Comparing Language Similarity across Genetic and Typologically-Based Groupings

Ryan Georgi
University of Washington
rgeorgi@uw.edu

Fei Xia
University of Washington
fxia@uw.edu

William Lewis
Microsoft Research
wilewis@microsoft.com

Abstract

Recent studies have shown the potential benefits of leveraging resources for resource-rich languages to build tools for similar, but resource-poor languages. We examine what constitutes "similarity" by comparing traditional phylogenetic language groups, which are motivated largely by genetic relationships, with language groupings formed by clustering methods using typological features only. Using data from the World Atlas of Language Structures (WALS), our preliminary experiments show that typologically-based clusters look quite different from genetic groups, but perform as good or better when used to predict feature values of member languages.

1 Introduction

While there are more than six thousand languages in the world, only a small portion of these languages have received substantial attention in the field of NLP. With the increase in use of data-driven methods, languages with few or no electronic resources have been difficult to process with current methods. The morphological tagging of Russian using Czech resources as done by (Hana et al., 2004) shows the potential benefit for using the resources of resource-rich languages to bootstrap NLP tools for related languages. Projecting syntactic structures across languages (Yarowsky and Ngai, 2001; Xia and Lewis, 2007) is another possible way to harness existing tools, though such projection is more reliable among languages with similar syntax.

Studies such as these show the possible benefits of working with similar languages. A crucial question is how we should define similarity between languages. While genetically related languages tend to have similar typological features as they could inherit the features from their common ancestor, they could also differ a lot due to language change over time. On the other hand, languages with no common ancestor could share many features due to language contact and other factors.

It is worth noting that the goals of historical linguistics differ from those of language typology in that while historical linguistics focuses primarily on diachronic language change, typology is more focused on a synchronic survey of features found in the world's languages: what typological features exist, where they are found, and why a language has a feature.

These differences between the concepts of genetic relatedness and language similarities lead us to the following questions:

Q1. If we cluster languages based only on their typological features, how do the induced clusters compare to phylogenetic groupings?

Q2. How well do induced clusters and genetic families perform in predicting values for typological features?

Q3. What typological features tend to stay the same within language families, and what features are likely to differ?

These questions are the focus of this study, and for the experiments, we use information from World Atlas of Language Structures (Haspelmath et al., 2005), or WALS.

Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 385–393,
Beijing, August 2010

ID#	Feature Name	Category	Feature Values
1	Consonant Inventories	Phonology (19)	{1:Large, 2:Small, 3:Moderately Small, 4:Moderately Large, 5:Average}
23	Locus of Marking in the Clause	Morphology (10)	{1:Head, 2:None, 3:Dependent, 4:Double, 5:Other}
30	Number of Genders	Nominal Categories (28)	{1:Three, 2:None, 3:Two, 4:Four, 5:Five or More}
58	Obligatory Possessive Inflection	Nominal Syntax (7)	{1:Absent, 2:Exists}
66	The Perfect	Verbal Categories (16)	{1:None, 2:Other, 3:From 'finish' or 'already', 4:From Possessive}
81	Order of Subject, Object and Verb	Word Order (17)	{1:SVO, 2:SOV, 3:No Dominant Order, 4:VSO, 5:VOS, 6:OVS, 7:OSV}
121	Comparative Constructions	Simple Clauses (24)	{1:Conjoined, 2:Locational, 3:Particle, 4:Exceed}
125	Purpose Clauses	Complex Sentences (7)	{1:Balanced/deranked, 2:Deranked, 3:Balanced}
138	Tea	Lexicon (10)	{1:Other, 2:Derived from Sinitic 'cha', 3:Derived from Chinese 'te'}
140	Question Particles in Sign Languages	Sign Languages (2)	{1:None, 2:One, 3:More than one}
142	Para-Linguistic Usages of Clicks	Other (2)	{1:Logical meanings, 2:Affective meanings, 3:Other or none}

Table 1: Sample features and their values used in the WALS database. There are eleven feature categories in WALS, one feature from each is given here. The numbers in parentheses in the 'Category' column are the total number of features in that category. Feature values are given with both the integers that represent them in the database and their description in the form {#:description}.

2 WALS

The WALS project consists of a database that catalogs linguistic features for over 2,556 languages in 208 language families, using 142 features in 11 different categories.[1] Table 1 shows a small sample of features, one feature from each category in WALS. Listed are the ID number for each example, the feature category, and the possible values for that feature.

WALS as a resource, however, is primarily designed for surveying the distribution of particular typological features worldwide, not comparing languages. The authors of WALS compiled their data from a wide array of primary sources, but these sources do not always cover the same sets of features or languages.

If we conceive of the WALS database as a two-dimensional matrix with languages along one dimension and features along the other, then only 16% of the cells in that matrix are filled. An empty cell in the matrix means the feature value for the (language, feature) pair is *not-specified (NS)*. Even well-studied languages could have many empty cells in WALS, and this kind of data sparsity presents serious problems to clustering algorithms that cannot handle unknown values. To address the data sparsity problem, we experiment with different pruning criteria to create a new matrix that is reasonably dense for our study.

[1]Our copy of the database was downloaded from `http://wals.info` in June of 2009 and appears to differ slightly from the statistics given on the website at the time of writing. Currently, the WALS website reports 2,650 languages, with 141 features in use.

2.1 Pruning Methods

Answering questions Q1–Q3 is difficult if there are too many empty cells in the data. Pruning the data to produce a smaller but denser subset can be done by one or more of the following methods.

Prune Languages by Minimum Features

Perhaps the most straightforward method of pruning is to eliminate languages that fail to contain some minimum number of features. Following Daumé (2009), we require languages to have a minimum of 25 features for the whole-world set, or 10 features for comparing across subfamilies. This eliminates many languages that simply do not have enough features to be adequately represented.

Prune Features by Minimum Coverage

The values for some features, such as those specific to sign languages, are provided only for a very small number of languages. Taking this into account, in addition to removing languages with a small number of features, it is also helpful to remove features that only cover a small portion of languages. Again we choose the thresholds selected by Daumé (2009) for pruning features that do not cover more than 10% of the selected languages in the whole-world set, and 25% in comparisons across subfamilies.

Use a Dense Language Family

Finally, using a well-studied family with a number of subfamilies can produce data sets with less sparsity. When clustering methods are used with this data, the groups correspond to subfamilies

Data Set	Min Features	Min Coverage	Grouped By	# Langs	# Groups	# Features	Density
Unpruned	0	0%	Family	2556	208	142	16.0%
Whole-World	25	10%	Family	735	121	139	39.7%
Indo-European	10	25%	Subfamily	87	10	64	44.9%
Sino-Tibetan	10	25%	Subfamily	96	14	64	38.6%

Table 2: Data sets and pruning options used for this paper. Density $= \frac{|Filled\ Cells|}{|Total\ Cells|} \cdot 100$

rather than families. In this study, we choose two families: Indo-European and Sino-Tibetan.

The resulting data sets after various methods of pruning can be seen in Table 2.

2.2 Features and Feature Values

Besides dealing with the sparsity of the features, the actual representation of the features in WALS needs to be taken into account. As can be seen in Table 1, features are represented with a range of discrete integer values. Some features, such as #58–Obligatory Possessive Inflection–are essentially binary features with values "`Absent`" or "`Exists`". Others, such as #1–Consonant Inventories–appear to be indices along some dimension related to size, ranging from small to large. Features such as these might conceivably be viewed as on a continuum where closer distances between values suggests closer relationship between languages.

Still other features, such as #81–Order of Subject, Object, and Verb–have multiple values but cannot be clearly be treated using distance measures. It's unclear how such a distance would vary between an SOV language and either VSO or VOS languages.

Binarization

Clustering algorithms use similarity functions, and some functions may simply check whether two languages have the same value for a feature. In these cases, no feature binarization is needed. If a clustering algorithm requires each data point (a language in this case) to be presented as a feature vector, features with more than two categorical values should be binarized. We simply treat a feature with k possible values as k binary features. There are other ways to binarize features. For instance, Daumé (2009) chose one feature value as the "canonical" value and grouped the other values into the second value (personal communication). We did not use this approach as it is not clear to us which values should be selected as the "canonical" ones.

3 Experimental Setup

To get a picture of how clustering methods compare to genetic groupings, we looked at three elements: cluster similarity, prediction capability, and feature selection.

3.1 Clustering

Our first experiment is designed to address question Q1: how do induced clusters compare to phylogenetic groupings?

Clustering Methods

For clustering, two clustering packages were used. First, we implemented the k-medoids algorithm, a partitional algorithm similar to k-means, but using median instead of mean distance for cluster centers (Estivill-Castro and Yang, 2000).

Second, we used a variety of methods from the CLUTO (Steinbach et al., 2000) clustering toolkit: repeated-bisection (`rb`), a k-means implementation (`direct`), an agglomerative algorithm (`agglo`) using UPGMA to produce hierarchical clusters, and `bagglo`, a variant of `agglo`, which biases the agglomerative algorithm using partitional clusters.

Similarity Measures

For similarity measures, we used CLUTO's default cosine similarity measure (`cos`), but also implemented another similarity measure `shared_overlap` designed to handle empty cells. Given two languages A and B, $shared\_overlap(A, B)$ is defined to be $\frac{\#\text{ Of Features with Same Values}}{\#\text{ Features Both Filled Out in WALS}}$. This measure can handle language pairs with many empty cells in WALS as it uses only features with cells

a is the number of language pairs found in the same set in both clusterings.
b is the number of language pairs found in different sets in C_1, and different sets in C_2.
c is the number of language pairs found in the same set in C_1, but in different sets in C_2.
d is the number of language pairs found in different sets in C_1, but the same set in C_2.

(a) Variables Used In Calculations

$$R(C_1, C_2) = \frac{a+b}{a+b+c+d}$$

(b) Rand Index

$$Precision(C_1, C_2) = \frac{a}{a+c}$$

(c) Cluster precision

$$Recall(C_1, C_2) = \frac{a}{a+d}$$

(d) Cluster recall

$$Fscore(C_1, C_2) = \frac{2 \cdot (Precision \cdot Recall)}{Precision + Recall}$$

(e) Cluster f-score

Figure 1: Formulas for calculating the Rand Index, cluster precision, recall, and f-score of two clusterings C_1 and C_2. C_1 is the system output, C_2 is the gold standard.

filled out for both languages, and calculates the percentage of features with the same values.

3.2 Clustering Performance Metrics

To measure clustering performance, we treat the genetic families specified in WALS as the gold standard, although we are not strictly aiming to recreate them.

Rand Index

The Rand Index (Rand, 1971) is one of the standard metrics for evaluating clustering results. It compares pairwise assignments of data points across two clusterings. For every pair of points there are four possibilities, as given in Figure 1. The Rand index is calculated by dividing the number of matching pairs ($a + b$) by the number of all pairs. This results in a number between 0 and 1 where 1 represents an identical clustering. Unfortunately, as noted by (Daumé and Marcu, 2005), the Rand Index tends to give disproportionately greater scores to clusterings with a greater number of clusters. For example, the Rand Index will always be 1.0 when each data point belongs to its own cluster. As a result, we have chosen to calculate metrics other than the Rand index: cluster precision, recall, and f-score.

Cluster Precision, Recall, and F-Score

Extending the notation in Figure 1, precision is defined as the proportion of same-set pairs in the target cluster C_1 that are correctly identified as being in the same set in the gold cluster C_2, while recall is the proportion of all same-set pairs in the gold cluster C_2 that are identified in the target cluster C_1. F-score is calculated as the usual harmonic mean of precision and recall. As it gives a more accurate representation of cluster similarity across varying amounts of clusters, we will report cluster similarity using cluster F-score.

3.3 Prediction Accuracy

Our second experiment was to answer the question posed in Q2: how do induced clusters and genetic families compare in predicting the values of features for languages in the same group?

To answer this question, we measure the accuracy of the prediction when both types of groups are used to predict the values of "empty" cells. We used 90% of the filled cells to build clusters, and then predicted the values of the remaining 10% of filled cells. The missing cells are filled with the value that occurs the most times among languages in the same group. If there are no other languages in the cluster, or the other languages have no values for this feature, then the cell is filled with the most common values for that feature across all languages in the dataset. Finally, the accuracy is calculated by comparing these predicted values with the actual values in the gold standard. We run 10-fold cross validation and report the average accuracy.

In addition to the prediction accuracy for each method of producing groupings, we calculate the baseline result where an empty cell is filled with the most frequent value for that feature across all the languages in the training data.

3.4 Determining Feature Stability

Finally, we look to answer Q3: what typological features tend to stay the same within related families? To find an answer, we look again to prediction accuracy. While prediction accuracy can be averaged across all features, it can also be broken down feature-by-feature to rank features according to how accurately they can be predicted

by language families. Features that can be predicted with high accuracy implies that these features are more likely to remain *stable* within a language family than others.

Using prediction accuracies based on the genetic families, we rank features according to their accuracy and then perform clustering using the top features to determine if the cluster similarity to the genetic groups increases when using only the stable features.

4 Results & Analysis

4.1 Cluster Similarity

The graph in Figure 2(a) shows f-scores of clustering methods with the whole-world set. None achieve an f-score greater than 0.15, and most perform even worse when the number of clusters matches the number of genetic families or subfamilies. This indicates that the induced clusters based on typological features are very different from genetic groupings.

The question of similarity between these induced clusters and the genetic families is however a separate one from how those clusters perform in predicting typological feature values.

4.2 Prediction Accuracy

To determine the amount of similarity between languages within clusters, we instead look at prediction accuracy across clustering methods and the genetic groups. These scores are similar to those given in Daumé (2009), though not directly comparable due to small discrepancies in the size of the data set. As can be seen by the numbers in Table 3 and the graph in 2(b), despite the lack of similarity between clustering methods and the genetic groups, the clustering methods produce as good or better prediction accuracies. Furthermore, the `agglo` and `bagglo` hierarchical clustering methods which are favored for producing phylogenetically motivated clusters do indeed result in higher f-score similarity to the genetic clusters than the partitional `rb` and `direct` methods, but produce poorer prediction-accuracy results.

In fact, it is not surprising that some induced clusters outperform the genetic groupings in prediction accuracy, considering that clustering algorithms often want to maximize the similarity between languages in the same clusters. Now that we know similarity between languages does not necessarily mirror language family membership, the next question is what features tend to stay the same among languages in the same language families.

4.3 Feature Selection

Our final experiment was to examine the features in WALS themselves, and look for features that appear to vary the least within families, and act as better predictors of family membership.

In order to do this, we again looked at prediction accuracy information on a feature-by-feature basis. The results from this experiment are shown in Table 4, which gives a breakdown of how features rank both individually and by category.

Since this table is built upon genetic relationships, it is not surprising that the category for "Lexicon" appears to be the most reliably stable category. As noted in (McMahon, 1994), lexical cognates are often used as good evidence for determining a shared ancestry. We also find that word order is rather stable within a family.

We ran one further experiment where, using the `agglo` clustering method that provided clusters most similar to the genetic families previously, only features that showed accuracies above 50%. This eliminated 28 features, leaving 111 higher-scoring features for the whole-world set. Pruning the features to use only these selected for their stability within the genetic groupings yielded a very small increase in f-score similarity, as can be seen in Figure 3. Although this increase is small, it suggests that more advanced feature selection methods may be able to reveal language features that are more resistant to language contact and language change.

5 Error Analysis

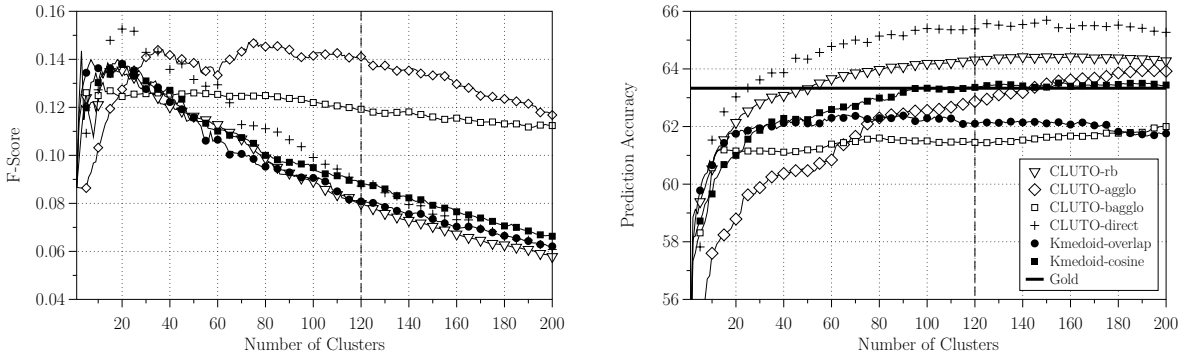
There are two main reasons for the differences between induced clusters and genetic groupings.

5.1 Language Similarity vs. Genetic Relatedness

As mentioned before, language similarity and genetic relatedness are two different concepts. Simi-

	baseline	gold	rb	agglo	bagglo	direct	k-medoids with similarity overlap	k-medoids with cosine similarity
Whole-World-Set (121 Clusters)								
F-Score	0.087	–	0.080	**0.140**	0.119	0.089	0.081	0.088
Acc (%)	53.72	63.43	64.33	62.86	61.44	**65.47**	62.11	63.36
Indo-European Subset (10 Clusters)								
F-Score	0.319	–	0.365	0.377	**0.391**	0.355	0.352	0.331
Acc (%)	64.27	74.1	71.12	72.26	70.62	**74.13**	73.36	72.12
Sino-Tibetan Subset (14 Clusters)								
F-Score	0.305	–	0.224	**0.340**	0.333	0.220	0.285	0.251
Acc (%)	58.08	61.71	63.93	63.74	63.06	**65.31**	64.55	63.94

Table 3: Comparison of clustering algorithms when the number of clusters is set to the same number of genetic groupings. The highest number in each row is in boldface.



(a) F-scores of clustering results

(b) Prediction accuracy

Figure 2: Comparison of the performances of different clustering methods using the whole-world data set. The number of groups in the gold standard (i.e., genetic grouping) is shown as a vertical dashed line in 2(a) and 2(b), and the prediction accuracy of the gold standard as a horizontal solid line in 2(b).

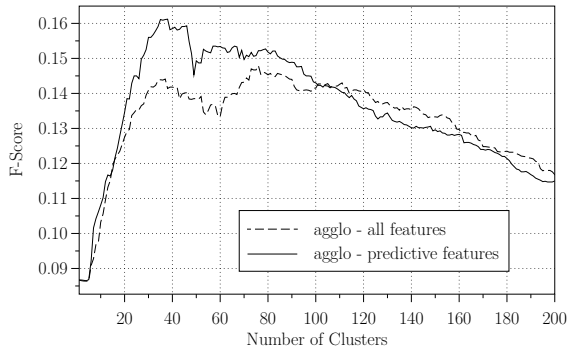


Figure 3: F-scores of the `agglo` clustering method when using all the features vs. only features whose prediction accuracy by the genetic grouping is higher than 50%.

lar languages might not be genetically related and dissimilar languages might be genetically related. An example is given in Table 5. Persian and English are both Indo-European languages, but look very different typologically; in contrast, Finnish and English are not genetically related but they look more similar typologically. While English and Persian are related, they have been diverging in geographically distant areas for thousands of years. Thus, the fact that English appears to share more features with a geographically closer Finnish is expected.

5.2 WALS as the Dataset

Perhaps the biggest challenge we encounter in this project has been the dataset itself. WALS has certain properties that complicate the task.

Data Sparsity and Shared Features

While the previous example shows unrelated languages can be quite similar typologically, our clustering methods put two closely related languages, Eastern and Western Armenian, into dif-

Breakdown by Feature Category		Breakdown By Feature: Top 10				Breakdown by Feature: Bottom 10			
Category	Accuracy	Feature	Acc	C	V	Feature	Acc	C	V
		Whole-World Set							
Lexicon	75.0%	(136) M-T Pronouns	94.0%	230	3	(1) Consonant Inventories	32.6%	561	5
Word Order	68.6%	(18) Absence of Common Consonants	93.7%	565	6	(133) Number of Basic Color Categories	33.3%	119	7
Phonology	65.9%	(11) Front Rounded Vowels	91.1%	560	4	(23) Locus of Marking in the Clause	33.9%	236	5
Complex Sentences	64.0%	(73) The Optative	89.6%	319	2	(71) The Prohibitive	34.6%	495	4
Nominal Syntax	63.2%	(137) N-M Pronouns	87.9%	230	3	(22) Inflectional Synthesis of the Verb	35.1%	145	7
Verbal Categories	61.9%	(6) Uvular Consonants	85.0%	565	4	(56) Conjunctions and Universal Quantifiers	38.2%	116	3
Simple Clauses	60.5%	(130) Finger and Hand	84.4%	591	2	(117) Predicative Possession	39.4%	240	5
Nominal Categories	59.1%	(115) Negative Indefinite Pronouns	84.2%	206	4	(92) Position of Polar Question Particles	40.0%	775	6
Morphology	53.9%	(19) Presence of Uncommon Consonants	83.0%	565	7	(38) Indefinite Articles	40.4%	473	5
Other	41.3%	(58) Obligatory Possessive Inflection	81.4%	244	2	(50) Asymmetrical Case-Marking	40.7%	261	6
		Indo-European Subset							
Lexicon	86.4%	(130) Finger and Hand	100.0%	35	2	(3) Consonant-Vowel Ratio	30.6%	31	5
Morphology	83.1%	(118) Predicative Adjectives	100.0%	29	3	(92) Position of Polar Question Particles	34.6%	47	6
Word Order	79.6%	(18) Absence of Common Consonants	100.0%	31	6	(78) Coding of Evidentiality	36.0%	23	6
Simple Clauses	76.6%	(107) Passive Constructions	100.0%	19	2	(1) Consonant Inventories	42.4%	31	5
Nominal Categories	70.4%	(88) Order of Demonstrative and Noun	97.2%	66	6	(2) Vowel Quality Inventories	44.4%	31	3
Phonology	66.7%	(89) Order of Numeral and Noun	95.7%	64	4	(84) Order of Object, Oblique, and Verb	47.8%	20	6
Verbal Categories	62.1%	(27) Reduplication	95.2%	20	3	(16) Weight Factors in Weight-Sensitive Stress Systems	51.1%	53	7
		(7) Glottalized Consonants	93.9%	31	8	(70) The Morphological Imperative	55.3%	53	5
		(93) Position of Interrogative Phrases in Content Questions	93.9%	44	3	(44) Gender Distinctions in Independent Personal Pronouns	56.5%	19	6
		(5) Voicing and Gaps in Plosive Systems	93.8%	31	5	(37) Definite Articles	59.2%	46	5
		Sino-Tibetan Subset							
Lexicon	100.0%	(130) Finger and Hand	100.0%	8	2	(77) Semantic Distinctions of Evidentiality	9.1%	18	3
Word Order	67.7%	(82) Order of Subject and Verb	100.0%	99	3	(78) Coding of Evidentiality	17.7%	18	6
Morphology	63.8%	(119) Nominal and Locational Predication	100.0%	13	2	(4) Voicing in Plosives and Fricatives	20.7%	26	4
Simple Clauses	60.9%	(86) Order of Genitive and Noun	100.0%	73	3	(1) Consonant Inventories	22.2%	26	5
Verbal Categories	60.7%	(129) Hand and Arm	100.0%	8	2	(14) Fixed Stress Locations	25.0%	4	7
Nominal Categories	55.8%	(18) Absence of Common Consonants	100.0%	26	6	(15) Weight-Sensitive Stress	25.0%	4	8
Phonology	50.7%	(93) Pos. of Interr. Phrases in Content Q's	100.0%	79	3	(38) Indefinite Articles	31.7%	36	5
		(85) Order of Adposition and Noun Phrase	97.5%	79	5	(120) Zero Copula for Predicate Nominals	37.5%	13	2
		(95) Relationship b/t Object and Verb and Adposition and Noun Phrase	96.3%	76	5	(2) Vowel Quality Inventories	42.9%	26	3
		(48) Person Marking on Adpositions	93.3%	14	4	(3) Consonant-Vowel Ratio	46.7%	26	5

Table 4: Prediction accuracy figures derived from genetic groupings for each dataset and broken down by WALS feature category and feature. Ordering is by descending accuracy for the top 10 features, and by increasing accuracy for the bottom 10 features. The 'C' and 'V' columns give the number of languages in the set that a feature appears in, and the number of possible values for that feature, respectively.

ferent clusters. A quick review shows that the reason for this mistake is due to a lack of shared features in WALS. Table 6 shows that very few features are specified for both languages. The data sparsity problem and the distribution of empty cells adversely affect clustering results.

Notice that in this example, the features whose values are filled for both languages actually have identical feature values. While using shared overlap as a similarity measure can capture the similarity between these two languages, this measure biases clustering toward features with fewer cells filled out. The only way out of errors like this, it seems, is to obtain more data.

There are a few other typological databases that might be drawn upon to define a more complete set of data: PHOIBLE, (Moran and Wright, 2009), ODIN (Lewis, 2006), and the AUTOTYP database (Nichols and Bickel, 2009). Using these databases to fill in the gaps in data may be the only way to fully address these issues.

The Feature Set in WALS

The features in WALS are not systematically chosen for full typological coverage; rather, the contributors to WALS decide what features they want to work on based on their expertise. Also, some features in WALS overlap; for example, one WALS feature looks at the order between subject, verb, and object, and another feature checks the order between verb and object. As a result, the feature set in WALS might not be a good representative of the properties of the languages covered in the database.

6 Conclusion & Further Work

By comparing clusters derived from typological features to genetic groups in the world's languages, we have found two interesting results. First, the induced clusters look very different from genetic grouping and this is partly due to the design of WALS. Second, despite the differences, induced clusters show similar, or even greater levels

ID: Feature Name	English	Finnish	Persian
2: Vowel Quality Inventories	**Large (7-14)**	**Large (7-14)**	Average (5-6)
6: Uvular Consonants	**None**	**None**	Uvular stops only
11: Front Rounded Vowels	**None**	High and Mid	**None**
27: Reduplication	**No productive reduplication**	**No productive reduplication**	Productive full and partial reduplication
37: Definite Articles	Definite word distinct from demonstrative	No definite or indefinite article	No definite, but indefinite article
53: Ordinal Numerals	**First, second, three-th**	**First, second, three-th**	First/one-th, two-th, three-th
81: Order of Subject, Object and Verb	**SVO**	**SVO**	SOV
85: Order of Adposition and Noun Phrase	**Prepositions**	Postpositions	**Prepositions**
87: Order of Adjective and Noun	**Adjective-Noun**	**Adjective-Noun**	Noun-Adjective
124: 'Want' Complement Subjects	**Subject left implicit**	**Subject left implicit**	Subject expressed overtly
Number of Features	139	135	128
Cosine Similarity to Eng	1.00	0.56	0.42
Shared Overlap with Eng	1.00	0.56	0.44

Table 5: A selection of ten features from English, Finnish, and Persian. Same feature values in each row are in boldface. Despite the genetic relation between English and Persian, similarity metrics place English closer to Finnish than Persian.

ID#	Feature Name	Armenian (Eastern)	Armenian (Western)
1	Consonant Inventories	Small	–
27	Reduplication	**Full Reduplication Only**	**Full Reduplication Only**
33	Coding of Nominal Plurality	–	Plural suffix
48	Person Marking on Adj.	None	–
81	Order of Subj. Obj., and V	–	SOV
86	Order of Adposition and Noun Phrase	**Postpositions**	**Postpositions**
100	Alignment of Verbal Person Marking	Accusative	–
129	Hand and Arm	–	Identical
	Number of Features	85	33
	Cosine Similarity	0.22	
	Shared Overlap	1.00	

Table 6: Comparison of features between Eastern and Western Armenian. Same feature values in each row are in boldface. Empty cells are shown as '–'.

of typological similarity than genetic grouping as indicated by the prediction accuracy.

While these initial findings are interesting, using WALS as a dataset for this purpose leaves a lot to be desired. Subsequent work that supplements the typological data in WALS with the databases mentioned in §5.2 would help alleviate the data sparsity and feature selection problems.

Another useful follow-up would be to perform application-oriented evaluations. For instance, evaluating the performance of syntactic projection methods between languages determined to have similar syntactic patterns, or using similar morphological induction techniques on morphologically similar languages. With the development of large typological databases such as WALS, we hope to see more studies that take advantage of resources for resource-rich languages when developing tools for typologically similar, but resource-poor languages.

Acknowledgment This work is supported by the National Science Foundation Grant BCS-0748919. We would also like to thank Emily Bender, Tim Baldwin, and three anonymous reviewers for helpful comments.

References

Daumé, III, Hal and Daniel Marcu. 2005. A Bayesian Model for Supervised Clustering with the Dirichlet Process Prior. *Journal of Machine Learning Research*, 6:1551–1577.

Daumé, III, Hal. 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 (HLT/NAACL)*, pages 593–601, Boulder, Colorado, June.

Estivill-Castro, Vladimir and Jianhua Yang. 2000. A fast and robust general purpose clustering algorithm. In *Proc. of Pacific Rim International Conference on Artificial Intelligence*, pages 208–218. Springer.

Hana, Jiri, Anna Feldman, and Chris Brew. 2004. A Resource-light Approach to Russian Morphology: Tagging Russian using Czech resources. In *Proceedings of EMNLP 2004*, Barcelona, Spain.

Haspelmath, Martin, Matthew S. Dryer, David Gil, and Bernard Comrie. 2005. *The World Atlas of Language Structures*. Oxford University Press, Oxford, England.

Lewis, William D. 2006. ODIN: A Model for Adapting and Enriching Legacy Infrastructure. In *Proceedings of the e-Humanities Workshop, held in cooperation with e-Science 2006: 2nd IEEE International Conference on e-Science and Grid Computing*, Amsterdam.

McMahon, April M. S. 1994. *Understanding language change*. Cambridge University Press, Cambridge; New York, NY, USA.

Moran, Steven and Richard Wright. 2009. Phonetics Information Base and Lexicon (PHOIBLE). Online: `http://phoible.org`.

Nichols, Johanna and Balthasar Bickel. 2009. The AUTOTYP genealogy and geography database: 2009 release. `http://www.uni-leipzig.de/~autotyp`.

Rand, William M. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850.

Steinbach, Michael, George Karypis, and Vipin Kumar. 2000. A comparison of document clustering techniques. In *Proceedings of Workshop at KDD 2000 on Text Mining*.

Xia, Fei and William D. Lewis. 2007. Multilingual structural projection across interlinear text. In *Proc. of the Conference on Human Language Technologies (HLT/NAACL 2007)*, pages 452–459, Rochester, New York.

Yarowsky, David and Grace Ngai. 2001. Inducing multilingual pos taggers and np bracketers via robust projection across aligned corpora. In *Proc. of the Second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies (NAACL-2001)*, pages 1–8, Morristown, NJ, USA.