How Good are Typological Distances for Determining Genealogical Relationships among Languages?

Taraka Rama[1] *KOLACHINA Prasanth*[2]
(1) Språkbanken, Department of Swedish Language, University of Gothenburg, Gothenburg, Sweden
(2) Language Technologies Research Centre, IIIT-Hyderabad, Hyderabad, India
taraka.rama.kasicheyanula@gu.se, prasanth_k@research.iiit.ac.in

ABSTRACT
The recent availability of typological databases such as World Atlas of Language Structures (WALS) has spurred investigations regarding their utility for language classification, the stability of typological features in genetic linguistics and typological universals across the language families of the world. Existing work on building NLP resources such as parallel corpora, treebanks for under-resourced languages has a lot to gain by taking into consideration insights about inter-language relationships. Since Yarowsky et al. (2001), there have been a number of attempts to create resources for resource-poor languages by projecting information from resource-rich languages using comparable corpora. An important intuition in such work is that syntactic information can be transferred with higher accuracy between languages if they are similar. In this paper, we compare typological distances derived from fifteen vector similarity measures with family internal classifications and also lexical divergence. These results are only a first step towards the use of WALS database in the projection of NLP resources for typologically or genetically similar, yet resource-poor languages.

KEYWORDS: WALS, ASJP, Vector similarity, Internal classification, Typological features.

1 Introduction

There are around 7000 languages in this world (Lewis, 2009) which fall into more than 140 genetic families having descended from a common ancestor. The aim of traditional historical linguistics is to trace the evolutionary path, a tree of extant languages to their extinct common ancestor. Genealogical relationship is not the only characteristic which relates languages; languages can also share structurally common features such as *word order*, *similar phoneme inventory size* and *morphology*. It would be a grave error to posit that two languages are genetically related due to a single common structural feature. There have been attempts in the past (Nichols, 1995) to rank the stability of structural features. Stability implies the resistance of a structural feature to change across space and time. For instance, Dravidian languages have adhered to subject-object-verb (SOV) word order for the last two thousand years (Krishnamurti, 2003; Dunn et al.). Hence, it can be claimed that the structural feature SOV is very stable in the Dravidian language family. Also, structural features have recently been used for inferring the evolutionary tree of a small group of Papuan languages of the Pacific (Dunn et al., 2005).

In the area of computational linguistics, existing work on building NLP resources such as parallel corpora, treebanks for under-resourced languages has a lot to gain by taking into consideration insights about inter-language relationships. For example, Birch et al. (2008) is an interesting example of a work that uses genealogical distances between two language families to predict the difficulty of machine translation. However, the use of typological distances in the development of various NLP tools largely remains unexplored. In this paper, we feed such research by providing robust estimates of inter-language distances and comparing them with family internal classification and also within-family lexical divergence.

The paper is structured as followed. In Section 2, we summarize the related work. Section 3 lists the contributions of this work. Section 4 describes the typological database, lexical database and the criteria for preparing the final dataset. Section 5 presents the different vector similarity measures and the evaluation procedure. The results of our experiments are given in Section 6.

2 Related Work

Dunn et al. (2005) were the first to apply a well-tested computational phylogenetic method (from computational biology), Maximum Parsimony (MP; Felsenstein 2004) to typological features (phonological, syntactic and morphological). They use MP to classify a set of unrelated languages – in Oceania – belonging to two different families. In another related work, Wichmann and Saunders (2007) apply three different phylogenetic algorithms – Neighbor Joining (Saitou and Nei, 1987), MP and Bayesian inference (Huelsenbeck et al., 2001) – to the typological features (from WALS) of 63 native American languages. They also ranked the typological features in terms of stability. Nichols and Warnow (2008) survey the use of typological features for language classification in computational historical linguistics. In a novel work, Bakker et al. (2009) combine typological distances with lexical similarity to boost the language classification accuracy. As a first step, they compute the pair-wise typological distances for 355 languages, obtained through the application of length normalized Hamming distance to 85 typological features (ranked by Wichmann and Holman 2009). They combine the typological distances with lexical divergence, derived from lexicostatistical lists, to boost language classification accuracy. Unfortunately, these works seem to have gone unnoticed in computational linguistics.

Typological feature such as phoneme inventory size (extracted from WALS database; Haspelmath et al. 2011) was used by Atkinson (2011) to claim that the phoneme inventory size shows a

negative correlation as one moves away from Africa[1]. In another work, Dunn et al. (2011) make an effort towards demonstrating that there are lineage specific trends in the word order universals across the families of the world.

In computational linguistics, Daume III (2009) and Georgi et al. (2010) use typological features from WALS for investigating relation between phylogenetic groups and feature stability. Georgi et al. (2010) motivate the use of typological features for projecting linguistic resources such as treebanks and bootstrapping NLP tools from "resource-rich" to "low-resource" languages which are genetically unrelated yet, share similar syntactic features due to contact (ex., Swedish to Finnish or vice-versa). Georgi et al. (2010) compute pair-wise distances from typological feature vectors using cosine similarity and a shared overlap measure (ratio of number of shared features to the total number of features, between a pair of feature vectors). They apply three different clustering algorithms – k-means, partitional, agglomerative – to the WALS dataset with number of clusters as testing parameter and observe that the clustering performance measure (in terms of F-score) is not the best when the number of clusters agree with the exact number of families (121) in the whole-world dataset. They find that the simplest clustering algorithm, k-means, wins across all the three datasets. However, the authors do not correct for geographical bias in the dataset.

3 Contributions

In this article, we do not investigate the topic of feature stability or prediction accuracy of clustering methods discussed in Georgi et al. (2010). Instead, we try to answer the following questions:

- Do we really need a clustering algorithm to measure the internal classification accuracy of a language family?
- How well do the typological distances within a family correlate with the lexical distances derived from lexicostatistical lists (Swadesh, 1952; Wichmann et al., 2011b), originally proposed for language classification?
- Given that there are more than dozen vector similarity measures, which vector similarity measure is best suited for the above mentioned tasks?

4 Database

In this section, we describe a database of typological features, referred to as WALS and a lexicostatistical database called *Automated Similarity Judgment Program* (ASJP), which are used in our experiments.

4.1 WALS

The WALS database[2] has 144 feature classes for 2676 languages distributed across the world. As noted in Hammarström (2009), the WALS database is sparse across many language families of the world and the dataset needs to be pruned before it is used for further investigations. The database is represented as matrix of languages vs. features. The pruning of the dataset has to be done in both the directions to avoid sparsity when computing the pair-wise distances between languages. Following Georgi et al. (2010), we remove all the languages which have less than 25 attested features. We also remove features with less than 10% attestations. This leaves the

[1]Assuming a mono-genesis hypothesis of language similar to the mono-genesis hypothesis of *homo sapiens*.
[2]Accessed on 2011-09-22.

dataset with 1159 languages and 193 features. Our dataset includes only those families having more than 10 languages (following Wichmann et al. 2010), shown in Table 1. Georgi et al. (2010) work with a pruned dataset of 735 languages and two major families Indo-European and Sino-Tibetan whereas, we stick to investigating the questions in Section 3 for the well-defined language families – Austronesian, Afro-Asiatic – given in Table 1.

Family	Count	Family	Count
Austronesian	150 (141)	Austro-Asiatic	22 (21)
Niger-Congo	143 (123)	Oto-Manguean	18 (14)
Sino-Tibetan	81 (68)	Arawakan	17 (17)
Australian	73 (65)	Uralic	15 (12)
Nilo-Saharan	69 (62)	Penutian	14 (11)
Afro-Asiatic	68 (57)	Nakh-Daghestanian	13 (13)
Indo-European	60 (56)	Tupian	13 (12)
Trans-New Guinea	43 (33)	Hokan	12 (12)
Uto-Aztecan	28 (26)	Dravidian	10 (9)
Altaic	27 (26)	Mayan	10 (7)

Table 1: Number of languages in each family. The number in parenthesis for each family gives the number of languages present in the database after mapping with ASJP database.

4.2 ASJP

A international consortium of scholars (calling themselves ASJP; Brown et al. 2008) started collecting Swadesh word lists (Swadesh, 1952) (a short concept meaning list usually ranging from 40–200) for most of the world's languages (more than 58%), in the hope of automatizing the language classification of world's languages [3]. The ASJP lexical items are transcribed using a broad phonetic transcription called ASJP Code (Brown et al., 2008). The ASJP Code collapses distinctions in vowel length, stress, tone and reduces all click sounds to a single click symbol. This database has word lists for a language (given by its unique ISO 693-3 code as well as WALS code) and its dialects. We use the WALS code to map the languages in WALS database with that of ASJP database. Whenever a language with a WALS code has more than one word list in ASJP database, we chose to retain the first language for our experiments. An excerpt of word list for Russian is shown in Figure 1. The first line consists of name of language, WALS classification (Indo-European family and Slavic genus), followed by Ethnologue classification (informing that Russian belongs to Eastern Slavic subgroup of Indo-European family). The second line consists of the latitude, longitude, number of speakers, WALS code and ISO 639-3 code. Lexical items begin from the third line.

4.3 Binarization

Each feature in the WALS dataset is either a binary feature (presence or absence of the feature in a language) or a multi-valued feature, coded as a discrete integers over a finite range. Georgi et al. (2010) binarize the feature values by recording the presence or absence of a feature value in a language. This binarization greatly expands the length of the feature vector for a language but allows to represent a wide-ranged feature such as *word order* (which has 7 feature values) in terms of a sequence of 1's and 0's. The issue of binary vs. multi-valued features has been a

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

```
RUSSIAN{IE.SLAVIC|Indo-European,Slavic,East}
 1   56.00   38.00   145031551   rus   rus
1 I      ya //
2 you    t3, v3 //
3 we     m3 //
4 this   iEt3 //
5 that   to //
6 who    kto //
7 what   tato //
8 not    ny~E //
9 all    fsy~e //
10 many imnogy~i //
```

Figure 1: 10 lexical items in Russian.

point of debate in genetic linguistics and has been shown to not give very different results for the Indo-European classification (Atkinson and Gray, 2006).

5 Measures

In this section, we list the 15 vector similarity measures (shown in Table 2), followed by a description of the evaluation measure used in our work to compare the typological distances to WALS classification. We also describe the procedure used to compute lexical divergence from the ASJP lists.

Vector similarity	
euclidean	$\sqrt[2]{\Sigma_{i=1}^{n}(v_1^i - v_2^i)^2}$
seuclidean	$\Sigma_{i=1}^{n}(v_1^i - v_2^i)^2$
nseuclidean	$\frac{\|\sigma_1 - \sigma_2\|}{2 * \|\sigma_1\| + \|\sigma_2\|}$
manhattan	$\Sigma_{i=1}^{n}\|v_1^i - v_2^i\|$
chessboard	$max((v_1^i - v_2^i) \forall i \in (1,n))$
braycurtis	$\frac{\Sigma_{i=1}^{n}\|v_1^i - v_2^i\|}{\Sigma_{i=1}^{n}\|v_1^i + v_2^i\|}$
cosine	$\frac{v_1 \cdot v_2}{\|v_1\| * \|v_2\|}$
correlation	$1 - \frac{\sigma_1 \cdot \sigma_2}{\|\sigma_1\| * \|\sigma_2\|}$

Boolean similarity	
hamming	$\#_{\neq 0}(v_1 \hat{} v_2)$
jaccard	$\frac{\#_{\neq 0}(v_1 \hat{} v_2)}{\#_{\neq 0}(v_1 \hat{} v_2) + \#_{\neq 0}(v_1 \& v_2)}$
tanimoto	$\frac{2 * \#_{\neq 0}(v_1 \hat{} v_2)}{\#_{\neq 0}(v_1 \& v_2) + \#_{=0}(v_1 \| v_2) + 2 * \#_{\neq 0}(v_1 \hat{} v_2)}$
matching	$\frac{\#_{\neq 0}(v_1 \hat{} v_2)}{\#v_1}$
dice	$\frac{\#_{\neq 0}(v_1 \hat{} v_2)}{\#_{\neq 0}(v_1 \hat{} v_2) + 2 * \#_{\neq 0}(v_1 \& v_2)}$
sokalsneath	$\frac{2 * \#_{\neq 0}(v_1 \hat{} v_2)}{2 * \#_{\neq 0}(v_1 \hat{} v_2) + \#_{\neq 0}(v_1 \& v_2)}$
russellrao	$\frac{\#_{\neq 0}(v_1 \hat{} v_2) + \#_{=0}(v_1 \| v_2)}{\#v_1}$
yule	$\frac{2 * \#_{\neq 0}(v_1 - v_2) * \#_{=0}(v_1 - v_2)}{\#_{\neq 0}(v_1 - v_2) * \#_{=0}(v_1 - v_2) + \#_{\neq 0}(v_1 \& v_2) * \#_{=0}(v_1 \| v_2)}$

Table 2: Different vector similarity measures used in our experiments (distance computed between v_1 and v_2). In vector similarity measures, $\|\|$ represents the L_2 norm of the vector, and σ represents the difference from mean of vector (μ_1) i.e. $(v_1 - \mu_1)$. Similarly, for the boolean similarity measures, $\hat{}$ stands for the logical XOR operation between bit vectors while & and $\|$ stand for logical AND and OR operations respectively. $\#_{\neq 0}(\cdot)$ stands for number of non-zero bits in a boolean vector.

5.1 Internal classification accuracy

Apart from typological information for the world's languages, WALS also provides a two-level classification of a language family. In the WALS classification, the top level is the family name, the next level is genus and a language rests at the bottom. For instance, Indo-European family has 10 genera. Genus is a consensually defined unit and not a rigorously established genealogical unit (Hammarström, 2009). Rather, a genus corresponds to a group of languages

which are supposed to have descended from a proto-language which is about 3500 to 4000 years old. For instance, WALS lists Indic and Iranian languages as separate genera whereas, both the genera are actually descendants of Proto-Indo-Iranian which in turn descended from Proto-Indo-European – a fact well-known in historical linguistics (Campbell and Poser, 2008).

The WALS classification for each language family listed in Table 1, can be represented as a 2D-matrix with languages along both rows and columns. Each cell of such a matrix represents the WALS relationship in a language pair in the family. A cell has 0 if a language pair belong to the same genus and 1 if they belong to different genera. The pair-wise distance matrix obtained from each vector similarity measure is compared to the 2D-matrix using a special case of pearson's r, called point-biserial correlation [4].

5.2 Lexical distance

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). The performance of LDND distance matrices was evaluated against two expert classifications of world's languages in at least two recent works (Pompei et al., 2011; Wichmann et al., 2011a). Their findings confirm that the LDND matrices largely agree with the classification given by historical linguists. This result puts us on a strong ground to use ASJP's LDND as a measure of lexical divergence within a family.

The distribution of the languages included in this study is plotted in Figure 2.

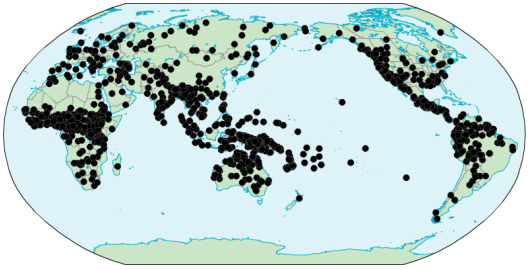

Figure 2: Visual representation of world's languages in the final dataset.

The correlation between typological distances and lexical distances is (within a family) computed as the spearman's rank correlation ρ between the typological and lexical distances for all language pairs in the family. It is worth noting that Bakker et al. (2009) also compare LDND distance matrices with WALS distance matrices for 355 languages from various families using a pearson's r whereas, we compare within-family LDND matrices with WALS distance matrices derived from 15 similarity measures.

6 Results

In this section, we present and discuss the results of our experiments in internal classification and correlation with lexical divergence. We use heat maps to visualize the correlation matrices resulting from both experiments.

[4] http://en.wikipedia.org/wiki/Point-biserial_correlation_coefficient

6.1 Internal classification

The point bi-serial correlation, r, introduced in Section 5, lies in the range of -1 to $+1$. The value of r is blank for Arawakan and Mayan families since both families have a single genus in their respective WALS classifications. Subsequently, r is shown in white for both of these families. Chessboard measure is blank across all language families since it gives a single score of 1 between two different binary vectors. Interestingly, all vector similarity measures perform well for Australian, Austro-Asiatic, Indo-European and Sino-Tibetan language families, except for 'russellrao'. We take this result to be encouraging since they consist of more than 33% of the total languages in the sample given in Table 1. Among the measures, 'matching', 'seuclidean', 'tanimoto', 'euclidean', 'hamming' and 'manhattan' perform the best across the four families. Interestingly, the widely used 'cosine' measure does not perform as well as 'hamming'. None of the vector similarity measures seem to perform well for Austronesian and Niger-Congo families which have more than 14% and 11% of the world's languages respectively. The worst performing language family is Tupian. This does not come as a surprise, since Tupian has 5 genera with one language in each and a single genus comprising the rest of family. Australian and Austro-Asiatic families shows the maximum correlation across 'seuclidean', 'tanimoto', 'euclidean', 'hamming' and 'manhattan'.

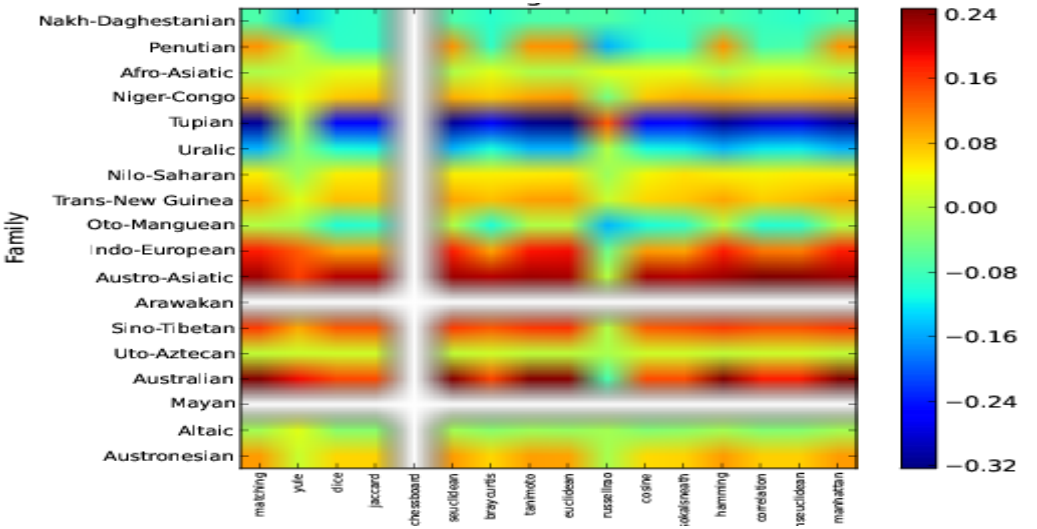

Figure 3: Heatmap showing the gradience of r across different language families and vector similarity measures.

6.2 Lexical divergence

The rank correlation between LDND and vector similarity measures is high across Australian, Sino-Tibetan, Uralic, Indo-European and Niger-Congo families. The 'Russel-Rao' measure works the best for families – Arawakan, Austro-Asiatic, Tupian and Afro-Asiatic – which otherwise have poor correlation scores for the rest of measures. The maximum correlation is for 'yule' measure in Uralic family. Indo-European, the well-studied family, shows a correlation from 0.08 to the maximum possible correlation across all measures, except for 'Russell-Rao' and 'Bray-Curtis' distances. The Hokan family shows the lowest amount of correlations across all

distance measures. One possible reason for this could be the controversial nature of the family, with a lack of proper consensus among historical linguistics regarding its status as a seperate language family.

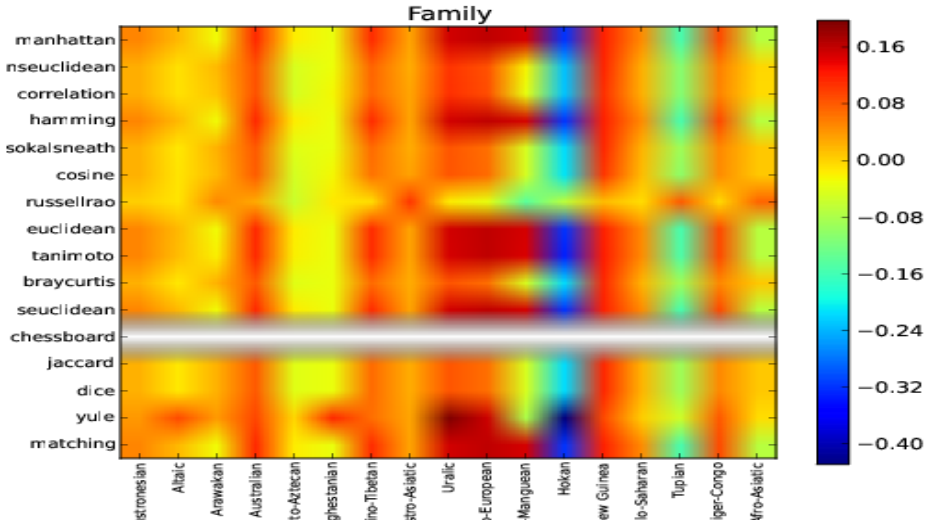

Figure 4: Heatmap showing the gradience of ρ across different families and vector similarity measures.

Conclusion

In summary, choosing the right vector similarity measure when calculating typological distances makes a difference in the internal classification accuracy. The choice of similarity measure does not influence the correlation between WALS distances and LDND distances within a family. The internal classification accuracies are similar to the accuracies reported in Bakker et al. (2009). Our correlation matrix suggests that internal classification accuracies of LDND matrices (reported in Bakker et al. 2009) can be boosted through the right combination of typological distances and lexical distances. In our experiments, we did not control for feature stability and experimented on all available features. By choosing a smaller set of typological features (from the ranking of Wichmann and Holman (2009)) and right similarity measure one might achieve higher accuracies. The current rate of language extinction is unprecedented in human history. Our findings might be helpful in speeding up the language classification of many small dying families by serving as a springboard for traditional historical linguists.

Acknowledgments

The research presented here was supported by the Swedish Research Council (the project Digital areal linguistics, VR dnr 2009-1448) and by the University of Gothenburg through its support of the Centre for Language Technology and of Språkbanken. We would like to thank Harald Hammarström, Lars Borin and Søren Wichmann for the discussions and their insights into this work. We would also like to thank the anonymous reviewers for their comments on the paper.

References

Atkinson, Q. and Gray, R. (2006). How old is the Indo-European language family? Progress or more moths to the flame. *Phylogenetic Methods and the Prehistory of Languages (Forster P, Renfrew C, eds)*, pages 91–109.

Atkinson, Q. D. (2011). Phonemic diversity supports a serial founder effect model of language expansion from Africa. *Science*, 332(6027):346.

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

Birch, A., Osborne, M., and Koehn, P. (2008). Predicting Success in Machine Translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 745–754, Honolulu, Hawaii. Association for Computational Linguistics.

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

Campbell, L. and Poser, W. (2008). *Language classification: history and method*. Cambridge University Press.

Daume III, H. (2009). Non-parametric bayesian areal linguistics. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 593–601. Association for Computational Linguistics.

Dunn, M., Greenhill, S., Levinson, S., and Gray, R. (2011). Evolved structure of language shows lineage-specific trends in word-order universals. *Nature*, 473(7345):79–82.

Dunn, M., Levinson, S., and Lindström, E. Ger reesink and angela terrill 2008. structural phylogeny in historical linguistics: Methodological explorations applied in island melanesia. *Language*, 84(4):710–59.

Dunn, M., Terrill, A., Reesink, G., Foley, R., and Levinson, S. (2005). Structural phylogenetics and the reconstruction of ancient language history. *Science*, 309(5743):2072–2075.

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

Georgi, R., Xia, F., and Lewis, W. (2010). Comparing language similarity across genetic and typologically-based groupings. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 385–393. Association for Computational Linguistics.

Hammarström, H. (2009). Sampling and genealogical coverage in the wals. *Linguistic Typology*, 13(1):105–119. Plus 198pp appendix.

Haspelmath, M., Dryer, M. S., Gil, D., and 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., and Bakker, D. (2008). Advances in automated language classification. In Arppe, A., Sinnemäki, K., and Nikanne, U., editors, *Quantitative Investigations in Theoretical Linguistics*, pages 40–43, Helsinki: University of Helsinki.

Huelsenbeck, J., Ronquist, F., et al. (2001). Mrbayes: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8):754–755.

Krishnamurti, B. (2003). *The Dravidian languages*. Cambridge Univ. Press.

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

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

Nichols, J. (1995). Diachronically stable structural features. *Historical Linguistics 1993. Selected Papers from the 11th International Conference on Historical Linguistics*, pages 337–355.

Nichols, J. and Warnow, T. (2008). Tutorial on computational linguistic phylogeny. *Language and Linguistics Compass*, 2(5):760–820.

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

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

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.

Wichmann, S. and Holman, E. (2009). Assessing temporal stability for linguistic typological features. *München: LINCOM Europa*.

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

Wichmann, S., Holman, E. W., Rama, T., and Walker, R. S. (2011a). Correlates of reticulation in linguistic phylogenies. *Language Dynamics and Change*, 2. In press.

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., and Hammarström, H. (2011b). The ASJP database (version 14). http://email.eva.mpg.de/ wichmann/listss14.zip.

Wichmann, S. and Saunders, A. (2007). How to use typological databases in historical linguistic research. *Diachronica*, 24(2):373–404.

Yarowsky, D., Ngai, G., and Wicentowski, R. (2001). Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the first international conference on Human language technology research*, pages 1–8. Association for Computational Linguistics.