Dependency-sensitive typological distance

Harald Hammarström and Loretta O'Connor

1. Introduction

Increasing amounts of data in machine readable form are becoming available for the study of linguistic typology, especially with the appearance of WALS.[1] Most such databases come in the form of a matrix of languages and features, where each language, for each feature, is given a value from a discrete set of possible values. If we fix a particular ordering of the list of features, we may denote a language X with its feature vector $L_X = [v_1, \ldots, v_n]$, meaning that it has value v_i for feature F_i. Similarly, we may use $L_X[i]$ to denote the feature value v_i of feature F_i in language X.

For a variety of purposes, researchers make use of a distance measure between two languages (cf. Chiswick and Miller 2004; Cysouw 2007; Holman et al. 2007; Dahl 2008; Polyakov et al. 2009; Wichmann and Holman 2010) that says how similar two languages are from 0.0 (identical) to 1.0 (totally different). Traditionally, such a distance measure has taken the form of the Gower coefficient (also known as relative Hamming distance):

$$G(L_X, L_Y) = \frac{\#_{i \in DEF(L_X, L_Y)} L_x[i] \neq L_Y[i]}{|\, DEF(L_X, L_Y)\,|}$$

Where $DEF(L_X, L_Y) = \{i | L_X[i]$ and $L_Y[i]$ are defined$\}$, i.e., the set of features which are defined for both languages (since, in general, there may be missing values. The Gower coefficient simply counts the number of features where the languages have a different value, divided by the total number of features compared. Therefore, using the Gower coefficient makes sense if the features are all independent and of equal weight. These assumptions appear to be largely well-founded if the features in question indicate, for example, the presence/absence of a lexical cognate for a particular meaning. However, in the case of typological features, many functional dependencies have been established (Dryer 1992) and linguists have intuitions about many further dependencies.

In this paper, we will develop two kinds of dependency-sensitive distance metrics.[2] The first captures the idea that if it can be shown that one feature can be (partly) predicted by another, then the predictable feature should be (partly) "discounted". This strategy tackles dependencies between features as a whole, not between specific values of features. The second dependency-sensitive metric addresses the significance of similarities between specific values of features. Globally, a specific combination of values may be very predictable, or, on the other end of the scale, a combination of values may be extremely unusual. Accordingly, when comparing two specific languages, scores may be weighted as to whether they share something predictable or something quirky.

The dependency-sensitive metrics will be illustrated on a dataset of typological features for Chibchan and neighboring languages developed by Constenla Umaña (1991). The database is dense (almost all features are defined for almost all languages) and published (thus publicly available). We are interested in potential differences the metrics may give as regards the classification of these particular languages. On the one hand, it could be that removing the dependencies merely has the effect of concentrating the distances uniformly across the languages. On the other hand, there could be drastic effects involving particular pairs of languages. A pair of languages that looked very similar when inflated by dependencies might become as distant as random languages when the dependencies are removed. Alternatively, two languages that do not share an impressive number of features may share them in such an unusual, "quirky" way (within this data sample) that the dependency value-sensitive metric singles out that pair of languages for, e.g., a borrowing scenario or a genealogical relation to explain the quirk.

2. The Isthmo-Colombian Area dataset

The Isthmo-Colombian Area dataset used in the present paper represents languages of Central America and the northwest corner of South America. We extracted the information for 34 languages from Constenla Umaña (1991) and added one more language, Damana, using sources (Williams 1993; Trillos Amaya 2005) that appeared after Constenla Umaña's compilation. The dataset consists of a total of 35 languages and 81 structural features, where all features are defined for all languages except for Damana, in which two feature values are undefined. Fifteen of the languages are Chibchan, with smaller representations from nine other families plus three isolates. More

languages belonging to each family are attested, but they are either outside the geographic sphere of interest or are insufficiently documented to be included.

The 35 languages in question are listed in table 1 (in bold, with ISO codes) and mapped in figure 1.

Table 1. Languages in the Isthmo-Colombian Area dataset: names and classifications are adapted from the following sources: Chibchan (Constenla Umaña 2012), Guajiro (Captain 2005), Quiché (Campbell 1997), Barbacoan (Adelaar and Muysken 2004: 141-151), Chocoan (Aguirre Licht 2006), Quechua (Cerrón-Palomino 2003) and Misumalpan (Pineda 2005).

- Chibchan
  - Core Chibchan
    - Isthmic
      - Eastern Isthmic
        - Guaymiic
          - **Movere [gym]**
          - **Bocotá [sab]**
        - Kuna
          - **Cuna [cuk]**
      - Western Isthmic
        - Viceitic
          - **Bribri [bzd]**
          - **Cabécar [cjp]**
        - **Boruca [brn]**
        - **Teribe [tfr]**
    - Magdalenic
      - Northern Magdalenic
        - Arhuacic
          - Eastern-Southern Arhuacic
            - **Bintucua [arh]**
            - Eastern Arhuacic
              - **Damana [mbp]**
          - **Cágaba [kog]**
      - Southern Magdalenic
        - Chibcha
          - **Chibcha [chb]**
        - Tunebo
          - **Central Tunebo [tuf]**
    - Votic

Guatuso [gut]
Rama [rma]
Paya [pay]
Chocoan
Embera
Atrato
Katio
Katío [cto]
Sambú [emp]
Woun Meu
Huaunana [noa]
Paez
Páez [pbb]
Cofan
Cofán [con]
Arawak
Arawak
Maipuran
Northern Maipuran
Lokono-Guajiro
Guajiro [guc]
Barbacoan
Cayapa-Colorado
Cayapa [cbi]
Colorado [cof]
Coconucan
Guambiano [gum]
Unclassified Barbacoan
Awa-Cuaiquer [kwi]
Jicaquean
Jicaque [jic]
Kamsa
Camsá [kbh]
Lencan
Lenca [iso-code missing]
Mayan
Yucatecan-Core Mayan
Core Mayan
Quichean-Mamean
Greater Quichean
Poqom-Quichean
Core Quichean
Quiche-Achi
K'iche' [quc]

Misumalpan
  Sumalpan
    Matagalpan
      Cacaopera [ccr]
    Sumo-Mayangna
      Sumo [sum]
      Ulua [sum]
  Mískito [miq]
Quechuan
  Quechua II
    Quechua II.B
      Imbabura Highland/Lowland Napo Quichua [qvi/qvo]
Xincan
  Xinca-Guazacapan [iso-code missing]

Non-trivial identifications based on Constenla Umaña's sources (1991: 190–192) are:

Constenla Name	Variety	ISO-639-3
Lenca	Lenca of El Salvador	-
Quechua	Amalgam of Imbabura Quechua and Lowland Napo Quechua	[qvi] and [qvo]
Xinca	Xinca of Guazacapán	-
Quiche	Central Quiché	[quc]
Cuna	San Blas Kuna	[cuk]
Tunebo	Central Tunebo	[tuf]

The ISO-639-3 codes for Xincan and Lencan languages are erroneous in that they lump together all varieties of each as one language. Both families are divergent enough to constitute families of different languages (Campbell 1997). Since the codes are thereby indeterminate, we chose not to use them at all. Note that Sumo and Ulua are distinguished by Constenla but considered the same language in ISO-639-3.

There are reasons why the Isthmo-Colombian Area dataset is of more than casual interest for the present experiment. The languages are spoken on and around the land bridge that unites the two American continents. Once considered an ever-changing transit region for people and goods moving between powerful civilizations north and south, the region has recently come to

Figure 1. Locations of languages in the Isthmo-Colombian Area dataset (language polygons from Eriksen 2011: 12–15).

be recognized as the site of long-time settlement by small, sedentary groups (Quesada 2007: 22–26 and references therein). This suggests long-term interaction *in situ* between particular groups of speakers and increases the likelihood of shared changes through language contact with each other and with common visitor groups. In addition, historical linguistics has seen a meshwork of genealogical proposals involving the Chibchan, Chocoan, Barbacoan, and Paezan languages present in this dataset (see Adelaar and Muysken

[2004: 22–34, 36–38, 41–45] for relevant discussion) and overlapping proposals of areal relationship that include three regional linguistic subareas for the entire dataset (Constenla Umaña 1991: 121–131) and a two-way division between Chibchan languages of Central America and those of Colombia (Quesada 2007: 44–45). Human genetics research with present-day speakers suggests clear differences between Chibchan and Chocoan populations of the Isthmus (Kolman and Berminghan 1997) and indicates different networks of possible relationship between and among the two regional groupings of Chibchan languages, Emberá, Guajiro, and Quiché (Melton et al. 2007). Any or all such historical events might yield pairs of languages with salient (dis)similiarity involving feature dependencies.

The list of 81 features is given in the appendix. It consists of 42 morphosyntactic features (e.g. Is there VO order in transitive clauses? Is there a distinction between inclusive and exclusive for personal pronouns?) and 39 phonological features (e.g. Is there a nasality contrast for vowels? Is there an aspiration contrast for plosives?). All features are binary, although the methods employed in the present paper do not require them to be binary. (Features do, however, need to be discrete-valued).

3. Computational approaches to dependencies

There are good reasons to expect that abstract grammatical features of language should show functional dependencies, even when features are logically independent. Grammatical features may (partly) overlap in function, and constraints on communicative efficiency may favor certain configurations of features and functions over others. For example, case-marking and strict constituent order may be both used to signal who did what to whom in a basic declarative clause. It is logically possible to use both (and indeed, some languages do), but it is nevertheless conceivable that there is some pressure from communicative principles that causes the redundancy to go away (Sinnemäki 2010). Perspectives in the linguistic literature vary on the types of dependencies present in language and particularly on the motivations for dependencies: explanations range from "nativist", i.e., inborn constraints on grammar (e.g. Chomsky 1981), to constraints from cognition, social interaction and/or efficient communication (cf. papers in Christiansen et al. 2009).

For the purposes of this paper, it makes no difference whether functional dependencies are innate or environmental in origin. In either case, dependencies that come from (hypothetical) communicative principles or (hypothetical) inborn constraints must:

a. be universal (in the sense of being common to all natural languages), since by definition natural languages are used for communication between humans, and,

b. concern the whole feature in question (not just some particular values), since all values of a feature have the same domain, and the hypothesized dependencies stem from overlapping domains.

Universal dependencies are not the *only* factors shaping the features landscape – random, areal and genealogical effects can, in fact, overshadow those that come from universal principles (Dunn et al. 2011). Universal dependencies (in the view taken in this paper) exist if (and only if) there are non-random dependencies in the languages of the world that cannot be explained areally or genealogically. Because of overshadowing effects, they do not have to be present in every (sub-)trajectory of history as long as they appear more often than random.

3.1 Factoring out feature dependencies

Assuming here that universal dependencies exist and that we are given a large and balanced enough sample of languages and their features, how can we find and factor out the dependencies?

One possibility would be to apply Principal Component Analysis (PCA) to the language-feature matrix (Pearson 1901). In essence, PCA breaks a given matrix with column dependencies into a smaller matrix without column dependencies which account for as much of the variability in the data as possible. The constraints are a) on the number of components (the number of new uncorrelated columns), b) that the new columns have to be linear combinations of the old ones, and c) that the new columns have to be independent. PCA would seem to be a well-suited technique for the data we have at hand except there is no natural way to know the appropriate number of components. Presumably, this number would vary across different datasets of linguistic features, and giving one number of components for all datasets seems arbitrary. Therefore we choose not to use PCA as a general approach to factoring out dependencies between linguistic features.

Another possibility,[3] which also has the advantage of being more easily interpretable to linguists, is as follows. Dependencies can be captured as (probabilistic) implications from *n* features to one other feature. We will make the simplifying assumption that the essential implications are where

n=1 (the extension to other values of n is relatively straightforward). We first go through the matrix to collect all such implications, creating a dependency graph. The dependency graph has features as nodes, and directed edges between the nodes reflect the implications. Potentially, every feature depends on every other feature to some degree, including circular dependencies, so it is not obvious how to go from a dependency graph between features to a distance metric. We then assume that the core dependencies can be captured by the Chu-Liu tree (Chu and Liu 1965) of the dependency graph. The Chu-Liu algorithm creates a maximum spanning tree of a directed graph, meaning it keeps the single strongest incoming dependency for each feature, and it removes epiphenomenal and circular dependencies. Now, using the Chu-Liu tree, we can modify the Gower coefficient to get a dependency-weighted distance metric. Essentially, instead of scoring 0 or 1 we score an amount proportional to how predictable the feature is. A more detailed description follows.

3.1.1 Finding feature implications

A general method for quantifying the predictive relationship of one frequency distribution from another which has the same domain (in this case, languages) is to calculate how much of the entropy (measure of uncertainty) of one variable can be predicted from knowing the other. This technique has already been used in linguistics by Bickel (2010) and, in a specialized form, by Daumé and Campbell (2007). Formally, for two features A and B we quantify 'A predicts B' as follows:

$$A \rightarrow B = MI(A,B)/H(B)$$

where $MI(A,B)=H(A)+H(B)-H(A,B)$ is the mutual information of A and B, and $H(B)$ is the Shannon entropy[4] of B.

A toy example is shown in table 2. Intuitively speaking, we can say that F_1 is some help in predicting F_2 (if F_1 is 1 then guess 'a'; if F_1 is 0 then guess 'b' or 'c' but not 'a'), that predicting F_1 from F_2 is easier, but that F_1 is no help at all in predicting F_3.

Table 2. Toy example of languages, features, and feature implication calculations.

	F_1	F_2	F_3	F_4
L_1	1	a	1	a
L_2	1	a	0	b
L_3	1	a	1	?
L_4	1	b	0	?
L_5	0	b	1	?
L_6	0	b	0	?
L_7	0	c	1	?
L_8	0	c	0	?
$H(A)$	1.00	1.56	1.00	0.81

	P(A, B)				MI(A,B)	$MI(A,B)/H(B)$	→
$F_1 \rightarrow F_2$	$P(1,a)=3/8$	$P(1,b)=1/8$	$P(0,b)=2/8$	$P(0,c)=2/8$	0.65	0.65/1.56	0.41
$F_2 \rightarrow F_1$	$P(1,a)=3/8$	$P(1,b)=1/8$	$P(0,b)=2/8$	$P(0,c)=2/8$	0.65	0.65/1.00	0.65
$F_1 \rightarrow F_3$	$P(0,0)=2/8$	$P(0,1)=2/8$	$P(1,0)=2/8$	$P(1,1)=2/8$	0.00	0.00/1.00	0.00

These intuitions are reflected accordingly in the spelled-out calculations of table 2. A final note concerns missing values: they are treated as distinct separate values, erring on the safe side (example F4 in table 2); otherwise, sparse data might exhibit strong random correlations.

Applied to all pairs of features in the Isthmo-Colombian Area dataset, some sample implications are shown in table 3.

Table 3. Some sample feature implications from the Isthmo-Colombian Area dataset.

Rank	Implication	Strength
1	13 → 12	1.000
649	39 → 67	0.180
1297	77 → 71	0.113
1945	37 → 6	0.079
2593	50 → 19	0.055
3241	14 → 27	0.037
3889	54 → 45	0.026
4537	38 → 29	0.015
5185	10 → 47	0.005
5833	28 → 42	0.000

3.1.2 Feature implication distillation

When checking statistical implications between all pairs of features this way, one risks finding epiphenomenal implications. For example, if *A* predicts *B* and *B* predicts *C* then we will also find that *A* predicts *C*, but this information is redundant if we already know the underlying first two implications. Bickel (2010) suggests removing the weakest dependency in all such chains, which is sufficient for a number of purposes. In our case, we are interested in creating a transparent similarity metric and need a slightly stronger method of purging, keeping only the strongest implications in chains and allowing maximally one (strongest) predictor for a feature. In other words, from the complete dependency graph, we compute the maximal directed spanning tree, also known as the Chu-Liu tree (Chu and Liu 1965). For the definition and proof of correctness of the Chu-Liu algorithm, the reader is referred to the more accessible treatment by Georgiadis (2003). Figure 2 shows the Chu-Liu tree for the Isthmo-Colombian Area dataset. The sum "predictability" in the tree is 35.02 (out of a total of 81 features). This can be taken to mean that approximately 35.02/81≈43.2% of the feature mass is redundant.

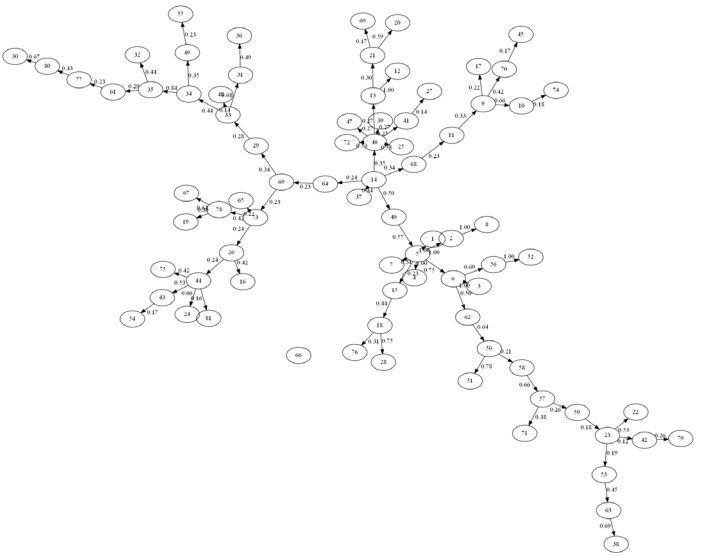

Figure 2. The maximal directed spanning tree (or Chu-Liu tree) for the Isthmo-Colombian Area dataset. (The one edge reaching feature 66 with value 0.0 has been graphically removed, but is strictly speaking part of the tree.)

3.1.3 A dependency-sensitive distance

Using the Chu-Liu tree, we can modify the Gower coefficient to get a dependency-weighted distance metric:

$$G_d(X,Y) = \frac{\sum_{i \in DEF(L_X,L_Y) \text{ and } L_X[i] \neq L_Y[i]} 1.0 - W(i)}{\sum_{i \in DEF(L_X,L_Y)} 1.0 - W(i)}$$

where $W(A)$ is the weight of the incoming edge predicting A (or 0.0 if there is no such edge). As with the Gower coefficient, only features for which both languages have a value are considered. For each feature, instead of a penalty of 1 for mismatches, we penalize the appropriate amount as to how predictable the feature in question is. The mismatch score is then relativized as to the maximum amount of penalty possible.

In the toy example of table 2, L_2 and L_3 differ only as to feature F_3 out of three features for which both are defined, so their (unmodified) Gower coefficient would be $G(L_2,L_3) = 1/3 \approx 0.33$. The Chu-Liu tree for the toy example in table 2 is $F_2 \rightarrow F_1 \approx 0.66$, $F_2 \rightarrow F_4 \approx 0.44$, $F_4 \rightarrow F_3 \approx 0.25$, so $W(1) \approx 0.66$, $W(2)=0.0$, $W(3) \approx 0.25$ and $W(4) \approx 0.44$. Thus the dependency-sensitive modified Gower coefficient for L_2 and L_3 is $G_d(L_2,L_3) = (1–0.25)/(1–0.66 + 1–0.00 + 1–0.25) \approx 0.75/2.09 \approx 0.36$.

3.2 Feature dependencies and quirky values

The dependency-sensitive distance metric G_d, as described, eliminates redundancy between features as a whole. It does *not* differentiate between common and uncommon feature *value* constellations as a result of feature dependencies.

To see the difference, consider the following analogy. Suppose there are creatures which have legs and arms. Some functional pressure, such as access to fruits in tall trees, favors tall creatures over short, and as a result nearly all creatures are tall. Suppose further that the creatures who are tall tend to have both long legs and long arms (as opposed to their having only long arms or only long legs), perhaps because the growth hormone in the creature is the same for both arms and legs. In this world, we will find a correlation between long arms and long legs, and the dependency-sensitive distance metric would tell us that, because of this, the distance between short and tall creatures is on the order of one unit (legs or arms) rather than two units (legs plus

arms). The distance metric would also, as designed, show that if we find two short creatures they are just as alike as if we find two tall creatures – in both cases, the distance is zero – even though, because of the functional pressure, finding two short creatures is very unusual. What this illustrates is that, given functional pressure, one might also want to distinguish, already within the distance metric itself, between significant (dis)similarity and insignificant (dis)similarity. For example, if we find two short creatures, or some other unusual constellation such as two creatures both with short legs and long arms, it would be much more likely that they share a common history than it would be for two creatures who share the usual constellation of features. This intuition has surfaced in linguistics under the name 'shared quirk' (Gensler 2003), or in other words, a match against the preference of a dependency.

Potential quirks are all constellations of feature values. In the present study we will restrict ourselves to unary and binary quirks, in other words, feature value constellations involving one or two variables. The quirkiness of a feature constellation (here, a binary constellation) can be defined as:

$$Q(f_i = u, f_j = v) = \frac{\text{The number of languages with values } f_i\text{=u and } f_j\text{=v}}{\text{Total number of languages with } f_i \text{ and } f_j \text{ defined}}$$

It is relatively straightforward to enumerate potential unary and binary quirks and, when comparing two languages, score their matches proportionately to their quirkiness. Again, we make a modified version of the Gower coefficient (the binary case):

$$G_q^2(X,Y) = 1.0 \frac{\sum_{i<j \in DEF(L_x,L_y) \text{ and } L_x[i]=L_y[i] \text{ and } L_x[j]=L_y[j]} 1.0 - Q(i = L_x[i], j = L_x[j])}{|\{(i,j) | i < j \in DEF(L_x,L_y)\}|}$$

Again, let us look at the toy example of table 2. As to unary quirks, L_2 and L_3 match their values for features $F_1 = 1$ and $F_2 = a$. $Q(f_1 = 1) = 4/8$ and $Q(f_2 = a_2) = 3/8$, so their distance based on unary quirks is 1.0– (1–4/8+1–3/8)/3 ≈ 0.71. As to binary quirks, the only match between the two languages is $G_q^2(L_2,L_3) = Q(f_1 = 1, f_2 = a_2) = 3/8$, so their distance based on unary quirks is $G_q^1(L_2,L_3) = 1.0–(1–3/8)/(3!/1!2!) \approx 0.79$. (Counting both unary and binary quirks at the same time yields 1.0–(1–4/8+1–3/8+1–3/8)/(3+3) ≈ 0.71.)

The quirk-based measure is strictly speaking not a distance measure because $G_q(X, X)$ is not necessarily 0: whether it is 0 or not depends on how *significant* the feature values of X are.

4. Experimental results

We are now ready to apply the new metrics to the Isthmo-Colombian Area dataset to see if there are language pairs that behave drastically differently. In all experiments, we use the Isthmo-Colombian Area dataset itself to extract dependencies and quirkiness rates. Ideally, we would use a large enough world-wide database with the same features, but such a database is not available for most of the features as defined in the Isthmo-Colombian Area dataset. Also, ideally, the dataset used for extracting dependencies and quirkiness rates should be genealogically and areally stratified to make sure that any skewed rates are the result of universal dependencies rather than historical relationships. The experiments reported here are relative to the assumption that the Isthmo-Colombian Area dataset as a whole contains sufficient evidence for universal dependencies.

We first look at the G_d-distances versus the traditional Gower coefficient (G). As an orientation, table 4 shows the top-5 and bottom-5 distances before (G) and after (G_d) dependency. As those distances suggest, in general the differences are slight, both in the actual values and in their relative rank.

Table 4. The top-5 and bottom-5 language pairs in terms of unmodified G-distance and the dependency-sensitive G_d-distance.

Rank	G		G_d	
1	Ulua-Sumo	0.00	Ulua-Sumo	0.00
2	Sumo-Misquito	0.01	Sumo-Misquito	0.02
3	Ulua-Misquito	0.01	Ulua-Misquito	0.02
4	Cabecar-Bribri	0.04	Cabecar-Bribri	0.05
5	Sambu-Catio	0.05	Sambu-Catio	0.07
…	…	…	…	…
591	Quiche-Bocota	0.58	Quiche-Bocota	0.54
592	Quiche-Cabecar	0.58	Xinca-Cabecar	0.55
593	Xinca-Cabecar	0.58	Xinca-Teribe	0.56
594	Teribe-Quiche	0.59	Teribe-Quiche	0.57
595	Quiche-Movere	0.60	Quiche-Movere	0.59

Table 5 shows the pairs whose distances change the most as dependencies are factored out.

Table 5. Language pairs that became more distant (left) or became closer (right) as a result of applying the dependency-sensitive Gower coefficient.

	G_d-G	G	G_d		G_d-G	G	G_d
Sambu-Cayapa	0.10	0.38	0.48	Quiche-Lenca	-0.08	0.36	0.28
Paya-Bintucua	0.09	0.26	0.35	Quiche-Cayapa	-0.07	0.43	0.36
Paya-Cagaba	0.09	0.22	0.31	Quiche-Paez	-0.06	0.49	0.43
Ulua-Paez	0.09	0.36	0.45	Quiche-Cuna	-0.06	0.41	0.35
Sumo-Paez	0.09	0.36	0.45	Quiche-Boruca	-0.06	0.47	0.41
Cuna-Boruca	0.09	0.26	0.35	Xinca-Camsa	-0.06	0.36	0.30
Paya-Muisca	0.09	0.25	0.33	Xinca-Cofan	-0.06	0.44	0.39
Paez-Bintucua	0.09	0.41	0.49	Quiche-Huaunana	-0.06	0.46	0.40
Huaunana-Boruca	0.08	0.23	0.32	Xinca-Boruca	-0.06	0.44	0.39
Paez-Misquito	0.08	0.35	0.43	Quiche-Colorado	-0.05	0.43	0.38

The pairs that become more distant include Paya, a Chibchan language of Honduras, now more distant from three Chibchan languages of Colombia; Chocoan languages Sambú and Huaunana, now more distant from a Barbacoan and a Chibchan neighbor, respectively, and Paez, now more distant from a genealogically heterogeneous group of languages. The pairs that become closer involve the two northernmost languages in the sample, Xinca and Quiché, each now closer to a large number of languages.

To get a feeling for the relative contribution of these changes, Neighbor-Joining trees (Saitou and Nei 1987) for the two distance matrices are shown in the leftmost columns of figure 3.[5]

Comparing the first two trees we see that Xinca and Quiché, as a result of having become closer to many other languages, cease to group together exclusively and move one short step up the tree along with their immediate neighbors. There is, however, little appreciable difference between the trees overall.

The high level of dependency in the feature set as a whole (recall the 43% estimate above) constitutes strong evidence that the dependency-induced inflation is uniformly distributed in the dataset. If this were to prove typical of feature sets in typology in general, there would be little need to consider feature dependencies in typological comparison.

Next we turn to applications of the quirkiness-sensitive metric. It is not meaningful to compare the quirk-based distances directly to the G and G_d since quirk-based distances count significant value-matches and the others do not, but there may be interesting relative effects between trees generated

by the quirk- vs. non-quirk-based distances. Table 6 shows the top-5 and bottom-5 language pairs, comparing unmodified distances to distances based on unary and binary quirks.

Table 6. The top-5 and bottom-5 language pairs in terms of G-distance (unmodified Gower) and G_q-distances (modified for unary quirks and binary quirks).

Rank	G		G_q^1		G_q^2	
1	Ulua-Sumo	0.00	Cabecar-Bribri	0.49	Cabecar-Bribri	0.48
2	Sumo-Misquito	0.01	Ulua-Sumo	0.53	Ulua-Sumo	0.53
3	Ulua-Misquito	0.01	Sumo-Misquito	0.55	Sumo-Misquito	0.54
4	Cabecar-Bribri	0.04	Ulua-Misquito	0.55	Ulua-Misquito	0.54
5	Sambu-Catio	0.05	Sambu-Catio	0.58	Sambu-Catio	0.58
…	…		…			
591	Quiche-Bocota	0.58	Xinca-Cabecar	0.93	Xinca-Cabecar	0.93
592	Quiche-Cabecar	0.58	Xinca-Teribe	0.93	Xinca-Teribe	0.93
593	Xinca-Cabecar	0.58	Quiche-Bocota	0.93	Quiche-Bocota	0.93
593	Teribe-Quiche	0.59	Quiche-Movere	0.94	Quiche-Movere	0.94
595	Quiche-Movere	0.60	Teribe-Quiche	0.94	Teribe-Quiche	0.94

Ulua-Sumo – the pair that shares every feature – is no longer the pair with the smallest distance in the quirkiness-sensitive distances, as what the Cabecar-Bribri pair shares is more significant. In general, as in the earlier tests, differences among language pairs are slight; nevertheless, there may still be specific pairs with drastic relative changes.

To check this, we again compute Neighbor-Joining trees (Saitou and Nei 1987) for the quirk-distances, shown in the rightmost trees in figure 3, and comb the outcome trees for reshuffled languages. There is very little difference compared to trees already commented on; in fact, the unary-quirk tree (third column) and the dependency-sensitive tree (second column) are topologically identical. A clear trend in the case of the binary-quirk distance (tree shown in fourth column) is that differences between all or any of the languages are de-emphasized, which presumably reflects that most quirks are not shared (a situation that dilutes the impact of any few that are shared). If this were to prove typical of feature sets in typology in general, quirks would be either insignificant details when it comes to typological profiles, or they would need to be projected on a different scale (artificially enhanced?) to have ramifications.

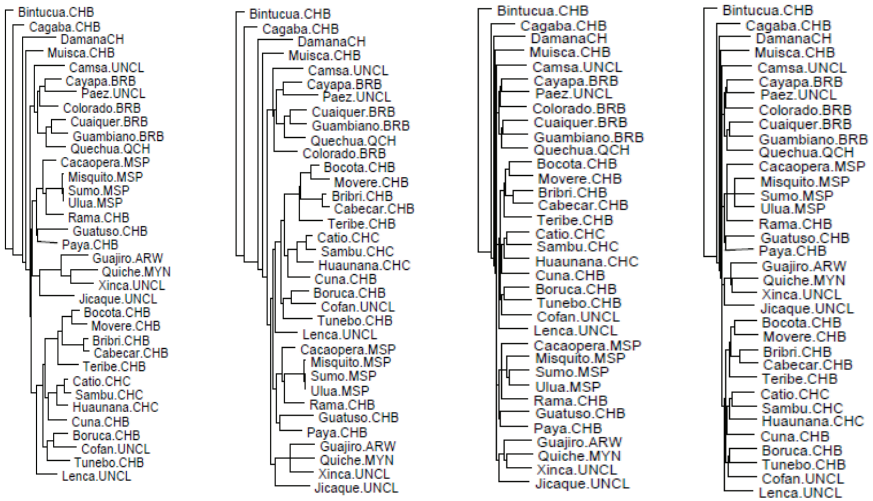

Figure 3: Neighbor-Joining trees based on distances calculated with the normal Gower coefficient G (far left), the modified dependency-sensitive Gower coefficient G_d (mid-left), the unary quirk-based distance G_q^1 (mid-right) and the binary quirk-based distance G_q^2 (far right). All have been arbitrarily rooted with Bintucua as the outlier.

5. Discussion and conclusions

In this paper we have presented two approaches to factoring out functional dependencies from datasets of typological features for natural languages. We first addressed the *presence of dependencies* among features and eliminated these in a dependency-sensitive version of the distance metric. The second approach considered distance with respect to the *value of features* in a modified metric that addressed the relative quirkiness of the features shared by a pair of languages. Both dependency-sensitive metrics make the assumption that the dependencies are of low order (binary or unary) in order to get a tractable approximation of arbitrary-order dependencies

Experiments on a dataset of Chibchan and neighboring languages revealed numerous dependencies between features. However, the impact of this dependency was found to be minimal for typological comparison between languages. When comparing any two languages, there was little difference between blind and dependency-sensitive distance metrics.

There is still the possibility that striking dependency-related effects between certain languages do exist, but on a more micro-level than the full typological profiles of the languages question. To test something in this direction, we re-ran the modified algorithms against the two meaningful subsets in the original 81-feature database: one data matrix with 42 morphosyntactic features, and a second data matrix with 39 phonological features. While there were a few changes using the dependency-sensitive metrics, these changes were as minor as those reported for the full typological profiles, and no valuable insights at all were gained from the quirkiness-sensitive measures.

Thus our final result is that dependencies inhabit the language-feature matrix of the Isthmo-Colombian Area uniformly and not as surprising contingencies between particular geographical neighbors or particular pairs of unrelated languages.

Acknowledgements

The authors are grateful to Michael Dunn and Annemarie Verkerk for digitizing the Constenla Isthmo-Colombian Area data and translating the list of features, and to Devdatt Dubhashi and Vinay Jethava for discussion and literature pointers regarding computational approaches to feature dependencies. We thank Pieter Muysken and Bernard Comrie for comments on an earlier draft; Muysken, Comrie, Mily Crevels, and Östen Dahl for comments on the final version, and the participants of the October 2011 Approaches to Measuring Linguistic Difference workshop at University of Gothenburg for useful discussion and suggestions. The usual disclaimers apply. This study was conducted with support from ERC Advanced Grant 230310 'Traces of Contact'.

Appendix: The set of typological features of the Isthmo-Colombian Area dataset, adapted from Constenla Umaña (1991: 88–120, 179–185).

1	Is there VO order in transitive clauses?
2	Is there VS (verb-agent) order in transitive clauses?
3	Is there OS (patient-agent) order in transitive sentences?
4	Is there VS (S may be agent or object) order in intransitive sentences?
5	Is the order of adpositions and nouns as follows: preposition - noun?
6	Is the order of adpositions and nouns as follows: noun - postposition or noun - case suffixes?

7 Is the order of the noun that is possessed and the noun that indicates the possessor (the genitive) N - Gen?

8 Is the order of the noun that is possessed and the noun that indicates the possessor (the genitive) Gen - N?

9 Is the order of the adjective and the noun A - N?

10 Is the order of the adjective and the noun N - A?

11 Is the order is numerals with respect to the indefinite nominal phrase Num - N?

12 Is the order of the demonstrative and the noun Dem - N?

13 Is the order of the demonstrative and the noun N - Dem?

14 Is the order of the interrogative word and the clause obligatorily question word - clause (i.e., are question words positioned initially)?

15 Is there a passive voice (only consider whether the language has a passive voice when it is possible to include a nominal phrase which indicates the semantic agent in the passive clause)?

16 Is there an anti-passive?

17 Is there an distinction between inclusive and exclusive for personal pronouns?

18 Is there an opposition between masculine and feminine personal pronouns?

19 Is there an opposition between formal and informal personal pronouns?

20 Is it the case that a negational element, which may be a particle or a prefix, constantly and obligatorily precedes the declarative clause?

21 Is it the case that a negational element, which may be a particle or a suffix, constantly and obligatorily follows the declarative clause?

22 Is there a morpheme that marks genitive case in inalienable possession?

23 Is there a morpheme that marks genitive case in alienable possession?

24 Is there a morpheme that marks accusative case?

25 Does the language have a case system that distinguishes between the agent of an intransitive action verb and the patient of intransitive process verbs?

26 Does the language have a case system that does not distinguish between the agent or patient of an intransitive verb and the patient of an transitive verb?

27 Are non-verbal predicates inflected for tense, aspect or person?

28 Are there gender oppositions (animate/inanimate, masculine/feminine or both) expressed in the inflection of any major word class (generally on verbs, on adjectives, or both types of word classes)?

29 Are there inflectional prefixes?

30 Is there inflection expressed with or marked through replacement of segmental or non-segmental phonemes?

31 Are there prefixes which indicate (grammatical) person?

32 Are there suffixes which indicate (grammatical) person?
33 Is there inflection for indicating the (grammatical) person of the possessor on the noun (personal possessive inflection)?
34 Is there inflection for indicating the (grammatical) person on intransitive verbs?
35 Is there inflection for indicating the (grammatical) person of the agent on transitive verbs?
36 Is there inflection for indicating the (grammatical) person of the object on the transitive verb?
37 Are there prefixes to indicate tense or aspect?
38 Are there suffixes to indicate tense or aspect?
39 Are there directional elements attached to the verb?
40 Is there a distinction in the shape of (all or some) nouns for possessed/non-possessed?
41 Are there numeral classifiers?
42 Is there definite marking distinct from demonstratives?
43 Is there an opposition between high plus mid vowels versus front vowels?
44 Is there an opposition between high plus mid vowels versus back vowels?
45 Is there a rounding contrast for non-front (central or back) vowels of the same height?
46 Is there a open/closed contrast for vowels of the same height and the same ‚series'?
47 Is there a nasality contrast for vowels?
48 Is there a quantity contrast for vowels?
49 Are there tonal contrasts?
50 Is there a non-labial glottalized occlusive?
51 Is there a labial glottalized occlusive?
52 Is there at least one implosive?
53 Is there an aspiration contrast for occlusives?
54 Is there a phonemic condition for the glottal occlusive?
55 Is there a phoneme /p/ (simple, bilabial, voiceless occlusive)?
56 Is there one or more uvular occlusive phoneme?
57 Is there at least one obstruent phoneme (occlusive or fricative) which is labial and voiced or weak/lenis?
58 Is there at least one obstruent phoneme (occlusive or fricative) which is dental or alveolar and voiced or weak/lenis?
59 Is there at least one obstruent phoneme (occlusive or fricative) which is velar and voiced or weak/lenis?
60 Is there at least one fricative phoneme that is hissing and voiced or weak/lenis?

61 Is there a sonority or lenition contrast for affricates?
62 Is there a glottalization contrast for affricates?
63 Is there an aspiration contrast for affricates?
64 Is there at least one alveolar affricate?
65 Is there at least one prepalatal affricate?
66 Is there a lateral affricate?
67 Is there at least one retroflex affricate?
68 Is there a voiceless labial fricative phoneme (/ɸ/ or /f/)?
69 Is there a voiceless prepalatal fricative phoneme (/ʃ/)?
70 Is there a voiceless retroflex fricative phoneme (/ʂ/)?
71 Is there a voiceless lateral fricative phoneme (/ɬ/)?
72 Are there the consonantal nasal phonemes bilabial /m/ and alveolar /n/?
73 Is there a mediopalatal nasal phoneme (/ɲ /)?
74 Is there a velar nasal phoneme /ŋ/?
75 Are there voiceless nasal phones as realizations of nasal phonemes before /h/ or before sequences of nasal phonemes with /h/?
76 Are there word initial consonant clusters that consist of consist of one nasal and another consonant?
77 Is there a voiced lateral approximant (/1/)?
78 Is there a voiced mediopalatal approximant /ʎ/?
79 Is there a simple central vibrant phoneme /ɾ/?
80 Is there a simple lateral vibrant phoneme /ɺ/?
81 Is there a multiple vibrant phoneme /r/?

Notes

1. Available online at http://wals.info accessed 1 June 2011.
2. The other assumption underlying the use of the Gower coefficient – that fully independent features should carry equal weight when calculating distance – is not explored in this paper.
3. Though there is a fair amount of work in the field of structured data prediction (cf. Getoor and Taskar 2007) we are not aware of any previous work that develops modified distance measures based on feature dependencies.
4. $H(B)=-\sum P(b_i)$ log $P(b_i)$ where P is the underlying probability/frequency distribution for the values b_i of feature B.
5. We are not concerned here with the validity of these trees for historical linguistics in the region – the interest is whether there is rearrangement of some kind in the distance-based clustering.

References

Adelaar, Willem F. H., and Pieter C. Muysken
2004 *The Languages of the Andes* (Cambridge Language Surveys). Cambridge: Cambridge University Press.

Aguirre Licht, Daniel
2006 Choco Languages. In Keith Brown (ed.), *Encyclopedia of Language and Linguistics,* Volume 2, 367–381. 2d ed. Amsterdam: Elsevier.

Bickel, Balthasar
2010 Capturing particulars and universals in clause linkage: a multivariate analysis. In Isabelle Bril (ed.), *Clause-hierarchy and clause-linking: the syntax and pragmatics interface*, 51–101. Amsterdam: John Benjamins.

Campbell, Lyle
1997 *American Indian Languages: the Historical Linguistics of Native America.* Oxford: Oxford University Press.

Captain, David
2005 Proto Lokono-Guajiro. *Revista Latinoamericana de Estudios Etnolingüísticos* 10: 137–172.

Cerrón-Palomino, Rodolfo
2003 *Lingüística quechua* (Monumenta Lingüística Andina 10). 2d ed. Cuzco: Centro de Estudios Regionales Andinos "Bartolomé de las Casas".

Chiswick, Barry R., and Paul W. Miller
2004 Linguistic Distance: A Quantitative Measure of the Distance Between English and Other Languages. Institute for the Study of Labor (IZA DP No. 1246). Bonn: IZA.

Christiansen, Morten H, Chris Collins, and Shimon Edelman (eds)
2009 *Language Universals*. Oxford: Oxford University Press.

Chomsky, Noam
1981 *Lectures on Government and Binding.* Dordrecht: Foris Publications.

Chu, Yoeng-Jin, and Tseng-Hong Liu
1965 On the Shortest Arborescence of a Directed Graph. *Scientia Sinica* 4: 1396–1400.

Constenla Umaña, Adolfo
1991 *Las lenguas del área intermedia: introducción a su estudio areal*. San José: Universidad de Costa Rica.

Constenla Umaña, Adolfo
2012 Chibchan languages. In Lyle Campbell and Verónica Grondona (eds.), *The Indigenous Languages of South America: A Comprehensive Guide* (The World of Linguistics 2), 391–440. Berlin: De Gruyter Mouton.

Cysouw, Michael
2007 New Approaches to Cluster Analysis of Typological Indices. In Reinhard Köhler and Peter Grzybek (eds.), *Exact Methods in the Study of Language and Text,* 61–76. Berlin: Mouton de Gruyter.

Dahl, Östen
2008 An exercise in *a posteriori* sampling. *Sprachtypologie und Universalienforschung* 61(3): 208–220.

Daumé, Hal III, and Lyle Campbell
2007 A Bayesian Model for Discovering Typological Implications. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 65–72. Prague: Assn for Computational Linguistics.

Dryer, Matthew S.
1992 The Greenbergian Word Order Correlations. *Language* 68(1): 81–138.

Dunn, Michael, Simon J. Greenhill, Stephen C. Levinson, and Russell D. Gray
2011 Evolved structure of language shows lineage-specific trends in word-order universals. *Nature* 473: 79-82.

Eriksen, Love
2011 *Nature and Culture in Prehistoric Amazonia. Using G.I.S. to reconstruct ancient ethnogenetic processes from archaeology, linguistics, geography, and ethnohistory.* Lund: Lund Studies in Human Ecology 12.

Gensler, Orin D.
2003 Shared quirks: a methodology for "non-orthodox" historical linguistics. Paper presented to the 17th International Conference of Historical Linguistics, Prague, 29 July.

Georgiadis, Leonidas
2003 Arborescence optimization problems solvable by Edmonds' algorithm. *Theoretical Computer Science* 71: 233–240.

Getoor, Lise, and Ben Taskar
2007 Introduction. In Lise Getoor and Ben Taskar (eds.), *Introduction to Statistical Relational Learning* (Adaptive Computation and Machine Learning), 1–11. MIT Press.

Holman, Eric W., Christian Schulze, Dietrich Stauffer, and Søren Wichmann
2007 On the relation between structural diversity and geographical distance among languages: Observations and computer simulations. *Linguistic Typology* 11(2): 395–423.

Kolman, Connie J., and Eldridge Bermingham
1997 Mitochondrial and Nuclear DNA Diversity in the Chocó and Chibcha Amerinds of Panamá. *Genetics Society of America* 147: 1289–1302.

Melton, P. E., Briceño, I., Gomez, A., Devor, E. J., Bernal, J. E., and Crawford, M.
2007 Biological Relationship Between Central and South American Chibchan Speaking Populations: Evidence from mtDNA. *American Journal of Physical Anthropology* 133: 753–770.

Pearson, Karl
1901 On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine* 2(6): 559–572.

Pineda, Baron
2005 Miskito and Misumalpan Languages. In Philipp Strazny (ed.), *Encyclopedia of linguistics* volume 2, 693–695. New York: Fitzroy Dearborn.

Polyakov, Vladimir N., Valery D. Solovyev, Søren Wichmann, and Oleg Belyaev
2009 Using WALS and Jazyki Mira. *Linguistic Typology* 13: 137–167.

Quesada, J. Diego
2007 *The Chibchan Languages*. Cartago: Editorial Tecnológica de Costa Rica.

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

Sinnemäki, Kaius
2010 Word order in zero-marking languages. *Studies in Language* 34(4): 869–912.

Trillos Amaya, Maria
2005 *Lenguas Chibchas de la Sierra Nevada de Santa María: Una Perspectiva Histórico-Comparativa.* Bogotá: Universidad de los Andes.

Wichmann, Søren, and Eric W. Holman
2010 Pairwise comparisons of typological profiles. In Jan Wohlgemuth and Michael Cysouw (eds.), *Rethinking Universals: How rarities affect linguistic theory* (Empirical Approaches to Language Typology 45), 241–254. Berlin: Mouton de Gruyter.

Williams, Cindy
1993 A grammar sketch of Dəməna. University of North Dakota MA thesis.