Predicting word order universals

Paola Merlo
University of Geneva, Department of Linguistics,
Geneva, Switzerland

ABSTRACT

This paper shows a computational learning paradigm to compare and test theories about language 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. In so doing, this method uncovers hidden dependencies between theoretical dimensions and primitives that were considered independent and independently motivated.

Keywords: frequency, word order universals, Greenberg's Universal 20, computational modelling, probabilistic models, Naive Bayes

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 multilingual 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 – whether defined as linguistic properties, observed or very abstract, that are exhibited by all languages or as statistical implications of pairs of linguistic properties – are at the moment a topic of great debate. Their nature and even their existence has been called into question (Dunn *et al.* 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 *et al.* 2012; Culbertson and Smolensky 2012; Futrell *et al.* 2015).

We will specifically concentrate on the quantitative properties of word order 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 not only at the level of language type, but also at the level of linguistic token, for each individual language.

This paper concentrates on a central methodological point. It will illustrate how to formalise some of the current proposals for the much debated Universal 20 (Greenberg 1966) – the universal governing the linear order of a noun and its modifiers – in such a way that they can be evaluated and compared quantitatively in a setting where their ability to generalise to new cases is properly tested. In this respect, this work shares the goals of Cysouw (2010a), but differently from these previous proposals of the same nature, the proposed theories are encoded as faithfully as possible, by using their defined primitives and operations as features in our models.

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**
> When any or all the items (demonstrative, numeral, and descriptive adjective) precede the noun, they are always found in this order. If they follow, the order is exactly the same or its exact opposite.

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. (See, for example, Dryer's and Cinque's large data collections in the cited work.) 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 the proposed counts, one has to assess the possible sources of errors induced by sampling. 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 inspections of larger samples have 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, for example, as generated by a posited probabilistic system. 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 hundreds. 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).[1]

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's (2006) counts by language and by genera; and the following column are Cinque's counts, as can be deduced from the 2005 paper. In the first column, the discretised frequencies are calculated according to Dryer's counts of genera. As can be observed, there are some discrepancies across the different counting methods and across authors, 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 aiming to predict the frequency rank, as opposed to exact frequency counts, would be more robust across theories and more robust to new observations. The numerical frequency data are then transformed into ordered data by a process of discretisation and then used by a discrete classifier. The discretisations can be done at different levels of granularity. Table 2 shows a two-way, four-way, and seven-way discretisation. More will be said about this discretisation later. In what follows, therefore, we investigate how different theories fare in explaining different levels of frequency of word orders and how well they generalise this prediction to previously unseen data.

[1] Dryer (2005, 584) provides the following definition: "A genus is a group of languages whose relatedness is fairly obvious without systematic comparative analysis and which even the most conservative "splitter" would accept.". (An explanation of genus is also available on WALS online at `http://wals.info/languoid/genealogy`.) Examples are such subfamilies of Indo-European as Germanic, Slavic, and Romance languages.

Table 1: Attested word orders of Universal 20 and their estimated frequencies. (See the text for more explanation.)

				D's	D's	D's	C's
				Discr	Lang	Gen	Freq
Dem	Num	Adj	N	V. Freq	74	44	V. 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	V. few (7)
Num	Dem	N	Adj	0	0	0	0
Num	Adj	N	Dem	Rare	4	3	V. 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	V. few (4)
Num	N	Dem	Adj	Rare	5	3	0
Num	N	Adj	Dem	Freq	38	21	Few (2)
Adj	N	Dem	Num	Rare	4	2	V. few (3)
Adj	N	Num	Dem	Rare	2	1	V. few
N	Dem	Num	Adj	Rare	4	3	Few (8)
N	Dem	Adj	Num	Rare	6	4	V. 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	V. Freq	108	57	V. many (27)

† The exact counts are not provided.

* Cinque mentions European languages and 13 others.

** Ten languages and alternative order for three more.

Table 2: Two-way (possible or 0), four-way (very frequent, frequent, rare, none, abbreviated as VF,F,R,0) and seven-way (57,44,22,11,6,3,0) discretisation and the observed counts based on genera from Dryer's.

				Two-way Discr	Four-way Discr	Seven-way Discr	Dryer's Genera
Dem	Num	Adj	N	Possible	VF	44	44
Dem	Adj	Num	N	Possible	R	3	2
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	Possible	F	22	17
Dem	Adj	N	Num	Possible	R	6	6
Num	Dem	N	Adj	0	0	0	0
Num	Adj	N	Dem	Possible	R	3	3
Adj	Dem	N	Num	0	0	0	0
Adj	Num	N	Dem	0	0	0	0
Dem	N	Adj	Num	Possible	F	22	22
Dem	N	Num	Adj	Possible	R	3	3
Num	N	Dem	Adj	Possible	R	3	3
Num	N	Adj	Dem	Possible	F	22	21
Adj	N	Dem	Num	Possible	R	3	2
Adj	N	Num	Dem	Possible	R	3	1
N	Dem	Num	Adj	Possible	R	3	3
N	Dem	Adj	Num	Possible	R	3	4
N	Num	Dem	Adj	Possible	R	3	1
N	Num	Adj	Dem	Possible	R	6	7
N	Adj	Dem	Num	Possible	F	11	11
N	Adj	Num	Dem	Possible	VF	57	57

3 SOME THEORIES

We will compare the descriptive and predictive adequacy of a few of the proposals that have been put forth to explain Greenberg's 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 as those shown in the fifth column of Table 1, Cinque remarks that there are 24 combinatorially possible orders of the four elements: N, Dem, Num, Adj. 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, Kayne 1994), 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. Three characteristics are used by all models of the 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).[2]

[2] Like Cinque, Cysouw is concerned with demonstrating that the proposed

This factorial explanation does not provide a generative 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. Figure 1 spells out the principles. 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

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. Recall Universal 1: "In declarative sentences with nominal subject and object, the dominant order is almost always one in which the subject precedes the object." This universal holds for 96% of the world's languages, but it does not model the finer-grained differences in frequency of the six word order types. Cysouw 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 word orders of the three S,O,V elements.

- **Symmetry Principle 1**
  The adjective and numeral tend to occur closer to the noun than the determiner, when they occur on the same side of the noun.
- **Symmetry Principle 2**
  The adjective tends to occur closer to the noun than the numeral, when they occur on the same side of the noun.
- **Asymmetry Principle**
  The symmetry principles tend to apply more strongly to prenominal modifiers than to postnominal modifiers; exceptions to the symmetry principles will occur only to the postnominal modifiers.
- **Greenberg's Universal 18**
  When the descriptive adjective precedes the noun, the demonstrative and the numeral, with overwhelmingly more than chance frequency do likewise.
- **Intra-categorial Harmony**
  The demonstrative, numeral, and adjective tend to all occurr on the same side of the noun.

Figure 1: The five principles used in Dryer's explanation of Universal 20.

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 some of the different explanations that have been proposed for Universal 20. The method will require transforming the three theories into a vectorial representation, as described below, and then automatically finding the relative weight of each element in the vector, a process of parameter fitting. We use the ability to classify new instances in a supervised learning setting as an indication of the generalising power of the theory. We compare the three theories described above.

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. (For a similar point with a different example, see also Abney 2011.) 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 weights on a subset of languages (the training data), and seeing if the quantitative model so developed predicts the frequency distribution for test data not seen during training.

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;
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. We describe here the process to reach this conversion in detail.

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. 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.

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 kinds, 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 (that is, movements that do not reach the left edge of the phrase) are reduced to one value.

The features and the possible values of Cinque's model are shown in Figure 2. *First*, *second* and *third* represent the three merge operations, and their values are the pairs of syntactic 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 of this feature are *not*, encoding the fact that no movement has taken place, *np*, encoding the movement of the NP alone, *of-who-pp*, encoding NP-movement with pied-piping of the *picture of who* type, and

Figure 2: Cinque's move and merge feature vectors. (See the text for explanation.)

- Template: < first, second, third, partial, complete, frequency >
- Attributes and Values
  - first: AN, DN, ND, NNum, NumN
  - second: AD, DA, DNum, NumA, NumD, NumN
  - third: AD, AN, ANum, DNum, NumA, NumD
  - 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)
- Vectors

AN	NumA	DNum	not	not	VF
NumN	DNum	AN	not	not	R
AN	DA	NumD	not	not	No
DN	AD	NumA	not	not	No
NumN	DNum	AD	not	not	No
DN	NumD	ANum	not	not	No
AN	NumA	DNum	whose-pp	not	F
AN	NumA	DNum	of-who-pp	not	R
AN	DA	NumD	whose-pp	not	No
AN	NumA	DNum	not	of-who-pp	R
NNum	DNum	AD	not	not	No
ND	NumN	ANum	not	not	No
AN	NumA	DNum	whose-pp	not	F
AN	NumA	DNum	np	not	R
AN	DA	NumD	not	not	R
AN	NumA	DNum	np	of-who-pp	F
AN	NumA	DNum	not	of-who-pp	R
AN	NumA	DNum	of-who-pp	whose-pp	R
AN	NumA	DNum	np	not	R
AN	NumA	DNum	whose-pp	np	R
AN	NumA	DNum	np	np	R
AN	NumA	DNum	np	whose-pp	R
AN	NumA	DNum	whose-pp	np	F
AN	NumA	DNum	whose-pp	whose-pp	VF

whose-pp, encoding NP-movement with pied-piping of the *whose picture* type.[3]

The values in the last column are the frequency property of the word order, the dependent variable we are trying to explain. We discuss them below.

Recall that Cysouw proposes a factorial explanation, where factors are preferences of directionality and surface proximity. Cysouw shows that three factors are sufficient to achieve a good fit to the data, and argues that a model with fewer parameters should be preferred to a model with more parameters: 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. The resulting features, feature values, and vectors are shown in Figure 3.

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 resulting features, feature values, and vectors are shown in Figure 4.

The goal attribute, the attribute we are trying to predict, is the frequency of a given word order. Since the actual counts of languages are still under discussion, and therefore are not entirely reliable, it is a better representation of the current state of reliability of the frequency counts to group them 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. Table 2 shows the different discretisaton values and how they compare to Dryer's counts based on genera. We defined four and seven discrete values, based on observation of the groupings of the actual numerical values. Figures 2, 3,

[3] Many instances of wh-movement involve pied-piping. Pied-piping occurs when a fronted wh-word pulls an entire encompassing phrase to the front of the clause. Cinque indicates that *picture of who* pied-piped movement moves a cluster of the form [XP[NP]], while the *whose picture* type moves [NP[XP]]. The names refer to the two constructions in questions such as *Whose pictures are you looking at?* and relative clauses such as *Mary, your picture of whom/whose picture Tom likes, is very nice*.

Figure 3: Cysouw's feature vectors. (See the text for explanation.)

- Template: < NA-adjacency, N-edge, Dem-edge, frequency >
- Attributes and Values
  - NA-adjacency: Y, N
  - N-edge: Y, N
  - Dem-edge: Y, N
  - frequency: very frequent, frequent, rare, none (VF,F,R,No)
- Vectors

Y	Y	Y	VF
Y	Y	Y	R
Y	Y	N	No
N	Y	N	No
N	Y	N	No
N	Y	N	No
Y	N	Y	F
Y	N	Y	R
Y	N	N	No
Y	N	N	R
N	N	N	No
N	N	N	No
Y	N	Y	F
N	N	Y	R
N	N	N	R
Y	N	N	F
Y	N	N	R
Y	N	N	R
N	Y	N	R
N	Y	N	R
N	Y	N	R
N	Y	Y	R
Y	Y	N	F
Y	Y	Y	VF

Figure 4: Dryer's feature vectors. (See the text for explanation.)

- Template:
  <symmetry1, symmetry2, asymmetry, U18, harmony, frequency>
- Attributes and Values
  - symmetry1: Y, N
  - symmetry2: Y, N
  - asymmetry: Y, N
  - U18: Y, N
  - harmony: Y, N
  - frequency: very frequent, frequent, rare, none (VF,F,R,No)
- Vectors

Y	Y	Y	Y	Y	VF
Y	N	N	Y	Y	R
N	Y	N	Y	Y	No
N	Y	N	Y	Y	No
N	N	N	Y	Y	No
N	N	N	Y	Y	No
Y	Y	Y	Y	N	F
Y	Y	Y	N	N	R
N	Y	N	Y	N	No
Y	Y	Y	N	N	R
N	Y	N	N	N	No
Y	N	N	N	N	No
Y	Y	Y	Y	N	F
Y	N	Y	Y	N	R
N	Y	Y	Y	N	R
Y	Y	Y	Y	N	F
N	Y	Y	N	N	R
Y	Y	Y	N	N	R
N	N	Y	Y	Y	R
N	Y	Y	Y	Y	R
N	N	Y	Y	Y	R
Y	N	Y	Y	Y	R
N	Y	Y	Y	Y	F
Y	Y	Y	Y	Y	VF

and 4 show a four-way discretisation into very frequent (VF), frequent (F), rare (R), and unattested (No). Notice that the fact that we also encode unattested word orders means we explicitly represent negative data.

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 points, 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. 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. These differential frequencies are represented in the training by repeating each example the number of times indicated in Dryer's frequency counts by genera. So, for example, the vector that represents the word order N Dem Adj Num, attested in four genera, is repeated four times. Unattested word orders will be explicitly represented as negative data. (That is, unattested word orders are explicitly represented by one training exemplar.)[4] This set up has many more data points (214 in total) and it could happen that the test set contains examples of word orders that have also been seen at training time.

Figure 5 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 *Models*

Once the data are 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

[4] This is a representational choice that allows us to represent negative data, as is common in supervised learning. Conceptually, this amounts to giving unattested word orders a (negative) observation in the training set. This means that we consider unattested data as data that we have not yet seen and that belong to a qualitatively different frequency class from rare data.

- **Type-based encoding:** each language type as positive or negative piece of data, possible or impossible word order.
- **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 5: Summary of materials and method.

data. Cinque's and Cysouw's methods consist, manually or automatically, in assigning weights to reproduce the observed frequencies of possible and impossible values, with as close a fit as possible.

We will then test the predictive ability of these weighted explanations on data not seen at training time. 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 order 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 with examples and 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

Figure 6: Naive Bayes classifier.

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

$$v = \operatorname*{arg\,max}_{v_j \in V} P(v_j | a_1, a_2 \ldots a_n), \tag{1}$$

$$v = \operatorname*{arg\,max}_{v_j \in V} \frac{P(a_1, a_2 \ldots a_n | v_j) P(v_j)}{P(a_1, a_2 \ldots a_n)} \tag{2}$$

$$= \operatorname*{arg\,max}_{v_j \in V} P(a_1, a_2 \ldots a_n | v_j) P(v_j). \tag{3}$$

Naive Bayes assumption:

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

Naive Bayes classifier: $v_{NB} = \operatorname*{arg\,max}_{v_j \in V} P(v_j) \prod_i P(a_i | v_j)$

training and testing protocol (Russel and Norvig 1995; Webb *et al.* 2005).

The Naive Bayes algorithm is based on Bayes theorem and is defined in Figure 6. 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 6). 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 6).

This method is chosen because it is a simple generative probabilistic model. Its generative probabilistic aspect provides a mathematically well-founded framework to predict frequencies and combine attributes. In a generative probabilistic setting, the typological frequencies are the expression of an underlying generative probabilistic model – the probabilistic independent variables – that give rise to the observed dependent variable – the frequency. The simplicity of the model has two justifications: on the one hand, the simplest models provide the strongest theories, by Occam's razor; on the other hand, a simple model allows a clear interpretation of the outputs and of the results.

In a classification task, we want to predict the class, in our case the frequency of the word order (for example, very frequent, frequent, rare, 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 6). 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. This is a strong assumption that has important theoretical consequences. To verify its validity, then, 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 data is randomly partitioned into n parts, and then the learner is run 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, the 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. We use the widely-used, open-source Weka data mining software.[5] Table 3 shows the results of the experiment, as the proportion of correct answers (percent accuracy).[6] In comparing these numbers, the discussion in the introduction should be borne in mind which indicated that models

[5] http://weka.wikispaces.com/

[6] As usual, accuracy is defined as the number of correctly classified items over the total number of items.

Table 3: Percent (rounded accuracy) of languages or language types classified in the right frequency class. Italics indicate lower than baseline results.

	Naive Bayes					
	Type (24)			Token (214)		
	Two	Four	Seven	Two	Four	Seven
Cinque	88	58	42	97	87	89
Cysouw	*67*	*21*	66	*93*	90	68
Dryer	92	54	63	97	93	71
Baseline	71	50	38	97	47	28

with more parameters have more degrees of freedom and can fit the data better, but at the cost of greater complexity. At comparable performance levels, then, smaller models are usually preferred. By the same reasoning, small models that achieve lower performance than their competitors can often improve results by adding factors. As can be seen by the accuracy results, the models' generalisation is 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 are incorrectly classified. See for example the results on two-way type-based classification of both Cinque and Dryer. Some of the models of type-based classification have performances below or equal to the baseline: the model does not learn. This result illustrates the lesson that models need to be tested on external data; conclusions based on the data used to develop the models are often overly optimistic. Token-based classification yields better results, especially in the four-way classification, with a small number of factors.

5.1 *Analysis of results*

We concentrate now on a more detailed analysis of the models, starting with Cinque's model. The aggregated accuracy results shown in Table 3 can be disaggregated into more informative subcases, by looking at precision and recall by frequency type and by looking at confusion matrices.[7] All the mistakes, as indicated by the results per class and by the confusion matrix, shown in Tables 4 and 5, fall in the *frequent, rare*, and *none* category. Interestingly, most mistakes tend to

[7] As usual, we use the measures of precision and recall. Precision is the number of correctly classified items over the total number of items proposed by the algorithm as belonging to a given class; recall is the number of correctly classified items over the total 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.

	Naive Bayes Results		
	Precision	Recall	F
Very Frequent	91	100	95
Frequent	85	86	85
Rare	91	57	70
None	56	71	62

Table 4:
Percent precision, recall and F measure by frequency class of token-based Naive Bayes classifier for Cinque's model.

	Confusion Matrix			
	Very Frequent	Frequent	Rare	None
Very Frequent	101	0	0	0
Frequent	10	61	0	0
Rare	0	11	20	4
None	0	0	2	5

Table 5:
Confusion Matrix of token-based Naive Bayes classifier for Cinque's model.

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 shows that the attributes associated with frequent events dominate the classification.

The Naive Bayes confusion matrix by frequency class indicates that very frequent orders and unattested word orders are overestimated (Recall > Precision), while frequent and rare word orders are underestimated (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, shown in Table 6, we can see that there is a very strong association among one value of the three merge attributes (first, second, and third) and one class of frequency: *first:AN*, *second:NumA*, and *third:DNum* are indicators of the difference between all three attested frequency classes and the unattested one. The attributes *complete* and *partial* are not as informative about the frequency distinctions.

We can also 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

Table 6: Cinque's joint probability Naive Bayes tables.

		Very Frequent	Frequent	Rare	None
First	AN	0.96	0.95	0.85	0.25
	DN	0.01	0.013	0.025	0.25
	ND	0.01	0.013	0.025	0.17
	NNum	0.01	0.013	0.025	0.17
	NumN	0.01	0.013	0.075	0.17
Second	AD	0.01	0.012	0.24	0.15
	DA	0.01	0.012	0.097	0.23
	DNum	0.01	0.012	0.073	0.23
	NA	0.95	0.93	0.76	0.08
	NumD	0.01	0.012	0.24	0.15
	NumN	0.01	0.012	0.24	0.15
Third	AD	0.01	0.012	0.24	0.23
	AN	0.01	0.012	0.073	0.08
	ANum	0.01	0.012	0.24	0.23
	DNum	0.95	0.93	0.76	0.08
	NumA	0.01	0.012	0.24	0.15
	NumD	0.01	0.012	0.097	0.23
Partial	not	0.43	0.013	0.28	0.64
	np	0.009	0.29	0.38	0.09
	of-who-pp	0.009	0.013	0.20	0.09
	whose-pp	0.55	0.68	0.13	0.18
Complete	not	0.43	0.53	0.46	0.73
	np	0.009	0.16	0.15	0.09
	of-who-pp	0.009	0.29	0.15	0.09
	whose-pp	0.55	0.013	0.23	0.09

Table 7: Probability distributions of feature values by type of movement.

	not	np	of-who-pp	whose-pp
Partial	0.34	0.19	0.08	0.39
Complete	0.54	0.10	0.13	0.22

in Table 7.[8] If we sum up the probabilities and compare all types of movement operations (the last three columns) to no movement, we find that the partial movement operation is twice as probable as no movement, while complete movement is a little less probable than no complete movement. This shows that while no movement is preferred to complete movement, as predicted by Cinque's theory, partial movement is more probable than no partial movement, and also more probable than complete movement. These two results

[8] Recall that movement of the *pictures of who* type is coded as *of-who-pp* and *whose picture* is coded as *whose-pp*.

	Very Frequent		Frequent		Rare		None	
	Y	N	Y	N	Y	N	Y	N
NA-adjacency	0.99	0.01	0.99	0.01	0.40	0.60	0.33	0.67
N-edge	0.99	0.01	0.16	0.84	0.49	0.51	0.55	0.45
Dem-edge	0.99	0.01	0.55	0.45	0.51	0.49	0.11	0.89

Table 8: Cysouw's joint probability Naive Bayes tables.

	Naive Bayes Results		
	Precision	Recall	F
Very Frequent	99	100	99
Frequent	83	100	91
Rare	81	60	69
None	0	0	0

Table 9:
Percent precision, recall and F measure by frequency class of token-based Naive Bayes classifier for Cysouw's model.

are not expected, as complete movement is supposed to be easier than partial movement, 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 the two central columns (Frequent, Rare) of the last two sets of rows (Partial, Complete) in Table 6. Partial movement is not always more frequent and complete movement is not always less frequent. The noticeable differences in distributions indicate that all these distinctions (partial, complete) and their four levels are needed for accurate classification.

The same analysis of results can be applied to Cysouw's model. In Table 8, we can see that *NA-adjacency* distinguishes very frequent and frequent word orders from rare and unattested word orders, but does not distinguish within these two groups; *N-edge* distinguishes all four classes; *Dem-edge* makes a three-way distinction, it distinguishes very frequent, from frequent and rare, from unattested. The most prominent results shown in the disaggregated precision and recall measures by class concerns unattested word orders, as indicated in Table 9. Cysouw's model does not appear to be able to predict this frequency class. The confusion matrix, shown in Table 10, indicates that unattested word orders are confused with rare word orders, but also with frequent word orders. Rare word orders also show several errors, confused with frequent and, in two cases, with very frequent word orders.

Table 10: Confusion Matrix of Naive Bayes classifier for Cysouw's model.

	Very frequent	Frequent	Rare	None
Very frequent	101	0	0	0
Frequent	0	71	0	0
Rare	2	12	21	0
None	0	2	5	0

Table 11: Dryer's joint probability Naive Bayes tables.

	Very Frequent		Frequent		Rare		None	
	Y	N	Y	N	Y	N	Y	N
Symmetry1	0.99	0.01	0.84	0.16	0.62	0.38	0.22	0.78
Symmetry2	0.99	0.01	0.99	0.01	0.54	0.46	0.55	0.45
Asymmetry	0.99	0.01	0.99	0.01	0.92	0.08	0.11	0.89
U18	0.99	0.01	0.99	0.01	0.65	0.35	0.67	0.33
Harmony	0.99	0.01	0.16	0.84	0.49	0.51	0.55	0.45

This model makes fewer mistakes, but appears to have a higher degree of confusion across frequency types than Cinque's model.

The analysis of Dryer's model shows different patterns of distributions and errors from the other two models. If we look at the joint probability distributions associated with the attributes in Dryer's model, shown in Table 11, we can observe that the principle *Symmetry1* discriminates all frequency classes, while neither the principles *Symmetry2*, *Asymmetry* nor *U18* make a clear distinction between very frequent and frequent word orders, and between rare and unattested word orders. The *Harmony* principle, on the other hand, does discriminate among all frequency classes, often in the opposite direction from the principle *Symmetry1*. The most surprising observation that emerges from these probabilites is that frequent word orders are observed to be frequent, despite the fact that they are disharmonic ($P = 0.17$ for the probability of exhibiting the *Harmony* property for frequent word orders, compared to $P = 0.87$ for those not exhibiting this property). Table 12 shows that this model is affected by frequency effects, as shown by the fact that frequent word orders are overestimated (Precision > Recall), while rare word orders are underestimated (Precision < Recall). Table 13 shows that there are twice as many errors confusing more frequent with less frequent word orders than the reverse (11 vs. 5). The table also shows that frequent and rare orders are confused, and that rare and unattested orders are also confused.

These analyses of the errors show that, once tested in a precise learner, the attributes that define these three theories do not always

	Naive Bayes Results		
	Precision	Recall	F
Very Frequent	100	100	100
Frequent	94	84	90
Rare	73	86	79
None	86	86	86

Table 12:
Percent precision, recall, and F measure by frequency class of token-based Naive Bayes classifier for Dryer's model.

	Very frequent	Frequent	Rare	None
Very frequent	101	0	0	0
Frequent	0	61	10	0
Rare	0	4	30	1
None	0	0	1	6

Table 13:
Confusion Matrix of Naive Bayes classifier for Dryer's model.

behave as expected. For example, in Cinque's model, complete movement is less likely than partial movement, while in Dryer's model some of the attributes do not discriminate the typological frequency classes.[9] Also, all the models make mistakes when used predictively. Because the Naive Bayes model is predicated on a strong independence assumption of the attributes, we turn to verifying if this assumption is valid for our data.

5.2 *Validating the independence assumption*

As a control of the independence assumption in the Naive 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 dependencies. 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. For Cinque's and Dryer's models, results are much better, as shown in Table 14. In particular, the classifiers no longer mistake systematically the frequent word orders, as shown in Tables 15, 16, and 17, 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

[9] I thank one of the reviewers for pointing out, correctly, that this result actually means that Dryer's model could have fewer attributes, hence could be made more economical, without loss in predictive power.

Table 14: Percent of languages classified in the right frequency class, for a token-based four-way classification.

	WAODE classifier				Naive Bayes
	Precision	Recall	F	Acc	Acc
Cinque	94	93	93	93	87
Cysouw	87	90	88	90	90
Dryer	96	96	96	96	93

Table 15: Confusion Matrix of WAODE classifier for Cinque's model.

	Very frequent	Frequent	Rare	None
Very frequent	101	0	0	0
Frequent	0	71	0	0
Rare	0	10	23	1
None	0	0	2	5

Table 16: Confusion Matrix of WAODE classifier for Dryer's model.

	Very frequent	Frequent	Rare	None
Very frequent	101	0	0	0
Frequent	0	71	0	0
Rare	0	7	28	0
None	0	0	1	6

Table 17: Confusion Matrix of WAODE classifier for Cysouw's model.

	Very frequent	Frequent	Rare	None
Very frequent	101	0	0	0
Frequent	0	71	0	0
Rare	1	12	21	0
None	0	2	5	0

proposed models. Cysow's model, on the other hand, has the same accuracy (and same confusion matrix) in the two models, which shows that the parameters in this model are indeed independent.

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 different 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 makes mathematically precise the assignment of 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, with sometimes interesting results. For example, we uncover the fact that the properties of the models are interdependent, and hence not theoretically fully independently motivated. Future work lies in developing more accurate models for more complex or more comprehensive problems.

7 REFERENCES

Steven ABNEY (2011), Data-intensive experimental linguistics, *Linguistic Issues in Language Technology (LILT)*, 6(2):1–27.

Guglielmo CINQUE (2005), Deriving Greenberg's Universal 20 and its exceptions, *Linguistic Inquiry*, 36(3):315–332.

Jennifer CULBERTSON and Paul SMOLENSKY (2012), A Bayesian model of biases in artificial language learning: The case of a word-order universal, *Cognitive Science*, 36(8):1468-1498.

Jennifer CULBERTSON, Paul SMOLENSKY, and Geraldine LEGENDRE (2012), Learning biases predict a word order universal, *Cognition*, 122(3):306–329.

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

Michael CYSOUW (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, pp. 29–35, Springer-Verlag, Berlin, Heidelberg.

Matthew DRYER (2006), The order demonstrative, numeral, adjective and noun: an alternative to Cinque, `http://attach.matita.net/caterinamauri/sitovecchio/1898313034_cinqueH09.pdf`. Accessed on 19th August, 2015.

Matthew DRYER (2005), "Genealogical language list", in Haspelmath, Martin and Dryer, Matthew and Gil, David and Comrie, Bernard (eds.), *The World Atlas of Language Structures*, Oxford University Press, Oxford, 584-644.

Matthew S. DRYER (1992), The Greenbergian word order correlations, *Language*, 68:81–138, doi:10.2307/416370.

Michael DUNN, 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.

Richard FUTRELL, Kyle MAHOWALD, and Edward GIBSON Large-scale evidence of dependency length minimization in 37 languages, *Proceedings of the National Academy of Sciences of the United States of America*, 112(33):10336-10341, doi:10.1073/pnas.1502134112.

Joseph H. GREENBERG (1966), *Language universals*, Mouton, The Hague, Paris.

Richard KAYNE (1994), *The antisymmetry of syntax*, MIT Press, Cambridge, MA.

Stuart RUSSEL and Peter NORVIG (1995), *Artificial intelligence: a modern approach*, Prentice Hall Series in Artificial Intelligence, Prentice Hall, Upper Saddle River, NJ.

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

G. I. WEBB, J. BOUGHTON, and Z. WANG (2005), Not so Naive Bayes: aggregating one-dependence estimators, *Machine Learning,* 58(1):5–24.