results

Acknowledgements

References
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 is a simple code using QWERTY keyboard

1. 34 symbols for consonants
2. 7 symbols for vowels
3. Two modifiers ~ and $ for combining the previous segments
4. ” indicates glottalization

For instance, “kwy” is a labialized velar with a palatal offglide
ASJP code II

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).
Figure: Language distribution across world
Outline

Introduction

Related Work

ASJP

SR as proxy

Definitions

Method

Properties

Results

Acknowledgements

References
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

- 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 excluding the rare phonemes.
SR as proxy IV

Figure: Pearson’s $r = .61$
Outline

Introduction
Related Work
ASJP
SR as proxy
Definitions
Method
Properties
Results
Acknowledgements
References
Phoneme N-grams

- $n$-grams defined over the Swadesh list of a language $L$.
- Sample Space $\Omega = \{\phi | \phi$ is a phoneme$\}^n$.
- Phoneme $n$-gram $P \in \Omega_n = \Omega \times \Omega \times \ldots \times \Omega^n$.
- 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 $|\Omega_N|$.
Outline

Introduction

Related Work

ASJP

SR as proxy

Definitions

Method

Properties

Results

Acknowledgements

References
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.
Idea II

- 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.
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 an item in a language family.
Computing $N$-gram frequency

$$rf_{ngram}^i = \frac{f_{ngram}^i}{\sum_{i=1}^{S} f_{ngram}^i} \quad (1)$$

$$H_{item}^k = -\sum_{i=1}^{S} rf_{ngram}^i \cdot \log(rf_{ngram}^i) \quad (2)$$
Outline

Introduction

Related Work

ASJP

SR as proxy

Definitions

Method

Properties

Results

Acknowledgements

References
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
Figure: Power Law for the size vs rank for WALS families.
## Results I

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

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

- $\rho$ between the ranks given in Table 1 and the ranks given in Holman et al. (2008) is 0.35 ($\rho = 0.028$).

- The inter-hemisphere correlation $\rho$ is 0.41, which is in the range of 0.37 reported by Holman et al. (2008).

- $\rho$ 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.
<table>
<thead>
<tr>
<th>Outline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
</tr>
<tr>
<td>Related Work</td>
</tr>
<tr>
<td>ASJP</td>
</tr>
<tr>
<td>SR as proxy</td>
</tr>
<tr>
<td>Definitions</td>
</tr>
<tr>
<td>Method</td>
</tr>
<tr>
<td>Properties</td>
</tr>
<tr>
<td>Results</td>
</tr>
<tr>
<td>Acknowledgements</td>
</tr>
<tr>
<td>References</td>
</tr>
</tbody>
</table>
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.
References


