Exploring Word Order Universals: a Probabilistic Graphical Model Approach

Xia Lu
Department of Linguistics
University at Buffalo
Buffalo, NY USA
xialu@buffalo.edu

Abstract

In this paper we propose a probabilistic graphical model as an innovative framework for studying typological universals. We view language as a system and linguistic features as its components whose relationships are encoded in a Directed Acyclic Graph (DAG). Taking discovery of the word order universals as a knowledge discovery task we learn the graphical representation of a word order sub-system which reveals a finer structure such as direct and indirect dependencies among word order features. Then probabilistic inference enables us to see the strength of such relationships: given the observed value of one feature (or combination of features), the probabilities of values of other features can be calculated. Our model is not restricted to using only two values of a feature. Using imputation technique and EM algorithm it can handle missing values well. Model averaging technique solves the problem of limited data. In addition the incremental and divide-and-conquer method addresses the areal and genetic effects simultaneously instead of separately as in Daumé III and Campbell (2007).

1 Introduction

Ever since Greenberg (1963) proposed 45 universals of language based on a sample of 30 languages, typologists have been pursuing this topic actively for the past half century. Since some of them do not agree with the term (or concept) of "universal" they use other terminology such as "correlation", "co-occurrence", "dependency", "interaction" and "implication" to refer to the relationships between/among linguistic feature pairs most of which concern morpheme and word order. Indeed the definition of "universals" has never been clear until recently, when most typologists agreed that such universals should be statistical universals which are "statistical tendencies" discovered from data samples by using statistical methods as used in any other science. Only those tendencies that can be extrapolated to make general conclusions about the population can be claimed to be "universals" since they reflect the global preferences of value distribution of linguistic features across genealogical hierarchy and geographical areas.

Previous statistical methods in the research of word order universals have yielded interesting results but they have to make strong assumptions and do a considerable amount of data preprocessing to make the data fit the statistical model (Greenberg, 1963; Hawkins, 1982; Dryer, 1989; Nichols, 1986; Justeson & Stephens, 1990). Recent studies using probabilistic models are much more flexible and can handle noise and uncertainty better (Daumé III & Campbell, 2007; Dunn et al., 2011). However these models still rely on strong theoretic assumptions and heavy data treatment, such as using only two values of word order pairs while discarding other values, purposefully selecting a subset of the languages to study, or selecting partial data with complete values. In this paper we introduce a novel approach of using a probabilistic graphical model to study word order universals. Using this model we can have a graphic representation of the structure of language as a complex system composed of linguistic features. Then the relationship among these features can be quantified as probabilities. Such a model does not rely on strong assumptions and has little constraint on data.

The paper is organized as follows: in Section 2 we discuss the rationale of using a probabilistic graphic model to study word order universals and introduce our two models; Section 3 is about learning structures and parameters for the two models. Section 4 discusses the quantitative analysis while Section 5 gives qualitative analysis of the results. Section 6 is about inference such as MAP query and in Section 6 we discuss the advantage of using PGM to study word order universals.

Proceedings of the ACL Student Research Workshop, pages 150–157,
Sofia, Bulgaria, August 4-9 2013.

2 A new approach: probabilistic graphical modeling

2.1 Rationale for using PGM in word order study

The probabilistic graphical model is the marriage of probabilistic theory and graph theory. It combines a graphical representation with a complex distribution over a high-dimensional space. There are two major types of graphical representations of distributions. One is a Directed Acyclic Graph (DAG) which is also known as a Bayesian network with all edges having a source and a target. The other is an Undirected Acyclic Graph, which is also called a Markov network with all edges undirected. A mixture of these two types is also possible (Koller & Friedman, 2009).

There are two advantages of using this model to study word order universals. First the graphical structure can reveal much finer structure of language as a complex system. Most studies on word order correlations examine the pairwise relationship, for example, how the order of verb and object correlates with the order of noun and adjective. However linguists have also noticed other possible interactions among the word order features, like chains of overlapping implications: Prep ⊃ ((NAdj ⊃ NGen) & (NGen ⊃ NRel)) proposed by Hawkins (1983); multi-conditional implications (Daumé III, 2007); correlations among six word order pairs and three-way interactions (Justeson & Stephens, 1990); spurious word order correlations (Croft et al., 2011); chains of associations, e.g. if C predicts B and B predicts A, then C predicts A redundantly (Bickel, 2010b). These claims about the possible interactions among word order features imply complex relationships among the features. The study of word order correlations started with pairwise comparison, probably because that was what typologists could do given the limited resources of statistical methods. However when we study the properties of a language, by knowing just several word orders such as order of verb and object, noun and adpositions, etc., we are unable to say anything about the language as a whole. Here we want to introduce a new perspective of seeing language as a complex system. We assume there is a meta-language that has the universal properties of all languages in the world. We want a model that can represent this meta-language and make inferences about linguistic properties of new languages. This system is composed of multiple sub-systems such as phonology, morphology, syntax, etc. which correspond to the subfields in linguistics. In this paper we focus on the subsystem of word order only.

The other advantage of PGM is that it enables us to quantify the relationships among word order features. Justeson & Stephens (1990) mentioned the notion of "correlation strength" when they found out that N/A order appears less strongly related to basic V/S/O order and/or adposition type than is N/G order. This is the best a log-linear model can do, to indicate whether a correlation is "strong", "less strong", "weak" or "less weak". Dunn et al. (2011) used Bayes factor value to quantify the relationships between the word order pairs but they mistook the strength of evidence for an effect as the strength of the effect itself (Levy & Daumé III, 2011). A PGM model for a word order subsystem encodes a joint probabilistic distribution of all word order feature pairs. Using probability we can describe the degree of confidence about the uncertain nature of word order correlations. For example, if we set the specific value as evidence, then we can get the values of other features using an inference method. Such values can be seen as quantified strength of relationship between values of features.

2.2 Our model

In our word order universal modeling we will use DAG structure since we think the direction of influence matters when talking about the relationship among features. In Greenberg (1966a) most of the universals are unidirectional, such as "If a language has object-verb order, then it also has subject-verb order" while few are bidirectional universals. The term "directionality" does not capture the full nature of the different statuses word order features have in the complex language system. We notice in all the word order studies the order of SOV or OV was given special attention. In Dryer's study VO order is the dominant one which determines the set of word order pairs correlated with it (or not). We assume word order features have different statuses in the language system and such differences should be manifested by directionality of relationships between feature pairs. Therefore we choose DAG structure as our current model framework.

Another issue is the sampling problem. Some typologists (Dryer 1989, Croft 2003) have argued that the language samples in the WALS database (Haspelmath et al., 2005) are not independent and identically distributed (i.i.d.) because languages can share the same feature values due to either genetic or areal effect. While

others (Maslova, 2010) argue that languages within a family have developed into distinct ones through the long history. We notice that even we can control the areal and genetic factors there are still many other factors that can influence the typological data distribution, such as 1) language speakers: cognitive, physiological, social, and communicative factors; 2) data collection: difficulty in identifying features; political biases (some languages are well documented); 3) random noise such as historical accidents. Here we do not make any assumption about the i.i.d property of the language samples and propose two models: one is FLAT, which assumes samples are independent and identically distributed (i.i.d.); the other is UNIV, which takes care of the possible dependencies among the samples. By comparing the predictive power of these two models we hope to find one that is closer to the real distribution.

3 Learning

To build our models we need to learn both structure and parameters for the two models. We used Murphy (2001)'s Bayesian Network Toolbox (BNT) and Leray & Francois (2004)'s BNT Structure Learning Package (BNT_SLP) for this purpose.

3.1 Data

As we mentioned earlier we will restrict our attention to the domain of word order only in this paper. In the WALS database there are 56 features belonging to the "Word Order" category. Because some of the features are redundant, we chose 15 sets of word order features which are: S_O_V[1] (order of subject, object and verb) [7[2]], S_V (order of subject and verb) [3], O_V (order of object and verb) [3], O_Obl_V (order of Object, Oblique, and Verb) [6], ADP_NP (order of adposition and noun phrase) [5], G_N (order of genitive and noun) [3], A_N (order of adjective and noun) [4], Dem_N (order of demonstrative and noun) [4], Num_N (order of numeral and noun) [4], R_N (order of relative clause and noun) [7], Deg_A (order of degree word and adjective) [3], PoQPar (position of polar question particles) [6], IntPhr (position of interrogative phrases in content questions) [3], AdSub_Cl (order of adverbial subordinator and clause) [5], Neg_V (order of negative morpheme and verb) [4]. We did some minimal treatment of data. For Neg_V which has 17 values we collapsed its values 7-17 to 6 ("Mixed"). For Dem_N and Neg_V, we treat word and suffix as the same and collapsed values 1 and 3 to 1, and values 2 and 4 to 2. After deleting those languages with no value for all 15 word order features we have 1646 data entries. This database is very sparse: in overall the percentage of missing values is 31%. For seven features more than 50% of the languages have values missing.

3.2 Learning the FLAT model

There are two big problems in learning DAG structure for the FLAT model. One is caused by large number of missing values. Because EM method for structures from incomplete data takes very long time to converge due to the large parameter space of our model, we decided to use imputation method to handle the missing data problem (Singh, 1997). The other difficulty is caused by limited data. To solve this problem we used model averaging by using bootstrap replicates (Friedman et al., 1999). We use GES (greedy search in the space of equivalent classes) algorithm in BNT_SLP to learn structure from a bootstrap dataset because it uses CPDAGs to represent Markov equivalent classes which makes graph fusion easier. The algorithm is as follows:

1) Use nearest-neighbor method to impute missing values in the original dataset D and create a complete dataset D_s.
2) Create T=200 bootstrap resamples by resampling the same number of instances as the original dataset with replacement from D_s. Then for each resample D_s^t learn the highest scoring structure G_s^t.
3) Fuse the 200 graphs into a single graph G_s using the "Intergroup Undirected Networks Integration" method (Liu et al., 2007). Then use *cpdag_to_dag.m* in BNT_SLP to change G_s into a directed graph G_s'.
4) Compute the BIC scores of G_s' using the 200 resamples and choose the highest one. If the convergence criterion (change of BIC is less than 10^{-4} compared with the previous iteration) is met, stop. Otherwise go to Step 5.
5) Learn 200 sets of parameters θ_s^t for G_s' using the 200 resamples and take a weighted-average as the final parameters θ_s'. Also use EM algorithm and dataset D to learn parameters θ_{EM} for G_s'. Choose the parameters θ between θ_s' and θ_{EM} that gives the highest BIC score. Use MAP estimation to fill in the missing values in D and generate a complete dataset D_{s+1}. Go to Step 2.

[1] The detailed descriptions of these word order features and values can be found at http://wals.info/.

[2] The number in the square brackets indicates the number of values for that feature.

The structure for the FLAT model is shown in Figure 1.

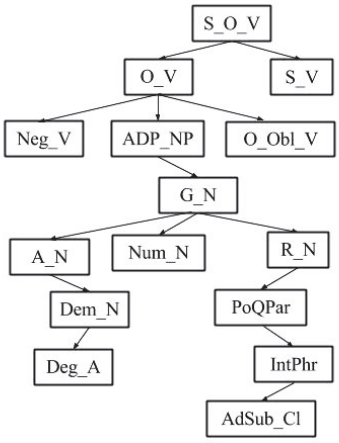

Figure 1. DAG structure of the FLAT model

3.3 Learning the UNIV model

As discussed in Section 2.2, the possible dependencies among language samples pose difficulty for statistical methods using the WALS data. Daumé III & Campbell (2007)'s hierarchical models provided a good solution to this problem; however their two models LINGHIER and DISTHIER dealt with genetic and areal influences separately and the two separate results still do not tell us what the "true universals" are.

Instead of trying to control the areal and genetic and other factors, we propose a different perspective here. As we have mentioned, the kind of universals we care about are the stable properties of language, which means they can be found across all subsets of languages. Therefore to solve the problem of dependence among the languages we take an incremental and divide-and-conquer approach. Using clustering algorithm we identified five clusters in the WALS data. In each cluster we picked *1/n* of the data and combine them to make a subset. In this way we can have *n* subsets of data which have decreased degree of dependencies among the samples. We learn a structure for each subset and fuse the *n* graphs into one single graph. The algorithm is as follows:

1) Use nearest-neighbor method to impute missing values and create M complete datasets D_m $(1 \leq m \leq M)$.
2) For each D_m divide the samples into n subsets. Then for each subset D_m^n learn the highest scoring structure G_m^n.
3) Fuse the n graphs into a single graph G_m using the "Intragroup Undirected Networks Integration" method (Liu et al., 2007).
4) Fuse the M graphs to make a single directed graph $G_M{}'$ as in Step 3 in the previous section.
5) Compute the BIC score of $G_M{}'$ using datasets D_m $(1 \leq m \leq M)$ and choose the highest score. If the convergence criterion (same as in the previous section) is met, stop. Otherwise go to Step 6.
6) Learn parameters θ_m for $G_M{}'$ using datasets D_m $(1 \leq m \leq M)$ and take a weighted-average as the final parameters $\theta_M{}'$. Also use EM algorithm and original dataset to learn parameters θ_{EM} for $G_M{}'$. Choose the parameters θ among $\theta_M{}'$ and θ_{EM} that gives the highest BIC score. Use MAP estimation to fill in the missing values in D and generate another M complete dataset. Go to Step 2.

The final structure for the UNIV model is shown in Figure 2.

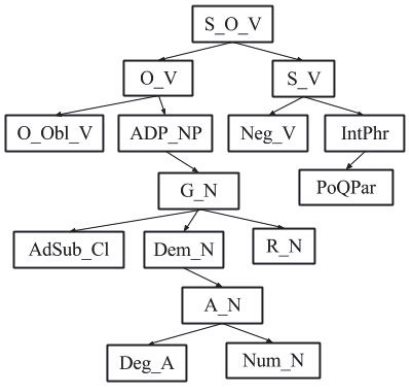

Figure 2. DAG structure of the UNIV model

The semantics of a DAG structure cannot be simply interpreted as causality (Koller & Friedman, 2009). From this graph we can see word order features are on different tiers in the hierarchy. The root S_O_V seems to "dominate" all the other features; noun modifiers and noun are in the middle tier while O_Obl_V, AdSub_Cl, Deg_A, Num_N, R, Neg_V and PoQPar are the leaf nodes which might indicate their smallest contribution to the word order properties of a language. O_V seems to be an important node since most paths start from it indicating its influence can flow to many other nodes.

We can also see there are two types of connections among the nodes: 1) direct connection: any two nodes connected with an arc directly have influence on each other. This construction induces a correlation between the two features regardless of the evidence. This type of dependency was the one most explored in the previous literatures. 2) three cases of indirect connections: a. indirect causal effect: e.g. O_V does not influence G_N directly, but via ADP_NP; b. indirect evidential effect: knowing G_N will change our belief about O_V indirectly; c. common cause: e.g. ADP_NP and O_Obl_V can influence each other without O_V being observed. Our model reveals a much finer structure of the word order

sub-system by distinguishing different types of dependencies that might have been categorized simply as "correlation" in the traditional statistical methods.

4 Quantitative Analysis of Results

The word order universal results are difficult to evaluate because we do not know the correct answers. Nonetheless we did a quantitative evaluation following Daumé III and Campbell (2007)'s method. The results are shown in Figure 3.

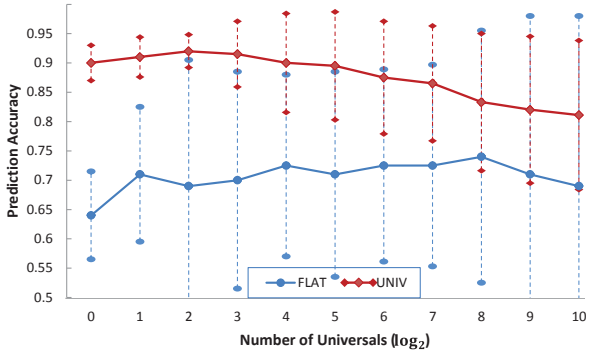

Figure 3. Results of Quantitative Evaluation

As we can see the predictive power of the UNIV model is much better than that of the FLAT model. The accuracy of our both models is lower than those of Daumé III and Campbell's. But this does not mean our models are worse considering the complexity in model learning. Instead our UNIV model shows steady accurate prediction for the top ten universals and has more stable performance compared with other models.

Using the UNIV model we can do many types of computation. Besides pairwise feature values, we can calculate the probability of any combination of word order feature values. If we want to know how value "GN" of feature "G_N" is dependent on value "POST" of feature "ADP_NP" we set POST to be evidence (probability=100%) and get the probability of having "GN". Such a probability can be taken as a measurement of dependence strength between these two values. We need more evidence for setting a threshold value to define a word order universal but for now we just use 0.5. We calculated the probabilities of all pairwise feature values in the UNIV model which can found at http://www.acsu.buffalo.edu/~xialu/univ.html.

5 Qualitative Analysis of Results

We also did qualitative evaluation through comparison with the well-known findings in word order correlation studies. We compared our results with three major works: those of Greenberg's, Dryer's, and Daumé III and Campbell's.

5.1 Evaluation: compare with Greenberg's and Dryer's work

Comparison with Greenberg's work is shown in Table 1 (in Appendix A). If the probability is above 0.5 we say it is a universal and mark it red. We think values like 0.4-0.5 can also give us some suggestive estimates therefore we mark these green. For Universal 2, 3, 4, 5, 10, 18 and 19, our results conform to Greenberg's. But for others there are discrepancies of different degrees. For example, for U12 our results show that "VSO" can predict "Initial" but not very strongly compared with "SOV" predicting "Not_Initial".

Table 2 (in Appendix A) shows our comparison with Dryer (1992)'s work. We noticed there is an asymmetry in terms of V_O's influence on other word order pairs, which was not discussed in previous work. In the correlated pairs, only ADP_NP and G_N show bidirectional correlation with O_V while PoQPar becomes a non-correlated pair. In the non-correlated pairs, Dem_N becomes a correlated pair and other pairs also show correlation of weak strength. Most of our results therefore do not confirm Dryer's findings.

5.2 Evaluation: compare with Daumé III and Campbell's work

We compared the probabilities of single value pairs of the top ten word order universals with Daumé III and Campbell's results, which are shown in the following figures.

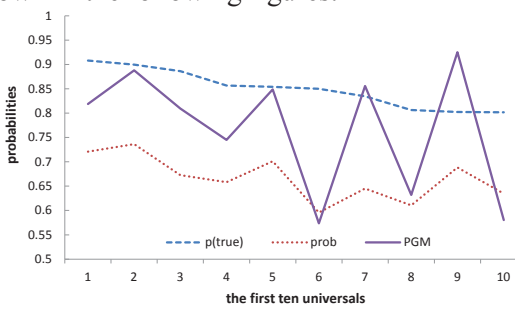

Figure 4. Compare with Daumé III and Campbell's HIER model

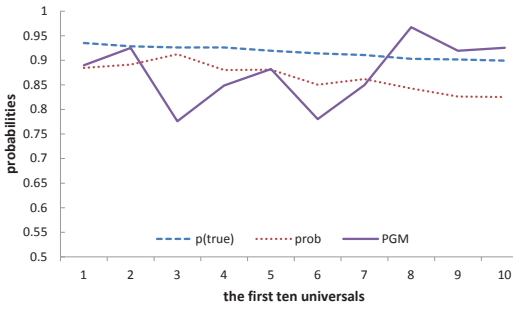

Figure 5. Compare with Daumé III and Campbell's DIST model

P(true) is the probability of having the particular implication; *prob* is the probability calculated in a different way which is not specified in Daumé III and Campbell's work. *PGM* is our model. It can be seen that our model provides moderate numbers which fall between the two probabilities in Daumé III and Campbell's results. In Figure 4 the two universals that have the biggest gaps are: 9) Prepositions ->VO and 10) Adjective-Noun->Demonstrative-Noun. In Figure 5 the three universals that have the biggest gaps are: 3) Noun-Genitive->Initial subordinator word, 6) Noun-Genitive->Prepositions and 8) OV->SV. It is hard to tell which model does a better job just by doing comparison like this. Daumé III and Campbell's model computes the probabilities of 3442 feature pairs separately. Their model with two values as nodes does not consider the more complex dependencies among more than two features. Our model provides a better solution by trying to maximize the joint probabilities of all word order feature pairs.

6 Inference

Besides discovering word order universals, our model can reveal more properties of word order sub-system through various inference queries. At present we use SamIam[3] for inference because it has an easy-to-use interface for probabilistic inference queries. Figure 6 (in Appendix B) gives an example: when we know the language is subject preceding verb and negative morpheme preceding verb, then we know the probability for this language to have postpositions is 0.5349, as well as the probabilities for the values of all other features.

The other type of query is MAP which aims to find the most likely assignments to all of the unobserved variables. For example, when we only know that language is VO, we can use MAP query to find the combination of values which has the highest probability (0.0032 as shown in Table 3 in Appendix C).

One more useful function is to calculate the likelihood of a language in terms of word order properties. If all values of 13 features of a language are known, then the probability (likelihood) of having such a language can be calculated. We calculated the likelihood of eight languages and got the results as shown in Figure 7 (in Appendix C). As we can see, English has the highest likelihood to be a language while Hakka Chinese has the lowest. German and French have similar likelihood; Portuguese and Spanish are similar but are less than German and French. In other words English is a typical language regarding word order properties while Hakka Chinese is an atypical one.

[3] SamIam is a tool for modeling and reasoning with Bayesian networks (http://reasoning.cs.ucla.edu/samiam/).

7 Discussion

Probabilistic graphic modeling provides solutions to the problems we noticed in the previous studies of word order universals. By modeling language as a complex system we shift our attention to the language itself instead of just features. Using PGM we can infer properties about a language given the known values and we can also infer the likelihood of a language given all the values. In the future if we include other domains, such as phonology, morphology and syntax, we will be able to discover more properties about language as a whole complex system.

Regarding the relationships among the features since PGM can give a finer structure we are able to see how the features are related directly or indirectly. By using probability theory we overcome the shortcomings of traditional statistical methods based on NHST. Probabilities capture our uncertainty about word order correlations. Instead of saying "A is correlated with B", we can say "A is correlated with B to a certain extent". PGM enables us to quantify our knowledge about the word order properties of languages.

Regarding the data treatment, we did very little preprocessing of data, therefore reducing the possibility of bringing in additional bias from other processes such as family construction in Dunn et al.'s experiment. In addition we did not remove most of the values so that we can make inferences based on values such as "no determinant order" and "both orders". In this way we retain the information in our data to the largest extent.

We think PGM has the potential to become a new methodology for studying word order universals. It also opens up many new possibilities for studying linguistic typology as well:

- It can include other domains to build a more complex network and to discover more typological properties of languages.
- It can be used in field work for linguists to make predictions about properties of unknown languages.

References

Bickel, B. 2010a. *Absolute and statistical universals*. In Hogan, P. C. (ed.) The Cambridge Encyclopedia of the Language Sciences, 77-79. Cambridge: Cambridge University Press.

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

Croft, William. 2003. *Typology and universals*. 2nd edn. Cambridge: Cambridge University Press.

Daphne Koller and Nir Friedman. 2009. *Probabilistic Graphical Models: Principles and Techniques (Adaptive Computation and Machine Learning series)*. MIT Press, Aug 31, 2009

Daumé, H., & Campbell, L. (2007). A Bayesian model for discovering typological implications. In *Annual Meeting –Association For Computational Linguistics* (Vol. 45, No. 1, p. 65).

D.M. Chickering, D. Heckerman, and C. Meek. 1997. A Bayesian approach to learning Bayesian networks with local structure. *Proceeding UAI'97 Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence*.

Dryer, M. S. 1989. Large linguistic areas and language sampling. *Studies in Language 13*, 257 – 292.

Dryer, Matthew S. & Martin Haspelmath (eds.). 2011. *The world atlas of language structures online*. München: Max Planck Digital Library.

Dryer, Matthew S. 2011. The evidence for word order correlations. *Linguistic Typology 15*. 335–380.

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

E. T. Jaynes. 2003. *Probability Theory: The Logic of Science*. Cambridge University Press, Apr 10, 2003.

Friedman, N. (1998, July). The Bayesian structural EM algorithm. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence* (pp. 129-138). Morgan Kaufmann Publishers Inc.

Friedman, N., Nachman, I., & Peér, D. (1999, July). Learning bayesian network structure from massive datasets: the "sparse candidate" algorithm. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence* (pp. 206-215). Morgan Kaufmann Publishers Inc.

Greenberg, J. H. 1963. *Some universals of grammar with particular reference to the order of meaningful elements*. In Universals of Language, J. H. Greenberg, Ed. MIT Press, Cambridge, MA, 73-113.

Greenberg, Joseph H. 1966. Synchronic and diachronic universals in phonology. *Language 42*. 508–517.

Greenberg, J. H. (1969). Some methods of dynamic comparison in linguistics. *Substance and structure of language*, 147-203.

Hawkins, John A. 1983. *Word Order Universals.* Academic Press, 1983.

Justeson, J. S., & Stephens, L. D. (1990). Explanations for word order universals: a log-linear analysis. In *Proceedings of the XIV International Congress of Linguists* (Vol. 3, pp. 2372-76).

Leray, P., & Francois, O. (2004). BNT structure learning package: Documentation and experiments.

Levy, R., & Daumé III, H. (2011). Computational methods are invaluable for typology, but the models must match the questions: Commentary on Dunn et al.(2011).Linguistic Typology.(To appear).

Liu, F., Tian, F., & Zhu, Q. (2007). Bayesian network structure ensemble learning. In *Advanced Data Mining and Applications* (pp. 454-465). Springer Berlin Heidelberg.

Maslova, Elena & Tatiana Nikitina. 2010. Language universals and stochastic regularity of language change: Evidence from cross-linguistic distributions of case marking patterns. Manuscript.

Murphy, K. (2001). The bayes net toolbox for matlab. *Computing science and statistics, 33*(2), 1024-1034.

Perkins, Revere D. 1989. Statistical techniques for determining language sample size. *Studies in Language 13*. 293–315.

Singh, M. (1997, July). Learning Bayesian networks from incomplete data. In *Proceedings of the National conference on Artificial Intelligence* (pp. 534-539). JOHN WILEY & SONS LTD.

William Croft, Tanmoy Bhattacharya, Dave Kleinschmidt, D. Eric Smith and T. Florian Jaeger. 2011. Greenbergian universals, diachrony and statistical analyses [commentary on Dunn et al., Evolved structure of language shows lineage-specific trends in word order universals*]*. *Linguistic Typology 15*.433-53.

Appendices

A. Comparison with others' work

Universals	Dependencies	UNIV
U2: ADP_NP<=>N_G	POST->GN PRE->NG GN->POST NG->PRE	83.59 70.29 78.45 81.91
U3: VSO->PRE	VSO->PRE	74.41
U4: SOV->POST	SOV->POST	85.28
U5: SOV&NG->NA	SOV&NG->NA	68.95
U9: PoQPar<=>ADP_NP	Initial->PRE Final->POST PRE->Initial POST->Final	41.87 49.67 15.80 31.73
U10: PoQPar<=> VSO	all values of PoQPar: VSO below 10%	below 10%
U11: IntPhr->VS	Initial->VS	24.12
U12: VSO->IntPhr	VSO->Initial SOV->Initial SOV->Not_Initial	50.54 28.52 60.41
U17: VSO->A_N	VSO->A_N	24.86
U18&19: A_N<=>Num_N<=>Dem_N	AN->NumN AN->DemN NA->NNum NA->NDem	68.86 73.74 61.74 61.00
U24: RN->POST (or AN)	RN->POST RN->AN	65.73 29.23

Table 1. Comparison with Greenberg's work

OV	UNIV	VO	UNIV
correlated pairs			
ADP_NP(POST)	90.48	ADP_NP(PRE)	82.72
G_N(GN)	79.38	G_N(NG)	61.49
R_N(RN)	19.66	R_N(NR)	75.17
PoQPar(Final)	31.89	PoQPar(Initial)	15.79
AdSub_Cl (Final)	20.90	AdSub_Cl (Initial)	49.22
IntPhr(Not_Initial)	58.74	IntPhr(Initial)	34.36
non-correlated pairs			
A_N(AN)	29.48	A_N(NA)	65.00
Dem_N(Dem_N)	52.27	Dem_N(N_Dem)	54.25
Num_N(NumN)	41.6	Num_N(NNum)	49.25
Deg_A(Deg_A)	43.48	Deg_A(A_Deg)	38.44
Neg_V(NegV)	48.06	Neg_V(VNeg)	25.13

Table 2. Comparison with Dryer's work

B. Probabilistic query example in SamIam

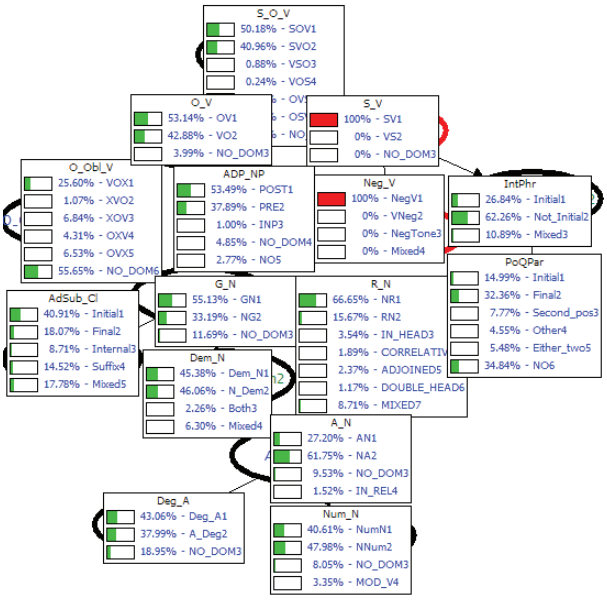

Figure 6. One query example

C. Inference examples

P(MAP,e)=0.0015052949102098631 P(MAP|e)=0.003213814742532023	
Variable	**Value**
A_N	NA
ADP_NP	PRE
AdSub_Cl	Initial
Deg_A	Deg_A
Dem_N	N_Dem
G_N	NG
IntPhr	Not_Initial
Neg_V	NegV
Num_N	NNum
O_Obl_V	VOX
PoQPar	Final
R_N	NR
S_O_V	SVO
S_V	SV

Table 3. MAP query example

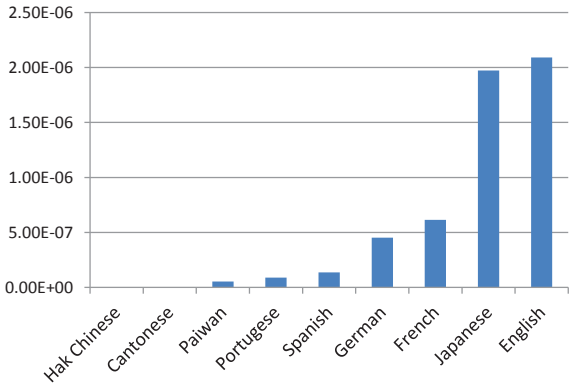

Figure 7. Likelihood of eight languages in terms of word order properties