TESTING MODELS OF WORD ORDER UNIVERSALS*

Paola Merlo (Paola.Merlo@unige.ch)

Abstract

This paper shows in detail an example of how computational learning paradigms can help test and compare theories about universals. Its main contribution lies in the illustration of the encoding and comparison of theories about typological universals to measure the generalisation ability of these theories.

1. MULTILINGUAL COMPUTATIONAL MODELLING OF LANGUAGE

Current computational linguistic work shows great interest in extending successful probabilistic modelling to multilingual approaches. Many tasks and applications, such as tagging or parsing, are being investigated in a multi-lingual perspective. The final-goal of this line of work is to uncover cross-linguistic regularities to automatically extend new techniques and technologies to new languages, and to make use of large amounts of data.

Computational modelling can interact with large-scale linguistic work at other interesting levels. From the point of view of the theory, the properties of the computational models might shed light on some of the properties of the generative processes underlying natural language. From the point of view of the data, computational models can be used to develop and test correlations between different aspects of the data on a large scale. Methodologically, computational models and machine learning techniques provide robust tools to test the predictive power of the proposed generalisation.

Language universals — linguistic properties, observed or very abstract, that are exhibited by all languages — are at the moment a topic of great debate. Their nature and even their existence has been called into question (Dunn, Greenhill, Levinson, and Gray 2011) and their general nature and distribution are being investigated from a formal and cognitive point of view (Cinque 2005; Cysouw 2010a; Steedman 2011; Culbertson, Smolensky, and Legendre 2012; Culbertson and Smolensky 2012), among many others.

We will specifically concentrate on the quantitative properties of language universals (Dryer 1992; Cysouw 2010b; Steedman 2011). In this debate, it is of great interest to attempt to explain not only the possible or impossible word orders as attested by typological traditions, but also their distribution. Data-driven computational models can help cast light on this question in two main ways. First, through their formal nature, they can make the assumptions in the proposals explicit and operational. Second, through the large-scale that is inherently possible with automatic methods, claims can be quantified and verified both at the level of language type, but also at the level of linguistic token, for each individual language.

This paper concentrates on a simple but fairly central methodological point. It will illustrate how to formalise current proposals for the much debated Universal 20 in such a way that they

*The intersection of generative grammar, corpus work and computational modelling is a lonely place. Warmest thanks to Jacques, for sharing some of these interests and for his contagious enthusiasm that has made this work more fun.

can be evaluated and compared quantitatively in a setting where their ability to generalise to new cases is properly tested.

2. THE FACTS

One of the most easily observable distinguishing features of human languages is the order of words: the order of the main grammatical functions in the sentence, the position of the verb in the sentence and the respective order of the modifiers of a noun, among others. While there is great variety in the orders, most languages have very strong preferences for a few or only one order, and, across languages, not all orders are equally preferred (Greenberg 1966; Dryer 1992). Greenberg's universal 20 describes the cross-linguistic preferences for the word order of elements inside the noun phrase.

> **Greenberg's Universal 20**
>
> If Dem, Num, Adj precede the Noun, they are found in this order; if they follow the Noun, they are either in this order or in reverse order.

We can reformulate universal 20 more explicitly (Cinque 2005):

(a) In prenominal position the order of demonstrative, numeral, and adjective is Dem>Num>A.

(b) In postnominal position the order is either Dem>Num>A or A>Num>Dem.

Some aspects of Greenberg's formulation have withstood the test of time, but some others have been found to be too strong. On the one hand, a larger sample of languages has shown that two of the three orders indicated by Greenberg's as the only possible orders are indeed among the most frequent ones. On the other hand, larger samples have also shown that many more orders are possible than stated in Greenberg's universal, but with different frequencies (Cinque 2005; Dryer 2006).

Establishing the actual basic facts is not so simple. We will concentrate here only on the quantitative aspects and will assume without argument the results described in the literature that assign certain languages to certain word orders. In assessing the reliability of proposed counts, one has to assess the possible sources of errors induced by sampling. First of all, there is the problem of the representativity of the sample of languages that is used to collect the observations. Sampling, in general, is subject to random error and to bias error. Random error occurs when the size of the sample is not adequate to the complexity of the problem, so that some possible events are not observed. Greenberg's sample of languages was probably too small, and looking at a larger sample has discovered some orders that looked impossible. Bias error occurs when the nature of the sample is biased with respect to the conclusions one wants to draw. To draw conclusions on language universals, it is therefore crucial that the sample be representative of the true underlying linguistic diversity. The remedy to random error is to have a sufficiently large number of data points: Dryer's and Cinque's current language collections range in the thousands. To address the problem of bias error, Dryer suggests counting language genera and not individual languages, since some genera are much more densely populated, and better studied, than others (Dryer 2006).

Table 1 reports the 24 combinatorially possible orders of the four elements: N, Dem, Num, Adj and the actual counts that have been proposed in several publications: the first column shows discretised frequencies; the following two columns are Dryer (2006)'s counts by language and by genera; and the following column are Cinque's counts, as can be deduced from

				D's Discrete	D's Lang Freq	D's Gen Freq	C's2005 Freq
Dem	Num	Adj	N	Very Freq	74	44	Very many†
Dem	Adj	Num	N	Rare	3	2	0
Num	Dem	Adj	N	0	0	0	0
Num	Adj	Dem	N	0	0	0	0
Adj	Dem	Num	N	0	0	0	0
Adj	Num	Dem	N	0	0	0	0
Dem	Num	N	Adj	Freq	22	17	Many*
Dem	Adj	N	Num	Rare	11	6	Very few (7)
Num	Dem	N	Adj	0	0	0	0
Num	Adj	N	Dem	Rare	4	3	Very few (8)
Adj	Dem	N	Num	0	0	0	0
Adj	Num	N	Dem	0	0	0	0
Dem	N	Adj	Num	Freq	28	22	Many**
Dem	N	Num	Adj	Rare	3	3	Very few (4)
Num	N	Dem	Adj	Rare	5	3	0
Num	N	Adj	Dem	Freq	4	2	Few (2)
Adj	N	Dem	Num	Rare	2	1	Very few (3)
Adj	N	Num	Dem	Rare			
N	Dem	Num	Adj	Rare	4	3	Few (8)
N	Dem	Adj	Num	Rare	6	4	Very few (3)
N	Num	Dem	Adj	Rare	1	1	0
N	Num	Adj	Dem	Rare	9	7	Few (7)
N	Adj	Dem	Num	Freq	19	11	Few (8)
N	Adj	Num	Dem	Very Freq	108	57	Very many (27)

Table 1: Attested word order of Universal 20 and their estimated frequencies. The third and fourth column show counts according to Dryer (2006), both in number of languages and number of genera involved, and the fifth column shows counts according to Cinque (2005). In the first column, the discretised frequencies are calculated according to Dryer (2006)'s counts of genera. Since the exact number of genera is not likely to be relevant, but only their relative order of magnitude, we show a discretisation of these counts into four levels: very frequent, frequent, rare, zero (unattested). † the exact counts are not provided; * Cinque mentions European languages and 13 others; ** ten languages and alternative order for three more.

the 2005 paper. As can be observed, there are some discrepancies, which have been discussed in detail in the related publications, but also many points of agreement. In particular, while the exact numbers sometimes vary, the rank of languages or genera based on frequencies is almost identical. This observation indicates that distinguishing possible from impossible is not robust to new observations. To distinguish common from rare is less brittle and should be our main explanatory goal. In what follows, therefore, we investigate how different theories fare in explaining rare from common word orders and how well they generalise this prediction to unseen data.

3. SOME THEORIES

We will compare the descriptive and predictive adequacy of a few of the proposals that have been put forth to explain universal 20, choosing a few theories that have different properties.

In a paper that has received much commentary (Cinque 2005), Greenberg's Universal 20 is derived from independently motivated principles of syntax organised in a derivational explanation. Based on data indicated in the fifth column of Table 1, Cinque remarks that there are 24 combinatorially possible orders of the four elements: N, Dem, Num, A. According to Cinque, only 14 of them are attested in the languages of the world (but see Dryer's counts in the same table, Table 1). Some of the 14 orders are unexpected under Universal 20. Cinque proposes that the actually attested orders, and none of the unattested ones, are derivable from a single universal order of the basic constructive syntactic operator (the Linear Correspondence Axiom), and from independent conditions on phrasal movement. The Linear Correspondence Axiom first combines Nouns and Adjectives, then adds Numerals and finally adds Demonstratives. Different types of movement can move the merged elements to different positions in the phrase: all the way to the beginning of the phrase or only partially. These conditions enable one to consider some forms of movement as more costly than others and no movement as the preferred unmarked option. In this way, Cinque's proposal also derives the exceptions, and the different degrees of markedness of the various orders.

In a different proposal, a factorial, but not derivational, explanation is proposed (Cysouw 2010a). Statistical models are used and an explanation of typological frequencies is produced by the cumulative combination of various interacting characteristics. The author experiments with various models to see which one better predicts the attested frequencies. Generalized linear models fit the data best. Three characteristics are used by all models of NP-internal word order: hierarchical structure, noun-adjective order, and whether the noun is at the phrase boundary. In a further simplification of the model, the hierarchical structure can be broken down into less complex features (noun-adjective co-occurrence, demonstrative at the edge of the phrase, and noun at the edge of the phrase).[1] This factorial explanation does not provide a generative

[1]Like Cinque, Cysouw is concerned with demonstrating that the proposed principles are not limited to explaining Universal 20. To strengthen the generality of the proposed method, Cysouw discusses how it can also be used to explain the typology of sentence word order, as it is captured by Greenberg's Universal 1. This universal holds for 96% of the world's languages, but it does not model the more fine-grained differences in frequency of the six word order types. The author proposes a more complex three-feature model. The first feature is pairwise order: whether the order is SO or OS, VO or OV, SV or VS. The second feature is pairwise adjacency: for instance, whether S and O are adjacent or not. The third feature is individual position: for instance, whether S is first, medial, or final. Cysouw shows that the first two features are less important than the third and that overall the model has a better fit than universal 1. However, notice that this model comprises two three-valued features and one binary feature, so it has five degrees of freedom. These are enough degrees of freedom to simply list all the six possible

process that explains how the different word orders could arise from a common grammar, but it identifies the predictive properties of the frequency distributions of word order and their relative importance.

Dryer proposes a factorial explanation based on general principles of symmetry and harmony (Dryer 2006). Differently from Cinque's and Cysouw's this proposal does not assign any weights to the factors. The factors comprise two symmetry principles that describe the closeness of the modifiers to the noun; a principle of asymmetry that captures the main observation that prenominal modifiers exhibit fewer alternatives than post-nominal modifiers (also observed by Cinque); a principle of intra-categorial harmony; and universal 18. What is really very important in Dryer's contribution are the provided observed frequencies. On the one hand, Dryer shows that a few of the word orders that Cinque had declared impossible are actually attested, one of them quite frequently. On the other hand, it provides frequency counts based on genera and not simply on languages, based on an independently justified sampling procedure that factors out influences of language family. These genus-based counts are used in our study, and are shown in Table 1.

In conclusion, all these theories attempt to describe the very different frequency counts of types of languages by proposing factors that favour harmonic orders, and that derive the asymmetry between prenominal and post-nominal modifiers. They all try to fit the frequency distribution of the languages to the models and to compare to other proposals. In the rest of the paper, we illustrate an encoding and an automatic learning method to test how well these models predict the observed distributions of word orders.

4. BUILDING PREDICTIVE MODELS

In this section, we test the generalising ability of the different explanations that have been proposed for universal 20. We use the ability to classify new instances in a supervised learning setting as an indication of the generalising power of the theory.

To explain the frequency distribution of the word orders, Cinque affects markedness weights to the different types of move operations. In the computational terminology that will be used below, these weights are the parameters of Cinque's model, and this process is a process of parameter fitting on the available data. Fitting parameters to a model based on available data gives us a measure of the descriptive fit of the model to the data, an interesting measure in itself, but it does not test the power of generalisation of the model. This is because it is always possible to fit the data if the number of parameters in the model is sufficiently large given the amount of variation to explain. So the true test of generalisation of a model cannot lie in showing that all the data is explained if that data was actually used to determine some aspects of the model. In explaining Universal 20, what needs to be shown is that the same set of operations and markedness weights that capture the observed data also predicts new data to a good degree. In practice, the proper procedure requires fitting the markedness weights on a subset of languages (the training data), and see if the quantitative model so developed predicts the frequency distribution for new unseen test data, the different word orders on new languages, and, for rare word orders, maybe even predict the rarity of new word orders never seen before.

The steps of the simple formalisation that we propose here, therefore, are as follows:

1. Formalise the properties and operations of a model of word order as simple primitive features with a set of associated values;

word orders of the three S,O,V elements.

2. Encode each word order as a vector of instantiated primitives defined by the model;

3. Learn the model through a learning algorithm on a subset of the data;

4. Run the model on previously unseen data to test generalisation ability.

In the rest of the section, we briefly illustrate the feature-based formalisation of the linguistic proposals, and describe the experimental materials and method.

4.1. Materials

The different linguistic proposals are translated into a feature-based summary description of each of the word orders. This vectorial representation of the data is compatible with many different training regimes and algorithms. Two proposals (Cysouw's and Dryer's) are declarative, and therefore easily transferred in the simple declarative feature-based framework. Cinque's model is derivational and requires the most interpretation to be formalised and translated into features. In the simplest set up, we code the principles and operations proposed by Cinque for each word order as a vector of properties, a summary that describes each language and its word order.

Recall that the salient property of Cinque's explanation is the interaction between a fixed universal word order (the Linear Correspondence Axiom) and structure movement operations, with different markedness weights. A simplified specification of Cinque's explanation for each word order can be encoded as the values of three merge operations and the values of two types of move operations, partial and complete movement. The three merge operations build the structure linearly, corresponding to the word order. Some word orders that require merge sequences not allowed by the Linear Correspondence Axiom are encoded as negative data. The move operations can move elements one step, two steps (that is they can be partial movement) or all the way to the beginning of the phrase, as complete movement. These two types of move operations can be of several types, np-movement, pied-piping, among others. It is crucial to point out that this is only a *model* of Cinque's explanation, limited only to the discriminating features. For example, the fact that there are two movement types in the description of each word order does not imply that there are necessarily two movement steps. There could be more than one partial movement or none. In the vectorial representation, all partial movements (i.e. movements that do not reach the left edge of the phrase) are reduced to one value.

Recall that Cysouw proposes a factorial explanation, where factors are preferences of directionality and surface proximity. Cysouw shows that three factors are enough to explain the variation: whether the Noun is near the edge of the Noun phrase or not, whether the Demonstrative is near the edge or not, and whether the Adjective is near the Noun. These are surface observed properties that can be encoded directly in the vector of features that describes each word order.

Dryer's factorial explanation is based on general principles of symmetry and harmony, and does not use any weighing coefficients. Again, these are observed properties that can be encoded directly in the vector of features that describes each word order.

The features and the possible values of Cinque's model are shown in Figure 1. First, second and third represent the three merge operations, and their values are the pairs of synctactic part-of-speech-tags of heads that are being merged (we assume a dependency representation for the trees). Partial and complete are the two features representing the two movements, and their possible values, which encode the types of movement that Cinque postulates. The values in the last column are the frequency property of the word order, the dependent variable we are trying

- Template: < first, second,third, partial, complete >
- Attributes and Values
  - first: adjn, demn, ndem, nnum, numn
  - second: adjdem, demadj, demnum, numadj, numdem, numn
  - third: adjdem, adjn, adjnum, demnum, numadj, numdem
  - partial: not, np, of-who-pp, whose-pp
  - complete: not, np, of-who-pp, whose-pp
  - frequency: very frequent, frequent, rare, none (VF,F,R,No)
- Example Vectors
  - adjn,numadj,demnum,not,not,VF
  - adjn,numadj,demnum,not,not,VF
  - numn,demnum,adjn,not,not,R
  - adjn,demadj,numdem,not,not,No
  - demn,adjdem,numadj,not,not,No
  - numn,demnum,adjdem,not,not,No
  - demn,numdem,adjnum,not,not,No
  - adjn,numadj,demnum,whose-pp,not,F

Figure 1: A sample of Cinque's move and merge feature vectors. (See text for explanation.)

- **Type-based encoding:** each language type as positive or negative piece of data.
- **Token-based encoding:** token-based classification encodes frequency of languages (notion of markedness), following Dryer's frequency counts based on genera, as size of sample in the training set.
- **Ten-fold cross-validation**
- **Three predictive regimes:**
  - two-way: (possible, impossible);
  - four-way: very frequent, frequent, rare, unattested;
  - seven-way: two levels of very frequent, two levels of frequent, two levels of rare; one for unattested.

Figure 2: Summary of materials and method.

to explain. Since the actual counts of languages are still under discussion, it is better to represent frequency in frequency classes. We can group the languages in different frequency groups, by discretising the frequencies in different ways: either as simply possible or impossible (two values), or as having different levels of frequency. Figure 1 shows a four-way discretisation into very frequent (VF), frequent (F), rare (R), and unattested (No). We defined four and seven discrete values, based on observation of the groupings of the actual numerical values. These differential frequencies are then represented in the training by repeating each example the number of times indicated in Dryer's frequency counts by genera. So, for examples the vector that represents the word order N Dem Adj Num, attested in four genera, is repeated four times. Notice that the fact that we also encode unattested word orders means we explicitly represent negative data. The encodings of the other approches are done analogously.

We can define the problem in two slightly different ways, as a classification of types or a classification of tokens. In classifying language types, we try to assign each language type to a correct frequency value. Each type to be classified is unique, which yields 24 data point, for this universal. In developing a model based on a subset of the data, we are guaranteed that the new test data will be completely unseen.

In classifying tokens, we construct an experimental situation which corresponds to the real sampling situation. Each language type is represented by a variable number of languages. Some of the types are represented by many languages (those that are frequent), in our representation many instances of a given feature vector, other types will be represented by fewer languages. Unattested word orders will be explicitly represented as negative data. This set up has many more data points and it could happen that the test set contains examples of word orders that have also been seen at training time.

Figure 2 summarises the experimental setup. The three predictive regimes, ten-fold cross validation and the learning methods will be explained in the next section.

4.2. The Models

Once the data is encoded in an appropriate way, we need to reproduce Cinque's way of assigning markedness values (fitting the weights), done by hand, or Cysouw's way of fitting the model to the data. Cinque's and Cysouw's method consist, manually or automatically, in assigning

Assume target function $f : X \rightarrow V$, where each instance x described by attributes $\langle a_1, a_2 \ldots a_n \rangle$.
Most probable value of $f(x)$ is:

$$v = \underset{v_j \in V}{\operatorname{argmax}} P(v_j | a_1, a_2 \ldots a_n) \quad (1)$$

$$v = \underset{v_j \in V}{\operatorname{argmax}} \frac{P(a_1, a_2 \ldots a_n | v_j) P(v_j)}{P(a_1, a_2 \ldots a_n)} \quad (2)$$

$$= \underset{v_j \in V}{\operatorname{argmax}} P(a_1, a_2 \ldots a_n | v_j) P(v_j) \quad (3)$$

Naive Bayes assumption:

$$P(a_1, a_2 \ldots a_n | v_j) = \prod_i P(a_i | v_j) \quad (4)$$

Naive Bayes classifier: $v_{NB} = \underset{v_j \in V}{\operatorname{argmax}} P(v_j) \prod_i P(a_i | v_j)$

Figure 3: Naive Bayes Classifier

weights to reproduce the observed frequencies of possible and impossible values as well as possible.

We will then test the predictive ability of these weighted explanations on new data. This is a supervised learning setting. In this set up, formally, we say that a computer program learns from experience E with respect to some task T and performance measure P, if its performance at task T, as measured by P, improves with E. In our case, the training experience E will be provided by a database of correctly classified language types or tokens; the task T consists in classifying word orders types or tokens unseen in E into predetermined frequency classes; and the performance measure P will be defined as the percentage of types or tokens correctly classified. This learning paradigm is called supervised learning, because of the training phase, in which the algorithm is provided examples with the correct answers. In the testing phase, these rules or probabilities are applied to additional data, not included in the training phase. The accuracy of classification on the test set indicates whether the rules or probabilities developed in the training phase are general enough, yielding good test accuracy, or are too specific to the training set to generalise well to other data.

There are numerous algorithms for learning the weights of a model in a supervised setting, and many regimes for training and testing such algorithms. In the following experiments, we use two probabilistic learning algorithms — Naive Bayes and the Weighted Average One-dependence Estimator —, and n-fold cross-validation as the training and testing protocol.

The Naive Bayes algorithm is based on Bayes theorem and is defined in Figure 3. In this method, the objective of training is to learn the most probable word order type given the probability of each vector of features (see equation (1) in Figure 3). This probability is decomposed, according to Bayes rule, into the probability of the features given the word order and the prior probability of the word order itself (see equations (2) and (3) in Figure 3).

This method is chosen because it is probabilistic and simple. Its probabilistic aspect provides a mathematically well-founded framework to predict frequencies and to combine attributes. The simplicity of the algorithm allows a clear interpreation of the outputs and the

	Naive Bayes					
	Type (24)			Token (213)		
	Two	Four	Seven	Two	Four	Seven
Cinque	88	58	42	98	82	90
Cysouw	*67*	*21*	38	*96*	91	38
Dryer	92	54	63	99	92	71
Baseline	71	50	38	97	47	28

Table 2: Percent of languages or language types classified in the right frequency class. Results (Rounded Accuracy) using Naive Bayes Classifier. Italics indicate lower than baseline results.

results.

In a classification task, we want to predict the class, in our case the frequency of the word order (very many, many, few, none), based on some descriptively pertinent features of the problem. The most noticeable feature of Naive Bayes is the very strong conditional independence assumption across features (see equation (4) in Figure 3). In our case, this assumption represents the intuition that the principles used to build word orders are independently motivated, and therefore they should be able to combine freely. We also experimented with a more complex model where properties are not assumed to be independent of each other. The model, called an averaged weighted one-dependence estimator (WAODE), assumes dependence from only one attribute at a time, taking the weighted average of the results of all the attributes.

To avoid excessive dependence of the results on a specific partition of the data, we use cross-validation. Cross-validation is a training and testing protocol in which the system randomly divides the data into n parts, and then runs the learner n times, using $n - 1$ partitions for training and the remaining one for testing. At each run of the learner, a different partition is chosen for testing. The performance measure is averaged over all n experiments.

Finally, results will be compared to an uninformed baseline which consists in assuming that all word orders belong to the most frequent class. The baseline is helpful in indicating whether the models learn anything beyond mere frequency effects.

5. RESULTS AND DISCUSSION

We are now in a position to run the experiment. We run a 10-fold cross-validation, using a Naive Bayes classifier. Table 2 shows the results of the experiment, as the proportion of correct answers (percent accuracy). As can be seen by the accuracy results, the models' generalisation are far from perfect, at the level of language types (shown in the left panel). In the binary classification, possible or impossible languages, almost 10% of the data is incorrectly classified. Several of the models of type-based classification have performances below or equal to the baseline: the model does not learn.

Token-based classification yields better results: Dryer's and Cinque's models, in particular, achieve good results with the same small number of factors; the three factors of Cysouw's model are clearly insufficient.

We concentrate now on a more detailed analysis of Cinque's model; analyses of the other two models can be done analogously. All the mistakes, as indicated by the accuracy per class and by the confusion matrix, shown in Tables 3 and 4, fall in the *many, few* and *none* category.[2]

[2]As usual, we use the measures of precision and recall. Precision is the number of correctly classified items over the total number of items in a given class; recall is the number of correctly classified items over the total

	Naive Bayes Results		
	Prec	Rec	F
Very Freq	83	100	90
Freq	82	70	76
Rare	90	56	69
None	56	71	62
W Avg	83	82	81

Table 3: Percent precision, recall and F measure by frequency class. Results (Rounded Accuracy) using Naive Bayes Classifier for Cinque's model in a token classification setting.

	Confusion Matrix			
	Very Freq	Freq	Rare	None
Very Freq	101	0	0	0
Freq	21	50	0	0
Rare	0	11	19	4
None	0	0	2	5

Table 4: Confusion Matrix of Naive Bayes Classifier for Cinque's model in a token classification setting.

	not	np	of-who-pp	whose-pp
partial	63	39	11	116
complete	110	20	30	69

Table 5: Distribution of counts of word order instances by type of movement.

Interestingly, most mistakes tend to classify the tokens in a class of higher frequency than the correct one; only four of the rare cases are mistakenly classified as unattested. This is because the attributes associated with frequency events dominate the classification.

The Naive Bayes confusion matrix by frequency class indicates that very frequent orders and unattested word orders are over-estimated (Recall $>$ Precision), while frequent and rare word orders are under-estimated (Precision $>$ Recall). The fact that the F-measure decreases with the frequency of the class indicates that the model is not a good predictor of cases that are rarely attested in the training data.

Even more informative are the actual probabilities learnt by the model. If we look at the joint probability distribution of the attributes and their values, we can calculate the probabilities of different aspects of the model by marginalising out some of the details of the distribution. If we marginalise out the values by frequency we find that partial and complete movement have very different distributions, as shown in Table 5.[3]

If we sum up the counts and compare all types of movement operations to no movement, we find that in 166 cases there is a partial movement operation and in 63 cases no movement, while in 110 there is no movement and in 119 there is complete movement. This shows that while no movement is preferred in both cases — as expected under the movement explanation of the prenominal, postnominal asymmetry proposed by Cinque — there are many more cases of partial movement than of complete movement. This is not expected as complete movement is supposed to be easier, so that one could expect it to occur more often.

We can also observe how partial and complete movement types pattern across frequency levels. There are different types of frequent word orders, and even more types of rare word orders. If we look at the distribution of types of movement for frequent and rare word orders, we see the patterns shown in Table 6. Partial movement is not always more frequent and complete

number of items that should have been found in a given class; and F is their harmonic mean. Confusion matrices indicate the correct output by rows and the model's predictions by columns.

[3]Movement of the *pictures of who* type is coded as *of-who* and *whose picture* is coded as *whose-pp*.

	Frequent		Rare	
	Partial	Complete	Partial	Complete
not	1	40	10	17
np	22	12	15	6
of-who-pp	1	22	8	6
whose-pp	51	1	5	9

Table 6: Distribution of counts of word order instances by type of movement in frequent and rare word orders.

	WAODE Classifier				Naive Bayes
	Precision	Recall	F	Acc	Acc
Cinque	94	94	94	94	82
Cysouw	88	91	89	91	91
Dryer	96	97	96	96	92

Table 7: Percent of languages or language types classified in the right frequency class, for a token-based four-way classification, averaged over a 10-fold cross-validation.

movement is not always less frequent. The noticeable differences in distributions indicates that all these distinctions (partial, complete) and their four levels are needed to make the necessary distinctions.

As a control of the independence assumption in the Bayes model, we also learn the data with a probabilistic classifier that relaxes the strong independence assumption. The model, called an averaged weighted one-dependence estimator (WAODE), assumes dependence from only one attribute at a time, taking the weighted average of all the possible dependences. What is relevant here is that this constitutes a minimally different model from a Naive Bayes classifier, so that only the assumption of independence of attributes is changed across the two models. Results are much better, as shown in Table 7. In particular, the classifier no longer mistakes systematically the frequent word orders, as shown in Tables 8 to 10, reporting the confusion matrices. However, here again the accuracy, while very high, is not perfect. This demonstrates that a true separate test set is needed to assess the real generality of the proposed models.

The fact that a classifier that makes weaker independence assumptions about its attributes yields better performance than Naive Bayes, which assumes conditional independence of the attributes, indicates that the attributes are not independent. These attributes are supposed to be the primitive, independently motivated, — in a differente sense of the word *independent* — operations and properties of the different linguistic proposals that give rise to the different word orders. Finding a statistical dependence among them indicates that part of the explanation of the data is given by the interaction of the factors, interaction that cannot be independently motivated, as it is specific to these data. This means that part of the explanation provided by the linguistic models rests on interactions other than those operations that can be justified on general theoretical grounds.

6. CONCLUSIONS

This paper has shown in detail how simple computational learning paradigms can help test and compare theories about universals. The process of finding probabilities automates and

	VF	F	R	No
VF	101	0	0	0
F	0	71	0	0
R	0	10	23	1
No	0	0	2	5

Table 8: Confusion Matrix of WAODE Classifier for Cinque's model.

	VF	F	R	No
VF	101	0	0	0
F	0	71	0	0
R	1	12	21	0
No	0	5	5	0

Table 9: Confusion Matrix of WAODE Classifier for Cysouw's model.

	VF	F	R	No
VF	101	0	0	0
F	0	71	0	0
R	0	7	27	0
No	0	0	1	6

Table 10: Confusion Matrix of WAODE Classifier for Dryer's model.

makes mathematically precise the assignment weights that we find in proposals about language universals, but does not change the logic of these proposals. The added value of this procedure is two-fold. On the one hand, we use a mathematically well-defined probabilistic framework, so that combination of factors, ranking and optimisation processes are well-defined. On the other hand, the evaluation rests on the use of unseen data, so that the quantitative results are a measure of generalisation. This method, then, constitutes a well-defined procedure to estimate the weights of the operations and aspects of the models and to compare their generalisation capabilities. Future work lies in developing more accurate models for more complex or more comprehensive problems.

REFERENCES

Cinque, G. (2005). Deriving Greenberg's universal 20 and its exceptions. *Linguistic Inquiry 36*(3), 315–332.

Culbertson, J. and P. Smolensky (2012). A Bayesian model of biases in artificial language learning: The case of a word-order universal. *Cognitive Science*, 1–31.

Culbertson, J., P. Smolensky, and G. Legendre (2012). Learning biases predict a word order universal. *Cognition*, 306–329.

Cysouw, M. (2010a). Dealing with diversity: towards an explanation of NP word order frequencies. *Linguistic Typology 14*(2), 253–287.

Cysouw, M. (2010b). On the probability distribution of typological frequencies. In *Proceedings of the 10th and 11th Biennial conference on The mathematics of language*, MOL'07/09, Berlin, Heidelberg, pp. 29–35. Springer-Verlag.

Dryer, M. (2006). The order demonstrative, numeral, adjuctive and noun: an alternative to cinque. http://exadmin.matita.net/uploads/pagine/1898313034_cinqueH09.pdf.

Dryer, M. S. (1992). The Greenbergian word order correlations. *Language 68*, 81–138.

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

Greenberg, J. H. (1966). *Language Universals*. The Hague, Paris: Mouton.

Steedman, M. (2011). Greenberg's 20th: The view from the long tail. unpublished manuscript, University of Edinburgh.