# Dealing with diversity: Towards an explanation of NP-internal word order frequencies

MICHAEL CYSOUW

*Abstract*

*The world's linguistic diversity is large, probably much larger than many linguists would want to admit. Dealing with this diversity is a central objective for worldwide crosslinguistic investigations. This article argues that to deal with diversity it is extremely fruitful to work with probable structures instead of possible structures, with models instead of theories, and with levels of justification instead of right or wrong. This is illustrated with the order of demonstrative, numeral, adjective, and noun within a complex noun phrase. Different NP-internal orders have strongly differing frequencies among the world's languages. Various models to capture these frequencies are proposed and compared to each other, and it will be argued that very simple models are sufficient. For example, a highly adequate model only refers to the fact that noun and adjective tend to occur together, nouns and demonstratives prefer to occur at the phrase boundary, and noun-adjective order is slightly more frequent than adjective-noun order. The same approach will also be used to model sentence word order frequencies, including areal preferences as random effects. Using such probabilistic models allows for a new take on typological explanations. In and of itself, a probabilistic model is no explanation. However, a well-fitting model instantiates a reformulation of the original phenomenon to be explained into smaller, more tractable phenomena.*

*Keywords:* *crosslinguistic frequency, methodology, noun phrase, syntax, word order*

## 1. Introduction

Some kind of comparison between languages is at the very heart of the field of general linguistics. Or to put it differently, for linguistic insights to be called "general" they have to surpass the details of the analysis of individual lan-

*Linguistic Typology* 14 (2010), 253–286
DOI 10.1515/LITY.2010.010
1430–0532/2010/014-0253

guages and argue for a more widespread relevance. General relevance of linguistic insights can take many different forms. For example, insights can be called general when they help to shed light on problems or results from other fields, like psychology, sociology, or neurology. However, there is also a recurrent conception of general relevance to apply inside the field of linguistics itself. Such general relevance can only mean that the insights are relevant for many, or even all, human languages. Only by actually comparing a wide array of languages is it possible to evaluate the extent of the relevance or generality of a particular insight.

Yet, when considering the world's languages, the extent of linguistic diversity is baffling. One of the central consequences of the acknowledgment of extensive diversity in the field of linguistic typology is the frequent presence of modesty among its practitioners concerning the impact of any generalization proposed. Such modesty can take many forms. One rather unproductive variant of modesty is hedging, for example in the form of long-winding discussions to "explain away" possible counterexamples. As more productive proposals of modesty, I will discuss three notions that appear to be fruitful in dealing with the large diversity of linguistic phenomena attested among the world's languages: probable structures instead of possible structures, model instead of theory, and level of justification instead of right or wrong.

The concrete case to exemplify these notions in the practice of linguistic typology will be the order of elements in a complex noun phrase, including demonstratives (Dem), numerals (Num), and adjectives (A) alongside the head noun (N).[1] The question is how a language arranges these four constituents when they co-occur in a single noun phrase. For example, an English noun phrase like *those four red books* uses the order [Dem-Num-A-N]. Recently, Cinque (2005) proposed an explanation for the crosslinguistic diversity of NP-internal word orders within a movement framework of linguistic structure. In a reply, Dryer (2006) discussed some shortcomings of Cinque's approach and proposes alternative explanations based on purely surface characteristics. In this article, I will compare these approaches – though in my own interpretation, which might not necessarily be exactly the same interpretation as intended by these authors themselves; see the Appendix for the basic data.

1. The crosslinguistic identification of the categories "demonstrative", "numeral", "adjective", and "noun" are far from obvious for many languages, and for the current article I completely rely on the judgments used by Dryer (2006). Further, there are many more NP constituents that one might consider, see, for example, Lahiri & Plank 2008: 45–48 for discussion of the typological variation concerning the order of value, size, and color predicates relative to a noun, like in the English NP *a beautiful big red book* (i.e., VALUE-SIZE-COLOR-N order).

This article starts with a short note on the notion of explanation in linguistics (Section 2), followed by a plea for a greater role of probabilities instead of (im)possibilities when reasoning about language variation (Section 3). Sections 4 and 5 introduce a distinction between the notions "generalization" and "model" using the example of NP-internal word order frequencies. Most results of typological research (like implicational universals or semantic maps) are generalizations that are in need of an explanation. A model is a combination of various (independent) generalizations that conspire to describe typological observations. A suitable model offers the possibility to disentangle complex observations into more limited generalizations. The improvement with regard to explaining typological observations is that only the more limited generalizations will have to be explained, instead of the whole complex observation at once. Another advantage of conceptualizing explanations in the form of models is that different models can be compared with each other quantitatively to find the most suitable one (Section 6). In Section 7 Cinque's model for word order frequencies is opposed to the other models presented in this article, arguing that Cinque's model does a reasonable job, though it does not seem to be an optimal proposal. Section 8 uses the same approach to model frequencies of sentence word order, additionally introducing a method to include areal preferences into the model. Section 9 summarizes the main claims and proposals made in this article.

## 2. Explaining typological frequencies

The kind of approach to the explanation of typological frequencies as proposed by Cinque (2005) is innovative for two reasons, both of which will be followed and extended in this article. First, Cinque does not just attempt to explain why one specific word order type is frequently attested, or unattested, but the frequencies of all theoretical possibilities are explained in one single model.[2] In his article, Cinque characterizes the frequency of each of the 24 possible different NP word order types on a five-point scale ranging from "zero" over "very few", "few", and "many" to "very many". The explanation offered accounts for all 24 empirical frequencies. This approach will be extended in this article by using actual numerical frequencies to be matched for all types.

The second important aspect of Cinque's approach is that he attempts an explanation of the typological frequencies by the cumulative combination of various interacting characteristics. In this way, the difficult issue of explaining an elaborate topic, like the order of complex NPs, can be divided into separate smaller, easier-to-explain characteristics. Cinque's interacting characteristics

2. Compare Hawkins 1990 and subsequent works for another example of trying to account for the full range of possibilities, though using rather different data.

are movement rules in a generative model of syntax. However, there is no necessity to restrict this approach to movement rules, as will be shown in this article.

The principle to explain typological frequencies by a combination of various smaller interacting factors could be dubbed "conspiring motivations" in contrast to "competing motivations" (Du Bois 1985, Jäger 2007). In the explanations put forward in this article, there are various conspiring motivations that work together to produce an adequate model of typological frequencies. Both these "reductionistic" approaches to explanation can be opposed to more holistic structural modeling widespread in modern linguistic (e.g., Harley & Ritter 2002 on person marking; cf. the criticism in Cysouw 2010b).

## 3. Probabilities instead of possibilities

There is a longstanding effort in linguistics to try to distinguish possible structures of human language from impossible ones; cf. Newmeyer 2005 for a recent plea in favor of this approach. However, when dealing with the actual variation as attested among the world's languages, the empirical evidence for distinguishing possible from impossible structures of human languages is meager. There might very well be other methods to provide evidence for distinguishing possible from impossible human language structures. My point here is that investigating the world's linguistic diversity is not a suitable approach to do so.

On the basis of the investigation of linguistic diversity, the closest equivalence to the possible/impossible opposition is a distinction between attested and unattested. However, being unattested clearly does not imply being impossible, and when something is possible this does not mean that it has to be attested. Even more pressing for typological practice, the distinction between attested and unattested appears to be very unreliable as it is highly dependent on the selection of languages investigated. A particular sample of languages might yield an example of a particular structure, but another sample will find it unattested. And even complete sampling of all the world's languages will be of no avail because a particular linguistic structure might not be found among the current world's languages, but might have existed in the past, or might possibly exist in an unknown future.[3]

The current example of NP-internal word order is exemplary for the problematic empirical distinction between attested and unattested. In Figure 1, a plot is shown of the frequencies of all possible NP word orders as reported by

3. The same point is made by Lahiri & Plank (2008: 32). See also Cysouw 2007 for a more extensive discussion of the insignificance of the attested vs. unattested distinction in typology, and Maslova 2000 for an argument that the current world's languages might indeed not represent the full array of possible languages.

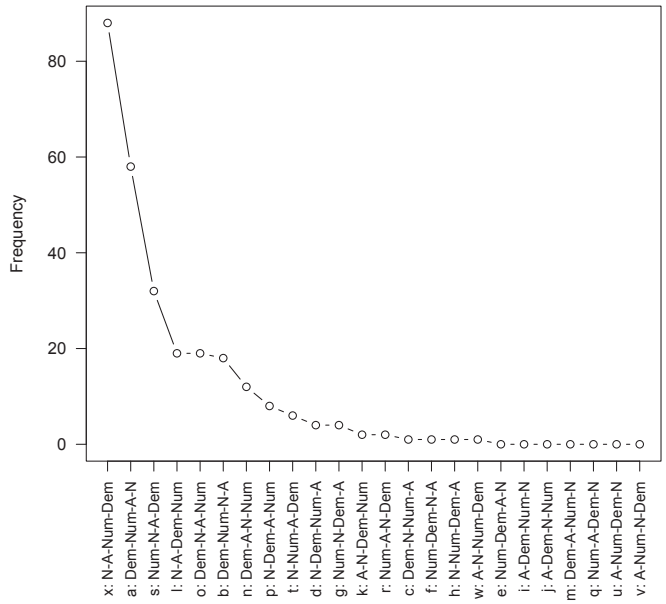


Figure 1. *Distribution of frequencies of NP-internal word order types according to Dryer (sorted left to right by decreasing frequency)*

Dryer (2006). The word orders are shown on the x-axis together with the letter that is used by Cinque to identify the various orders. The different possible types are ordered by decreasing frequency, showing a (negative) exponential distribution. This kind of distribution is commonly found for structural variation among the world's languages (cf. Cysouw 2005, 2010a; Maslova 2008). Such a distribution suggests that the distinction between unattested (the last seven cases) and attested (the rest) is epiphenomenal. Adding more languages might result in one of the unattested orders being attested, but it will not change the form of the distribution. Adding languages will only raise all frequencies wholesale. Likewise, more than one third of the types attested by Dryer (six types, to be precise) are attested in less than one percent of the languages investigated. Many of these types would probably not even be attested in only slightly smaller samples. For example, the word order [Num-N-Dem-A] (called *g* by Cinque) is unattested by Cinque, but found to exist in four languages by Dryer.

To distinguish attested from unattested is thus practically meaningless. Conversely, to distinguish common from rare is highly important. Common types will always be common, even in the most apocryphal samples (provided that

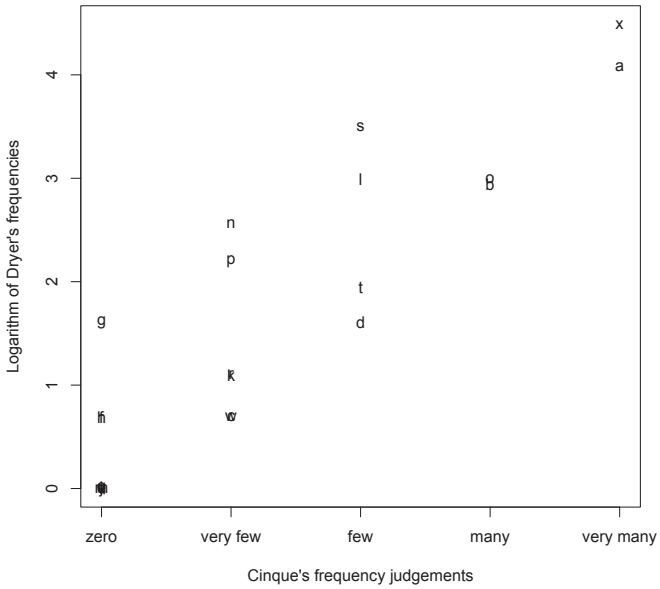


Figure 2. *Correlation between Cinque's assessment of commonality and the logarithm of Dryer's frequencies for the 24 types of NP word order (the letters in the figure are Cinque's codes for the various NP word orders, see Appendix)*

the samples are not deliberately manipulated to avoid a particular common structure). More generally, the approximate probability of each linguistic type can very well be estimated and will show a good between-sample correlation. Normally, different researchers will roughly agree on a type's prevalence, though they might disagree on details. As an example of such inter-sample correlation, consider the strong correlation as shown in Figure 2 between the assessments of prevalence as given by Cinque (2005) and the logarithm of the frequencies as presented by Dryer (2006).[4] Highly frequent language types are judged to be frequent by both, and rare types are considered to be rare by both. There are some slight disagreements. For example, according to Cinque the word order [Num-N-A-Dem] (Cingue's type *s*, attested in "few" cases) is less common than [Dem-Num-N-A] (Cinque's type *b*, attested in "many" cases). However, according to Dryer's counts the situation is reversed as these types are attested in 32 and 18 languages respectively.

4. The fact that there is an almost linear relation between the assessments of Cinque and the LOGARITHM of the frequencies of Dryer, as shown in Figure 2, indicates that typological frequencies should probably better be taken logarithmically. Cf. Cysouw 2010a for an in-depth discussion of this proposal.

The field of linguistic typology would be well advised to shift attention away from distinguishing possible from impossible language types – or attested from unattested types – to discussing estimates of prevalence for each type.[5] Such estimates can of course be very high or very low, suggesting that some types are practically universal or impossible, respectively. However, it is not possibility that we should care for, but probability.

## 4. Probability of generalizations

My plea for abandoning essentialistic notions, like possible vs. impossible, when dealing with linguistic diversity does not stop with structural types. Also any generalization that is made on the basis of the variation attested is preferably not to be interpreted essentialistically – as either right or wrong – but as applicable to a certain degree. As an example, consider the following semi-random selection of (binary) surface characteristics for NP-internal word order:

(i) What is the order of noun and adjective?
(ii) Are noun and adjective always adjacent, or are any other NP elements possibly interfering?
(iii) Is the noun always found at the boundary of the NP (either at the complete beginning or at the complete end), or not?
(iv) Is the demonstrative always found at the boundary of the NP (either at the complete beginning or at the complete end), or not?
(v) Do adjective and demonstrative occur at the same side of the noun (either both before the noun or both after the noun), or not (i.e., the equivalence AN↔DemN)?
(vi) Does the order adjective-noun imply the order demonstrative-noun, or not (i.e., Greenberg's Universal 18: AN→DemN, Greenberg 1963: 86)?
(vii) Does NP-internal word order conform to the hierarchical structure [Dem[Num[A[N]A]Num]Dem], or not (Rijkhoff 2002: 224)?

Table 1 lists the percentage of languages (according to Dryer's counts) for which these characteristics apply. The frequencies of these characteristics almost span the complete spectrum, from nearly universal to approaching a 50/50 distribution. Some of these characteristics are clearly more widespread than others, but overall there is just a continuous cline from higher to lower applicability. The frequencies of occurrence alone do not provide a clear dividing

5. Considering the current state of the typological art, any estimated prevalence of a type is of course only valid for the current world's languages. The most pressing question to be answered by typological research is to determine to which extent the current distribution of a typological variable is influenced by (pre)historical coincidences (cf. Maslova 2000). In this article I will simply ignore this issue, which means that any results obtained only hold for the present state of the world's languages, which might – or might not – be representative of all possible human languages.

Table 1. *Occurrence of selected surface characteristics of NP-internal word order*

| Characteristic | Languages included (%) | Types excluded | Weight |
|---|---|---|---|
| Adjective-noun implies demonstrative-noun order | 98.2 | 4 | 2 |
| Noun and adjective adjacent | 91.3 | 12 | 1 |
| Demonstrative at phrase boundary | 85.9 | 12 | 1 |
| Adjective and demonstrative at same side of noun | 84.1 | 8 | 2 |
| Hierarchical structure | 83.3 | 16 | 4 |
| Noun-adjective order | 72.8 | 12 | 1 |
| Noun at phrase boundary | 66.7 | 12 | 1 |

line between characteristics that are commonly found among human languages (in which case they would probably be called "generalizations" or even "universals") and characteristics that divide human languages into roughly equally sized groups (in which case they would most likely be called "parameters"). Just as there is no obvious division into attested vs. unattested structures but only probabilities, there is also no clear distinction between GENERALIZATIONS ABOUT variation and PARAMETERS OF variation. The only observation attainable is the probability of occurrence of linguistic characteristics.

Notwithstanding the importance of probabilities, frequencies alone do not make a typologist's day (although they are a crucial part of day-to-day business). Arguably, the proportions of occurrence as shown in Table 1 only in part guide the typologist's intuition about which of these characteristics is the most interesting generalization. For example, it is also important to consider how many different types are excluded (cf. the third column of Table 1). Characteristics that only exclude few types might be considered less interesting. For example, the implication AN→DemN captures more than 98 % of the attested languages, but it only excludes 4 orders from the possible 24 orders, namely [A-N-Dem-Num], [Num-A-N-Dem], [A-Num-N-Dem], and [A-N-Num-Dem]. The bidirectional variant AN↔DemN (i.e., adjective and demonstrative always occur at the same side of the noun) is a less accurate generalization, as it captures only 84.1 % of the attested languages. However, it is still a good generalization, because it reaches this 84.1 % while excluding many more types, namely eight. Even stronger, the characteristic "hierarchical structure" captures 83.3 % of the attested languages, but it does so by excluding 16 out of the 24 orders. From this perspective, hierarchical structure is the much stronger generalization than the implication AN→DemN or the equivalence AN↔DemN.

Another argument for the importance of a generalization is something that I will provisionally call "weight" here: the more different linguistic notions are combined into one characteristic, the "heavier" it is (cf. the rightmost column of Table 1). For the purpose of this article, the weight of a characteristic is roughly related to the number of pairwise comparisons needed to establish whether the characteristic is attested or not in a language. This notion of weight as used here is completely pre-theoretical and only has an exemplary character. For the characteristic "noun at phrase boundary" it is necessary to check the position of the noun in the phrase, disregarding any finer grained distinctions within these other NP elements. I count this as a single pairwise comparison, which consequently has a weight of one. In contrast, to assess the characteristic "hierarchical structure" the precise relationships between all pairs of NP elements has to be established, hence there is maximally a weight of six. However, because not always all comparisons are necessary, I give this characteristic a weight of four here.[6] In a sense, "heavier" characteristics are more interesting generalizations as they bring together various (independent) aspects of linguistic structure. So, although the fact that noun and adjective are adjacent is the better generalization judging by pure frequencies alone, hierarchical structure is probably felt by many typologists to be the more interesting generalization, because of its high weight and still fairly high percentage of occurrence. Finding good generalizations then becomes a problem of optimizing both weight and fit.

One consequence of interpreting a generalization as a highly probable characteristic is that we can finally lay off the discussion of counterexamples. There are no counterexamples to a highly probable characteristic, there are just a few languages that do not have the characteristic – as there should be for something being highly probable, but not universal. There is no need to "explain away" languages that do not fit the generalization. Such languages are anyway only unsuitable from the perspective of linguistic theory. For the speakers of such languages, crosslinguistically "exceptional" languages are just as perfectly functional human languages as every other.

## 5. Modeling variation

Whatever kind of theory of language structure one prefers, for linguistic typology it is crucial that the theory helps one to understand why some types are common and other types are rare. Ideally, a theory should be able to predict the

6. My intuitions leading to the specification of these weights are highly debatable, and when challenged I will probably not hold on to the precise numbers as listed in Table 1. More discussion and research is needed to flesh out this notion of "weight" of a typological characteristic.

observed type-frequencies like those in Figure 1 (cf. the Appendix for the frequencies observed by Dryer 2006). The accuracy of such a prediction can then be used as a measure for the relevance of the theory for linguistic typology. For example, consider the presence of hierarchical structure in NP word order. As a GENERALIZATION about NP-internal word order, this is a good one: it makes detailed structural claims (it has a high weight), it excludes many possible orders, and it is valid for a very large proportion of the world's languages. In contrast, as a PREDICTION of the probability of all 24 word order types this generalization is of limited value. On its own, it predicts the existence of just two different kinds of languages, namely those that have hierarchical structure (8 out of 24 NP order types) and those that do not have this structure (the rest). So, this generalization alone predicts that the types with hierarchical structure are common, and that the others are rare. In total, the eight common types account for 83.3 % of the languages sampled. Divided by eight types, this is 10.4 % per type on average, implying a statistical prediction of about 29 languages for each type in Dryer's 276-language sample. However, the observed frequencies of these eight types in Dryer's sample deviate widely from this average (the actual values are 88, 58, 32, 19, 18, 12, 2, and 1, i.e., a standard deviation of 30.1, with two hierarchical types actually being extremely rare, viz. [Num-A-N-Dem] and [A-N-Num-Dem]). Hierarchical structure alone is thus a rather bad predictor for the frequencies of NP orders.

The accuracy of a prediction can be improved by combining various characteristics into a MODEL of linguistic structure. Such a model can be seen as a miniature theory, claiming that only those characteristics included in the model are relevant to explain the world's linguistic variation. The simplest form of such a model consists of an unstructured set of independent characteristics, the cross-section of which predicts the type frequencies. As an example, consider the following model for NP-internal word order, which consists of three characteristics about NP word order. First, hierarchical structure is expected to occur in 83.3 % of the world's languages; second, noun-adjective order is expected to occur in 72.8 % of the languages; and third, the noun is expected to occur at the phrase boundary in 66.7 % of the languages. Roughly speaking, a prediction of this model can be calculated by taking the product of the percentages of the different characteristics (cf. Cysouw 2008). Consider, for example, the order [N-A-Num-Dem]. This order has the following characteristics: hierarchical structure, noun-adjective order, and the noun is at the boundary of the phrase. The predicted number of languages having this order is then the product of the probabilities of the individual characteristics in the model. This amounts to the expected proportion of 40.4 % languages ($0.833 \times 0.728 \times 0.667 = 0.404$). Given that there are 276 languages in Dryer's sample, this proportion corresponds to a prediction of 111.6 languages for the order [N-A-Num-Dem]. This prediction somewhat overestimates the actually attested numbers of 88 lan-

guages in Dryer's count, but the prediction has the right order of magnitude.

The quantitative implementation of such a model can be strongly improved. The problem with the simple quantification just discussed is that the observed frequency of the characteristics is taken as a measure of importance of the characteristic in the model. This is not necessarily the case. Fitting suitable values for the impact of each characteristics is much better performed by GENERALIZED LINEAR MODELING (Baayen 2010: 165ff.).[7] The fitted formula for Dryer's counts of NP-internal word order with the model containing the characteristics "hierarchical structure", "noun at edge", and "noun-adjective order" is shown in (1). The constant at the end of the formula ($-0.7939$) transforms the prediction to the number of languages appropriate for the current sample of 276 languages. This can easily be transformed into a prediction for the fraction of languages by subtracting $log(276) = 5.62$, leading to the model shown in (2). To make this model somewhat easier to interpret, the model can be approximated by rounding the parameters as shown in (3).

(1) Number of languages =
$$e^{2.858 \times Hierarchical + 1.595 \times Noun\text{-}at\text{-}edge + 0.986 \times NA\text{-}order - 0.794}$$

(2) Fraction of languages =
$$e^{2.858 \times Hierarchical + 1.595 \times Noun\text{-}at\text{-}edge + 0.986 \times NA\text{-}order - 6.414}$$

(3) Fraction of languages =
$$e^{3 \times Hierarchical + 1.5 \times Noun\text{-}at\text{-}edge + 1 \times NA\text{-}order - 6.5}$$

The frequency of all NP word orders is predicted by this formula. For example, the order [N-A-Num-Dem] has the following character-settings: hierarchical structure $= 1$, noun-at-edge $= 1$ and NA-order $= 1$. So, this model predicts a fraction of 36.8 % of languages of this type (viz. $e^{3\times1+1.5\times1+1\times1-6.5} = 2.72^{-1} = 0.368$), which amounts to a prediction of 101.6 languages in a 276 language sample. This is still not perfect, though quite close to the observed 88 languages. In Figure 3, the frequencies for all 24 possible word orders as predicted by this model are compared with the observed frequencies as reported by Dryer (2006). Keeping with the exponential distribution of frequencies, Figure 3 shows a comparisons of the logarithms of both the predicted and the observed frequencies.

7. For the computation of generalized linear models in this article I used the function *glm* as provided in the statistical environment R (R Development Core Team 2010). Given that we are dealing with count data, and that the data is expected to be exponentially distributed (Cysouw 2010a), I will use a *glm* with a Poisson error distribution and a logarithmic link function.

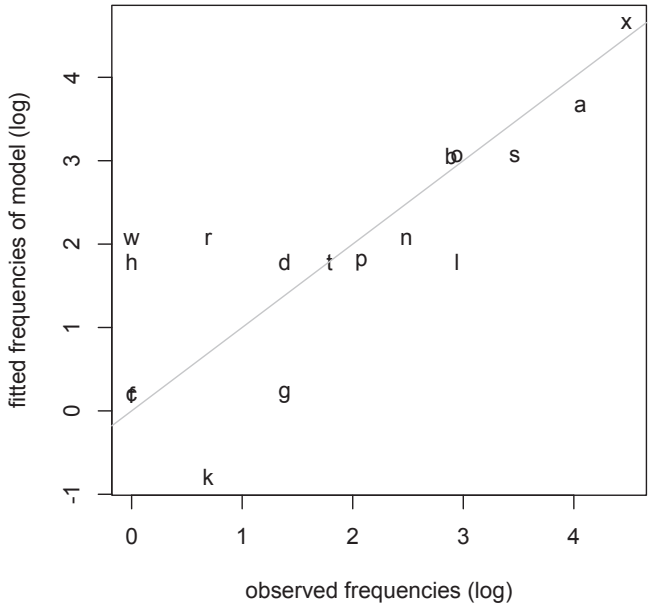


Figure 3. *Double-logarithmic comparison of Dryer's observed frequencies with the predictions by the model in (2) (the letters in the figure are Cinque's codes for the various NP word orders, see Appendix)*

The match between the predicted and the observed frequencies looks promising: this model appears to be a good attempt at predicting the observed distribution of NP-internal word order among the world's languages. It most strongly overestimates the frequency of the following orders:
(i) *w* [A-N-Num-Dem]: observed 1, predicted 7.9;
(ii) *h* [N-Num-Dem-A]: observed 1, predicted 6.0;
(iii) *r* [Num-A-N-Dem]: observed 2, predicted 7.9;
and underestimates the frequency of the following orders:
(i) *k* [A-N-Dem-Num]: observed 2, predicted 0.5;
(ii) *g* [Num-N-Dem-A]: observed 4, predicted 1.2;
(iii) *l* [N-A-Dem-Num]: observed 19, predicted 6.0.
To improve the model, one could try to find characteristics in such errors of the prediction that might be good candidates for inclusion in future models.

Most importantly, such a well-fitted model allows for a new approach to explain typological frequencies. Instead of trying to explain case-by-case why NP word orders have the precise frequencies that are observed, the model as shown in (3) can be used to explain these frequencies. To make the meaning of this

formula more transparent, I have rewritten the formula slightly as shown in (4). Then, because the parameters "hierarchical", "noun at edge", and "NA order" are binary characteristics (i.e., they only can be true or false) this formula can be rewritten as shown in (5), including some rounding.[8]

(4) Fraction of languages =
$e^{-6.414} \times e^{2.858 \times \textit{Hierarchical}} \times e^{1.595 \times \textit{Noun-at-edge}} \times e^{0.986 \times \textit{NA}-\textit{order}}$

(5) Fraction of languages =
$0.0015 \times 17^{\textit{Hierarchical}} \times 5^{\textit{Noun-at-edge}} \times 3^{\textit{NA-order}}$

The model in (5) can be read as follows. To explain NP-internal word order frequencies among the world's languages, three characteristics are important. Language prefer to have their NPs to be organized hierarchically, to have the head noun at the boundary of the phrase, and to have noun-adjective order. Only a tiny fraction of the world's languages (viz. 0.0015, or 1.5 out of 1,000) will not have any of these three critical characteristics. This prediction of the model is fairly accurate, as there are three orders that do not have any of these characteristics, namely [A-Dem-N-Num], [A-N-Dem-Num], and [A-Num-N-Dem], and in Dryer's 276 language-sample only one of these ([A-N-Dem-Num]) is attested, although twice, i.e., on average $2/3 = 0.67$ attestations. So, the empirical fraction of languages without any of the three critical characteristics is $0.67/276 = 0.0024$ (or 2.4 out of 1,000).

All other orders are predicted to occur in frequencies according to the numbers shown in (5). When an order has hierarchical structure, then it will be 17 times as frequent as without; when it has the noun at the edge of the phrase, it will be 5 times as frequent as without; and when it has noun-adjective order, it will be 3 times as frequent as without. For example, the order [N-Num-A-Dem] has no hierarchical structure, the noun is at the edge of the phrase, and it has noun-adjective order. So, the model predicts a fraction of 2.3 % (viz. $0.0015 \times 17^0 \times 5^1 \times 3^1 = 0.023$) of the languages of a sample, which is 6.3 languages in a 276 language sample. Dryer actually observed 6 languages of this type.

The model in (5) predicts NP-internal word order frequencies by appealing to the five claims listed below. Then, instead of explaining the frequencies of all 24 word orders directly, an explanation can now be separated into five separate smaller explanations. This is the general approach using a model for explanation. By finding a suitable model for the data, the explanation can be

8. This last rewrite is not valid in general, but only in this special circumstance in which the parameters exclusively attain the values 1 and 0. Yet, using the form of the model as shown in (5) it is easier to explain how such a model leads to a new kind of explanation of typological frequencies.

broken down into smaller problems, the explanation of which can be tackled independently. For the model in (5) these claims are:

(i) there are three critical characteristics: hierarchical structure, head noun at the edge of the NP, and noun-adjective order;
(ii) the absence of all of these three characteristics will be attested in a fraction of 0.0015 of all languages;
(iii) the presence of hierarchical structure will multiply the number of attested languages by 17;
(iv) the position of the head noun at the edge of the NP will multiply the number of attested languages by 5;
(v) the presence of noun-adjective order will multiply the number of attested languages by 3.

The first claim is actually the most difficult to explain. Why are these three characteristics critical? As will be extensively discussed in the next section, these characteristics are actually not critical. There are many possible models, and there are various others that are at least equally good as this one. So, when some of the critical characteristics turn out to be difficult to explain, we could go and find another well fitting model based on characteristics that we do feel we can explain. There is normally not just one good model out there: there are (very) many possible models that can help to break down difficult questions into smaller, more practicable units.

The other four claims all involve a specific numerical value, obtained by fitting the model to the empirical data. For these values it is not necessary to explain the precise values, e.g., why 0.0015, and why not 0.0010 or 0.0020? The specific values obtained are simply such that they describe the data best, and they will change slightly for different samples. However, what is in need of an explanation is their general tendency. So, for example, why has the absence of all three critical characteristics such a low probability of occurrence? Why is the factor for hierarchical structure so much higher than the factor for noun-adjective order?

As for the low probability of the absence of all three characteristics, this is necessary in any model of NP-internal word order because there are some orders that are practically impossible (i.e., highly improbable), so every model has to start from a prediction close to zero. In contrast, a model for a distribution in which all types are abundantly attested (though not necessarily all equally frequent) will have a higher value for the situation in which all critical characteristics are absent. So, the explanation that is needed for NP-internal word orders is why there is such a strong skewing of their frequencies, including some types that are highly frequent and others that are almost unattested.[9]

9. The explanation for this skewing has to be irrespective of the specific orders that are so highly

As a pointer towards an explanation, I think that strong skewing is an inherent aspect of all typological distributions, a tendency that will be even more pronounced the more types are distinguished (cf. Cysouw 2010a for a detailed argument for this position). Probably a typology with 24 different types will thus always be strongly skewed.

The three values describing the strength of the preference of the three critical characteristics also need an approximate explanation. They are all statistical preferences, so the explanation only has to describe the relative strength of the preferences. First, the order of noun and adjective is slightly skewed towards noun-adjective, which is probably related to the fact that it is communicatively somewhat easier (both in comprehension and in production) to state the modifier after mentioning the thing to be modified. Second, the noun is clearly preferred to be on the edge of the NP, which might be explained by a preference for having all modification at the same side of the noun. Finally, there is a very strong preference for hierarchical structure. I actually find this preference difficult to explain – not because a preference for hierarchical structure is in any sense incomprehensible, but because this characteristic is so complex.[10] As I will argue in the next section, there are other suitable models that do not use this complex characteristic.

## 6. Comparing models

The real beauty of modeling variation is that different models can easily be compared with each other. This is the more important in a field like linguistics, where competing theories notoriously ignore each other. When theories are formulated in the form of predictive models, then their merits can be (quantitatively) compared. As an example, I will here discuss a comparison of various models of NP-internal word order, all based on simple surface word order characteristics, inspired by the discussion in Dryer 2006. In the next section, I will then compare the best examples of these surface structure models with the movement model as proposed by Cinque (2005). As it will turn out, the accuracy of Cinque's movement model plays in the same ballpark as those of the best surface structure models, though Cinque's model is relatively "heavy" in relation to its accuracy.

To be able to compare models, we need a measure to assess the accuracy of a model's prediction. Computational methods, like the generalized linear model

improbable empirically. The actual frequencies will be dealt with by explaining the nature and impact of the three crucial characteristics.

10. However, from the perspective of other theoretical backgrounds hierarchical structure might fit in nicely with available assumptions about the structure of language and would thus be a good candidate for being characteristic in a model (e.g., Rijkhoff 2002, specifically Chapter 7).

Table 2. *Analysis of deviance for the NP-internal word order model in (1)*

| Model | Deviance reduction | Residual deviance | Estimate |
|---|---|---|---|
| Null model | | 602.62 | −0.794 |
| +Hierarchical structure | 293.95 | 308.67 | 2.858 |
| +Noun at edge of phrase | 158.18 | 150.49 | 0.986 |
| +Noun-adjective order | 59.71 | 90.78 | 1.595 |

used here, normally include measures to assess the performance of a model. Basically, a fitted model will have a RESIDUAL DEVIANCE indicating how much variation cannot be explained by the model. The deviance values for the model (1) discussed in the previous section are shown in Table 2. The table starts with the null-model (i.e., no explaining factors are needed, only constants), which has a deviance of 602.62. The reduction of the deviance is shown subsequently for each characteristic added to the model, resulting in a residual deviance of 90.78. The lower this residual, the better the model. To compare this table to the formula in (1), the estimated coefficients of the model are shown in the last column of Table 2. These coefficients are the values used in the formula in (1).

Besides the residual deviance, there is a second aspect of a model that has to be included into the comparison, namely the weight of the model. Models can easily be improved by adding more characteristics. Ultimately, a perfect fit could be achieved by stipulating separate characteristics for every type, like saying that, for example, the order [Dem-Num-A-N] is found in 21 % of the cases and repeating this for all other orders. Such a model might nicely fit to the observations, but it is not very illuminating. Normally, there will be a trade-off between the accuracy of the prediction and the number of characteristics in a model: the fewer characteristics in the model, the worse will be the prediction. In addition, not only the number of characteristics is important, but also their weight, as discussed in Section 4. For example, hierarchical structure is a rather heavy characteristic as it involves a complex interaction between all NP elements. In general, the lower the total weight of a model, the better the model. Optimal models, then, are such models which find a good balance between the total weight of the characteristics and accuracy of the fit. To investigate the balance between weight and fit I will consider models containing only simple word order characteristics:[11]

11. The characteristics included in the modeling are restricted to structural variables. For more profound insights into crosslinguistic frequencies it would be highly desirable to also include genealogical and geographical characteristics. However, such characteristics will make the mathematical methods more involved, so I have decided to leave them out of the present discussion. A first indication of a possible approach is discussed in Section 8.

Table 3. *Analysis of deviance of the best model for NP-internal word order*

| Model | Deviance reduction | Residual deviance | Estimate |
|---|---|---|---|
| Null model | | 602.62 | 0.852 |
| +Noun-adjective order | 59.71 | 542.91 | 1.296 |
| +Noun-demonstrative order | 1.05 | 541.86 | 3.535 |
| +Adjective-numeral order | 15.62 | 526.25 | −0.358 |
| +Adjective-demonstrative order | 12.07 | 514.18 | −2.951 |
| +Numeral-demonstrative order | 0.05 | 514.13 | −1.068 |
| +Noun at edge of phrase | 24.53 | 489.60 | −0.628 |
| +Adjective at edge of phrase | 137.91 | 351.69 | −3.472 |
| +Numeral at edge of phrase | 123.64 | 228.05 | −2.526 |
| +Noun and adjective together | 194.62 | 33.43 | 3.767 |
| +Noun and numeral together | 4.61 | 28.82 | 0.561 |

(i) pairwise order: six characteristics concerning the relative order of noun-adjective, noun-numeral, noun-demonstrative, adjective-numeral, adjective-demonstrative, or numeral-demonstrative;
(ii) location at edge: four characteristics concerning whether the noun, adjective, numeral, or demonstrative occurs at the boundary of the NP;
(iii) uninterrupted occurrence: six characteristics concerning which kinds of constituent occur immediately adjacent to each other: noun+adjective, noun+numeral, noun+demonstrative, adjective+numeral, adjective+demonstrative, or numeral+demonstrative.

Including all these characteristic in one model results in a residual deviance of 16.97. One heuristic to find suitable models with less characteristics is (roughly speaking) to subsequently remove characteristics that do not seem to improve explanatory accuracy.[12] The result is a model as shown in Table 3, including 10 out of the 16 characteristics considered, with a residual deviance of 28.82. This is somewhat worse than the complete model, but still highly accurate. The comparison between the frequencies predicted by this model and the attested frequencies is shown in Figure 4. All the highly frequent types are predicted extremely accurately. There are only some slight misses for low frequent types, like *r* [Num-A-N-Dem] (predicted 5.0, attested 2) and *g* [Num-N-Dem-A] (predicted 0.4, attested 4).

Thus, the model as shown in Table 3 highly accurately predicts the attested frequencies, but it is rather "heavy". Very many characteristics are needed to reach such a good prediction, and the model is too complex to really provide

12. I used the function *step* as available in R (R Development Core Team 2010) for this heuristic search. I thank Balthasar Bickel for pointing out this possibility.

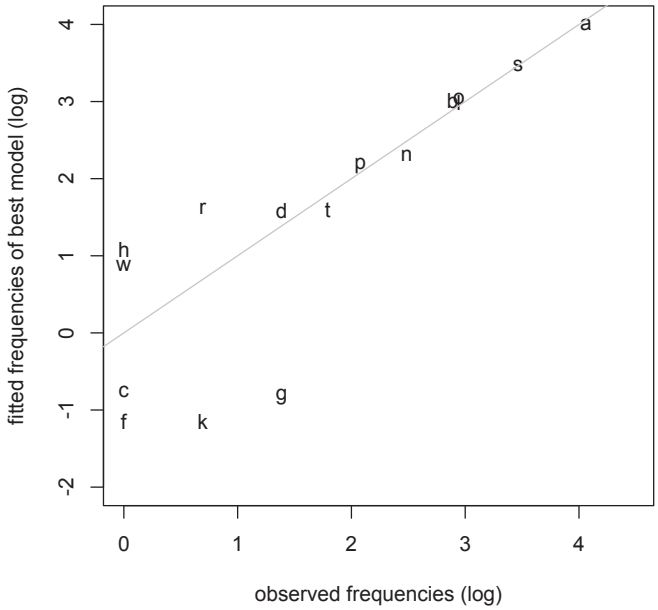


Figure 4. *Double-logarithmic comparison of Dryer's observed frequencies with the predictions by the model in Table 3*

new insights (other than that it is possible to make highly accurate predictions with only simple ordering characteristics). For the further development of linguistic theory it would be more interesting to have a slightly less accurate model that is more readily interpretable. One approach to find such a leaner model is to select the characteristics with the highest deviance reduction. Such a selection is shown in Table 4, taking only four of the characteristics with strong deviance reduction from Table 3. The residual deviance is clearly higher (89.23), but the model is much easier to interpret.

Note that two of the four characteristics in this model have a negative modifier meaning that these characteristics are dispreferred (this can be inferred from the negative estimates in the last column of Table 4). NP-internal word order thus disprefers to have the numeral and the adjective at the edge. This can be reformulated by saying that NP word order prefers noun and demonstrative to be at the edge of the phrase. This reformulated model is shown in Table 5, resulting in a roughly identical residual deviance.

The model in Table 5 results in the formula in (6), using the procedure as explained in the previous section. Comparing this formula with the one discussed previously in (5), we find great similarity. To be precise, the complex

Table 4. *Simple model of NP-internal word order, version 1*

| Model | Deviance reduction | Residual deviance | Estimate |
|---|---|---|---|
| Null model | | 602.62 | 1.339 |
| +Noun and adjective together | 219.53 | 383.09 | 2.271 |
| −Numeral at edge of phrase | 91.22 | 291.87 | −1.503 |
| −Adjective at edge of phrase | 142.93 | 148.94 | −1.889 |
| +Noun-adjective order | 59.71 | 89.23 | 0.986 |

Table 5. *Simple model of NP-internal word order, version 2*

| Model | Deviance reduction | Residual deviance | Estimate |
|---|---|---|---|
| Null model | | 602.62 | −1.984 |
| +Noun and adjective together | 219.53 | 383.09 | 2.335 |
| +Demonstrative at edge of phrase | 70.42 | 312.66 | 1.674 |
| +Noun at edge of phrase | 160.52 | 152.14 | 1.596 |
| +Noun-adjective order | 59.71 | 92.43 | 0.986 |

characteristic "hierarchical structure" is broken down into the much simpler characteristics: noun-adjective occurring together, demonstrative at the edge of the phrase, and noun at the edge. Together theses preferences provide a highly accurate model of the crosslinguistic frequencies of NP-internal word order, as illustrated in Figure 5.

(6) Fraction of languages =
$0.0005 \times 10^{NA\text{-}together} \times 5^{Noun\text{-}at\text{-}edge} \times 5^{Dem\text{-}at\text{-}edge} \times 3^{NA\text{-}order}$

The most important difference between this model and the model in (5) is that the complex characteristic "hierarchical structure" is not necessary to model NP-internal word order frequencies. This characteristic is a very interesting generalization about the worldwide variation of linguistic structure, but it is also difficult to explain. The more basic characteristics in the model in (6) seem much less complicated to explain. Further, the model in (6) better predicts the frequency of order *l* [N-A-Dem-Num], which is attested in 19 languages on Dryer's counts. The model with hierarchical structure in (5) predicts 6.6 languages for this order, but the model in (6) much more accurately predicts 20.7 languages. The model in (6) is thus "lighter" and easier to explain, and slightly more accurate in its prediction.

## 7. Cinque's movement model

Cinque (2005) also proposes a model for the worldwide variation of NP-internal word order, and it is now possible to compare the merits of his model to the

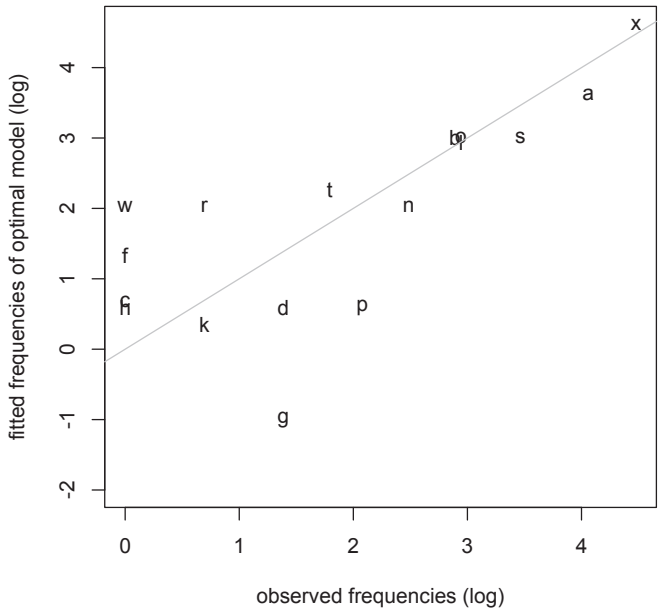


Figure 5. *Double-logarithmic comparison of Dryer's observed frequencies with the predictions by the model in (6)*

surface structure models as discussed in this article. His model is based on a basic order [Dem-Num-A-N] with various kinds of movement that can be used by a language to derive a particular surface structure. It is important to realize that the characteristics used by Cinque to model NP order frequencies are (mostly) kinds of movement that have been discussed before in the generative literature, and can thus be seen to be independently motivated. I will here reformulate Cinque's model using Dryer's frequencies to assess its accuracy (see Appendix for details).

In Cinque's model, the underlying order [Dem-Num-A-N] is proposed for the languages that show exactly this surface order. This same base order is also proposed for languages where the surface order can be derived from this order by using any combination of the available movement rules (including multiple usage of the same movement). Cinque chose the movement rules in such a way that the types which are unattested (in his sample of languages) coincide with the types that cannot be derived from the underlying [Dem-Num-A-N] order. For Cinque, this implies that [Dem-Num-A-N] basic order is universal for human language. However, Dryer (2006) reports on six languages that have word orders deemed to be impossible by Cinque, most importantly, four languages

Table 6. *Cinque's movement model for NP-internal word order*

| Characteristic | Occurrence (%) |
|---|---|
| Underlying [Dem-Num-A-N] base order | 97.8 |
| Movement with pied-piping of the "whose picture"-type | 67.0 |
| Movement without pied-piping | 11.6 |
| Movement with pied-piping of the "picture of who"-type | 19.9 |
| Partial movement | 30.4 |
| Extraction of the noun around the demonstrative | 2.9 |

with [Num-N-Dem-A]. Keeping with the "improbable instead of impossible" motto discussed in Section 3, the underlying base order [Dem-Num-A-N] has a probability of occurrence of $270/276 = 97.8\,\%$, which is of course still practically universal.

In his article, Cinque explicitly discusses three different kinds of movement, and the possibility that the movement is only partial (i.e., the noun does not end up all the way up in the tree after the end of movement). The first kind of movement – movement with pied-piping of the "whose picture"-type – is considered to be unmarked by Cinque. This makes sense, considering that this kind of movement is used in 67 % of the languages (using Dryer's counts), and often even several times per language. The other kinds of movement are all deemed marked by Cinque. And indeed, these kinds of movement are all used by much fewer languages (see Table 6), though note that the bandwidth of the marked proportions of occurrence is rather large (roughly between 12 and 30 percent). Finally, Cinque uses one extra movement rule, somewhat hidden in the list of derivations (2005: 323), namely extraction of the noun around the demonstrative. This movement rule is only used to derive the order [N-Dem-A-Num] and deemed to be extremely rare by Cinque (the eight languages with this order in Dryer's count amount to 2.9 %).

Reformulated in this way, Cinque's movement model can be used to make a prediction of the frequency for each NP-internal word order type in the same way as explained in Section 5. The formula is shown in (7). The impact of the base order is extremely high in Cinque's model. Also note that the various marked kinds of movement get a factor between zero and one, indicating that their presence lowers the predicted frequency. The predictions from this model are compared to Dryer's observed frequencies in Figure 6. Cinque's model shows a good match, comparable to the models shown in Figures 3 and 5. The residual deviance of Cinque's models is 121.20, slightly higher compared to the residuals of about 90 for the surface order models from the previous section. Further, Cinque's model is also clearly "heavier". If every movement rule would be given a weight of 1, and the underlying base order a weight of 3

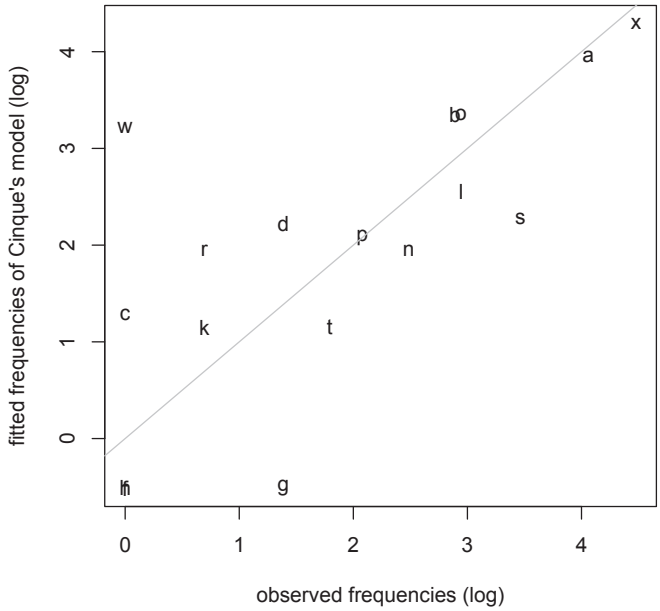


Figure 6. *Double-logarithmic comparison of Dryer's observed frequencies with the predictions by Cinque's movement model*

(comparable to the weight of hierarchical structure in Table 1), then Cinque's movement model has a total weight of 8.

(7) Fraction of languages =
$0.0022 \times 87^{Base} \times 1.4^{Whose} \times 0.2^{No\text{-}pp}$
$\times 0.3^{Of\text{-}who} \times 0.4^{Partial} \times 0.1^{Extract}$

To put the various models in perspective, I have randomly sampled a few hundred different models using the simple surface characteristics described in the previous section. The residual deviance for all these models is shown in Figure 7 as a series of box plots for the different weights of these models. The models are grouped by weight (shown on the *x*-axis) and within each weight the variation in residual deviance is shown as a boxplot. As can be clearly seen in this figure, the residual deviance becomes less the more characteristics are included in the model. The four models discussed in the article are indicated in the figure by lowercase letters. First, the model in (5) with hierarchical structure is indicated by an "a" in the figure. It has a really low residual deviance, and it is not too heavy, so this seems to be a very suitable model (given that it is possible to explain the strong crosslinguistic preferences for hierarchical structure). An

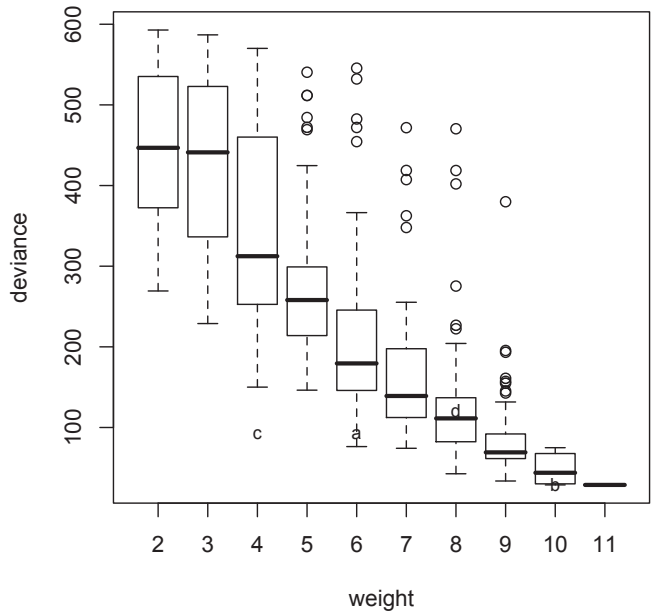


Figure 7. *Box plots of residual deviance for randomly sampled models using only simple surface characteristics (the letters indicate the models discussed in this article)*

even better model was discussed in Table 3, though this improvement came at the price of a much heavier model, indicated by "b" in Figure 7. Although this is the best model possible with the current characteristics, it is not the optimal model. An optimal model is the model in (6), which both has a low residual deviance and a low weight, shown as "c". For comparison, Cinque's model is shown as "d", indicating that it is neither very accurate nor very "light" in comparison to the surface structure models discussed here. Even when my judgment of the weight of Cinque's model is contested, the absolute residual deviance of Cinque's model is still higher than the models in (5) and (6).

## 8. Modeling sentence word order

The method proposed in this article is a general method, and in order to strengthen the point of its generality I will discuss in this section how the method can approach the explanation of the typology of basic sentence word order. Ever since Greenberg's seminal article on word order (1963) it is known that the six possible word order types of subject, object, and verb are not equally frequent among the world's languages. Greenberg's Universal 1 famously says that "in

Table 7. *Word order frequencies according to Dryer 2005*

| Word order | Number of languages | Fractions of languages (%) | Genus-corrected fractions (%) | Area-corrected fractions (%) |
|---|---|---|---|---|
| SOV | 497 | 47.1 | 58.5 | 52.9 |
| SVO | 435 | 41.2 | 28.5 | 33.2 |
| VSO | 85 | 8.0 | 8.1 | 9.2 |
| VOS | 26 | 2.5 | 3.1 | 3.0 |
| OVS | 9 | 0.9 | 1.4 | 1.3 |
| OSV | 4 | 0.4 | 0.4 | 0.5 |

declarative sentences with nominal subject and object, the dominant order is almost always one in which the subject precedes the object" (Greenberg 1963: 61). Using the counts of Dryer (2005b), this universal holds for more than 96 % of the world's languages, and it is thus a very good generalization.[13] However, interpreted as a simple model of variation, Greenberg's universal is of limited value, because it only predicts that the three SO orders (SOV, SVO, VSO) would be all frequent, while the three OS orders (VOS, OVS, OSV) are all rare. Although this prediction is roughly true, it is not sufficient to model the more fine-grained differences in frequency of the six word order types. More complex models are necessary if we want to get closer to an explanation of the actual frequencies attested (shown here in Table 7). In keeping with the proposal as developed in this article, the strategy to find an explanation will be to first look for a suitable model matching the attested frequencies. Ideally, such a suitable model consists of various simpler generalizations that are easier to explain. In this way, the explanation of a complex phenomenon can be tackled by individually explaining the simpler characteristics in the model.

To find suitable models of sentence word order frequency I considered the following characteristics and searched for the optimal model to match the language frequencies as listed in Table 7:

(i) pairwise order: characteristics concerning the relative order of constituents, i.e., whether the order is SO or OS, VO or OV, SV or VS (see Hawkins 2001, Cysouw 2008);

(ii) pairwise adjacency: characteristics concerning the direct adjacency of constituents, i.e., whether S and O are adjacent or not, whether S and V are adjacent or not, whether O and V are adjacent or not (see Ferrer i Cancho 2008);

13. For this section, I ignored the languages that are classified by Dryer as having no dominant word order.

(iii) individual position: characteristics concerning the sentence position of individual constituents, i.e., whether S is first, medial, or final, V is first, medial, or final, O is first, medial, or final.

A search through all possible models (using the approach described in Section 6) suggests that the optimal model with these characteristics consists of a model including subject-object order, object-verb order, and the filling of the first position of the sentence. Specifically, the estimated parameters are shown in (8). For example, SVO languages have SO order and S-first, but no OV order nor V-first. Using the formula in (8), this model correctly predicts 435.2 languages ($= 5.69 \times 3.27^{1} \times 1.14^{0} \times 23.39^{1} \times 4.57^{0}$) with SVO, exactly identical to the number of 435 observed languages from Table 7.

(8) Number of languages =
$5.69 \times 3.27^{SO\text{-}order} \times 1.14^{OV\text{-}order} \times 23.39^{S\text{-}first} \times 4.57^{V\text{-}first}$

As expected from typological intuition, SO order and OV order are important characteristics to model sentence word order. However, a somewhat unexpected result from the formula in (8) is that these characteristics are less important than the filling of the first position in the sentence. Having the subject first in the sentence makes a word order 23.39 times more probable than not having subject first in the sentence. Having the verb in first position makes a word order 4.57 times more probable than not having the verb first. Both these factors have a stronger impact than having SO order, which makes a word order "only" 3.27 times more probable than having OS order. Having OV order just barely has any impact at all. This result can be understood by realizing that the most common orders are S-first (SOV and SVO), followed by V-first (VSO and VOS), with O-first being the least common (OVS and OSV). Within this basic tripartite division, the difference between SO and OS order is only necessary to distinguish VSO from VOS, and the difference between VO and VO is only necessary to distinguish SOV from SVO.[14]

Thus, to explain the typological distribution of word order, the first and foremost characteristic to tackle is the preference for the subject being first in the sentence. Given the default topical prominence of subject this is a rather obvious typological preference. The preference for the verb being first in the sentence can be seen as part of a hierarchy of possible first sentence constituents: the subject has the highest probability to end up as the default first element, the verb is second, and the object is last. Next, the preference for SO order is easily explainable from an information structure point of view (similar to the preference for S-first). Subject and object, being the prototypical theme and rheme, preferably occur in theme-rheme order. The preference for OV order is

14. The model in (8) does not makes any prediction about different frequencies for OVS and OSV.

very small. This factor is mainly necessary to distinguish the slight preference for SOV over SVO among the world's languages. Given the small effect of this characteristic, it is worthwhile to look a bit further into the numbers as reported by Dryer to see whether this difference really is significant.

The frequencies of word orders as reported by Dryer and shown in Table 7 are raw numbers of languages. Although his sample of languages is not purposely biased, there are many closely related languages in his extremely large and impressive sample. To evaluate how strong the influence of these genealogical affiliations is on the worldwide frequencies, I computed genus-corrected frequencies for the word order types, using the division of languages into genera as described in Dryer 2005a. The idea of this genus-correction is that each genus should count equally in the establishment of worldwide frequencies. When more than one language from a genus is available, the impact of these languages is weighted down in correspondence to the number of languages available for that genus. For example, Dryer includes eight Tupi-Guaraní languages in his sample, four of which are SOV, two are SVO, one is VSO, and one is OVS (viz. Asuriní). In the correction, this genus counts for half SOV (4 languages out of 8), a quarter SVO (2 out of 8), an eighth VSO (1 out of 8), and an eighth OVS (1 out of 8).[15] The worldwide average of all these genus-corrected fractions is shown in the last column of Table 7. These fractions closely match the raw numbers of languages, except for an even stronger predominance of SOV over SVO.[16] Performing the same calculations as above with these genus-corrected frequencies results in basically the same optimal model, except for slight changes in the estimates coefficients, as shown in (9). Using the formula in (9), this model correctly predicts the genus-corrected fractions, e.g., a fraction of 28.5 % SVO languages ($0.285 = 0.0045 \times 2.59^{1} \times 2.05^{0} \times 24.44^{1} \times 6.92^{0}$).

(9) Fraction of languages =
$0.0045 \times 2.59^{\textit{SO-order}} \times 2.05^{\textit{OV-order}} \times 24.44^{\textit{S-first}} \times 6.92^{\textit{V-first}}$

Word orders are not equally distributed geographically among the world's languages. The distribution of word orders according to macro-areas is shown

15. Ideally, such an approach should be balancing word orders throughout the family tree (see Bickel 2008). However, given the currently incomplete knowledge of the structure of most family trees for the world's languages, I opted for the simpler approach used here.
16. Note that the expected negative exponential distribution of unordered typological parameters, mentioned in Section 3, seems to be disproved by the language counts in which SOV and SVO are almost equally frequent. However, the genus-corrected fractions again show the expected negative exponential distribution (cf. Cysouw 2010a).

Table 8. *Areal distribution of word order types (according to Dryer 2005)*

| Word order | Africa | Eurasia | South-East Asia & Oceania | Australia & New Guinea | North America | South America |
|---|---|---|---|---|---|---|
| SOV | 54 | 105 | 76 | 148 | 55 | 59 |
| SVO | 225 | 30 | 112 | 32 | 15 | 21 |
| VSO | 22 | 4 | 20 | 1 | 32 | 6 |
| VOS | 0 | 0 | 11 | 3 | 8 | 4 |
| OVS | 1 | 0 | 0 | 2 | 0 | 6 |
| OSV | 0 | 0 | 1 | 1 | 0 | 2 |

in Table 8.[17] These six macro-areas can be seen as random factors in a so-called MIXED MODEL (Baayen 2008: 241ff.). A mixed model is a further development of the kind of models used up till now in this article. In mixed models the influence of additional "random" factors is included in the model. The basic idea is that observed frequencies are both influenced by the main "fixed" factors and additional "random" factors, and that it is necessary to remove the influence of the random factors to establish the true impact of the fixed factors. In the current case the fixed factors are the various word order parameters (i.e., first position, SO order, OV order) and the random factor is the distribution of the orders in the various macro-areas. The question for typology is (see Dryer 1989, 1992): How strong is the influence of the word order parameters independent of the "random" (coincidental) distribution within the six macro-areas? The resulting coefficients of the fixed factors are shown in (10). Again, the coefficients are roughly in the same order of magnitude as in (8) and (9), arguing that the (large) variation between the macro-areas is indeed just a random effect relative to worldwide variation. Note the addition of the factor "AREA" at the end of this formula, which will contain the area-specific coefficients.[18]

(10) Number of languages =
$0.33 \times 3.49^{SO\text{-}order} \times 1.73^{OV\text{-}order} \times 39.07^{S\text{-}first} \times 7.81^{V\text{-}first} \times \text{AREA}$

The area-specific coefficients (i.e., the random factors) are shown in Table 9. To obtain the frequencies for each specific area, the main coefficient from (10) has to be combined with these area-specific coefficients. For example, the area-specific coefficients for Africa are shown in (11). Combining

17. An additional correction of the frequencies by genera (as above) does hardly change these differences, except for SVO in Africa, which goes down from 75 % of all African languages in Table 8 to 60 % when all African genera are counted equally, basically because of an overrepresentation of Bantu languages in Dryer's data.
18. For the calculation of this mixed model I used the function *glmer* ("generalized linearly mixed model") in the R-package *lme*4 (Bates & Maechler 2010).

Table 9. *Area-specific coefficients (i.e., random effects)*

| | Africa | Eurasia | South-East Asia & Oceania | Australia & New Guinea | North America | South America |
|---|---|---|---|---|---|---|
| Intercept | 5.01 | 0.26 | 2.16 | 0.91 | 0.19 | 3.59 |
| SO order | 8.05 | 1.34 | 0.57 | 0.17 | 1.11 | 0.52 |
| OV order | 0.14 | 1.95 | 0.38 | 2.63 | 1.89 | 1.71 |
| S-first | 0.12 | 2.01 | 2.01 | 4.72 | 1.79 | 0.26 |
| V-first | 0.06 | 1.3 | 1.76 | 1.11 | 16.46 | 0.42 |

the main coefficients with the area-specific coefficients for Africa results in a formula for the number of languages for the different word orders in Africa, as shown in (12).[19] For example, the number of VSO languages observed in Africa is 22, and the formula in (12) accurately predicts 21.7 languages ($= 1.64 \times 28.10^{1} \times 0.25^{0} \times 4.85^{0} \times 0.47^{1}$).

(11) Area-specific coefficients for Africa =
$5.01 \times 8.05^{SO\text{-}order} \times 0.14^{OV\text{-}order} \times 0.12^{S\text{-}first} \times 0.06^{V\text{-}first}$

(12) Number of languages in Africa =
$1.64 \times 28.10^{SO\text{-}order} \times 0.25^{OV\text{-}order} \times 4.85^{S\text{-}first} \times 0.47^{V\text{-}first}$

The same calculation for the frequencies of word order types in North American languages results in the formula shown in (13). For example, the number of VSO languages observed in North America is 32, and the formula predicts 30.26 ($= 0.061 \times 3.86^{1} \times 3.26^{0} \times 70.10^{0} \times 128.5^{1}$). Comparing the formulas for Africa (12) and North America (13) it is remarkable how strongly different the coefficients are, which is of course necessary to model the highly diverging frequencies observed (as shown in Table 8).

(13) Number of languages in North America =
$0.061 \times 3.86^{SO\text{-}order} \times 3.26^{OV\text{-}order} \times 70.10^{S\text{-}first} \times 128.5^{V\text{-}first}$

19. To obtain (12), the coefficients of (10) and (11) are simply multiplied. This only works because the characteristics are binary, i.e., they only can be present (i.e, 1) or absent (i.e., 0), see Footnote 8. This calculation does not work in different circumstances, and should thus only be taken as illustrative here.

## 9. Summary

Abstracting away from the concrete aspects of the examples discussed in this article, I have argued for the following approach to deal with diversity:

(i) Linguistic types are not possible or impossible, neither attested nor unattested; they have a particular probability of occurrence.
(ii) Generalizations about linguistic types are not right or wrong, they only have a level of justification ("fit").
(iii) Because generalizations are probabilistic, they do not have counterexamples.
(iv) Generalizations are more interesting the MORE possibilities are excluded and the more different linguistic notions are combined into a meaningful bond ("weight").
(v) Good generalizations should both have a high weight and a good fit; however, these two desiderata will mostly counteract each other, so a suitable balance should be searched for.

In addition to generalizations, I have argued for the usage of models for the explanation of typological observations:

(i) Various (independent) generalizations together form a model of linguistic structure.
(ii) Different models of linguistic structure can be compared as to its accuracy of predicting the observed frequencies of linguistic types ("fit of model").
(iii) A model is more interesting the FEWER generalizations are included ("weight of model").
(iv) Good models both have a low weight and a good fit; again, these desiderata normally counteract each other, so a balance is needed.
(v) Specific shortcomings of a model can be observed by inspecting the predictions of a model and comparing these predictions with the observed frequencies; on this base, new models can be developed to improve the prediction.
(vi) The choice of characteristics to be included in a model completely depends on what kind of characteristics one finds easiest to explain.
(vii) A good model offers a new approach to the explanation of typological frequencies by dividing up the problem into separate smaller problems that have to be explained independently.

Concerning the concrete example discussed in this article, the crosslinguistic frequencies of NP-internal word order types, the characteristic "hierarchical structure" is a very good generalization, combining a good fit with a high weight. However, this characteristic is not necessary to model NP word order frequencies. A good model was found using only the following basic characteristics:

(i) noun and adjective have a strong tendency to occur immediately adjacent to each other (a preference with a factor 10 to 1);
(ii) head nouns tend to occur at the boundary of the NP (a preference with a factor 5 to 1);
(iii) demonstratives also tend to occur at the boundary of the NP (likewise with a preference of 5 to 1);
(iv) adjectives tend to follow nouns (a preference with a factor 3 to 1).

Thus, to explain the complete range of typological frequencies as summarized in Figure 1 it is now "only" necessary to explain these four characteristics, which is of course still far from trivial. As for an explanation, it is important to realize that these characteristics are all statistical preferences, so the explanation only has to explain the relative strength of the preferences. First, the strongest preference is the fact that noun and adjective tend to occur adjacent to each other (see also Hawkins 2001: 16). In a sense, this is a linguistic truism, as evidenced by the fact that most syntactic analyses group noun and adjective as immediate sisters in their trees. From this perspective it is almost more interesting to reverse the characteristic and ask for an explanation why in about one of ten languages the noun and the adjective do NOT necessarily occur adjacent to each other. I currently do not have an explanation for this statistical preference. Second, the noun is clearly preferred to be on the edge of the NP, which might be explained by a preference for having all modification at the same side of the noun. Third, the demonstrative is also clearly preferred to be on the edge of the phrase, which might be related to its function as determiner. Finally, the order of noun and adjective is slightly skewed towards noun-adjective, which is probably related to the fact that it is communicatively somewhat easier (both in comprehension and in production) to state the modifier after mentioning the thing to be modified.

In a similar fashion, for the sentence word order of subject, object, and verb the following main effects were found (abstracting away from the random effects of macro-areas):

(i) the first element of the sentence is preferably the subject, probably a verb, but only rarely an object (a preference with factors 40 to 8 to 1, respectively);
(ii) subject-object order is preferred over object-subject order (a preference with a factor 3.5 to 1);
(iii) object-verb order is preferred over verb-object order (a preference with a factor of less than 2 to 1).

The most striking aspect of this result is the massive importance of the filling of the first position in the sentence. The preference for subject-first is of course not remarkable at all from the perspective of the information structure of a sentence. The two remaining preferences only act as minor typological adjustments. The preference for subject-object order over object-subject order

has been extensively discussed ever since Greenberg's (1963: 61) original observation as formulated in his Universal 1, quoted above. Interestingly, after taking into account the typological preferences for first position, as has been done in the current model, the remaining subject-object order preference is not very strong anymore. This reduction of the importance of the subject-object order preference can be elucidated by realizing that all subject-initial languages necessarily have subject-object order. Finally, there is a slight preference for object-verb order in comparison to verb-object. However, this factor is so close to one that it is questionable whether it is really significant. The precise interpretation of significances, which are also presented in the implementations of generalized linear models and mixed models used here, is a central question for further research.

*Received: 14 September 2009*
*Revised: 16 August 2010*

*Ludwig-Maximilians-Universität München*

*Correspondence address:* Ludwig-Maximilians-Universität München, Fakultät für Sprach- und Literaturwissenschaften, Research Unit "Quantitative Language Comparison", Geschwister-Scholl-Platz 1, 80539 München, Germany; e-mail: cysouw@lmu.de

*Acknowledgements:* This article arose originally as a reaction to a presentation by Matthew Dryer at the MPI for Evolutionary Anthropology in 2006. I thank Matthew Dryer for his consent to use his fascinating data collection on NP-internal word orders. Three anonymous referees for *Linguistic Typology* added various helpful comments and suggestions to improve the presentation of the concepts in this article.

## Appendix

| Cinque's codes | NP order | Dryer's frequencies | Hierarchical structure | Adjective-noun implies demonstrative-noun order | Noun and adjective at same side of noun | Noun-adjective order | Noun and adjective adjacent | Demonstrative at phrase boundary | Noun at phrase boundary | Prediction Figure 3 | Prediction Figure 5 | Cinque's frequency judgements | [Dem-Num-A-N] Base structure | Movement with pied piping of the 'whose picture"-type | Movement without pied piping | Movement with pied piping of the 'picture of who"-type | Partial movement | Extraction of N around Dem | Prediction Figure 6 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| a | Dem-Num-A-N | 58 | + | + | + | | + | + | + | 38.9 | 37.3 | very many | + | | | | | | 51.8 |
| b | Dem-Num-N-A | 18 | + | + | | + | + | + | | 21.1 | 20.3 | many | + | + | | | + | | 23.0 |
| c | Dem-N-Num-A | 1 | | + | | + | | + | | 1.2 | 2.0 | very few | + | | + | | + | | 3.0 |
| d | N-Dem-Num-A | 4 | | + | + | + | | | + | 6.0 | 1.8 | few | + | | + | | | | 6.8 |
| e | Num-Dem-A-N | 0 | | + | + | | + | | + | 2.2 | 7.0 | zero | | | | | | | 0.1 |
| f | Num-Dem-N-A | 1 | | + | | + | + | | | 1.2 | 3.8 | zero | | | | | | | 0.1 |
| g | Num-N-Dem-A | 4 | | + | + | + | | | | 1.2 | 0.4 | zero | | | | | | | 0.1 |
| h | N-Num-Dem-A | 1 | | + | + | + | | | + | 6.0 | 1.8 | zero | | | | | | | 0.1 |
| i | A-Dem-Num-N | 0 | | + | + | | | | + | 2.2 | 0.7 | zero | | | | | | | 0.1 |
| j | A-Dem-N-Num | 0 | | + | + | | | | | 0.5 | 0.1 | zero | | | | | | | 0.1 |
| k | A-N-Dem-Num | 2 | | | | | + | | | 0.5 | 1.4 | very few | + | | + | + | | | 0.8 |
| l | N-A-Dem-Num | 19 | | + | + | + | + | | + | 6.0 | 18.8 | few | + | + | + | | | | 13.8 |
| m | Dem-A-Num-N | 0 | | + | + | | | + | + | 2.2 | 3.6 | zero | | | | | | | 0.1 |
| n | Dem-A-N-Num | 12 | + | + | + | | + | + | | 7.9 | 7.6 | very few | + | | | + | + | | 2.8 |
| o | Dem-N-A-Num | 19 | + | + | | + | + | + | | 21.1 | 20.3 | many | + | + | | | + | | 23.0 |
| p | N-Dem-A-Num | 8 | | + | + | + | | | + | 6.0 | 1.8 | very few | + | + | | | | + | 3.1 |
| q | Num-A-Dem-N | 0 | | + | + | | | | + | 2.2 | 0.7 | zero | | | | | | | 0.1 |
| r | Num-A-N-Dem | 2 | + | | | | + | + | | 7.9 | 7.6 | very few | + | | | + | + | | 2.8 |
| s | Num-N-A-Dem | 32 | + | + | + | + | + | + | | 21.1 | 20.3 | few | + | + | | + | + | | 11.5 |
| t | N-Num-A-Dem | 6 | | + | + | + | | + | + | 6.0 | 9.7 | few | + | | + | + | | | 0.8 |
| u | A-Num-Dem-N | 0 | | + | + | | | | + | 2.2 | 0.7 | zero | | | | | | | 0.1 |
| v | A-Num-N-Dem | 0 | | | | | | + | | 0.5 | 0.7 | zero | | | | | | | 0.1 |
| w | A-N-Num-Dem | 1 | + | | | | + | + | | 7.9 | 7.6 | very few | + | + | | + | | | 26.2 |
| x | N-A-Num-Dem | 88 | + | + | + | + | + | + | + | 104.1 | 100.1 | very many | + | + | | | | | 105.3 |
| | Total languages | 276 | 230 | 271 | 232 | 201 | 252 | 237 | 184 | 276 | 276 | | 270 | 185 | 32 | 55 | 84 | 8 | 276 |
| | Proportion (%) | 100 | 83.3 | 98.2 | 84.1 | 72.8 | 91.3 | 85.9 | 66.7 | 100 | 100 | | 97.8 | 67.0 | 11.6 | 19.9 | 30.4 | 2.9 | 100 |

## References


Baayen, R. Harald. 2008. *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Bates, Douglas & Martin Maechler. 2010. lme4: Linear mixed-effects models using S4 classes. R package version available online at: http://r-forge.r-project.org/projects/lme4/

Bickel, Balthasar. 2008. A refined sampling method for genealogical control. *Sprachtypologie und Universalienforschung* 61. 221–233.

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

Cysouw, Michael. 2005. What it means to be rare: The case of person marking. In Zygmunt Frajzyngier, Adam Hodges & David S. Rood (eds.), *Linguistic diversity and language theories,* 235–258. Amsterdam: Benjamins.

Cysouw, Michael. 2007. Building semantic maps: The case of person marking. In Bernhard Wälchli & Matti Miestamo (eds.), *New challenges in typology,* 225–248. Berlin: Mouton de Gruyter.

Cysouw, Michael. 2008. Linear order as a predictor of word order regularities. *Advances in Complex Systems* 11. 415–420.

Cysouw, Michael. 2010a. On the probability distribution of typological frequencies. In Christian Ebert, Gerhard Jäger & Jens Michaelis (eds.), *The mathematics of language*, 29–35. Berlin: Springer.

Cysouw, Michael. 2010b. The expression of person and number: A typologist's perspective. *Morphology.* Online at doi 10.1007.s11525-010-9170-5

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

Dryer, Matthew S. 1992. The Greenbergian word order correlations. *Language* 68. 80–138.

Dryer, Matthew S. 2005a. Genealogical language list. In Haspelmath et al. (eds.) 2005, 582–642.

Dryer, Matthew S. 2005b. Order of subject, object and verb. In Haspelmath et al. (eds.) 2005, 330–333.

Dryer, Matthew S. 2006. On Cinque on Greenberg's universal 20. Paper presented at Max-Planck-Institute für evolutionäre Anthropologie, Leipzig.

Du Bois, John W. 1985. Competing motivations. In John Haiman (ed.), *Iconicity in syntax,* 343–365. Amsterdam: Benjamins.

Ferrer i Cancho, Ramon. 2008. Some word order biases from limited brain resources: A mathematical approach. *Advances in Complex Systems* 11. 393–414.

Greenberg, Joseph H. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In Joseph H. Greenberg (ed.), *Universals of language,* 73–113. Cambridge, MA: MIT Press.

Harley, Heidi & Elizabeth Ritter. 2002. Person and number in pronouns: A feature-geometric analysis. *Language* 78. 483–526.

Haspelmath, Martin, Matthew S. Dryer, David Gil, & Bernard Comrie (eds.). 2005. *World atlas of language structures*. Oxford: Oxford University Press.

Hawkins, John A. 1990. A parsing theory of word order universals. *Linguistic Inquiry* 21. 223–261.

Hawkins, John A. 2001. Why are categories adjacent? *Journal of Linguistics* 37. 1–34.

Jäger, Gerhard. 2007. Evolutionary game theory and typology: A case study. *Language* 83. 74–109.

Lahiri, Aditi & Frans Plank. 2008. What linguistic universals can be true of. In Sergio Scalise, Elisabetta Magni & Antonietta Bisetto (eds.), *Universals of language today,* 31–58. Berlin: Springer.

Maslova, Elena. 2000. A dynamic approach to the verification of distributional universals. *Linguistic Typology* 4. 307–333.

Maslova, Elena. 2008. Meta-typological distributions. *Sprachtypologie und Universalienforschung* 61. 199–207.

Newmeyer, Frederick J. 2005. *Possible and probable languages: A generative perspective on linguistic typology*. Oxford: Oxford University Press.
R Development Core Team. 2010. *R: A language and environment for statistical computing*. Wien: R Foundation for Statistical Computing. Online: http://www.r-project.org
Rijkhoff, Jan. 2002. *The noun phrase*. Oxford: Oxford University Press.