N-gram approaches to the historical dynamics of basic vocabulary

Taraka Rama

June 18, 2012

1 Introduction and Related Work

In traditional lexicostatistics, distances between languages are determined by human expert cognacy judgements of items in standardized word lists, e.g., the Swadesh lists (Swadesh 1955). Recently, some researchers have turned to approaches more amenable to automation, hoping that large-scale automatic (lexicostatistic) language classification will thus become feasible.

The Automated Similarity Judgement Program (ASJP)[1](Brown et al. 2008), as they call themselves, is a group of scholars who have embarked on an ambitious program of automating the computation of similarities between languages using lexical similarity distances. They have collected Swadesh lists, a short concept list of 100 lexical items (pruned to 40 after Holman et al. 2008*a*), which are supposed to be highly stable and therefore useful for estimating inter-language similarities. The ASJP program computes the distance between two languages as the average pair-wise length-normalized Levenshtein distance, called Levenshtein Distance Normalized (LDN), Levenshtein (1965). LDN is further modified to account for chance resemblance such as accidental phoneme inventory similarity between a pair of languages to yield LDND (Levenshtein Distance Normalized Divided; Holman et al. 2008*a*). Holman et al. (2008*b*) use 100–item word lists from 245 languages to determine the stability of items[2] and to evaluate the effect of the word-list size on automatic language classification by comparing the inter-language distances to the genetic classification given in World Atlas of Language Structures (WALS; Haspelmath et al. 2011) and Ethnologue (Lewis 2009).

This paper considers a different approach from that of ASJP to investigate the individual relationship of phonological similarity with item stability. The approach in this paper is inspired by the work of Cavnar & Trenkle (1994), who use character n-grams for text categorization. Cavnar & Trenkle (1994) observe that the n-grams for a particular document category follows a Zipfian distribution. The rank of a character n-gram varies across documents belonging to different categories. Building upon this work, Dunning (1994) motivates the use of these character n-grams for automatic language identification. Based on this, we can stipulate that languages belonging to a single language family (or genus) have similar phoneme n-gram distributions. In computational historical linguistics, there are at least two earlier works which use character n-grams for computing the pair-wise distances between languages. Huffman & Mentor-Loritz (1998) compute pair-wise language distances based on n-grams extracted from the texts of European and American Indian languages (mostly from the Mayan language family). In another work, 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).

[1] http://email.eva.mpg.de/∼wichmann/ASJPHomePage.htm

[2] Petroni & Serva (2010) apply LDND to two of the world's well-studied language families – Indo-European (Dyen et al. 1992) and Austronesian (Greenhill et al. 2008) – to rank the Swadesh items by their resistance to lexical replacement (stability).

Since Brown et al. (2008), the ASJP database has been going through an expansion, to include more than 5500 word lists representing well over one half of the languages of the world (Wichmann et al. 2011). A natural step would be to apply n-gram analysis to investigate item stability using phoneme n-grams for language families across the world. Interestingly, Wichmann, Rama & Holman (2011) show that the phoneme inventory sizes of 458 of the world's languages (Maddieson & Precoda 1990) have a robust correlation with the number of 1-grams extracted from the corresponding languages in the ASJP database. Given this result, it is reasonable to assume that the phoneme n-grams extracted from ASJP database can be used for investigating item stability.

The paper is structured as followed. Section 2 describes the dataset, a subset of the full ASJP database, used in our experiments. Section 3 describes the method for computing the item stability across language families of the world. In Section 4, we present the results obtained through the application of the method, described in Section 3 and compare our ranking of item stability with the rankings presented in Petroni & Serva (2010) and Holman et al. (2008*b*).

2 Dataset

All the experiments reported in this paper were performed on the subset of version 12 of ASJP database.[3] The database has 4169 word lists from languages that are not only extant but also extinct. The database also contains word lists for pidgins, creoles, mixed languages, artificial languages and, proto–languages. All these languages were excluded from the current study. Among the extinct languages, only those languages were included, which were extinct in the last three centuries. Also, any word list containing less than 28 words (70% of the 40-word list) was not included in the final dataset. Since we use the family names listed in WALS (Haspelmath et al. 2011) classification, any family with less than ten languages is excluded from our experiments. The final dataset has a total of 3730 word lists represented by 49 language families.

3 Method

Item stability is defined as the degree of resistance of an item to lexical replacement over time. In other words, a item is relatively stable when it has not been replaced by a lexical item from the same language or by a borrowed lexical item. Holman et al. (2008*b*) note that words for stable items yield higher number of cognates than the words for less stable items in closely related languages. To this end, Holman et al. (2008*b*) defined a measure – based on the phonological matches between words for a single item for closely related languages (as defined in terms of WALS genera of a family) – to rank items in a 100-item Swadesh list. Our method is closely related to the idea that words for highly stable items yield phonologically similar cognates but differs in the way the phonological similarity is measured. The motivation behind our method is that a phoneme n-gram profile, derived from words of a item across closely related languages (i.e., families) is defined in terms of lesser number of phoneme n-grams than the phoneme n-gram profile of lesser stable items. It is very straightforward to see that cognate words for an item tend to be more similar phonologically than the words for less stable items which have undergone more lexical replacement. One could imagine a scenario where words for a item are distributed across multiple unrelated cognate classes. Such a scenario would yield a larger phoneme n-gram profile, since 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

[3] Available on http://email.eva.mpg.de/ wichmann/listss12.zip

of phonological divergence in a phoneme n-gram profile for a item in a language family. We describe the computation of phoneme n-gram profile and self-entropy in rest of the section.

The phoneme n-gram profile for a language family is computed in the following manner. A phoneme n-gram is defined as the consecutive phonemes in the window of a fixed value of n. The value of n ranges from one to five. All the phoneme 1-grams to 5-grams are extracted for a lexical item in a item-list. All n-grams for a item, extracted from word-lists belonging to a family, are mixed and sorted in the descending order of their frequency. Usually, only the top N n-grams are retained and the rest of them are pruned. For our present experiments, we retain all the n-grams since, pruning rare n-grams would mean a loss of information when computing the phonological divergence within the n-gram profile for a item in a language family. In the next step, the relative frequency of each n-gram in a n-gram profile for an item is computed by normalizing the frequency of a phoneme n-gram by the sum of frequency of all the n-grams in a item's n-gram profile. This can be summarized in (1), where f^i_{ngram} denotes the frequency of the i^{th} n-gram and S denotes the size of the n-gram profile for a item.

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

Given this background, the self-entropy of k^{th} item's n-gram profile can defined as in (2):

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

Since self-entropy $H(\cdot)$ measures the amount of divergence in the phoneme n-gram profile for a item, the items can be ranked relatively in terms of the ascending order of self-entropy averaged across the families. In the next section, we present and discuss the results of our experiments.

4 Results

Figures: 1 and 2 show the frequency-rank plot of the phoneme n-grams for two geographically distant families, Indo-European and Khoisan. The two figures show that the frequency-rank plots of the phoneme n-grams follow a Zipfian distribution, also, the top phoneme n-grams differ in both the families in agreement with the hypothesis of Cavnar & Trenkle (1994).

We perform two experiments on two datasets: 1) on all the 3730 40-word lists representing 49 language families 2) on all the 100-word lists, extracted from the ASJP database, representing 30 language families. Table 1 shows the 40-items ranked in the decreasing order of stability, obtained through the application of self-entropy method. A Spearman's rank correlation ρ between the ranks given in Table 1 and the ranks given in Holman et al. (2008*b*) is 0.35 ($p = 0.028$). Although the correlation is low, it is still significant at a level of 0.05. We further tested if there is a correlation between the stability rankings, given by $H(\cdot)$, computed separately on the language families belonging to the Eastern hemisphere and Western hemisphere. The resulting correlation ρ is 0.41, which is in the range of 0.37 reported by Holman et al. (2008*b*). The item stability ranks derived from 100-word lists is quite similar to the results obtained from 40-word lists. We compare the item stability ranks obtained from 100-word list with that of Petroni & Serva (2010) and Holman et al. (2008*b*). The Spearman's ρ between the item stability rank of Holman et al. (2008*b*) and that of self-entropy is $\rho = 0.61$ and is significant at the level of 0.01. The correlation is robust and significant and suggests that the self-entropy method indeed agrees with the stability ranks given by Holman et al. (2008*b*). Moreover, the

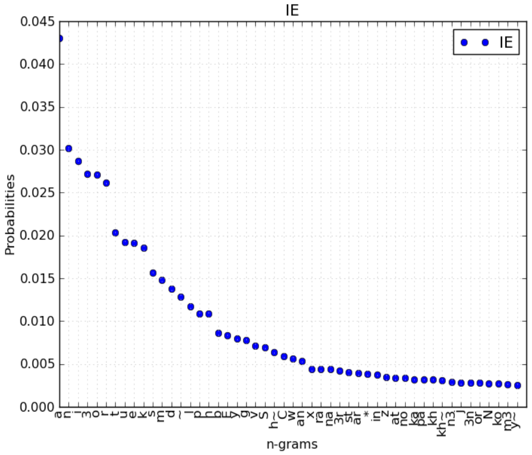

Figure 1: Indo-European

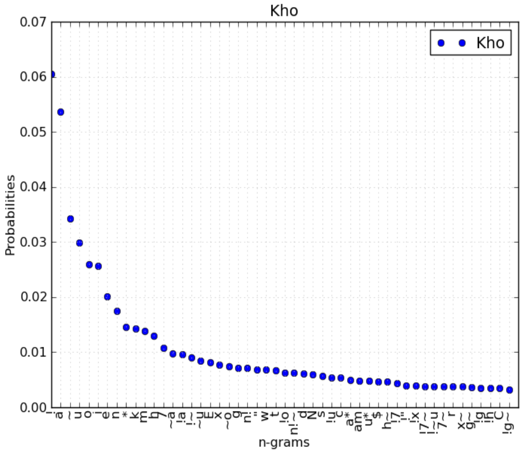

Figure 2: Khoisan

40-item list given by the self-entropy method and that of Holman et al. (2008*b*) has 28 items in common. In another work, Petroni & Serva (2010) rank the 200-item list by applying LDND to the publicly available Indo-European dataset of Dyen et al. (1992). Petroni & Serva (2010) gave a stability ranking for 200-items and found that there is not much improvement in the quality of the inferred tree of Indo-European family after the inclusion of items beyond the 100 most stable items. We compare our stability ranks with that of Petroni & Serva (2010) and find that the expected items such as "I, head, horn, ear and eye" are not present in their 100 most stable items. Nevertheless, Petroni & Serva (2010) has 54 items in common with our method.

Meaning	# in ASJP list	Stability $exp(H(\cdot))$	Meaning	# in ASJP list	Stability $exp(H(\cdot))$
I	1	1717.3609	new	96	3435.0336
you	2	2134.5054	nose	41	3446.6322
water	75	2150.416	breast	51	3458.9689
horn	34	2323.5502	tongue	44	3500.0106
louse	22	2735.9681	blood	30	3505.9971
hand	48	2837.8896	stone	77	3567.2699
tree	23	2868.8678	sun	72	3683.9486
we	3	2927.731	dog	21	3693.7477
name	100	2940.973	fish	19	3700.0209
drink	54	2998.3115	one	11	3820.584
bone	31	3066.0844	leaf	25	3834.6073
fire	82	3084.6197	full	95	3857.6387
liver	53	3098.0558	ear	39	3884.9767
person	18	3128.8495	skin	28	3887.211
tooth	43	3189.1238	mountain	86	4298.8018
eye	40	3202.9192	hear	58	4429.0253
die	61	3267.3181	see	57	4449.0301
path	85	3371.6788	night	92	4549.2087
come	66	3429.0297	star	74	4754.1568
two	12	3431.9033	knee	47	4967.5705

Table 1: Stability ranking for the reduced ASJP 40–item list. The stability value given against each item is the exponentiated $H(\cdot)$ value.

In summary, the item stability ranks derived from *n*-gram analysis largely agrees with the item stability ranks based on phonological matches of Holman et al. (2008*b*). This result suggests that phoneme *n*-grams could be used for investigating the individual relation of phonological similarity with geographical spread, word-list size and, typological similarity.

Acknowledgements

The author would like to thank Søren Wichmann for stimulating discussions and comments on the paper.

References

Brown, C. H., Holman, E. W., Wichmann, S. & Velupillai, V. (2008), ‘Automated classification of the world’ s languages: A description of the method and preliminary results’, *STUF-Language Typology and Universals* **61**(4), 285–308.

Cavnar, W. B. & Trenkle, J. M. (1994), ‘N-gram-based text categorization’, *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval* pp. 161–175.

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

Dyen, I., Kruskal, J. B. & Black, P. (1992), ‘An Indo-European classification: A lexicostatistical experiment’, *Transactions of the American Philosophical Society* **82**(5), 1–132.

Greenhill, S., Blust, R. & Gray, R. (2008), ‘The austronesian basic vocabulary database: from bioinformatics to lexomics’, *Evolutionary Bioinformatics Online* **4**, 271.

Haspelmath, M., Dryer, M. S., Gil, D. & Comrie, B. (2011), *WALS online*, Munich: Max Planck Digital Library. http://wals.info.

Holman, E. W., Wichmann, S., Brown, C. H., Velupillai, V., Müller, A. & Bakker, D. (2008*a*), Advances in automated language classification, *in* A. Arppe, K. Sinnemäki & U. Nikanne, eds, ‘Quantitative Investigations in Theoretical Linguistics’, Helsinki: University of Helsinki, pp. 40–43.

Holman, E. W., Wichmann, S., Brown, C. H., Velupillai, V., Müller, A. & Bakker, D. (2008*b*), ‘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.

Levenshtein, V. (1965), ‘Binary codes capable of correcting spurious insertions and reversals’, *Cybernetics and Control Theory* **10**, 707–710.

Lewis, P. M., ed. (2009), *Ethnologue: Languages of the World*, Sixteenth edn, SIL International, Dallas, TX, USA.

Maddieson, I. & Precoda, K. (1990), ‘N.d.’, *The UCLA Phonological Segment Inventory Database* . http://web.phonetik.uni-frankfurt.de/upsid.html.

Petroni, F. & Serva, M. (2010), 'Lexical evolution rates derived from automated stability measures', *Journal of Statistical Mechanics: Theory and Experiment* **2010**, P03015.

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.

Swadesh, M. (1955), 'Towards greater accuracy in lexicostatistic dating', *International Journal of American Linguistics* **21**(2), 121–137.

Wichmann, S., Müller, A., Velupillai, V., Wett, A., Brown, C. H., Molochieva, Z., Sauppe, S., Holman, E. W., Brown, P., Bishoffberger, J., Bakker, D., List, J.-M., Egorov, D., Belyaev, O., Urban, M., Mailhammer, R., Geyer, H., Beck, D., Korovina, E., Epps, P., Valenzuela, P., Grant, A. & Hammarström, H. (2011), 'The ASJP database (version 14)'. http://email.eva.mpg.de/ wichmann/listss14.zip.

Wichmann, S., Rama, T. & Holman, E. W. (2011), 'Phonological diversity, word length, and population sizes across languages: The ASJP evidence', *Linguistic Typology* **15**, 177–198.