Comparative evaluation of string similarity measures for automatic language classification.

Taraka Rama and Lars Borin

1 Introduction

Historical linguistics, the oldest branch of modern linguistics, deals with language-relatedness and language change across space and time. Historical linguists apply the widely-tested comparative method (Durie and Ross, 1996) to establish relationships between languages to posit a *language family* and to reconstruct the proto-language for a language family.[1] Although, historical linguistics has a *curious parallel origins* with biology (Atkinson and Gray, 2005), unlike the biologists, main-stream historical linguists have seldom been enthusiastic (Kroeber and Chrétien, 1937; Ellegård, 1959) about using quantitative methods for the discovery of language relationships or predicting the structure of a language family. However, an earlier shorter period of enthusiastic application of quantitative methods marked by Swadesh (1950) ended with the critique of Bergsland and Vogt, 1962. The field of computational historical linguistics did not receive much attention until the beginning of 90's with the exception of two note-worthy doctoral dissertations: Embleton, 1986; Sankoff, 1969.

In traditional lexicostatistics, as introduced by Swadesh (1952), distances between languages are based on human expert cognacy judgments of items in standardized word lists, e.g., the Swadesh lists (Swadesh, 1955).[2] Recently, some researchers have turned to approaches more amenable to automation, hoping that large-scale lexicostatistical language classification will thus become feasible. The ASJP (Automated Similarity Judgment Program) project[3] represents such an approach, where automatically estimated distances between languages are provided as an input to phylogenetic programs originally developed in computational biology (Felsenstein, 2004), for the purpose of inferring genetic relationships among organisms.

As noted above, traditional lexicostatistics assumes that the cognate judgments for a group of languages have been supplied before hand. Given a standardized word list, consisting of 40–100 items, the distance between a pair of languages is defined as the percentage of shared cognates subtracted from 100%. This procedure is applied to all pairs of languages, under consideration, to produce a pair-wise inter-language distance matrix. This inter-language distance matrix is then supplied to a tree-building algorithm

[1] The Indo-European family is a classical case of the successful application of comparative method which establishes a tree relationship between the most populous languages of the world

[2] Gilij (2001) is one of the earliest known attempts at using core vocabulary for positing inter-language relationships in the Americas.

[3] `http://email.eva.mpg.de/~wichmann/ASJPHomePage.htm`

such as Neighbor-Joining (NJ; Saitou and Nei, 1987) or a clustering algorithm such as UPGMA (Sokal and Michener, 1958) to infer a tree structure for the set of languages. One such attempt by Swadesh (1950), even before the discovery of the first clustering algorithm: UPGMA, for Salish languages is reproduced in figure 1.

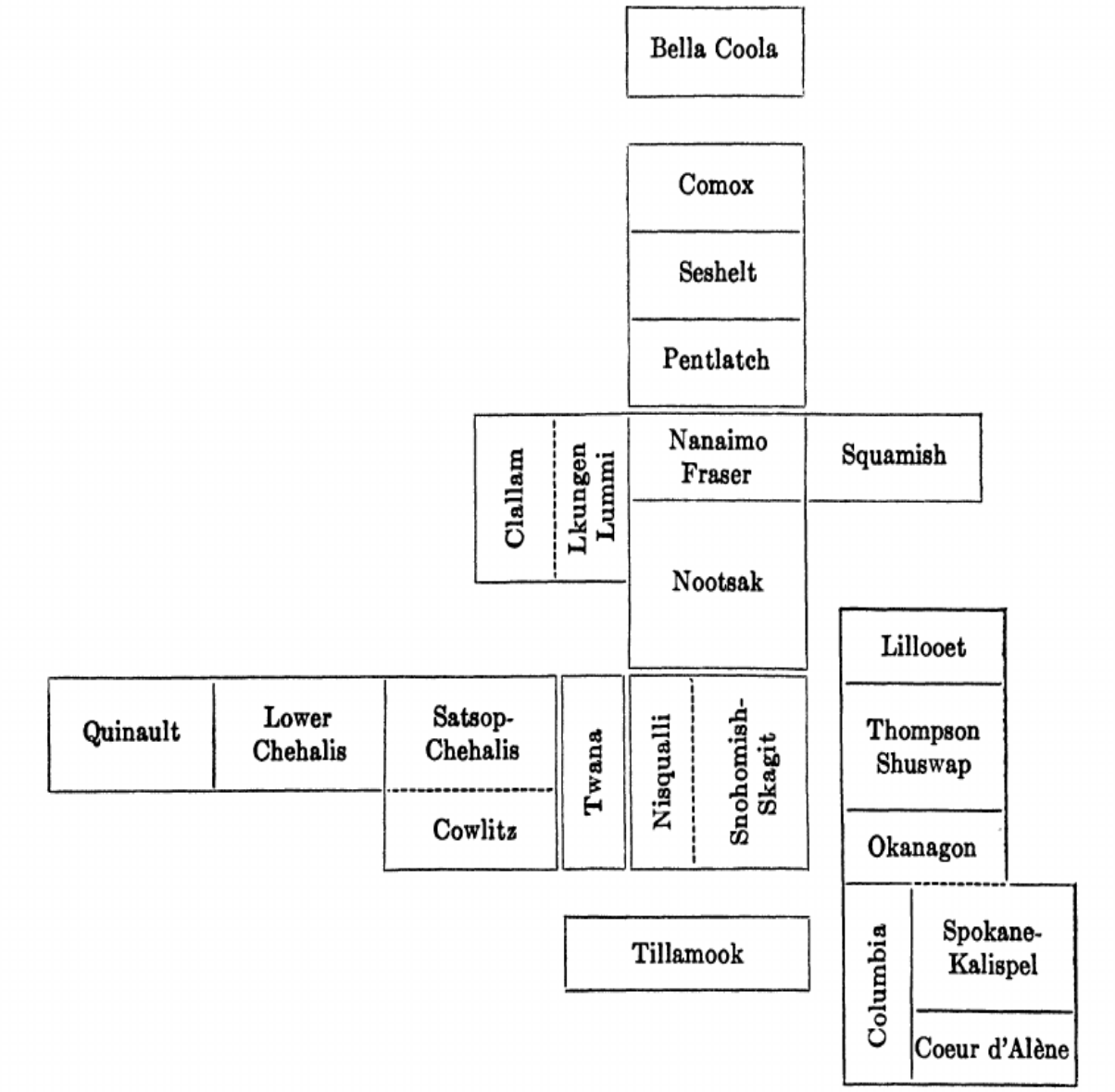

Figure 1: Salish box-diagram from Swadesh 1950.

In the terminology of historical linguistics, cognates are related words across languages that can be traced back to the proto-language. Cognates are identified through regular sound correspondences. Usually cognates have similar surface form and related meanings. Examples of such revealing kind of cognates are: English $\sim$ German *night* $\sim$ *Nacht* 'night' and hound $\sim$ *Hund* 'dog'. If a word has undergone many changes then the relatedness is not obvious from visual inspection and one needs to look into the history of the word to exactly understand the sound changes which resulted in the synchronic form. For

instance, the English ∼ Hindi *wheel* ∼ *chakra* 'wheel' are cognates and can be traced back to the proto-Indo-European root of k^wek^wlo-. In any case, the items such as 'wheel' do not appear on short word lists since they are not universal items and are culture-specific technological innovations. Usually, the cognate judgments are obtained from the expert historical linguist's judgments.

The crucial element in these automated approaches is the method used for determining the overall similarity between two word lists.[4] Often, this is some variant of the popular edit distance or Levenshtein distance (LD; Levenshtein, 1966). LD for a pair of strings is defined as the minimum number of symbol (character) additions, deletions and substitutions needed to transform one string into the other. A modified LD (named as LDND) is used by the ASJP consortium, as reported in their publications (e.g., Bakker et al. 2009 and Holman et al. 2008).

We describe the related work in the field of automatic language classification in section 2.

2 Related Work

In this section, we survey the earlier work in cognate identification and distributional similarity measures for computing inter-language distances. The tasks of cognate identification and tree-inference are closely related tasks in historical linguistics. Considering each task as a computational module would mean that each cognate set identified across a set of tentatively related languages feed into the refinement of the tree inferred at each step. In a critical article, Nichols (1996) points that the historical linguistics enterprise, since its beginning, always used a refinement procedure to posit relatedness and tree-structure for a set of tentatively related languages.[5] The inter-language distance approach to tree-building, is incidentally straight-forward and comparably accurate in comparison to the computationally intensive Bayesian-based tree-inference approach of Greenhill and Gray, 2009.[6]

The inter-language distances are either an aggregate score of the pair-wise item distances or based on a distributional similarity score. The string similarity measures used for the task of cognate identification can also be used for computing the similarity between two lexical items for a particular word sense.

2.1 Cognate identification

The task of identifying the genetically related words – cognates – has received a lot of attention in language technology. Kondrak (2002a) compares a number of algorithms based on phonetic and orthographical similarity for judging the cognateness of a word pair. His work surveys string similarity / distance measures such as *edit distance*, *dice coefficient*,

[4]At this point, we use word list and language interchangeably. Strictly speaking, a language, identified by its ISO 639-3 code, can have as many word lists as the number of dialects.

[5]This idea is quite similar to the famous paradigm of Expectation-Maximization in machine learning field. Kondrak (2002b) employs this paradigm for extracting sound correspondences from pair-wise word lists for the task of cognate identification. The recent paper of Bouchard-Côté et al. (2013) employs a feed-back procedure for the reconstruction of Proto-Austronesian with a great success.

[6]For a comparison of these methods, see Wichmann and Rama, 2014.

and *longest common subsequence ratio* (LCS) for the task of cognate identification. It has to be noted that, not until recently (Hauer and Kondrak, 2011; List, 2012), most of the work in cognate identification focused on determining the cognateness between a word pair and not among a set of words.

Kondrak (2000) developed a string matching algorithm based on articulatory features (called ALINE) for computing the similarity between a word pair. ALINE was evaluated for the task of cognate identification against machine learning algorithms such as Dynamic Bayesian Networks and Pair-wise HMMs for automatic cognate identification (Kondrak and Sherif, 2006). Even though, the approach is technically sound, it suffers due to the bare-boned phonetic transcription used in Dyen et al.'s Indo-European dataset.[7]

Inkpen et al. (2005) compared various string similarity measures for the task of automatic cognate identification for two closely related languages: English and French. The paper shows an impressive array of string similarity measures. However, the results are too language specific to be generalized for the rest of Indo-European family.

In another work, Ellison and Kirby (2006) use Scaled Edit Distance (SED)[8] for computing intra-lexical similarity for estimating language distances based on the dataset of Indo-European languages prepared by Dyen et al. (1992). The distance matrix is then given as input to the NJ algorithm as implemented in PHYLIP package (Felsenstein, 2002) to infer a tree for 87 languages of Indo-European family. They make a qualitative evaluation of the inferred tree with the standard Indo-European tree.

Petroni and Serva (2010) use a modified version of Levenshtein distance for inferring the trees of Indo-European and Austronesian language families. LD is usually normalized by the maximum of the lengths of the two words to account for length-bias. The length normalized LD (LDN) can then be used in computing distances between a pair of word lists in at least two ways: LDN and LDND. LDN is computed as the sum of the Levenshtein distance between the words occupying the same meaning slot, normalized by length. Similarity between phoneme inventories and chance similarity might cause a pair of not-so related languages to show up as related languages. This is compensated for by computing the length-normalized Levenshtein distance between all the pairs of words occupying different meaning slots and summing the different word-pair distances.

The summed Levenshtein distance between the words occupying the same meaning slots is divided by the sum of Levenshtein distances between different meaning slots. The intuition behind this idea is that if two languages are shown to be similar (small distance) due to accidental chance similarity then the denominator would also be small and the ratio would be high.

If the languages are not related and also share no accidental chance similarity, then the distance as computed in the numerator would be unaffected by the denominator. If the languages are related then the distance as computed in the numerator is small anyway, whereas the denominator would be large since the languages are similar due to genetic relationship and not from chance similarity. Hence, the final ratio would be smaller than the original distance given in the numerator.

Petroni and Serva (2010) claim that LDN is more suitable than LDND for measuring linguistic distances. In reply, Wichmann et al. (2010a) empirically show that LDND

[7]The dataset contains 200-word Swadesh lists for 95 speech varieties. Available on `http://www.wordgumbo.com/ie/cmp/index.htm`.

[8]SED is defined as the edit distance normalized by the average of the lengths of the pair of strings.

performs better than LDN for distinguishing the languages belonging to a same family from the languages of other families.

As noted by Jäger (2014), Levenshtein distance only distinguishes between identical and non-identical sound symbols whereas a graded notion of sound similarity would be a closer approximation to historical linguistics as well as achieving better results at the task of phylogenetic inference. Jäger (2014) uses empirically determined weight between a symbol pair (from computational dialectometry; Wieling et al. 2009) to compute distances between ASJP word lists and finds that there is an improvement over LDND at the task of internal classification of languages.

2.2 Distributional similarity measures

Huffman (1998) compute pair-wise language distances based on character n-grams extracted from Bible texts in European and American Indian languages (mostly from the Mayan language family). Singh and Surana (2007) use character n-grams extracted from raw comparable corpora of ten languages from the Indian subcontinent for computing the pair-wise language distances between languages belonging to two different language families (Indo-Aryan and Dravidian). Rama and Singh (2009) introduce a factored language model based on articulatory features to induce a articulatory feature level n-gram model from the dataset of Singh and Surana, 2007. The feature n-grams of each language pair are compared using a distributional similarity measure called cross-entropy to yield a single point distance between the language pair.

Taking cue from the development of tree similarity measures in computational biology, Pompei et al. (2011) evaluate the performance of LDN vs. LDND on the ASJP and Austronesian Basic Vocabulary databases (Greenhill et al., 2008). These authors compute NJ and Minimum Evolution trees[9] for LDN as well as LDND distance matrices. They compare the inferred trees to the classification given in Ethnologue (Lewis, 2009) using two different tree similarity measures: Generalized Robinson-Foulds distance (GRF; A generalized version of Robinson-Foulds [RF] distance; Robinson and Foulds 1979) and Generalized Quartet distance (GQD; Christiansen et al. 2006). GRF and GQD are specifically designed to account for the polytomous nature – a node having more than two children – of the Ethnologue trees. The Dravidian family tree given in figure 3 shows four branches radiating from the top node. Finally, Huff and Lonsdale (2011) compare the NJ trees from ALINE and LDND distance metrics to Ethnologue trees using RF distance. The authors did not find any significant improvement by using a linguistically well-informed similarity measure such as ALINE over LDND.

However, LD is only one of a number of string similarity measures used in fields such as language technology, information retrieval, and bio-informatics. Beyond the works cited above, to the best of our knowledge, there has been no study to compare different string similarity measures on the ASJP dataset in order to determine their relative suitability for genealogical classification.[10] In this paper we compare various string similarity measures[11] for the task of automatic language classification. We evaluate their effectiveness in

[9] A tree building algorithm closely related to NJ.

[10] One reason for this may be that the experiments are computationally demanding, requiring several days for computing a single measure over the whole ASJP dataset.

[11] A complete list of string similarity measures is available on: `http://www.coli.uni-saarland.de/`

language discrimination through distinctiveness measure; and genealogical classification by comparing the distance matrices to the language classifications provided in WALS (World Atlas of Language Structures; Haspelmath et al., 2011)[12] and Ethnologue.

3 Contributions

In this article, we ask the following questions:

- Out of the numerous string similarity measures given in section 5:
  - Which measure is best suited for the tasks of distinguishing related lanugages from unrelated languages?
  - Which is measure is best suited for the task of internal language classification?
  - Is there a statistical procedure for determining the best string similarity measure?
- What is the best way to length normalize a string similarity measure?[13]

4 Database and language classifications

In this section, we describe the ASJP database and the two classifications: WALS and *Ethnologue*.

4.1 Database

The ASJP database offers an attractive alternative to corpora as the basis for massive cross-linguistic investigations. The ASJP effort began with a small dataset of 100-word lists for 245 languages. These languages belong to more than 23 language families, as defined in WALS (Haspelmath et al., 2011). Since Brown et al. (2008), the ASJP database has been going through an expansion, to include in its latest version (v. 14)[14] more than 5500 word lists representing closer to half of the languages spoken in the world (Wichmann et al., 2011). Because of the findings reported by Holman et al. (2008), the later versions of the database aimed to cover only the 40-item most stable Swadesh sublist, and not the 100-item list.

Each lexical item in an ASJP word list is transcribed in a broad phonetic transcription known as ASJP Code (Brown et al., 2008). The ASJP code consists of 34 consonant symbols, 7 vowels, and four modifiers (∗, ", ˜, $), all rendered by characters available on the English version of the QWERTY keyboard. Tone, stress, and vowel length are ignored

courses/LT1/2011/slides/stringmetrics.pdf

[12]WALS does not provide a classification to all the languages of the world. The ASJP consortium gives a WALS-like classification to all the languages present in their database.

[13]Marzal and Vidal (1993) propose an alternate normalization based on the length of the editing path. Kondrak (2005) tests this claim on three different datasets and finds that there is no significant difference between the two normalizations.

[14]Version 16 as of 2014.

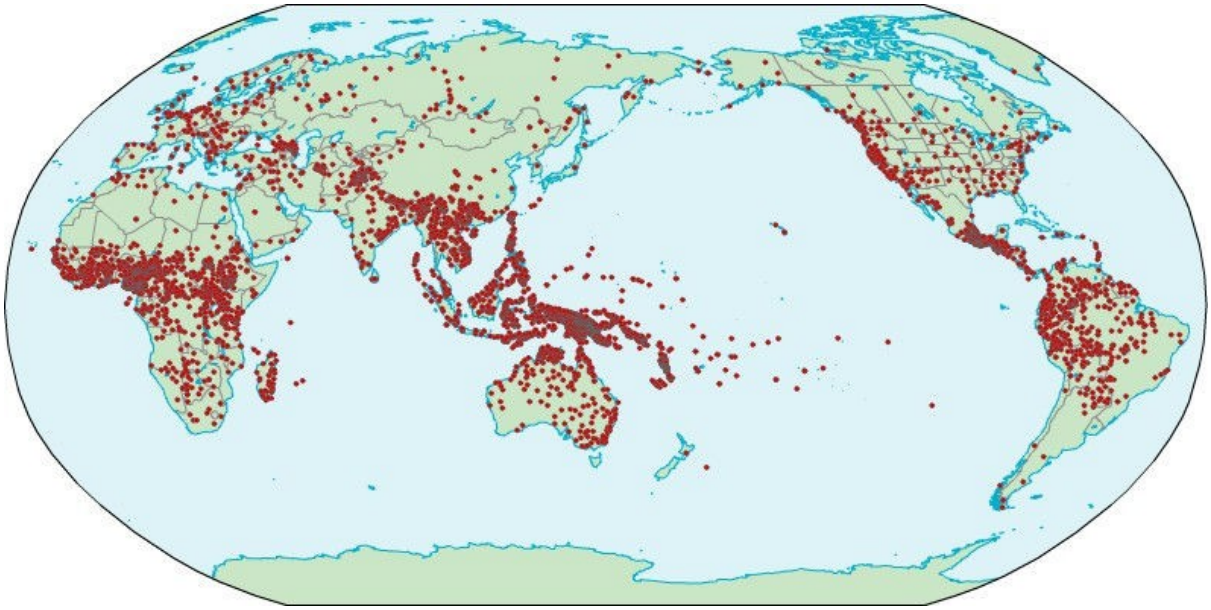

Figure 2: Distribution of languages in ASJP database (version 14).

in this transcription format. The three modifiers combine symbols to form phonologically complex segments (e.g., aspirated, glottalized, or nasalized segments).

In order to ascertain that our results would be comparable to those published by the ASJP group, we successfully replicated their experiments for LDN and LDND measures using the ASJP program and the ASJP dataset version 12 (Wichmann et al., 2010b).[15] This database comprises of reduced (40-item) Swadesh lists for 4169 word lists. All pidgins, creoles, mixed languages, artificial languages, proto-languages, and languages extinct before 1700 CE were excluded for the experiment, as were language families represented by less than 10 word lists. This leaves a dataset with 3730 word lists. It turned out that 60 word lists did not have English glosses for the items, which meant that they could not be processed by the program, so these languages were excluded from the analysis.

All the experiments reported in this paper were performed on a subset of version 14 of the ASJP database whose distribution is shown in figure 2.[16] The database has 5500 word lists, including not only living languages, but also extinct ones. The database also contains word lists for pidgins, creoles, mixed languages, artificial languages, and proto-languages, all of which have been excluded from the current study. Among the extinct languages, only those languages were included which have gone extinct less than three centuries ago. Also, any word list containing less than 28 words (70%) was not included in the final dataset. We use the family names of the WALS (Haspelmath et al., 2011) classification. Following Wichmann et al., 2010a, any family with less than ten languages is excluded from our experiments.[17] The final dataset for our experiments has 4743 word

[15]The original python program was created by Hagen Jung. We modified the program to handle the ASJP modifiers.

[16]Available on `http://email.eva.mpg.de/~wichmann/listss14.zip`.

[17]The reason behind this decision is that correlations resulting from smaller samples (less than 40 language pairs) tend to be unreliable and ten is the lower limit to the number of languages included in

lists for 50 language families.

Family Name	WN	# WLs	Family Name	WN	# WLs
Afro-Asiatic	AA	287	Mixe-Zoque	MZ	15
Algic	Alg	29	MoreheadU.Maro	MUM	15
Altaic	Alt	84	Na-Dene	NDe	23
Arwakan	Arw	58	Nakh-Daghestanian	NDa	32
Australian	Aus	194	Niger-Congo	NC	834
Austro-Asiatic	AuA	123	Nilo-Saharan	NS	157
Austronesian	An	1008	Otto-Manguean	OM	80
Border	Bor	16	Panoan	Pan	19
Bosavi	Bos	14	Penutian	Pen	21
Carib	Car	29	Quechuan	Que	41
Chibchan	Chi	20	Salish	Sal	28
Dravidian	Dra	31	Sepik	Sep	26
Eskimo-Aleut	EA	10	Sino-Tibetan	ST	205
Hmong-Mien	HM	32	Siouan	Sio	17
Hokan	Hok	25	Sko	Sko	14
Huitotoan	Hui	14	Tai-Kadai	TK	103
Indo-European	IE	269	Toricelli	Tor	27
Kadugli	Kad	11	Totonacan	Tot	14
Khoisan	Kho	17	Trans-NewGuinea	TNG	298
Kiwain	Kiw	14	Tucanoan	Tuc	32
LakesPlain	LP	26	Tupian	Tup	47
Lower-Sepik-Ramu	LSR	20	Uralic	Ura	29
Macro-Ge	MGe	24	Uto-Aztecan	UA	103
Marind	Mar	30	West-Papuan	WP	33
Mayan	May	107	WesternFly	WF	38

Table 1: Distribution of language families in ASJP database. WN and WLs stands for WALS Name and Word Lists.

WALS. WALS classification is a two-level classification where each language belongs to a genus and a family. A genus is a genetic classification unit given by Dryer (2000) and consists of set of languages supposedly descended from a common ancestor which is 3000 to 3500 years old. For instance, Indic languages are classified as a separate genus from Iranian languages although, it is quite well known that both Indic and Iranian languages descended from a common proto-Indo-Iranian ancestor.

Ethnologue. Ethnologue classification is a multi-level tree classification for a language family. The Ethnologue classification is produced by missionaries of the Summer Institute of Linguistics and is very opportunistic in the inclusion of languages or genetic relatedness. The highest node in a family tree is the family itself and languages form the lowest nodes (leaves). A internal node in the tree is not necessarily binary and can have more than two branches emanating from it. For instance, the Dravidian language family has four branches emanating from the top node (see figure 3 for the Ethnologue family tree of Dravidian languages).

the experiments.

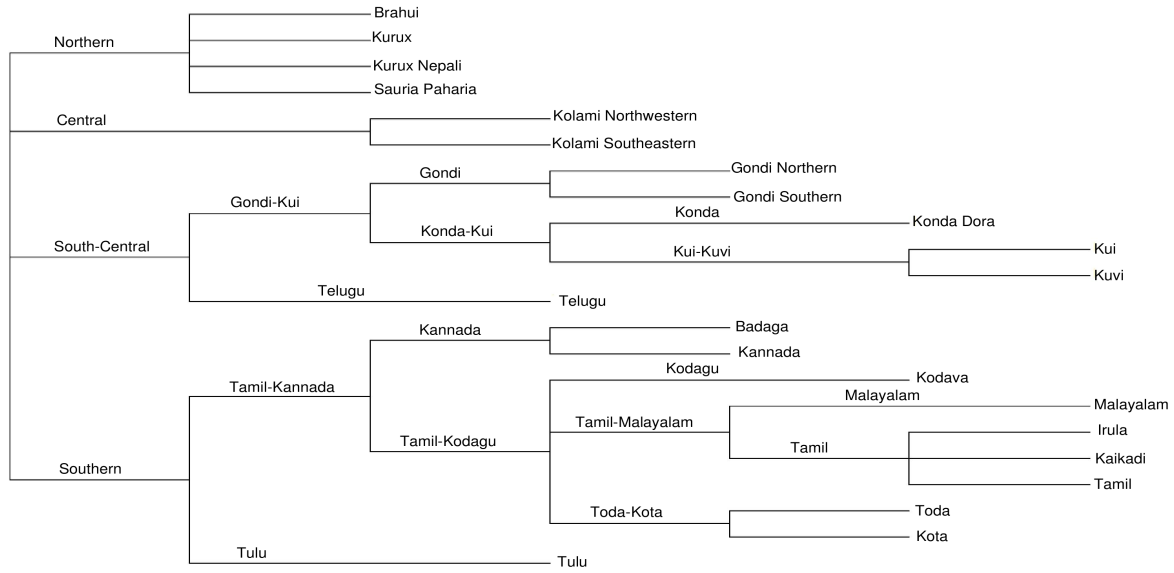

Figure 3: Ethnologue tree for Dravidian language family.

5 Methodology

In this section, we describe the various string and distributional similarity measures that are used to compute the distance between a pair of languages. As mentioned earlier, string similarity measures work at the level of word pairs and provide an aggregate score of the similarity between word pairs whereas distributional measures compare the n-gram profiles between a language pair to yield a distance score.

5.1 String similarity measures

The description of different string similarity measures for a word pair are as followed:[18]

- *IDENT* returns 1 if the words are identical, otherwise it returns 0.
- *PREFIX* returns the length of the longest common prefix divided by the length of the longer word.
- *DICE* is defined as the number of shared bigrams divided by the total number of bigrams in both the words.
- *LCS* is defined as the length of the longest common subsequence divided by the length of the longer word (Melamed, 1999).
- *TRIGRAM* is defined in the same way as *DICE* but uses trigrams for computing the similarity between a word pair.

[18]Some of these measures are employed by Inkpen et al. (2005) for the task of cognate identification between English-French language pair. We do not defined Levenshtein distance as it has been described earlier.

- *XDICE* is defined in the same way as *DICE* but uses "extended bigrams", which are trigrams without the middle letter (Brew and McKelvie, 1996).

- Jaccard's index, *JCD*, is a set cardinality measure that is defined as the ratio of the number of shared bigrams between the two words to the ratio of the size of the union of the bigrams between the two words.

Each word-pairs similarity score is converted to its distance counterpart by subtracting the score from 1.0.[19] Note that this conversion can sometimes result in a negative distance which is due to the double normalization involved in LDND.[20] The distance score resulting from a word-pair's string similarity measure is then used to compute the pair-wise distance between a language pair. The distance computation between a language pair is performed as described in section 2.1. Following the naming convention of LDND, a suffix "D" is added to the name of each measure to indicate its LDND distance variant.

5.2 N-gram similarity

This section describes the distributional similarity measures originally developed for automatic language identification in a multilingual document. This line of work started with the heavily cited paper of Cavnar and Trenkle, 1994 who used character n-grams for text categorization. They observed that each document category has a characteristic character n-gram profile. The rank of a character n-gram varies across different categories and documents belonging to the same category have similar character n-gram Zipfian distributions.

Building on this idea, Dunning (1994) postulates that each language has its own signature character (or phoneme; depending on the level of transcription) n-gram distribution. Comparing the character n-gram profiles of two languages can yield a single point distance between the language pair. The comparison procedure is usually accomplished through the use of one of the distance measures given in Singh 2006. The following steps are followed for extracting the phoneme N-gram profile for a language.

For a language, the n-gram extraction procedure is as follows:

- A n-gram is defined as the consecutive phonemes in a window of N. The value of N usually ranges from 1 to 5.

- All N-grams are extracted for a lexical item. This step is repeated for all the lexical items in a word list.

- All the extracted n-grams are mixed and sorted in the descending order of their frequency. The relative frequency of the n-grams are computed.

- Only the top G n-grams are retained and the rest of them are pruned. The value of G is determined empirically.

[19] Lin (1998) investigates three distance to similarity conversion techniques and motivates the results from an information-theoretical point. In this article, we do not investigate the effects of similarity to distance conversion. Rather, we stick to the traditional conversion technique.

[20] Thus, the resulting distance is not a true distance metric.

For a language pair, the n-gram profiles can be compared using Out-of-Rank measure, Jaccard's index, Dice distance, Overlap distance, Manhattan distance, and Euclidean distance. The distances are explained below:

1. Out-of-Rank measure is defined as the aggregate sum of the absolute difference in the rank of the shared n-grams between a pair of language. If there are no shared bigrams between a n-gram profile, then the difference in ranks is assigned a maximum out-of-place score.

2. Jaccard's index is a set cardinality measure. It is defined as the ratio of the cardinality of the intersection of the n-grams between the two languages to the cardinality of the union of the two languages.

3. Dice distance is related to Jaccard's Index. It is defined as the ratio of twice the number of shared n-grams to the total number of n-grams in both the language profiles.

4. Manhattan distance is defined as the sum of the absolute difference between the relative frequency of the shared n-grams.

5. Euclidean distance is defined in a similar fashion to Manhattan distance where the individual terms are squared.

While replicating the original ASJP experiments on the version 12 ASJP database, we tested if the above distributional measures, [1–4] perform as well as LDN. Unfortunately, the results are on the discouraging side of the spectrum and we do not repeat the experiments on the version 14 of the database. One main reason for this result is the relatively small size of ASJP word list. The relatively small word list size provides a poor estimates of the true language signatures. The next section describes the three different evaluation measures for comparing the different string similarity measures.

One possible set of measures, based on n-grams, and could be included are information theoretic based such as cross entropy and KL-divergence. These measures have been well-studied in natural language processing systems such as machine translation, natural language parsing, sentiment identification, and also in automatic language identification. The probability distributions required for using these measures are usually estimated through maximum likelihood estimation which require a fairly large amount of data.

6 Evaluation measures

In this section, we describe the three different measures for evaluating the performance of string similarity measures given in section 5:

1. The first measure *dist*, originally given by Wichmann et al. (2010a), tests if LDND is better than LDN at the task of distinguishing related languages from unrelated languages.

2. The second measure, *RW*, is a special case of Pearson's **r** – called point biserial correlation (Tate, 1954) – computes the agreement between a family's pair-wise distances and the family's WALS classification.

3. The third measure, γ, is related to the Goodman and Kruskal's Gamma (1954) and measures the strength of association between two ordinal variables. In this paper, it is used to compute the level of agreement between the pair-wise intra-language distances and the family's Ethnologue classification.

6.1 Distinctiveness measure (dist)

The *dist* measure for a family consists of three components: the mean of the pair-wise distances inside a language family (d_{in}); and mean of the pair-wise distances from each language in a family to the rest of the language families (d_{out}). sd_{out} is defined as the standard deviation of all the pair-wise distances used to compute d_{out}. Finally, *dist* is defined as $\frac{d_{in}-d_{out}}{sd_{out}}$. The resistance of a string similarity measure to other language families is reflected by the value of sd_{out}.

A comparatively higher *dist* value suggests that a string similarity measure is particularly resistant to random similarities between unrelated languages and performs well at distinguishing languages belonging to the same language family from the other language families.

6.2 Correlation with WALS

The WALS database has classification at three levels. The top level is the language family, second level is the genus and the lowest level is the language itself. Two languages that belong to different genera but same family have a distance of 2. If the languages fall in the same genus, they have a distance of 1. This allows us to define a distance matrix for each family based on WALS. The WALS distance matrix can be compared to the distance matrices of any string similarity measure using point biserial correlation – a special case of Pearson's **r**. If family in WALS classification has a single genus there is no computation of RW and the corresponding row for a family is empty in table 7.

6.3 Agreement with Ethnologue

Given a distance-matrix d of order $N \times N$, where each cell d_{ij} is the distance between two languages i and j; and an *Ethnologue* tree E, the computation of γ for a language family is defined as follows:

1. Enumerate all the triplets for a language family of size N. A triplet, t for a language family is defined as $\{i, j, k\}$, where $i \neq j \neq k$ are languages belonging to a family. A language family of size N has $\binom{N}{3}$ triplets.

2. For the members of each such triplet t, there are three lexical distances d_{ij}, d_{ik}, and d_{jk}. The expert classification tree E can treat the three languages $\{i, j, k\}$ in four possible ways (| denotes a partition): $\{i, j \mid k\}$, $\{i, k \mid j\}$, $\{j, k \mid i\}$ or can have a tie where all languages emanate from the same node. All ties are ignored in the computation of γ since, a tie in the gold standard indicates uncertainty in the classification.

3. A distance triplet d_{ij}, d_{ik}, and d_{jk} is said to agree completely with an Ethnologue partition $\{i, j \mid k\}$ when the following conditions are satisfied:

$$d_{ij} < d_{ik} \tag{1}$$

$$d_{ij} < d_{jk} \tag{2}$$

A triplet that satisfies these conditions is counted as a concordant comparison, C; else it is counted as a discordant comparison, D.

4. Steps 2 and 3 are repeated for all the $\binom{N}{3}$ triplets to yield γ for a family defined as $\gamma = \frac{C-D}{C+D}$. γ lies in the range $[-1, 1]$ where a score of -1 indicates perfect disagreement and a score of $+1$ indicates perfect agreement.

At this point, one might wonder about the decision for not using a off-the-shelf tree-building algorithm to infer a tree and compare the resulting tree with the Ethnologue classification. Although both the works of Pompei et al., 2011; Huff and Lonsdale, 2011 compare their inferred trees – based on Neighbor-Joining and Minimum Evolution algorithms – to Ethnologue trees using cleverly crafted tree-distance measures (GRF and GQD), they do not make a direct comparison of the distance matrices to the Ethnologue trees. A direct comparison of a family's distance matrix to the family's Ethnologue tree circumvents the choice of the tree inference algorithm. Whenever, the Ethnologue tree of a family is completely unresolved, it is shown by an empty row. For example, the family tree of Bosavi languages is a star structure. Hence, the corresponding row in the table 5 is left empty.

7 Item-item vs. length visualizations

A distance measure such as LDND has two components:

- The average LDN computed between pair-wise lexical items for the same meaning.
- The average LDN computed between pair-wise lexical items for different meaning pairs.

Although, there have been multitude of publications involving LDND, no effort has been put to look the base components of LDND through a magnifying glass. It is quite interesting to see the overall distribution between LD and the lengths of the lexical items under comparison. We proceed to see the variation of the pair-wise LDs vs. the pair-wise word lengths for lexical items sharing the same meaning and different meanings. We select a small (31 word lists) but well-studied language family called Dravidian family, spoken in South Asia (comprising of modern day India, Pakistan, and Nepal), for this study.

We compute the LD between all the word pairs for a same meaning. We plot the corresponding LD and the word-length pairs on a three-dimensional plot as shown in figure 4. We repeat the exercise now for the word-pairs for different meaning pairs and plot them in figure 5. A 3-dimensional scatterplot would has the advantage of showing the distribution of pair-wise LD distances against the length of the word-pairs. In the

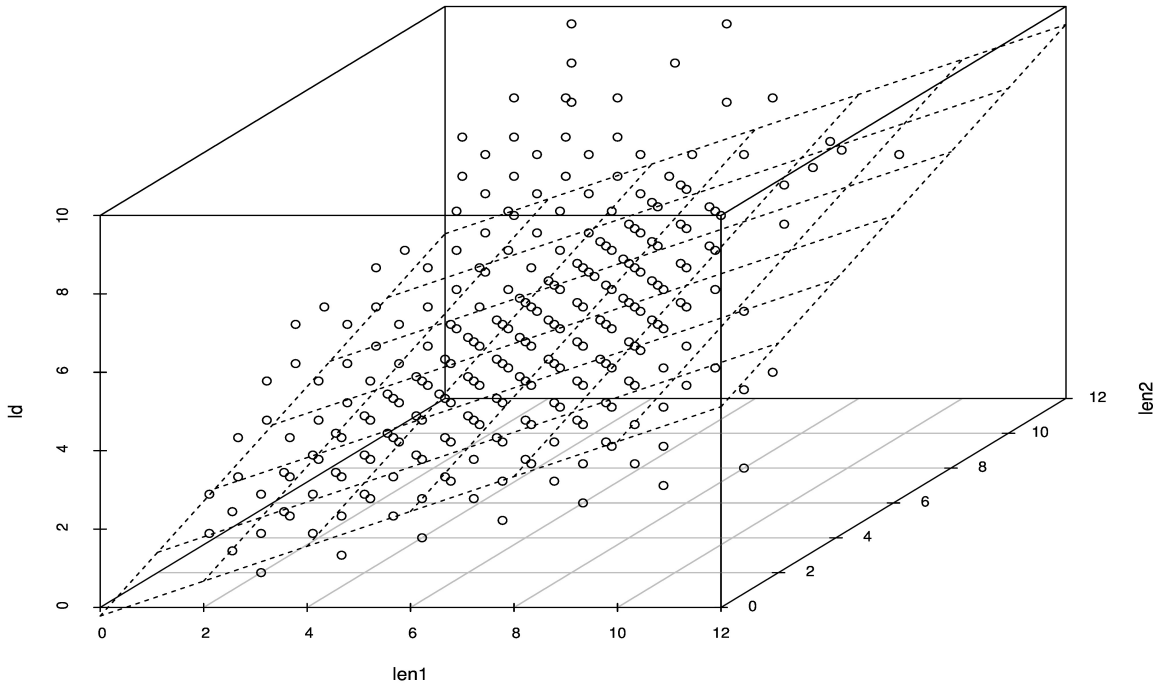

Figure 4: 3D scatterplot for Dravidian language family and same items. There are more than 21,000 points in this plot.

next step, we fit a linear regression plane to the distribution. This plane would show the number of points which fall below or above the regression plane. The regression shown in figure 4 is highly significant and has an adjusted-R^2 value of 0.2554 ($p < 2.2e-16$). The multiple regression for different meaning-meaning pairs shows an adjusted-R^2 value of 0.4953 ($p < 2.2e-16$). These multiple regressions support the hypothesis that there exists a linear relationship between pair-wise lengths and LD.

Given that, the multiple regressions support a linear function of pair-wise lengths, it would be interesting to make a scatterplot of LD vs. average length and LD vs. maximum length for the above two datasets. We employ the hexagonal binning technique (Carr et al., 2010) for showing the huge number of points. The size of a bin is depicted through the color intensity and a color legend shows the number of points in a bin. The results of this technique is shown in the figures 6, 7, 8, and 9. Each of the hexagonal plot is plotted using 100 bins. Average length vs. LD is more dispersed in same as well as different meaning pairs. The maximum length normalization seems to follow LD closely in both the plots. We take this as visualization as a support for preferring maximum length normalization over average length normalization.

The results of comparison of the various string similarity measures is described in the next section.

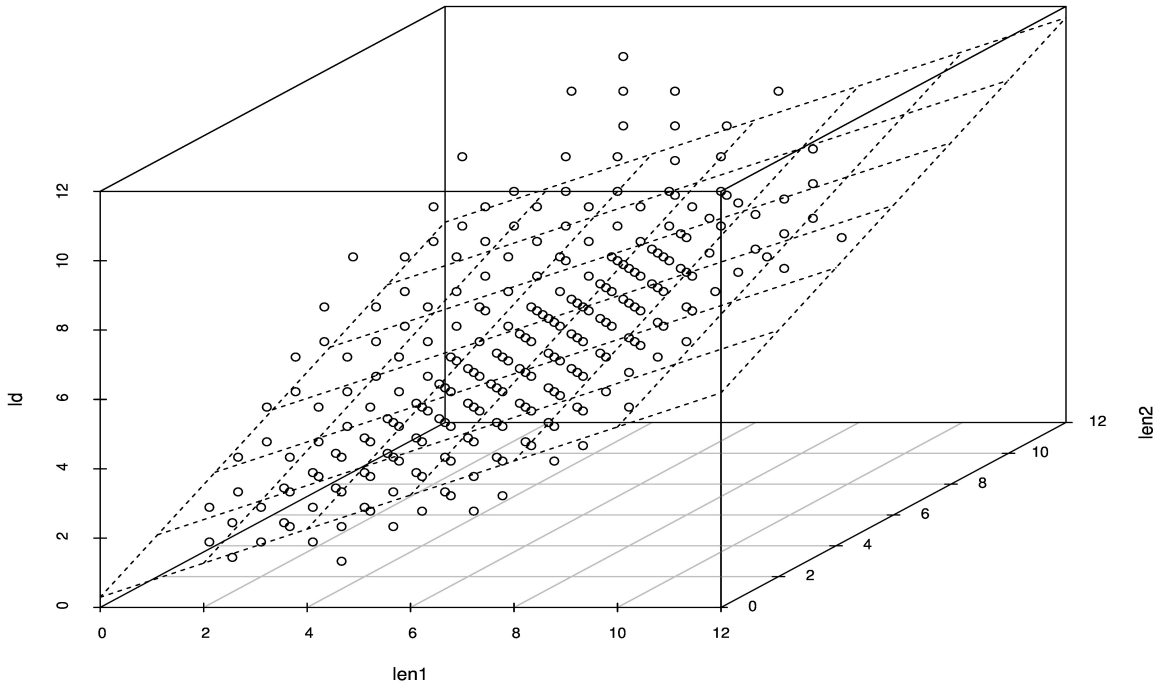

Figure 5: 3D scatterplot for Dravidian language family and different items. There are about 1 million data points resulting from this exercise. Hence, we randomly show 22,000 points in this plot.

8 Results and discussion

In table 2 we give the results of our experiments. We only report the average results for all measures across the families listed in table 1. Further, we check the correlation between the performance of the different string similarity measures across the three evaluation measures by computing a Spearman's ρ. The pair-wise ρ is given in table 3. The high correlation value of 0.95 between RW and γ suggests that all the measures agree roughly on the task of internal classification.

The average scores in each column suggests that the string similarity measures exhibit different degrees of performances. How does one decide which measure is the best in a column? What kind of statistical testing procedure should be adopted for deciding upon a measure? We address this questions through the following procedure:

1. For a column i, sort the average scores, s in descending order.

2. For a row index $1 \leq r \leq 16$, test the significance of $s_r \geq s_{r+1}$ through a sign test (Sheskin, 2003). This test yields a $p - value$.

The above significant tests are not independent by themselves. Hence, we cannot reject a null hypothesis H_0 at a significance level of $\alpha = 0.01$. The α needs to be corrected

Measure	Average Dist	Average RW	Average γ
DICE	3.3536	0.5449	0.6575
DICED	9.4416	0.5495	0.6607
IDENT	1.5851	0.4013	0.2345
IDENTD	8.163	0.4066	0.3082
JCD	13.9673	0.5322	0.655
JCDD	15.0501	0.5302	0.6622
LCS	3.4305	0.6069	0.6895
LCSD	6.7042	0.6151	0.6984
LDN	3.7943	0.6126	0.6984
LDND	7.3189	0.619	0.7068
PREFIX	3.5583	0.5784	0.6747
PREFIXD	7.5359	0.5859	0.6792
TRIGRAM	1.9888	0.4393	0.4161
TRIGRAMD	9.448	0.4495	0.5247
XDICE	0.4846	0.3085	0.433
XDICED	2.1547	0.4026	0.4838
Average	6.1237	0.5114	0.5739

Table 2: Average results for each string similarity measure across the 50 families. The rows are sorted by the name of the measure.

	Dist	RW
γ	0.30	0.95
Dist		0.32

Table 3: Spearman's ρ between γ, RW and Dist

for multiple tests. Unfortunately, the standard Bonferroni's multiple test correction or Fisher's Omnibus test works for a global null hypothesis and not at the level of a single test. We follow the procedure, called False Discovery Rate (FDR), given by Benjamini and Hochberg (1995) for adjusting the α value for multiple tests. Given $H_1 \ldots H_m$ null hypotheses and $P_1 \ldots P_m$ p-values, the procedure works as follows:

1. Sort the P_k, $1 \leq k \leq m$, values in ascending order. k is the rank of a p-value.
2. The adjusted α_k^* value for P_k is $\frac{k}{m}\alpha$.
3. Reject all the H_0s from $1, \ldots, k$ where $P_{k+1} > \alpha_k^*$.

The above procedure ensures that the chance of incorrectly rejecting a null hypothesis is 1 in 20 for $\alpha = 0.05$ and 1 in 100 for $\alpha = 0.01$. In this experimental context, this suggests that we erroneously reject 0.75 true null hypotheses out of 15 hypotheses for $\alpha = 0.05$ and 0.15 hypotheses for $\alpha = 0.01$. We report the Dist, γ, and RW for each

family in tables 5, 6, and 7. In each of these tables, only those measures which are above the average scores from table 2, are reported.

The FDR procedure for γ suggests that no sign test is significant. This is in agreement with the result of Wichmann et al., 2010a, who showed that the choice of LDN or LDND is quite unimportant for the task of internal classification. The FDR procedure for RW suggests that LDN > LCS, LCS > PREFIXD, DICE > JCD, and JCD > JCDD. Here $A > B$ denotes that A is significantly better than B. The FDR procedure for Dist suggests that JCDD > JCD, JCD > TRID, DICED > IDENTD, LDND > LCSD, and LCSD > LDN.

The results point towards an important direction in the task of building computational systems for automatic language classification. The pipeline for such a system consists of 1) distinguishing related languages from unrelated languages and 2) internal classification accuracy. JCDD performs the best with respect to Dist. Further, JCDD is derived from JCD and can be computed in $\mathcal{O}(m+n)$, for two strings of length m and n. In comparison, LDN is in the order of $\mathcal{O}(mn)$. In general, the computational complexity for computing distance between two word lists for all the significant measures is given in table 4. Based on the computational complexity and the significance scores, we propose that JCDD be used for step 1 and measure like LDN be used for internal classification.

Measure	Complexity
JCDD	$C\mathcal{O}(m+n+\min(m-1,n-1))$
JCD	$l\mathcal{O}(m+n+\min(m-1,n-1))$
LDND	$C\mathcal{O}(mn)$
LDN	$l\mathcal{O}(mn)$
PREFIXD	$C\mathcal{O}(\max(m,n))$
LCSD	$C\mathcal{O}(mn)$
LCS	$l\mathcal{O}(mn)$
DICED	$C\mathcal{O}(m+n+\min(m-2,n-2))$
DICE	$l\mathcal{O}(m+n+\min(m-2,n-2))$

Table 4: Computation complexity for top performing measures for computing distance between two word lists. Given two word lists each of length l. m and n denote the lengths of a word pair w_a and w_b and $C = l(l-1)/2$

9 Conclusion

We conclude the article by pointing that this is the first known attempt at applying more than 20 similarity (or distance) measures for closer to half of the world's languages. We examine various measures at two levels, namely, distinguishing related from unrelated languages and internal classification of related languages. We find that the choice of string similarity measures (among the tested pool of measures) is not very important for the task of internal classification whereas, the choice affects the results of discriminating related languages from unrelated ones.

Acknowledgments

The authors thank Søren Wichmann, Eric W. Holman, Harald Hammarström, and Roman Yangarber for the useful comments in improving the text. The string similarity experiments have been made possible through the use of ppss software[21] recommended by Leif-Jöran Olsson. The first author would like to thank Prasant Kolachina for the discussions on parallel implementations in Python.

[21] http://code.google.com/p/ppss/

Family	JCDD	JCD	TRIGRAMD	DICED	IDENTD	PREFIXD	LDND	LCSD	LDN
Bos	15.0643	14.436	7.5983	10.9145	14.4357	10.391	8.6767	8.2226	4.8419
NDe	19.8309	19.2611	8.0567	13.1777	9.5648	9.6538	10.1522	9.364	5.2419
NC	1.7703	1.6102	0.6324	1.1998	0.5368	1.0685	1.3978	1.3064	0.5132
Pan	24.7828	22.4921	18.5575	17.2441	12.2144	13.7351	12.7579	11.4257	6.8728
Hok	10.2645	9.826	3.6634	7.3298	4.0392	3.6563	4.84	4.6638	2.7096
Chi	4.165	4.0759	0.9642	2.8152	1.6258	2.8052	2.7234	2.5116	1.7753
Tup	15.492	14.4571	9.2908	10.4479	6.6263	8.0475	8.569	7.8533	4.4553
WP	8.1028	7.6086	6.9894	5.5301	7.0905	4.0984	4.2265	3.9029	2.4883
AuA	7.3013	6.7514	3.0446	4.5166	3.4781	4.1228	4.7953	4.3497	2.648
An	7.667	7.2367	4.7296	5.3313	2.5288	4.3066	4.6268	4.3107	2.4143
Que	62.227	53.7259	33.479	29.7032	27.1896	25.9791	23.7586	21.7254	10.8472
Kho	6.4615	6.7371	3.3425	4.4202	4.0611	3.96	3.8014	3.3776	2.1531
Dra	18.5943	17.2609	11.6611	12.4115	7.3739	10.2461	9.8216	8.595	4.8771
Aus	2.8967	3.7314	1.5668	2.0659	0.7709	1.8204	1.635	1.5775	1.4495
Tuc	25.9289	24.232	14.0369	16.8078	11.6435	12.5345	12.0163	11.0698	5.8166
Ura	6.5405	6.1048	0.2392	1.6473	-0.0108	3.4905	3.5156	3.1847	2.1715
Arw	6.1898	6.0316	4.0542	4.4878	1.7509	2.9965	3.5505	3.3439	2.1828
May	40.1516	37.7678	17.3924	22.8213	17.5961	14.4431	15.37	13.4738	7.6795
LP	7.5669	7.6686	3.0591	5.3684	5.108	4.8677	4.3565	4.2503	2.8572
OM	4.635	4.5088	2.8218	3.3448	2.437	2.6701	2.7328	2.4757	1.3643
Car	15.4411	14.6063	9.7376	10.6387	5.1435	7.7896	9.1164	8.2592	5.0205
TNG	1.073	1.216	0.4854	0.8259	0.5177	0.8292	0.8225	0.8258	0.4629
MZ	43.3479	40.0136	37.9344	30.3553	36.874	20.4933	18.2746	16.0774	9.661
Bor	9.6352	9.5691	5.011	6.5316	4.1559	6.5507	6.3216	5.9014	3.8474
Pen	5.4103	5.252	3.6884	3.8325	2.3022	3.2193	3.1645	2.8137	1.5862
MGe	4.2719	4.0058	1.0069	2.5482	1.6691	2.0545	2.4147	2.3168	1.1219
ST	4.1094	3.8635	0.9103	2.7825	2.173	2.7807	2.8974	2.7502	1.3482

Tor	3.2466	3.1546	2.2187	2.3101	1.7462	2.1128	2.0321	1.9072	1.0739
TK	15.0085	13.4365	5.331	7.7664	7.5326	8.1249	7.6679	6.9855	2.8723
IE	7.3831	6.7064	1.6767	2.8031	1.6917	4.1028	4.0256	3.6679	1.4322
Alg	6.8582	6.737	4.5117	5.2475	1.2071	4.5916	5.2534	4.5017	2.775
NS	2.4402	2.3163	1.1485	1.6505	1.1456	1.321	1.3681	1.3392	0.6085
Sko	6.7676	6.3721	2.5992	4.6468	4.7931	5.182	4.7014	4.5975	2.5371
AA	1.8054	1.6807	0.7924	1.2557	0.4923	1.37	1.3757	1.3883	0.6411
LSR	4.0791	4.3844	2.2048	2.641	1.5778	2.1808	2.1713	2.0826	1.6308
Mar	10.9265	10.0795	8.5836	7.1801	6.4301	5.0488	4.7739	4.5115	2.8612
Alt	18.929	17.9969	6.182	9.1747	7.2628	9.4017	8.8272	7.9513	4.1239
Sep	6.875	6.5934	2.8591	4.5782	4.6793	4.3683	4.1124	3.8471	2.0261
Hui	21.0961	19.8025	18.4869	14.7131	16.1439	12.4005	10.2317	9.2171	4.9648
NDa	7.6449	7.3732	3.2895	4.8035	2.7922	5.7799	5.1604	4.8233	2.3671
Sio	13.8571	12.8415	4.2685	9.444	7.3326	7.8548	7.9906	7.1145	4.0156
Kad	42.0614	40.0526	27.8429	25.6201	21.678	17.0677	17.5982	15.9751	9.426
MUM	7.9936	7.8812	6.1084	4.7539	4.7774	3.8622	3.4663	3.4324	2.1726
WF	22.211	20.5567	27.2757	15.8329	22.4019	12.516	11.2823	10.4454	5.665
Sal	13.1512	12.2212	11.3222	9.7777	5.2612	7.4423	7.5338	6.7944	3.4597
Kiw	43.2272	39.5467	46.018	30.1911	46.9148	20.2353	18.8007	17.3091	10.3285
UA	21.6334	19.6366	10.4644	11.6944	4.363	9.6858	9.4791	8.9058	4.9122
Tot	60.4364	51.2138	39.4131	33.0995	26.7875	23.5405	22.6512	21.3586	11.7915
HM	8.782	8.5212	1.6133	4.9056	4.0467	5.7944	5.3761	4.9898	2.8084
EA	27.1726	25.2088	24.2372	18.8923	14.1948	14.2023	13.7316	12.1348	6.8154
Average	15.0501	13.9673	9.448	9.4416	8.163	7.5359	7.3189	6.7042	3.7943

Table 5: Dist for families and measures above average

Family	LDND	LCSD	LDN	LCS	PREFIXD	PREFIX	JCDD	DICED	DICE	JCD
WF										
Tor	0.7638	0.734	0.7148	0.7177	0.7795	0.7458	0.7233	0.7193	0.7126	0.7216
Chi	0.7538	0.7387	0.7748	0.7508	0.6396	0.7057	0.7057	0.7057	0.7057	0.7477
HM	0.6131	0.6207	0.5799	0.5505	0.5359	0.5186	0.4576	0.429	0.4617	0.4384
Hok	0.5608	0.5763	0.5622	0.5378	0.5181	0.4922	0.5871	0.5712	0.5744	0.5782
Tot	1	1	1	1	0.9848	0.9899	0.9848	0.9899	0.9949	0.9848
Aus	0.4239	0.4003	0.4595	0.4619	0.4125	0.4668	0.4356	0.4232	0.398	0.4125
WP	0.7204	0.7274	0.7463	0.7467	0.6492	0.6643	0.6902	0.6946	0.7091	0.697
MUM	0.7003	0.6158	0.7493	0.7057	0.7302	0.6975	0.5477	0.5777	0.6594	0.6213
Sko	0.7708	0.816	0.7396	0.809	0.7847	0.7882	0.6632	0.6944	0.6458	0.6181
ST	0.6223	0.6274	0.6042	0.5991	0.5945	0.5789	0.5214	0.5213	0.5283	0.5114
Sio	0.8549	0.8221	0.81	0.7772	0.8359	0.8256	0.772	0.7599	0.7444	0.7668
Pan	0.3083	0.3167	0.2722	0.2639	0.275	0.2444	0.2361	0.2694	0.2611	0.2306
AuA	0.5625	0.5338	0.5875	0.548	0.476	0.4933	0.5311	0.5198	0.5054	0.5299
Mar	0.9553	0.9479	0.9337	0.9017	0.9256	0.9385	0.924	0.918	0.9024	0.9106
Kad										
May	0.7883	0.7895	0.7813	0.7859	0.7402	0.7245	0.8131	0.8039	0.7988	0.8121
NC	0.4193	0.4048	0.3856	0.3964	0.2929	0.2529	0.3612	0.3639	0.2875	0.2755
Kiw										
Hui	0.9435	0.9464	0.9435	0.9464	0.9464	0.9435	0.8958	0.9107	0.9137	0.8988
LSR	0.7984	0.7447	0.7234	0.6596	0.7144	0.692	0.7626	0.748	0.6484	0.6775
TK	0.7757	0.7698	0.7194	0.7158	0.7782	0.7239	0.6987	0.6991	0.6537	0.6705
LP	0.6878	0.6893	0.7237	0.7252	0.6746	0.7065	0.627	0.6594	0.6513	0.6235
Que	0.737	0.7319	0.758	0.7523	0.742	0.7535	0.7334	0.7335	0.7502	0.7347
NS	0.5264	0.4642	0.4859	0.4532	0.4365	0.3673	0.5216	0.5235	0.4882	0.4968
AA	0.6272	0.6053	0.517	0.459	0.6134	0.5254	0.5257	0.5175	0.4026	0.5162
Ura	0.598	0.5943	0.6763	0.6763	0.5392	0.6495	0.7155	0.479	0.6843	0.7003

MGe	0.6566	0.6659	0.6944	0.716	0.6011	0.662	0.7245	0.7099	0.7508	0.6983
Car	0.325	0.3092	0.3205	0.3108	0.2697	0.2677	0.313	0.3118	0.2952	0.316
Bor	0.7891	0.8027	0.7823	0.7914	0.7755	0.7619	0.7846	0.8005	0.7914	0.7823
Bos										
EA	0.844	0.8532	0.8349	0.8349	0.8716	0.8899	0.8716	0.8716	0.8899	0.8899
TNG	0.6684	0.6692	0.6433	0.6403	0.643	0.6177	0.5977	0.5946	0.5925	0.5972
Dra	0.6431	0.6175	0.6434	0.6288	0.6786	0.6688	0.6181	0.6351	0.655	0.6112
IE	0.7391	0.7199	0.7135	0.6915	0.737	0.7295	0.5619	0.5823	0.6255	0.5248
OM	0.9863	0.989	0.9755	0.9725	0.9527	0.9513	0.9459	0.9472	0.9403	0.9406
Tuc	0.6335	0.623	0.6187	0.6089	0.6189	0.6153	0.5937	0.5983	0.5917	0.5919
Arw	0.5079	0.4825	0.4876	0.4749	0.4475	0.4472	0.4739	0.4773	0.4565	0.4727
NDa	0.9458	0.9578	0.9415	0.9407	0.9094	0.9121	0.8071	0.8246	0.8304	0.8009
Alg	0.5301	0.5246	0.5543	0.5641	0.4883	0.5147	0.4677	0.4762	0.5169	0.5106
Sep	0.8958	0.8731	0.9366	0.9388	0.8852	0.9048	0.8535	0.8724	0.892	0.8701
NDe	0.7252	0.7086	0.7131	0.7017	0.7002	0.6828	0.6654	0.6737	0.6715	0.6639
Pen	0.8011	0.7851	0.8402	0.831	0.8092	0.8092	0.7115	0.7218	0.7667	0.7437
An	0.2692	0.2754	0.214	0.1953	0.2373	0.1764	0.207	0.2106	0.1469	0.2036
Tup	0.9113	0.9118	0.9116	0.9114	0.8884	0.8921	0.9129	0.9127	0.9123	0.9119
Kho	0.8558	0.8502	0.8071	0.7903	0.8801	0.8333	0.8052	0.8146	0.736	0.7378
Alt	0.8384	0.8366	0.85	0.8473	0.8354	0.8484	0.8183	0.8255	0.8308	0.8164
UA	0.8018	0.818	0.7865	0.8002	0.7816	0.7691	0.8292	0.8223	0.8119	0.8197
Sal	0.8788	0.8664	0.8628	0.8336	0.8793	0.8708	0.7941	0.798	0.7865	0.7843
MZ	0.7548	0.7692	0.7476	0.7524	0.7356	0.7212	0.6707	0.6779	0.6731	0.6683

Table 6: GE for families and measures above average.

Family	LDND	LCSD	LDN	LCS	PREFIXD	PREFIX	DICED	DICE	JCD	JCDD	TRIGRAMD
NDe	0.5761	0.5963	0.5556	0.5804	0.5006	0.4749	0.4417	0.4372	0.4089	0.412	0.2841
Bos											
NC	0.4569	0.4437	0.4545	0.4398	0.3384	0.3349	0.3833	0.3893	0.3538	0.3485	0.2925
Hok	0.8054	0.8047	0.8048	0.8124	0.6834	0.6715	0.7987	0.8032	0.7629	0.7592	0.5457
Pan											
Chi	0.5735	0.5775	0.555	0.5464	0.5659	0.5395	0.5616	0.5253	0.5593	0.5551	0.4752
Tup	0.7486	0.7462	0.7698	0.7608	0.6951	0.705	0.7381	0.7386	0.7136	0.7125	0.6818
WP	0.6317	0.6263	0.642	0.6291	0.5583	0.5543	0.5536	0.5535	0.5199	0.5198	0.5076
AuA	0.6385	0.6413	0.5763	0.5759	0.6056	0.538	0.5816	0.5176	0.5734	0.5732	0.5147
Que											
An	0.1799	0.1869	0.1198	0.1003	0.1643	0.0996	0.1432	0.0842	0.1423	0.1492	0.1094
Kho	0.7333	0.7335	0.732	0.7327	0.6826	0.6821	0.6138	0.6176	0.5858	0.582	0.4757
Dra	0.5548	0.5448	0.589	0.5831	0.5699	0.6006	0.5585	0.589	0.5462	0.5457	0.5206
Aus	0.2971	0.2718	0.3092	0.3023	0.2926	0.3063	0.2867	0.257	0.2618	0.2672	0.2487
Tuc											
Ura	0.4442	0.4356	0.6275	0.6184	0.4116	0.6104	0.2806	0.539	0.399	0.3951	0.1021
Arw											
May											
LP	0.41	0.4279	0.4492	0.4748	0.3864	0.4184	0.3323	0.336	0.3157	0.3093	0.1848
OM	0.8095	0.817	0.7996	0.7988	0.7857	0.7852	0.7261	0.7282	0.6941	0.6921	0.6033
Car											
MZ											
TNG	0.5264	0.5325	0.4633	0.4518	0.5	0.472	0.469	0.4579	0.4434	0.4493	0.3295
Bor											
Pen	0.8747	0.8609	0.8662	0.8466	0.8549	0.8505	0.8531	0.8536	0.8321	0.8308	0.7625
MGe	0.6833	0.6976	0.6886	0.6874	0.6086	0.6346	0.6187	0.6449	0.6054	0.6052	0.4518
ST	0.5647	0.5596	0.5435	0.5261	0.5558	0.5412	0.4896	0.4878	0.4788	0.478	0.3116

IE	0.6996	0.6961	0.6462	0.6392	0.6917	0.6363	0.557	0.5294	0.5259	0.5285	0.4541
TK	0.588	0.58	0.5004	0.4959	0.5777	0.4948	0.5366	0.4302	0.5341	0.535	0.4942
Tor	0.4688	0.4699	0.4818	0.483	0.4515	0.4602	0.4071	0.4127	0.375	0.3704	0.3153
Alg	0.3663	0.3459	0.4193	0.4385	0.3456	0.3715	0.2965	0.3328	0.291	0.2626	0.1986
NS	0.6118	0.6072	0.5728	0.5803	0.5587	0.5118	0.578	0.5434	0.5466	0.5429	0.4565
Sko	0.8107	0.8075	0.806	0.7999	0.7842	0.7825	0.6798	0.6766	0.6641	0.6664	0.5636
AA	0.6136	0.6001	0.4681	0.431	0.6031	0.4584	0.5148	0.3291	0.4993	0.4986	0.4123
LSR	0.5995	0.5911	0.6179	0.6153	0.5695	0.5749	0.5763	0.5939	0.5653	0.5529	0.5049
Mar	0.654	0.6306	0.6741	0.6547	0.6192	0.6278	0.568	0.5773	0.5433	0.5366	0.4847
Alt	0.8719	0.8644	0.8632	0.8546	0.8634	0.8533	0.7745	0.7608	0.75	0.7503	0.6492
Hui	0.6821	0.68	0.6832	0.6775	0.6519	0.6593	0.5955	0.597	0.5741	0.5726	0.538
Sep	0.6613	0.656	0.6662	0.6603	0.6587	0.6615	0.6241	0.6252	0.6085	0.6079	0.5769
NDa	0.6342	0.6463	0.6215	0.6151	0.6077	0.5937	0.501	0.5067	0.4884	0.4929	0.4312
Sio											
Kad											
WF											
MUM											
Sal	0.6637	0.642	0.6681	0.6463	0.6364	0.6425	0.5423	0.5467	0.5067	0.5031	0.4637
Kiw											
UA	0.9358	0.9332	0.9296	0.9261	0.9211	0.9135	0.9178	0.9148	0.8951	0.8945	0.8831
Tot											
EA	0.6771	0.6605	0.6639	0.6504	0.6211	0.6037	0.5829	0.5899	0.5317	0.5264	0.4566
HM											
Average	0.619	0.6151	0.6126	0.6069	0.5859	0.5784	0.5495	0.5449	0.5322	0.5302	0.4495

Table 7: RW for families and measures above average.

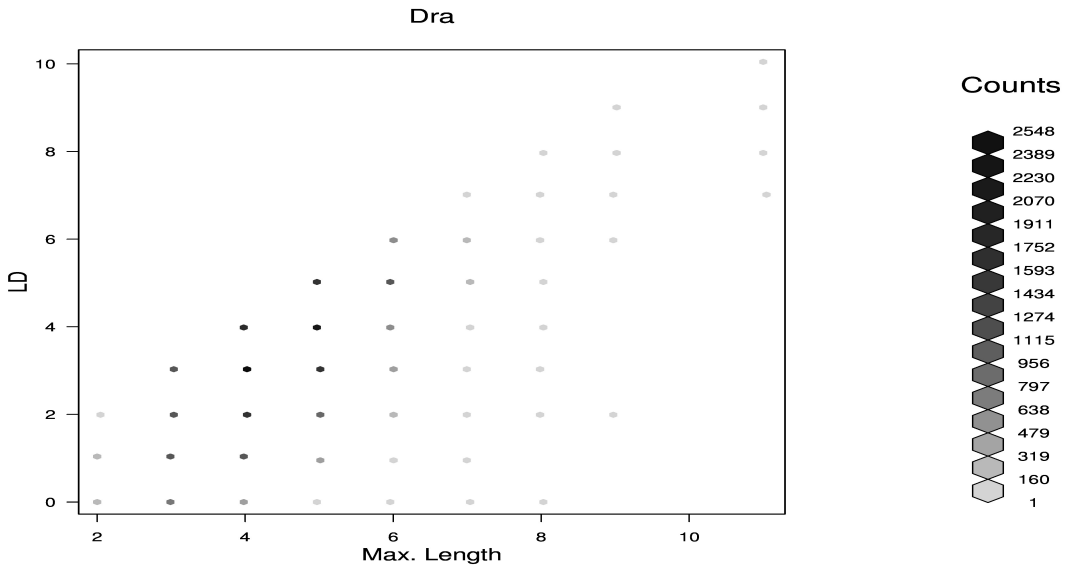

Figure 6: Hexagonally binned plot of same meaning LD and maximum length.

References

Atkinson, Q. D. and R. D. Gray (2005). Curious parallels and curious connections—phylogenetic thinking in biology and historical linguistics. *Systematic Biology*, **54**(4):513–526.

Bakker, D., A. Müller, V. Velupillai, S. Wichmann, C. H. Brown, P. Brown, D. Egorov, R. Mailhammer, A. Grant, and E. W. Holman (2009). Adding typology to lexicostatistics: A combined approach to language classification. *Linguistic Typology*, **13**(1):169–181. ISSN 1430-0532.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**(1):289–300.

Bergsland, K. and H. Vogt (1962). On the validity of glottochronology. *Current Anthropology*, **3**(2):115–153. ISSN 00113204.

Bouchard-Côté, A., D. Hall, T. L. Griffiths, and D. Klein (2013). Automated reconstruction of ancient languages using probabilistic models of sound change. *Proceedings of the National Academy of Sciences*, **110**(11):4224–4229.

Brew, C. and D. McKelvie (1996). Word-pair extraction for lexicography. In *Proceedings of the Second International Conference on New Methods in Language Processing*, pp. 45–55. Ankara.

Brown, C. H., E. W. Holman, S. Wichmann, and V. Velupillai (2008). Automated classification of the world's languages: A description of the method and preliminary results. *Sprachtypologie und Universalienforschung*, **61**(4):285–308.

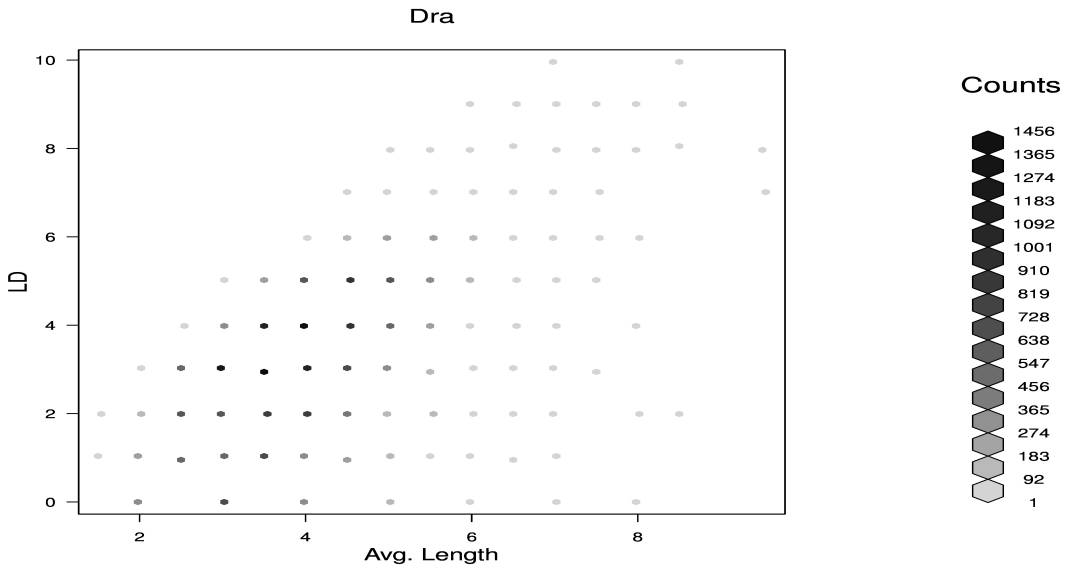

Figure 7: Hexagonally binned plot of same meaning LD and average length.

Carr, D., N. Lewin-Koh, and M. Maechler (2010). hexbin: Hexagonal binning routines. *R package version*, **1**(0).

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

Christiansen, C., T. Mailund, C. Pedersen, M. Randers, M. Stissing, et al. (2006). Fast calculation of the quartet distance between trees of arbitrary degrees. *Algorithms for Molecular Biology*, **1**(1):16.

Dryer, M. S. (2000). Counting genera vs. counting languages. *Linguistic Typology*, **4**:334–350.

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

Durie, M. and M. Ross, eds. (1996). *The comparative method reviewed: regularity and irregularity in language change*. Oxford University Press, USA.

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

Ellegård, A. (1959). Statistical measurement of linguistic relationship. *Language*, **35**(2):131–156.

Ellison, T. M. and S. Kirby (2006). Measuring language divergence by intra-lexical comparison. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pp. 273–280. Association for Computational Linguistics, Sydney, Australia.

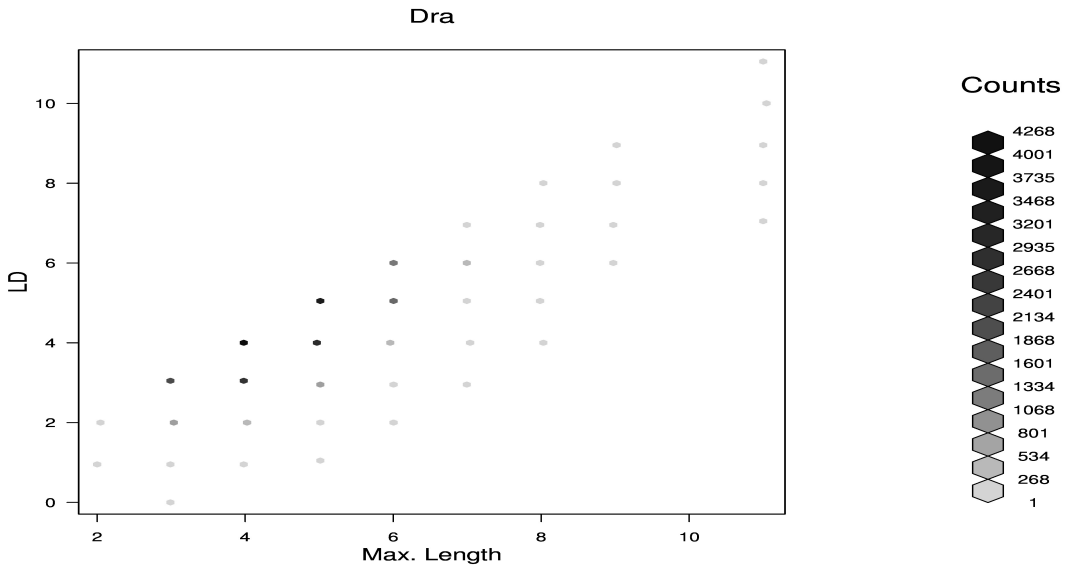

Figure 8: Hexagonally binned plot different meaning LD and maximum length.

Embleton, S. M. (1986). *Statistics in historical linguistics*, volume 30. Brockmeyer.

Felsenstein, J. (2002). PHYLIP (phylogeny inference package) version 3.6 a3. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Felsenstein, J. (2004). *Inferring phylogenies*. Sinauer Associates, Sunderland, Massachusetts.

Gilij, F. S. (2001). Saggio di storia americana, osia Storia naturale, ciuile, e sacra de regni, e delle provincie spagnuole di Terra-ferma nell'America meridional/descrita dall'abate filippo salvadore gilij.-roma: per luigi perego erede salvioni..., 1780-1784. In *Textos clásicos sobre la Historia de Venezuela:[recopilación de libros digitalizados]*, p. 11. MAPFRE.

Goodman, L. A. and W. H. Kruskal (1954). Measures of association for cross classifications. *Journal of the American Statistical Association*, pp. 732–764.

Greenhill, S. J., R. Blust, and R. D. Gray (2008). The Austronesian basic vocabulary database: from bioinformatics to lexomics. *Evolutionary Bioinformatics Online*, **4**:271–283.

Greenhill, S. J. and R. D. Gray (2009). Austronesian language phylogenies: Myths and misconceptions about Bayesian computational methods. *Austronesian Historical Linguistics and Culture History: A Festschrift for Robert Blust*, pp. 375–397.

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

Hauer, B. and G. Kondrak (2011). Clustering semantically equivalent words into cognate sets in multilingual lists. In *Proceedings of 5th International Joint Conference*

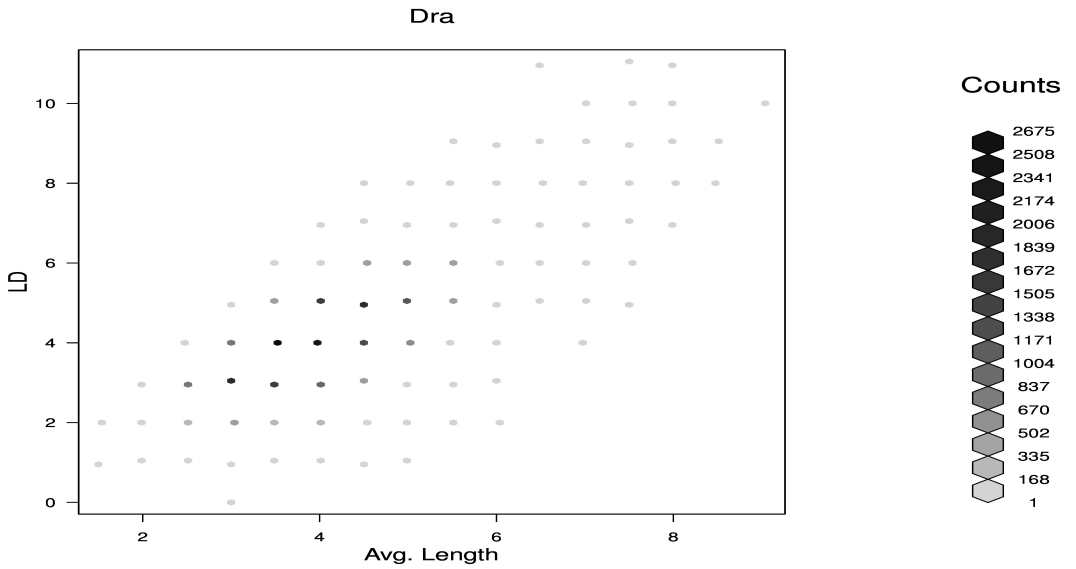

Figure 9: Hexagonally binned plot different meaning LD and average length.

on Natural Language Processing, pp. 865–873. Asian Federation of Natural Language Processing, Chiang Mai, Thailand.

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

Huff, P. and D. Lonsdale (2011). Positing language relationships using ALINE. *Language Dynamics and Change*, **1**(1):128–162.

Huffman, S. M. (1998). *The genetic classification of languages by n-gram analysis: A computational technique*. Ph.D. thesis, Georgetown University, Washington, DC, USA. AAI9839491.

Inkpen, D., O. Frunza, and G. Kondrak (2005). Automatic identification of cognates and false friends in French and English. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pp. 251–257.

Jäger, G. (2014). Phylogenetic inference from word lists using weighted alignment with empirically determined weights. Forthcoming.

Kondrak, G. (2000). A new algorithm for the alignment of phonetic sequences. In *Proceedings of the First Meeting of the North American Chapter of the Association for Computational Linguistics*, pp. 288–295.

Kondrak, G. (2002a). *Algorithms for language reconstruction*. Ph.D. thesis, University of Toronto, Ontario, Canada.

Kondrak, G. (2002b). Determining recurrent sound correspondences by inducing translation models. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pp. 1–7. Association for Computational Linguistics.

Kondrak, G. (2005). N-gram similarity and distance. In *String Processing and Information Retrieval*, pp. 115–126. Springer.

Kondrak, G. and T. Sherif (2006). Evaluation of several phonetic similarity algorithms on the task of cognate identification. In *Proceedings of ACL Workshop on Linguistic Distances*, pp. 43–50. Association for Computational Linguistics.

Kroeber, A. L. and C. D. Chrétien (1937). Quantitative classification of Indo-European languages. *Language*, **13**(2):83–103.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. In *Soviet physics doklady*, volume 10, p. 707.

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

Lin, D. (1998). An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, volume 1, pp. 296–304.

List, J.-M. (2012). LexStat: Automatic detection of cognates in multilingual wordlists. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, pp. 117–125. Association for Computational Linguistics, Avignon, France.

Marzal, A. and E. Vidal (1993). Computation of normalized edit distance and applications. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **15**(9):926–932.

Melamed, D. I. (1999). Bitext maps and alignment via pattern recognition. *Computational Linguistics*, **25**(1):107–130. ISSN 0891-2017.

Nichols, J. (1996). The comparative method as heuristic. In M. Durie and M. Ross, eds., *The comparative method revisited: Regularity and Irregularity in Language Change*, pp. 39–71. Oxford University Press, New York.

Petroni, F. and M. Serva (2010). Measures of lexical distance between languages. *Physica A: Statistical Mechanics and its Applications*, **389**(11):2280–2283. ISSN 0378-4371.

Pompei, S., V. Loreto, and F. Tria (2011). On the accuracy of language trees. *PloS one*, **6**(6):e20109.

Rama, T. and A. K. Singh (2009). From bag of languages to family trees from noisy corpus. In *Proceedings of the International Conference RANLP-2009*, pp. 355–359. Association for Computational Linguistics, Borovets, Bulgaria.

Robinson, D. and L. Foulds (1979). Comparison of weighted labelled trees. *Combinatorial mathematics VI*, pp. 119–126.

Saitou, N. and M. Nei (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, **4**(4):406–425.

Sankoff, D. (1969). *Historical linguistics as stochastic process*. Ph.D. thesis, McGill University.

Sheskin, D. J. (2003). *Handbook of parametric and nonparametric statistical procedures*. Chapman & Hall/CRC Press.

Singh, A. K. (2006). Study of some distance measures for language and encoding identification. In *Proceeding of ACL 2006 Workshop on Linguistic Distances*. Association for Computational Linguistics, Sydney, Australia.

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

Sokal, R. R. and C. D. Michener (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, **38**:1409–1438.

Swadesh, M. (1950). Salish internal relationships. *International Journal of American Linguistics*, **16**(4):157–167.

Swadesh, M. (1952). Lexico-statistic dating of prehistoric ethnic contacts: with special reference to North American Indians and Eskimos. *Proceedings of the American philosophical society*, **96**(4):452–463. ISSN 0003-049X.

Swadesh, M. (1955). Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics*, **21**(2):121–137. ISSN 0020-7071.

Tate, R. F. (1954). Correlation between a discrete and a continuous variable. Point-biserial correlation. *The Annals of mathematical statistics*, **25**(3):603–607.

Wichmann, S., E. W. Holman, D. Bakker, and C. H. Brown (2010a). Evaluating linguistic distance measures. *Physica A: Statistical Mechanics and its Applications*, **389**:3632–3639.

Wichmann, S., A. Müller, V. Velupillai, C. H. Brown, E. W. Holman, P. Brown, M. Urban, S. Sauppe, O. Belyaev, Z. Molochieva, Annkathrin Wett, Dik Bakker, Johann-Mattis List, Dmitry Egorov, Robert Mailhammer, David Beck, and H. Geyer (2010b). The ASJP database (version 12).

Wichmann, S., A. Müller, V. Velupillai, A. Wett, C. H. Brown, Z. Molochieva, S. Sauppe, Eric W. Holman, Pamela Brown, Julia Bishoffberger, Dik Bakker, Johann-Mattis List, Dmitry Egorov, Oleg Belyaev, Matthias Urban, Robert Mailhammer, Helen Geyer, David Beck, Evgenia Korovina, Pattie Epps, Pilar Valenzuela, Anthony Grant, and H. Hammarström (2011). The ASJP database (version 14). http://email.eva.mpg.de/ wichmann/listss14.zip.

Wichmann, S. and T. Rama (2014). Jackknifing the black sheep: Asjp classification performance and austronesian. Submitted to the proceedings of the symposium "Let's talk about trees", National Museum of Ethnology, Osaka, Febr. 9-10, 2013.

Wieling, M., J. Prokić, and J. Nerbonne (2009). Evaluating the pairwise string alignment of pronunciations. In *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education*, pp. 26–34. Association for Computational Linguistics.