Glossa a journal of general linguistics

Merlo, Paola and Sarah Ouwayda. 2018. Movement and structure effects on Universal 20 word order frequencies: A quantitative study. *Glossa: a journal of general linguistics* 3(1): 84. 1–35, DOI: https://doi.org/10.5334/gjgl.149

RESEARCH

Movement and structure effects on Universal 20 word order frequencies: A quantitative study

Paola Merlo[1] and Sarah Ouwayda[2]

[1] University of Geneva, 5 Rue de Candolle, 1204 Geneva, CH

[2] Independent Researcher, US

Corresponding author: Sarah Ouwayda (ouwaydasarah@gmail.com)

In this paper, we illustrate a novel method to translate a derivational explanation of Universal 20 into vectorial representations. We exploit this vectorial representation to answer a number of theoretical questions. First, we use linear regression to automatically rank the costs of different syntactic movements within this proposal and investigate some proposals on partial and complete movement. This investigation of movement suggests that the nature of the movement is important, while the importance of harmonic specification of functional categories, i.e. whether the movement is partial or complete, is more context-dependent. We then evaluate whether the base order DEM NUM ADJ N is the best predictor of the typological facts. We compare different syntactic proposals on the position of numerals in the noun phrase. We find that a merge position of numerals higher than adjectives has better results in both methods. We also show, using this method, that the independently motivated low merge position for numerals can only be semantically motivated, which results in intra-linguistic variation, and is not a parametric choice.

Keywords: Universal 20; cost of movement; computational modelling; merge positions; numerals

1 Introduction

Language universals, formal or statistical, absolute or implicational, linguistic properties exhibited by all languages, is one of the main topics in the study of language, and their existence, general nature and distribution are being investigated from a formal and cognitive point of view (Cinque 2005; Dunn et al. 2011; Culbertson et al. 2012; Culbertson & Smolensky 2012; Culbertson & Adger 2014).

We will concentrate on the quantitative properties of language universals (Dryer 1992; Cinque 2005; Cysouw 2010b; Merlo 2015) and will employ them to ask theoretical questions.

Specifically, we set out to address the following three questions:

- Can the ranking of Cinque's different kinds of movements be obtained automatically?
- Is movement always more costly than lack of movement?
- Is the base structure proposed by Cinque the best predictor of the typological frequency facts?

Data-driven computational models can help cast light on linguistic issues in two main ways. First, through their formal nature, they can make the linguistic assumptions in the proposals explicit and operational. Second, 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.

This paper uses a computational modelling methodology previously developed in Cysouw (2010a) and Merlo (2015) to illustrate how to formalise proposals about Universal 20, the universal governing the linear order of a noun and its modifiers, in such a way that their underlying assumptions can be evaluated quantitatively in a computational setting.

2 The facts and the questions

One of the most easily observable distinguishing features of human languages is the order of words: the position of the verb in the sentence or the respective order of the modifiers of a noun, for example. Word orders vary greatly cross-linguistically, but each language has very strong preferences for a few orders, 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.

A more explicit formulation is found in 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.

Currently, we have access to larger samples of languages than Greenberg did. (See, for example, Dryer's and Cinque's large data collections in the cited work). These larger samples have confirmed that two of the three orders indicated by Greenberg as the only possible orders are indeed among the most frequent ones. 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).

Table 1 reports the 24 combinatorially possible orders of the four elements mentioned in Universal 20 (N, DEM, NUM, ADJ) and the actual counts: the second and third columns show the counts reported in Dryer (2006) by language and by genera; the fourth and the last columns are Cinque's counts, as can be deduced from the database we use, provided to us by Cinque himself (September 2013). As can be observed, there are some discrepancies across the different counting methods and across authors, 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 rank correspondence across data collection methods, Dryer's and Cinque's, corroborates the actual counts, as it shows they are not artifacts of one methodology. Moreover, the correspondence between languages and genera indicates that the counts on which we work below are representative, at least to a good approximation, of the real distribution of word orders in the world and are controlled for historical influence or areal spread.

Many proposals have been put forth to identify the factors that could give rise to the distributions of different word orders of the noun phrase across languages of the world. These proposals range from general principles of symmetry and harmony (Dryer 2006), to a larger number of base word orders (Abels & Neeleman 2009), to observable preferences for head positions (Cysouw 2010a), to Optimality theory constraints (Steddy & Samek-Lodovici 2011), to primitive operations of CCG (Steedman 2011), to preferences

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

				Dryer's Languages	Dryer's Genera	Cinque's 05 Languages	Cinque's 13 Languages
DEM	NUM	ADJ	N	74	44	V. many	300
DEM	ADJ	NUM	N	3	2	0	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	22	17	Many	114
DEM	ADJ	N	NUM	11	6	V. few (7)	35
NUM	DEM	N	ADJ	0	0	0	0
NUM	ADJ	N	DEM	4	3	V. few (8)	40
ADJ	DEM	N	NUM	0	0	0	0
ADJ	NUM	N	DEM	0	0	0	0
DEM	N	ADJ	NUM	28	22	Many	125
DEM	N	NUM	ADJ	3	3	V. few (4)	37
NUM	N	DEM	ADJ	5	3	0	0
NUM	N	ADJ	DEM	38	21	Few (2)	180
ADJ	N	DEM	NUM	4	2	V. few (3)	14
ADJ	N	NUM	DEM	2	1	V. few	15
N	DEM	NUM	ADJ	4	3	Few (8)	48
N	DEM	ADJ	NUM	6	4	V. few (3)	24
N	NUM	DEM	ADJ	1	1	0	0
N	NUM	ADJ	DEM	9	7	Few (7)	35
N	ADJ	DEM	NUM	19	11	Few (8)	69
N	ADJ	NUM	DEM	108	57	V. many (27)	411

with alignment for semantic scope (Culbertson & Adger 2014). Three of these methods were compared in Merlo (2015), with an approach similar to what we use in some sections of this paper.

The trigger of the work on these proposals for universal 20 is the generative, derivational account proposed in Cinque's (2005), where several kinds of movements, of different costs, are applied to a fixed-base word order, DEM NUM ADJ N, to generate the possible word orders and to account for the asymmetry in typological frequency between the prenominal and the postnominal orders.

In this paper, the large-scale quantitative typological observations and the underlying generative process that is specifically proposed in Cinque (2005) will be investigated in detail to ask two types of questions. The first concerning the weight of different syntactic operations. Specifically, what is the cost of the proposed syntactic operations? And is lack of movement always less costly than movement? The second concerns the structure of the DP. Specifically: is the merge order assumed by Cinque (2005) the best predictor of the typological facts? More precisely, in the second question, we focus on determining

the structural position of numerals. These two kinds of questions are of a different nature and they are chosen to show that quantitative large-scale facts are useful both to study gradient processes, such as the cost of operations, but also to ask structural and categorical questions usually tackled by symbolic means.

The proposal in Cinque (2005) was specifically chosen because of its derivational nature and because it makes use of unobserved abstract operations, and is therefore the most complex theory to represent in a vectorial form, as we will illustrate later.

2.1 Weight of syntactic operations

A question of significant interest for syntacticians in the generative framework is which syntactic operations are possible, and which ones are not possible. And among those that are possible, which ones cost more than others. Cinque (2005; 2013) makes a proposal that assigns different weights to different syntactic operations, in order to derive the typological frequencies of different word orders in the DP. Experiments 1 and 2 of our study use linear regression to derive these weights automatically. Crucially, Cinque proposes that one kind of movement, recursive pied-piping of the NP, is cost-free. Our results confirms that this kind of movement is mostly cost-free, and that, in specific circumstances, it is more desirable than partial movement. In this way, we explain a fact that has remained unexplained in Cinque's accounts: word orders produced by derivations in which no movement has occurred are attested by fewer languages than those in which movement has occurred.[1]

2.2 Structural position of cardinal numerals

Having established the relative costs of movement to derive word orders in the DP, we turn to the category of some of the elements in the DP. Specifically, another question that has been the object of a lot of attention in theoretical linguistics in recent years is the syntactic category of cardinal numerals occurring in noun phrases, such as those shown in (1).

(1) a. Three apples
b. Many apples
c. Garden apples
d. Big apples

Since numerals can appear in various syntactic contexts, it is not clear what syntactic category fits them best. It has been argued that numerals are quantifiers, adjectives, nouns, or a combination of categories. Cinque's (2005) proposal to account for Universal 20 assumes a high merge position for numerals, much like quantifiers. Stavrou & Terzi (2008; 2009), who examine numerals in Greek, argue that indeed, simple numerals like *three* in (1-a) constitute a subclass of weak quantifiers, like *many* in (1-b). In contrast, other researchers treat all numerals as nouns, like *garden* in (1-c) (Hurford 1975; 1987; 2003; Ionin & Matushansky 2004; 2006). Ionin & Matushansky (2006) further propose that numerals have the semantics of modifiers, like adjectives. Landman (2000), in fact, argues that all indefinite DPs are predicates of type $<e, t>$, which requires numerals to be adjectival, like *big* in (1-d). This would allow numerals to merge either higher or lower than other adjectives.

Experiments 3 and 4 of the current study compare two of these proposals from a typological perspective, namely merging numerals high versus merging them low, as well as a number of intermediate views where they merge high in some languages and low in others. The results suggest that, assuming numerals constitute a uniform syntactic category

[1] The preference for recursive pied-piping of the NP among movement operations is also compatible with recent results on scope preferences for universal 20 reported in Culbertson & Adger (2014).

cross-linguistically, treating them as structurally high allows for a better prediction of the typological frequencies than treating them as adjectives, and structurally lower.

2.3 Roadmap

The rest of the paper will develop as follows. Section 3 presents the computational and experimental method. The following three sections each present an experimental question: section 4 presents Experiment 1, which tests the predictive power of Cinque's (2005) proposal for Universal 20 using linear regression, and automatically determines the weights of the different syntactic operations. Section 5 presents Experiment 2, which replicates Experiment 1, but changes the encoding to treat lack of movement as a uniform factor suggesting that the right kind of movement is not only free, but also desirable. Section 6 presents Experiment 3, which also uses linear regression, this time to evaluate the predictive powers of an alternative to Cinque's (2005) proposal where numerals are allowed to merge either low or high as a parametric choice, rather than having an exclusively high position as assumed in the proposal. Subsection 6.5 presents Experiment 4, which replicates Experiment 3 using discrete categories (very frequent, frequent, infrequent or zero) instead of raw frequencies to control for the inherent uncertainty about typological counts, and trains a Naive Bayes Classifier to classify different word orders into one of these classes. The results of Experiment 3 and Experiment 4 are consistent, and both point to the high merge position as the better choice.

3 Vectorial models of linguistic proposals

The method used in this paper will require transforming the linguistic proposals concerning Universal 20 into a vectorial representation, as described below. This process transforms a derivational account into a non-derivational, fixed-length vector of the important properties of the account. This vectorial representation is very abstract and is compatible with many statistical methods. It is possible then to find automatically the relative weight of each element in the vector, a process of parameter fitting that best describes and predicts the typological frequencies. This method of transforming a derivational linguistic theory into vectors and then using machine learning techniques has first been proposed and methodologically justified in Merlo (2015), where other theories of Universal 20 are compared. The steps of the formalisation that we propose here, therefore, are as follows:

i. Formalise the properties and operations of a model of word order as simple primitive features with a set of associated values;
ii. Encode each word order as a vector of instantiated primitives defined by the model;
iii. Learn the parameters of the model through a learning algorithm on (a subset of) the data;
iv. Run the model 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 method.

3.1 Encoding linguistic proposals as vectors of features

In 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 second to last column of Table 1, Cinque remarks that there are 24 combinatorially possible orders of the four elements: N, DEM, NUM, ADJ. Only 14 of them are attested in the languages of the world (see also Dryer's counts in the same table, Table 1, and Cinque's new counts, last column). Some of the 14 orders are unexpected under Universal 20. It is proposed 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 order of merger in (2) is assumed, where only the overt NP or phrases containing the overt NP can move. The allowed syntactic operations (movements) are given in (3), omitting intermediate projections for simplicity.

(2) $[_{WP}$ Dem $[_{XP}$ Num $[_{YP}$ Adj $[_{NP}$ N]]]]

(3) a. NP movement without pied-piping:
$[_{WP}\ [_{NP}\ N]_1$ Dem $[_{YP}$ Adj $t_1\ ]]]$

b. Movement of a constituent containing the NP with pied-piping of the *picture of who* type (moving of an [XP[NP]]):
$[_{WP}\ [_{YP}$ Adj $[_{NP}$ N] $]_1$ Dem $t_1\ ]]$

c. NP movement with pied-piping of the *whose picture* type (moving NP above XP, then moving [NP[XP]], and so on):
$[_{WP}\ [_{YP}\ [_{NP}$ N $]_1$ Adj $t_1\ ]_2$ Dem $t_2]$

d. Partial movement:[2]
$[_{WP}$ Dem $[_{XP}$ Num $[_{YP}\ [_{NP}$ N $]_1$ Adj $t_1\ ]]]$

e. Splitting the NP out of a moved element to move it to a higher position:
$[_{WP}\ [_{NP}$ N $]_3$ Dem $[_{XP}\ [_{YP}\ (\text{t}_{\text{NP}}$ ~~N]~~$_1)_3$ Adj $t_1\ ]_2$ Num $t_2]$

The 24 possible permutations of demonstrative, numeral, adjective and noun are derived using these movements. Some orders require no movements, other require derivations of different numbers of movement steps. These derivations are summarized in Tables 2 and 3. The tables show two lines for each word order, for each movement step. The first line describes the word order movement operation, the second line gives the name of the type of movement according to our formal encoding.

Since each language in the sample has one dominant word order, and that word order is assumed to be the result of a derivation produced by some of the syntactic operations in (3), we treat these syntactic operations as binary parameters that a language either has or does not have.

In addition to the allowed movements, we define one parameter to describe movements like (4), which are argued to be impossible by Cinque (2005), and where a phrase not containing the overt NP is moved. These parameters are those in (5).

(4) NPless-Move:
$[_{WP}\ [_{YP}$ Adj $t_1\ ]_2$ Dem $[_{XP}\ [_{NP}$ N $]_1$ Num $t_2\ ]]$

(5) a. Uses NP movement without pied piping
b. Uses NP movement with pied-piping of the [XP[NP]] type
c. Uses NP movement with pied-piping of the [NP[XP]] type
d. Involves partial movement
e. Uses NP-splitting movement
f. Requires movement of a phrase not containing the NP

[2] This is our interpretation of Cinque's partial movement. Partial movement occurs when there is movement, of any category, but nothing has moved above the demonstrative.

Table 2: Movements necessary for each word order in Cinque's proposal (continued in next table). The table shows two lines for each word order, for each movement step. The first line describes the word order movement operation, the second line gives the name of the type of movement according to our formal encoding.

Word Order	Step 1	Step 2	Step 3	Step 4
a. DEM NUM ADJ N	No movements			
b. DEM NUM N ADJ	NP above ADJ	No more mov'ts		
	[NP[XP]]-Move	Partial mov't		
c. DEM N NUM ADJ	NP above NUM	No more mov'ts		
	No-Pied-Piping	Partial mov't		
d. N DEM NUM ADJ	NP above DEM			
	No-Pied-Piping			
e. NUM DEM ADJ N	AP above NUM	NPless-NUM above DEM		
	[XP[NP]]-Move	NPless-Move		
f. NUM DEM N ADJ	NP above ADJ	AP above NUM	NPless-NUM above DEM	
	[NP[XP]]-Move	[NP[XP]]-Move	NPless-Move	
g. NUM N DEM ADJ	NP above ADJ	AP above NUM	NP splits, above DEM	NPless-NUM moves
	[NP[XP]]-Move	[NP[XP]]-Move	Split	NPless-Move
h. N NUM DEM ADJ	NP above ADJ	AP above NUM	NPless-NUM moves	NP splits, above DEM
	[NP[XP]]-Move	[NP[XP]]-Move	NPless-Move	Split
i. ADJ DEM NUM N	NP above ADJ	NPless-AP moves		
	[NP[XP]]-Move	NPless-Move		
j. ADJ DEM N NUM	NP above NUM	NPless-AP moves		
	No-Pied-Piping	NPless-Move		
k. ADJ N DEM NUM	AP above NUM	AP above DEM		
	{ [XP[NP]]-Move, No-Pied-Piping }			
l. N ADJ DEM NUM	NP above ADJ	AP above DEM		
	[NP[XP]]-Move	No-Pied-Piping		

Since these parameters are binary, we can now encode the different word orders as a vector of values, by assigning either 1 or 0 to each parameter. Importantly, since the frequency of a word order is not correlated with the number of movements necessary to reach it, but rather with the type of movement necessary to derive it, we do not count the different occurrences of a certain movement. Rather, a parameter has a positive value, if it needs its corresponding movement, and we assign it the value 1, regardless of how many times it is needed for a given derivation.

To illustrate the encoding of three different word orders, step by step, consider the English word order DEM NUM ADJ N, the second most frequent order. This order can be obtained immediately by linearizing the structure (2), with no movements at all. This means that to reach this word order, no movements are needed. So all the parameters in (5) are set to 0 for this word order. This gives us the values in Table 4, which is equivalent to row a of Table 7.

The mirror order of English N ADJ NUM DEM occurs in 411 languages, an example is Sudanese Arabic. This order involves NP moving above ADJ; after this movement, the

phrase [$_{YP}$ [$_{NP}$ N]] Adj] moves above NUM; finally, [$_{XP}$ [$_{YP}$ [$_{NP}$ N]] Adj] Num] moves above DEM. These movements are illustrated in (6).

(6) a. [$_{WP}$ Dem [$_{XP}$ Num [$_{YP}$ Adj [$_{NP}$ N]]]]
b. [$_{WP}$ Dem [$_{XP}$ Num [$_{YP}$ [$_{NP}$ N]$_1$ Adj t$_1$]]]
c. [$_{WP}$ Dem [$_{XP}$ [$_{YP}$ [$_{NP}$ N]$_1$ Adj t$_1$]$_2$ Num t$_2$]]
d. [$_{WP}$ [$_{XP}$ [$_{YP}$ [$_{NP}$ N]$_1$ Adj t$_1$]$_2$ Num t$_2$]$_3$ Dem t$_3$]

Table 3: Movements necessary for each word order in Cinque's proposal (continued from Table 2). The table shows two lines for each word order, for each movement step. The first line describes the word order movement operation, the second line gives the name of the type of movement according to our formal encoding.

Word Order	Step 1	Step 2	Step 3	Step 4
m. DEM ADJ NUM N	NP above NUM	NPless AP above NUM	No more mov't	
	[NP[XP]]-Move	NPless-Move	Partial mov't	
n. DEM ADJ N NUM	AP above NUM	No move mov't		
	[XP[NP]]-Move	Partial mov't		
o. DEM N ADJ NUM	NP above Adj	AP above NUM	No more mov't	
	[NP[XP]]-Move	[NP[XP]]-Move	Partial mov't	
p. N DEM ADJ NUM	NP above ADJ	AP above NUM	NP splits, above DEM	
	[NP[XP]]-Move	[NP[XP]]-Move	Split mov't	
q. NUM ADJ DEM N	NP above NUM	NPless-NUM above DEM		
	No-Pied-Piping	NPless-Move		
r. NUM ADJ N DEM	NumP above DEM			
	[XP[NP]]-Move			
s. NUM N ADJ DEM	NP above ADJ	NumP above DEM		
	[NP[XP]]-Move	[XP[NP]]-Move		
t. N NUM ADJ DEM	NP above NUM	NumP above DEM		
	No-Pied-Piping	[NP[XP]]-Move		
u. ADJ NUM DEM N	NP above ADJ	AP above NUM	NP splits to stay	NPless-NumP above DEM
	[NP[XP]]-Move	[NP[XP]]-Move	Split	NPless-Move
v. ADJ NUM N DEM	NP above ADJ	AP above NUM	NP splits above DEM	NPless-NUM above DEM
	[NP[XP]]-Move	[NP[XP]]-Move	{ Split, NPless-Move }	
w. ADJ N NUM DEM	AP above NUM	NumP above DEM		
	[XP[NP]]-Move	[NP[XP]]-Move		
x. N ADJ NUM DEM	NP above ADJ	AP above NUM	NumP above DEM	
	[NP[XP]]-Move	[NP[XP]]-Move	[NP[XP]]-Move	

Table 4: The order DEM NUM ADJ N is encoded as <0, 0, 0, 0, 0, 0>.

Feature	Value
a. Uses NP movement without pied-piping	0
b. Uses NP movement with pied-piping of the [XP[NP]] type	0
c. Uses NP movement with pied-piping of the [NP[XP]] type	0
d. Involves partial movement	0
e. Uses NP-splitting movement	0
f. Requires movement of a phrase not containing the NP	0

Table 5: The encoding of N ADJ NUM DEM is <0, 0, 1, 0, 0, 0>.

Feature	Value
a. Uses NP movement without pied-piping	0
b. Uses NP movement with pied-piping of the [XP[NP]] type	0
c. Uses NP movement with pied-piping of the [NP[XP]] type	1
d. Involves partial movement	0
e. Uses NP-splitting movement	0
f. Requires movement of a phrase not containing the NP	0

Table 6: The order DEM N NUM ADJ is encoded as <1, 0, 0, 1, 0, 0>.

Feature	Value
a. Uses NP movement without pied-piping	1
b. Uses NP movement with pied-piping of the [XP[NP]] type	0
c. Uses NP movement with pied-piping of the [NP[XP]] type	0
d. Involves partial movement	1
e. Uses NP-splitting movement	0
f. Requires movement of a phrase not containing the NP	0

Table 7: The orders and their encodings according to Cinque's 2005 operations. The columns of binary features correspond to the rows a-f of (5). Frequencies are provided to us by Cinque (p.c., September 2013).

	NP moves w/o pp	[XP[NP]] moves	[NP[XP]] moves	Partial move	Split move	NPless move	Freq.
a. DEM NUM ADJ N	0	0	0	0	0	0	300
b. DEM NUM N ADJ	0	0	1	1	0	0	114
c. DEM N NUM ADJ	1	0	0	1	0	0	37
d. N DEM NUM ADJ	1	0	0	0	0	0	48
e. NUM DEM ADJ N	0	1	0	0	0	1	0
f. NUM DEM N ADJ	0	0	1	0	0	1	0
g. NUM N DEM ADJ	0	0	1	0	1	1	0
h. N NUM DEM ADJ	0	0	1	0	1	1	0
i. ADJ DEM NUM N	0	0	1	0	0	1	0
j. ADJ DEM N NUM	1	0	0	0	0	1	0
k. ADJ N DEM NUM	1	1	0	0	0	0	14
l. N ADJ DEM NUM	1	0	1	0	0	0	69
m. DEM ADJ NUM N	0	0	1	1	0	1	0
n. DEM ADJ N NUM	0	1	0	1	0	0	35
o. DEM N ADJ NUM	0	0	1	1	0	0	125
p. N DEM ADJ NUM	0	0	1	0	1	0	24
q. NUM ADJ DEM N	1	0	0	0	0	1	0
r. NUM ADJ N DEM	0	1	0	0	0	0	40
s. NUM N ADJ DEM	0	1	1	0	0	0	180
t. N NUM ADJ DEM	1	0	1	0	0	0	35
u. ADJ NUM DEM N	0	0	1	0	1	1	0
v. ADJ NUM N DEM	0	0	1	0	1	1	0
w. ADJ N NUM DEM	0	1	1	0	0	0	23
x. N ADJ NUM DEM	0	0	1	0	0	0	411

Since all of these movements are movements of the NP with pied-piping (i.e. movements of the [NP[XP]] type), only the parameter (5-c) is set to 1, and all other ones are 0. This gives us Table 5, which is row x of Table 7.

Finally, to illustrate a less straightforward case, consider the word order DEM N NUM ADJ. This order is less frequent than the other two, it occurs in 37 languages, for instance Xhosa (Bantoid, Niger-Congo). In order to derive this order, the NP has to move above NUM, skipping the adjective, without pied-piping. This word order therefore requires movement of the NP without pied-piping, so the value of (5a) for this word order is 1. In addition, since the noun remains to the right of the demonstrative, the NP must stop after moving above NUM. This word order therefore requires partial movement (movement without going above DEM). So the value of (5-d) for this word order is 1 as well. Since no other movement is needed, all the remaining parameters are set to 0. So this word order is encoded as illustrated in Table 6, which corresponds to row c of Table 7.

Table 7 summarizes the encoding for each of the 24 possible word orders, as a vector of the six parameters defined in (5). Now that we have transformed the theory into a vectorial representation, we automatically find the relative weight of each element in the vector, a process of parameter fitting.

4 Experiment 1: The cost of moving

In this first experiment, we find the weights of the movement operations encoded vectorially, and compare them to Cinque's proposal.

4.1 Materials

The primary data used in all four experiments is Cinque's (2013) sample of languages containing the order of the four elements demonstrative, numeral, adjective, and noun in the DP in 1475 languages. The sample specifies the language, the word order, the genus, and the family. This data was collected by Cinque through examination of different grammars and resources, and converted into a table to allow computerized work.[3] Except for one word order that turns out to be more represented than previously known, the overall tendencies in this large sample of languages is similar to earlier published samples by Cinque (2005) and Dryer (2006), as shown in Table 1. As is often the case with linguistic distributional data, the word order frequencies are skewed according to Zipf's law. Only 5 of the 24 possible orders account for almost 70% of the world's languages, and the remaining 30% share another 9 word orders, and 10 orders are unattested. This is summarized in Table 8.

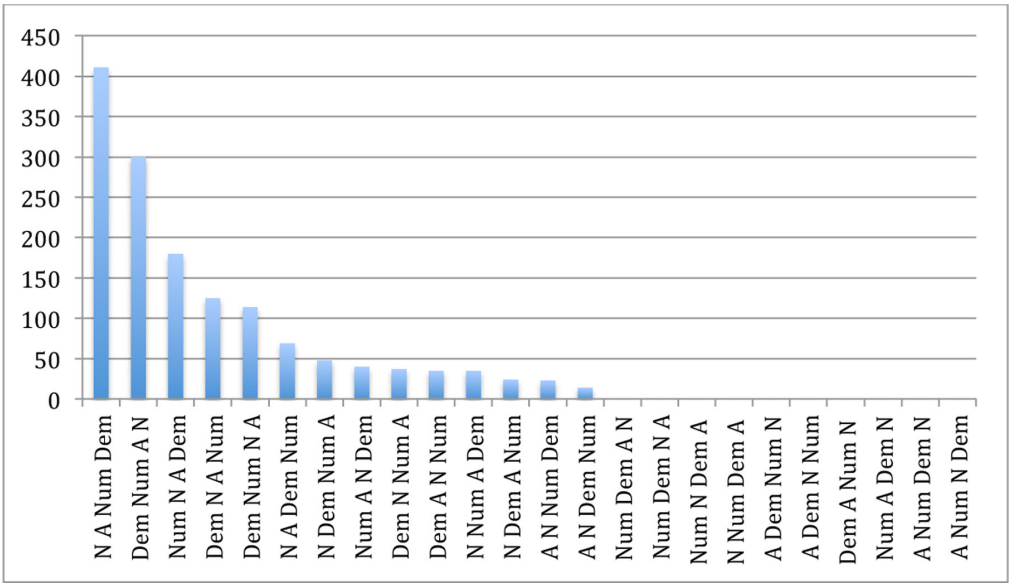
The different movement operations corresponding to each word order are shown in (5) and were discussed in the previous section. The features and the possible values of Cinque's model are shown in Table 7. The data in Table 7 are the data that we use in our experiments. The values in the last column are the frequency property of the word order, the dependent variable that we are trying to explain.

4.2 Method

Using the encoded data in Table 7, we used the Waikato Environment for Knowledge Analysis, WEKA (Hall et al. 2009) to derive the best linear regression model of this data. The linear regression expresses the frequency of the word order as a function of the different syntactic operations involved.

[3] We thank Guglielmo Cinque for giving us acces to this invaluable database.

Table 8: Zipfian distribution of word orders.



Each syntactic operation is encoded as an indicator variable: a variable that has the two values 0 and 1, and indicates if the property is present or not. These are nominal variables, while the dependent variable is numeric. In this setting our multi-variable linear regression gives us the positive and negative coefficients that are the difference from the predicted frequency value of the control group, the base order DEM NUM ADJ N, whose frequency is represented by the intercept of the function. As such, syntactic operations that are associated with frequencies lower than the frequency of the base order will get negative coefficients, and those associated with higher frequencies will have zero coefficients, or even positive coefficients.

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.

We used a leave-one-out cross-validation, automatically eliminating collinear attributes, to generate a linear regression model of the data.[4]

4.3 Results and discussion

The linear regression model that was generated was the function in (7), whose goodness of fit is indicated by the correlation coefficient of 0.52.

(7) Frequency = –129.0 × Uses NP movement without pied-piping
–115.6 × Uses NP movement, pied-piping [XP[NP]]
–37.8 × Uses NP movement with pied-piping of the [NP[XP]] type
–65.6 × Partial Move
–91.6 × Uses NP-splitting movement
–135.9 × Requires moving a phrase not containing NP
+242.8

[4] Collinearity refers to a linear relationship between two or more explanatory variables. Correlation between two variables increases the variance of the correlation coefficient and makes the prediction unstable.

The linear model in (7) can be read as a ranking of the different syntactic operations in terms of markedness. The weights are summarised in (8). Specifically, we note that large negative values of any of the parameters (5-a), (5-b), (5-e) and (5-f) are considered very costly.

(8) Weights of the different syntactic operations

a.	NP movement without pied-piping	–129
b.	NP movement with pied-piping of the [XP[NP]] type	–115
c.	NP movement with pied-piping of the [NP[XP]] type	–37
d.	Partial movement	–65
e.	NP-splitting movement	–91
f.	Movement of a phrase not containing the NP	–135

If we interpret the weights as costs, so that high negative weights are high costs, we can rank the different movements in a partial order, as in (9) (where the symbol "$<$" means "less costly" and we use abbreviated symbols):

(9) [NP[XP]] $<$ Partial $<$ Split $<$ [XP[NP]] $<$ NPw/oPied-P $<$ Movew/oNP.

A system of weights can also be inferred from Cinque's proposal (2005: 321), based on the markedness levels assigned to the movement types, as in (10):

(10) {NoMove, [NP[XP]], Total} $<$ Partial $<$ NPw/oPied-P $<$ [XP[NP]] $<$ {Split, Movew/oNP}.

Considering only the types of movements that are encoded in both accounts and simplifying labelling for readability, we have the two following partial orders:

(11) a. Cinque: [NP[XP]] = Partial $<$ NPw/oPied-P $<$ [XP[NP]] $<$ Split = Movew/oNP.
b. Us: [NP[XP]] $<$ Partial $<$ Split $<$ [XP[NP]] $<$ NPw/oPied-P $<$ Movew/oNP.

We can see that the two orders are well correlated: tied orders in Cinque's are ranked but adjacent by our ranking, and in only one case the rank is reversed (XPNP and Split) in the two orders. Kendall's Tau_b is a measure of rank correlation that allows ties. A comparison of our two orders yields a Kendall's Tau_b of 0.6, which is statistically significant at $p < 0.5$ even in such a small sample. We can conclude that our regression corroborates not only the nature of the movements but also, to some extent, the respective costs that were assigned in Cinque's work.

The low cost of partial movement does not confirm Cinque's assumption that partial movement is penalising. Rather, partial movement appears to be less costly than other operations. Partial movement was considered more marked than complete movement in Cinque's system to derive the fact that the mirror image of the base order is very frequent. This is an order where movement is complete. However, Cinque's system does not predict that the mirror image of the base order is more frequent than the base order. In Cinque's model, *no movement* is unmarked. Since there seems to be a weak correlation between the weights assigned to complete and partial movement and typological frequencies, we investigate the relative weights of partial and complete

movement. We develop therefore a new model, where we investigate different encodings of movement.

5 Experiment 2: The cost of staying put

In this three-fold experiment, we test whether different definitions of partial and complete movement make a difference in predicting word orders. The first question we ask is: how do we define partial movement? Do we have partial movement when something moves, of any category, but does not move all the way above DEM? Or do we have partial movement when the N moves, but without moving all the way up above DEM? Do these two definitions make a difference in predicting word order frequencies?

A second question we ask is: if partial movement is costly as Cinque proposes, is lack of movement free, as Cinque (2005) assumes? Or is it costly, like partial movement? This question compares two existing proposals about the cost of movement. The first proposal is that of Cinque, contending that movement can be free, or costly, depending on the kind of movement, and that lack of movement is by default free. The second existing proposal is proposed in Shlonsky (2012), whose work suggests that in some contexts (specifically when agreement is involved), movement is desirable, and, for our purposes, less costly than absence of movement.

The model tested in this experiment replaces partial movement with a NoMove parameter, to test if it is really the case that the best explanation of the typological frequency data is based on the assumption that *no movement* is unmarked and preferred to movement.

5.1 Materials

The same set of data was used for this experiment as for experiment 1. The encoding, however, involved a new parameter, which replaced *Partial Movement*. So the new set of parameters is shown in (12).

(12) a. Uses NP movement without pied-piping
b. Uses NP movement with pied-piping of the [XP[NP]] type
c. Uses NP movement with pied-piping of the [NP[XP]] type
d. Involves lack of movement (partial or complete)
e. Uses NP-splitting movement
f. Requires movement of a phrase not containing the NP

The materials are the same as those of the previous experiment, with modifications to the feature *Partial Movement*. In particular, in experiment 1 the partial movement feature was encoded as (13), such that complete movement and lack of movement were grouped together, and partial movement had a different specification.

(13) Complete movement of anything is better than partial movement.
1 if nothing moves above DEM (partial movement)
0 if nothing moved at all (no movement)
0 if anything moves, N or other, above DEM (complete movement)

In (13), partial movement is defined as involving any category, but receives a different value from no movement or complete movement. We also encode the two new models shown in (14) and (15). Both models consider that the distinction between complete and partial movement is relevant only for N. In these two models, the only distinguishing

Table 9: Partial movement encodings.

	Equation (13)	Equation (14)	Equation (15)
a. DEM NUM ADJ N	0	0	1
b. DEM NUM N ADJ	1	1	1
c. DEM N NUM ADJ	1	1	1
d. N DEM NUM ADJ	0	0	0
e. NUM DEM ADJ N	0	1	1
f. NUM DEM N ADJ	0	1	1
g. NUM N DEM ADJ	0	0	0
h. N NUM DEM ADJ	0	0	0
i. ADJ DEM NUM N	0	1	1
j. ADJ DEM N NUM	0	1	1
k. ADJ N DEM NUM	0	0	0
l. N ADJ DEM NUM	0	0	0
m. DEM ADJ NUM N	1	1	1
n. DEM ADJ N NUM	1	1	1
o. DEM N ADJ NUM	1	1	1
p. N DEM ADJ NUM	0	0	0
q. NUM ADJ DEM N	0	1	1
r. NUM ADJ N DEM	0	0	0
s. NUM N ADJ DEM	0	0	0
t. N NUM ADJ DEM	0	0	0
u. ADJ NUM DEM N	0	1	1
v. ADJ NUM N DEM	0	0	0
w. ADJ N NUM DEM	0	0	0
x. N ADJ NUM DEM	0	0	0

property is the fact that no movement receives the same value as complete movement, as in (14), and that in (15), no movement receives the same value as partial movement, defined for N only. The encodings of partial movements in the three sets of data is shown in detail in Table 9, and exemplified for one order in Table 10.

(14) Complete N movement is better than Partial N movement
1 if N moves, but not above DEM (partial movement)
0 if nothing moves at all (no movement)
0 if N moves above DEM (complete movement)

(15) Any lack of N movement, complete or partial, is costly.
1 if N moves, but not above DEM (partial movement)
1 if nothing moves at all (no movement)
0 if N moves above DEM (complete movement)

5.2 Methods

Hall et al.'s (2009) WEKA was used again, with the same procedure as experiment 1. The complete data set is shown in Table 11, corresponding to the model in (14). The table corresponding to the model in (15) is the same with the exception of the first line, which is

Table 10: Order DEM NUM ADJ N is encoded as <0, 0, 0, 1, 0, 0>.

Feature	Value
a. Uses NP movement without pied-piping	0
b. Uses NP movement with pied-piping of the [XP[NP]] type	0
c. Uses NP movement with pied-piping of the [NP[XP]] type	0
d. Involves lack of movement (partial or complete)	1
e. Uses NP-splitting movement	0
f. Requires movement of a phrase not containing the NP	0

Table 11: The orders and their encodings according to Cinque's (2005) operations corresponding to the model in (14). The table corresponding to the model in (15) is the same with the exception of the first line, which is <a′ DEM NUM ADJ N 0 0 0 1 0 0 300>. Frequencies are provided to us by Cinque (p.c.).

	No Pied piping	[XP[NP]] moves	NP[XP]] moves	No move	Split move	NPless move	Freq.
a. DEM NUM ADJ N	0	0	0	0	0	0	300
b. DEM NUM N ADJ	0	0	1	1	0	0	114
c. DEM N NUM ADJ	1	0	0	1	0	0	37
d. N DEM NUM ADJ	1	0	0	0	0	0	48
e. NUM DEM ADJ N	0	1	0	1	0	1	0
f. NUM DEM N ADJ	0	0	1	1	0	1	0
g. NUM N DEM ADJ	0	0	1	0	1	1	0
h. N NUM DEM ADJ	0	0	1	0	1	1	0
i. ADJ DEM NUM N	0	0	1	1	0	1	0
j. ADJ DEM N NUM	1	0	0	1	0	1	0
k. ADJ N DEM NUM	1	1	0	0	0	0	14
l. N ADJ DEM NUM	1	0	1	0	0	0	69
m. DEM ADJ NUM N	0	0	1	1	0	1	0
n. DEM ADJ N NUM	0	1	0	1	0	0	35
o. DEM N ADJ NUM	0	0	1	1	0	0	125
p. N DEM ADJ NUM	0	0	1	0	1	0	24
q. NUM ADJ DEM N	1	0	0	1	0	1	0
r. NUM ADJ N DEM	0	1	0	0	0	0	40
s. NUM N ADJ DEM	0	1	1	0	0	0	180
t. N NUM ADJ DEM	1	0	1	0	0	0	35
u. ADJ NUM DEM N	0	0	1	1	1	1	0
v. ADJ NUM N DEM	0	0	1	0	1	1	0
w. ADJ N NUM DEM	0	1	1	0	0	0	23
x. N ADJ NUM DEM	0	0	1	0	0	0	411

<a′ DEM NUM ADJ N 0 0 0 1 0 0 300>. WEKA was used to derive the best linear regression model of this data, assigning negative coefficients to the syntactic operations (or lack thereof) that are associated with lower frequencies, and zero or positive cofficients to those associated with high frequencies. As before, we also used a leave-one-out cross-validation, automatically eliminating collinear attributes, to generate a linear regression model of the data.

5.3 Results

The three models derived by the linear regression are shown in examples (16), (17), and (18).

(16) Frequency = –151.9 × NP movement without pied-piping
–139.3 × NP movement, pied-piping [XP[NP]]
–42.5 × NP movement, pied-piping [NP[XP]]
–92.2 × Involves lack of movement
–106.7 × Uses NP-splitting movement
–140.2 × Requires moving a phrase not containing NP
+266.2
Correlation coefficient 0.59

(17) Frequency = –148.0 × Uses NP movement without pied-piping
–138.3 × Uses NP movement, pied-piping [XP[NP]]
–50.8 × Uses NP movement, pied-piping [NP[XP]]
–90.0 × Involves lack of movement
–138.5 × Uses NP-splitting movement
–73.4 × Requires moving a phrase not containing NP
+270.8
Correlation coefficient 0.59

(18) Frequency = –167.3 × Uses NP movement without pied-piping
–157.2 × Uses NP movement, pied-piping [XP[NP]]
–71.5 × Uses NP movement, pied-piping [NP[XP]]
–81.9 × Involves lack of movement
–136.0 × Uses NP-splitting movement
–88.7 × Requires moving a phrase not containing NP
+299.6
Correlation coefficient 0.47

The first observation about these models is that the *NoMove* feature, in any of its encodings, is among the least penalising and it ranks second least penalising, like in the model of experiment 1. In comparison to experiment 1, though, this encoding yields, in two cases, a much better correlation to the frequency data.

The comparison of the first encoding of complete movement compared to the second encoding of complete movement, anything moves above DEM vs N moves above DEM, shown in (17) and (18), shows that the two models yield in some cases the same correlation with an almost identical ranking of cost of features, but that results are mixed, and on average, the more general definition of partial movement yields better fit to the data.[5]

The comparison of the first and second encoding of movement compared to the third encoding of movement (N moves above DEM is as costly as not moving, partial N moves,

[5] This conclusion is drawn based on a systematic comparison of correlation coefficients of linear regression in different settings of cross-validation (leave one out, or 10-fold cross-validation) and with or without attribute selection, as shown below.

Cross-val Method	Attribute Selection?	Equation (13)	Equation (14)	Equation (15)
Leave one out	no	0.59	0.59	0.47
Leave one out	yes	0.55	0.53	0.42
10-fold	no	0.56	0.60	0.53
10-fold	yes	0.57	0.54	0.48
Average		0.57	0.56	0.47

not above DEM, is as costly as not moving, where it is free) shown in (16), (17) and (18), and footnote 5 is much clearer. It shows that for these data no movement is always preferred. While lack of movement does not become particularly costly in the third model, the correlation coefficient is lower.

5.4 Discussion

The model in (14), where lack of movement is encoded in the same way as complete movement, as proposed in Cinque (2005), which considers both unmarked, yields better results than (15), where lack of movement and partial movement have the same value. Assuming that lack of movement is sometimes costly would explain the fact that the base order DEM NUM ADJ N is less frequent than the mirror order, N ADJ NUM DEM, which requires four movements. There are, in fact, theoretical precedents to this position. Indeed, Shlonsky (2012) proposes that movement of the noun (or the NP) over any agreeing DP components is obligatory. In other words, some languages (namely those with agreeing modifiers) cannot do without movement. This is encoded with the weights assigned in (15): If movement is required in a significant number of languages, and allowed in the rest, then lack of movement will be more penalized than the right kind of movement. The low correlation of model (15) does not appear to support this proposal.

This conclusion is corroborated by the fact that the model in (13) is better than the one in (14), which shows that no movement and complete movement appear to pattern together. This is what we find in the distribution of word orders: the most frequent word order is generated by roll-up movement of different kinds of phrases, and the third and fourth most frequent word orders are generated by partial movement of the N. These orders together, however, do not make movement a less costly option than no movement, exhibited by the unique base order. In other words, our model of cost of operations derives the number of languages that exhibit a given order in a complex configuration of values that involve both the type of movement and the category being moved.

5.5 Conclusions

The two experiments described in the last two sections use a novel method to address the question of which syntactic operations are possible, which ones are not possible; and among those that are possible, which ones cost more than others. The results confirm previous theoretical proposals, but also explore new possibilities.

In particular, the ranking of costs among operations proposed by previous studies is mostly confirmed (Cinque 2005; 2013). The similarity of correlation between the two definitions of partial movement, movement of any category or movement of N, confirms that partial movement is penalised in any definition, and leaves the door open for a hitherto unexplained fact, namely that between the two fully harmonic word orders, the one that according to Cinque's theory requires movement (N ADJ NUM DEM) is more frequently attested among the world's languages than the one that does not require any movement (DEM NUM ADJ N).

The modelling of movement operations in Cinque's theory is based on some basic assumptions on the structural dominance of the syntactic categories that occur in the DP/NP. Cinque's (2005) account places numerals below demonstratives and above adjectives. For a similar assumption, see also Stavrou & Terzi (2008; 2009). But since numerals can appear in various syntactic contexts, it is not clear what syntactic category fits them best or what syntactic category would best fit our data. In the next two experiments, we investigate what syntactic category better fits the typological distributional data, by studying different structural positions for numerals.

6 Experiment 3: The structural position of numerals

We move, now, to the question of the status of numerals, and where they merge in the DP. One possibility is that numerals merge higher than adjectives, and below the demonstrative (Cinque 2005). A similar position is also argued for in Stavrou & Terzi (2008; 2009), who provide evidence from Greek. It is also assumed in much work that numerals are adjectives. For example, Landman (2000) proposes that all indefinite DPs are predicates of type $<e, t>$, which requires numerals to be adjectives, like *big* in (1-d). The data we have allows us to compare these two theories, whether numerals are quantifiers or adjectives, by testing the predictions on word order given each assumed position of the numeral.[6]

This experiment compares two possibilities: That numerals are higher in the structure than all adjectives, and that numerals are themselves adjectives, and are therefore higher than some adjectives and lower than others. The second assumption, treating numerals as adjectives, is tantamount to assuming two base orders, (19-a) and (19-b), and that every word order can be derived from either of the two base orders in (19). So each word order would correspond to two vectors: one corresponding to the movements needed to derive it from (19-a), and one corresponding to the movements needed to derive it from (19-b).

(19) a. $[_{\text{WP}}$ Dem $[_{\text{YP2}}$ Num $[_{\text{YP1}}$ Adj $[_{\text{NP}}$ N]]]]
b. $[_{\text{WP}}$ Dem $[_{\text{YP2}}$ Adj $[_{\text{YP1}}$ Num $[_{\text{NP}}$ N]]]]

It is crucial at this point to distinguish parametric movements that occur in order to license agreement or case-marking, and semantically-motivated movements, like Quantifier Raising. A parametric movement would occur in a language regardless of the intended interpretation. In contrast, a semantically-motivated movement only occurs when a special kind of interpretation (scope, collective/distributive/etc.) is intended, and will not typically occur in the dominant order of any language. While the former will be visible in typological data, the latter will not. For this reason, we now limit ourselves to parametric movements, and not semantically-motivated ones.

To test these two options, we need to encode different base orders, and the sequence of movements that would generate this order. The encodings of the derivations from (19-b) are shown in Tables 12 and 13. The encodings of the derivations from (19-a) were shown in Tables 2 and 3.

6.1 Materials

Experiment 1 already details the encoding of the assumption that numerals merge always above adjectives. In order to encode the idea that a numeral can merge either below or above adjectives, we encode each word order as two vectors, one that corresponds to the movements assuming the base order [Dem [Num [Adj [N]]]], i.e. those in Table 7; and one that corresponds to the movements assuming the base order [Dem [Adj [Num [N]]]], i.e. those in Tables 16 and 17.

[6] There are other proposals about numerals. For example, (Hurford 1975; 1987; 2003); Ionin & Matushansky (2004; 2006) treat all numerals as nouns, like *garden* in (1-c), albeit with the semantics of modifiers. Others, including Corbett (1978); Shlonsky (2004); Corver & Zwarts (2006); Danon (2012) propose that numerals do not constitute a uniform syntactic category. Finally, Scha (1984); Zabbal (2005); Ouwayda (2013; 2014); Ouwayda & Shlonsky (2015; 2017) propose that numerals are their own syntactic and semantic category, and cannot be reduced to quantifiers, adjectives, or nouns. Unfortunately, none of these alternatives can be tested with these data, because they do not contain information concerning N-N combinations.

Table 12: Movements necessary for each word order in Cinque's proposal (continued in next table), assuming DEM ADJ NUM N as base order. The table shows two lines for each word order, for each movement step. The first line describes the word order movement operation, the second line gives the name of the type of movement according to our formal encoding.

Word Order	Step 1	Step 2	Step 3
a. DEM NUM ADJ N	NP above NUM	NPless NumP above ADJ	No more mov'ts
	[NP[XP]]-Move	NPless Move	Partial mov't
b. DEM NUM N ADJ	NumP above ADJ	No more mov'ts	
	[XP[NP]]-Move	Partial movement	
c. DEM N NUM ADJ	NP above NUM	NumP above ADJ	No more mov'ts
	No-Pied-Piping	[NP[XP]]-Move	Partial mov't
d. N DEM NUM ADJ	NP above NUM	NumP above ADJ	NP splits above DEM
	No-Pied-Piping	[NP[XP]]-Move	Split Move
e. NUM DEM ADJ N	NP above NUM	NP-less NumP above DEM	
	[NP[XP]]-Move	NPless-Move	
f. NUM DEM N ADJ	NP above ADJ	NPless NumP above DEM	
	No-Pied-Piping	NPless-Move	
g. NUM N DEM ADJ	NumP above DEM		
	[XP[NP]]-Move		
h. N NUM DEM ADJ	NP above NUM	NumP above DEM	
	No-Pied-Piping	[NP[XP]]-Move	
i. ADJ DEM NUM N	NumP above ADJ	NPless ADJ above DEM	
	[XP[NP]]-Move	NPless-Move	
j. ADJ DEM N NUM	NP above NUM	NPless ADJ above DEM	
	[NP[XP]]-Move	NPless-Move	
k. ADJ N DEM NUM	N above DEM	ADJ above N	
	No-Pied-Piping	NPless Move	
l. N ADJ DEM NUM	N above ADJ	NumP above ADJ	NP above DEM
	No-Pied-Piping	NP-less Move	[NP[XP]]-Move

To differentiate the two vectors for each word order, we add the binary parameter *Involves the numeral merging below the adjectives*, which is added to the list of parameters in (20).

(20) a. Uses NP movement without pied-piping
b. Uses NP movement with pied-piping of the [XP[NP]] type
c. Uses NP movement with pied-piping of the [NP[XP]] type
d. Involves lack of movement (partial or complete)
e. Uses NP-splitting movement
f. Requires movement of a phrase not containing the NP
g. Involves the numeral merging below the adjectives

Now, instead of having one vector per word order, with this new encoding, any given *derivation* of a word order has its own unique vector, which is formed by assigning values to the attributes in (20). And every word order has two derivations (and therefore two vectors). Going back to our example illustration, for this proposal, the word order DEM N NUM ADJ has two associated vectors. The first vector represents the movements involved

in deriving it from the base order [$_{WP}$ Dem [$_{YP2}$ Num [$_{YP1}$ Adj [$_{NP}$ N]]]], and is given in Table 14. The second vector represents the movements involved in deriving it from the base order [$_{WP}$ Dem [$_{YP2}$ Adj [$_{YP1}$ Num [$_{NP}$ N]]]], and is given in Table 15. This results in two sets of encodings, one for each base order, shown in Tables 16 and 17.

Table 13: Movements necessary for each word order in Cinque's proposal (continued from Table 12), assuming DEM ADJ NUM N as base order. The table shows two lines for each word order, for each movement step. The first line describes the word order movement operation, the second line gives the name of the type of movement according to our formal encoding.

Word Order	Step 1	Step 2	Step 3	Step 4
m. DEM ADJ NUM N	No mov't			
n. DEM ADJ N NUM	NP above NUM	No more mov't		
	[NP[XP]]-Move	Partial mov't		
o. DEM N ADJ NUM	NP above ADJ	No move mov't		
	No-Pied-Piping	Partial mov't		
p. N DEM ADJ NUM	NP above DEM			
	No-Pied-Piping			
q. NUM ADJ DEM N	N above NUM	NumP above ADJ	NPless AP abv DEM	NUM above ADJ
	[NP[XP]]-Move	[NP[XP]]-Move	{Split, NPless Move}	NPless Move
r. NUM ADJ N DEM	NP above NUM	NUM above ADJ	AdjP above DEM	NumP above AP
	[NP[XP]]-Move	NPless Move	[XP[NP]]-Move	[XP[NP]]-Move
s. NUM N ADJ DEM	NUM P above ADJ	AdjP above DEM		
	[NP[XP]]-Move	[NP [XP]]-Move		
t. N NUM ADJ DEM	NP above NUM	NumP abv ADJ	AdjP abv DEM	
	[NP[XP]]-Move	[NP[XP]]-Move	[NP[XP]]-Move	
u. ADJ NUM DEM N	NP above ADJ	NPless AdjP abv DEM		
	No-Pied-Piping	NPless-Move		
v. ADJ NUM N DEM	AdjP above DEM			
	[XP[NP]]-Move			
w. ADJ N NUM DEM	NP above NUM	AdjP above DEM		
	[NP[XP]]-Move	[XP[NP]]-Move		
x. N ADJ NUM DEM	NP above ADJ	AdjP above DEM		
	No Pied-Piping	[NP[XP]]-Move		

Table 14: Encoding of first derivation of DEM N NUM ADJ as <0, 1, 0, 0, 1, 0, 0> (row c of Table 16), from DEM NUM ADJ N base.

	Feature	Value
g.	Involves the numeral merging below the adjectives	0
a.	Uses NP movement without pied-piping	1
b.	Uses NP movement with pied-piping of the [XP[NP]] type	0
c.	Uses NP movement with pied-piping of the [NP[XP]] type	0
d.	Involves lack of movement (partial or complete)	1
e.	Uses NP-splitting movement	0
f.	Requires movement of a phrase not containing the NP	0

Table 15: Encoding of second derivation of DEM N NUM ADJ as <1, 1, 0, 1, 1, 0, 0> (row c′ of Table 17), from DEM ADJ NUM N base.

Feature	Value
g. Involves the numeral merging below the adjectives	1
a. Uses NP movement without pied-piping	1
b. Uses NP movement with pied-piping of the [XP[NP]] type	0
c. Uses NP movement with pied-piping of the [NP[XP]] type	1
d. Involves lack of movement (partial or complete)	1
e. Uses NP-splitting movement	0
f. Requires movement of a phrase not containing the NP	0

Table 16: Movements needed for each order with [Dem [Num [Adj [N]]]] as base order.

	Num merge below	No pied piping	XP[NP]] moves	[NP[XP]] moves	Partial move	Split move	NPless move
a. DEM NUM ADJ N	0	0	0	0	0	0	0
b. DEM NUM N ADJ	0	0	0	1	1	0	0
c. DEM N NUM ADJ	0	1	0	0	1	0	0
d. N DEM NUM ADJ	0	1	0	0	0	0	0
e. NUM DEM ADJ N	0	0	1	0	0	0	1
f. NUM DEM N ADJ	0	0	0	1	0	0	1
g. NUM N DEM ADJ	0	0	0	1	0	1	1
h. N NUM DEM ADJ	0	0	0	1	0	1	1
i. ADJ DEM NUM N	0	0	0	1	0	0	1
j. ADJ DEM N NUM	0	1	0	0	0	0	1
k. ADJ N DEM NUM	0	1	1	0	0	0	0
l. N ADJ DEM NUM	0	1	0	1	0	0	0
m. DEM ADJ NUM N	0	0	0	1	1	0	1
n. DEM ADJ N NUM	0	0	1	0	1	0	0
o. DEM N ADJ NUM	0	0	0	1	1	0	0
p. N DEM ADJ NUM	0	0	0	1	0	1	0
q. NUM ADJ DEM N	0	1	0	0	0	0	1
r. NUM ADJ N DEM	0	0	1	0	0	0	0
s. NUM N ADJ DEM	0	0	1	1	0	0	0
t. N NUM ADJ DEM	0	1	1	0	0	0	0
u. ADJ NUM DEM N	0	0	0	1	0	1	1
v. ADJ NUM N DEM	0	0	0	1	0	1	1
w. ADJ N NUM DEM	0	0	1	1	0	0	0
x. N ADJ NUM DEM	0	0	0	1	0	0	0

6.2 Method

If numerals are adjectives, there are two possible derivations for each word order, as detailed in Tables 16 and 17. It is, therefore, no longer possible to establish which of these two derivations took place by simply looking at the resulting word order in any given

Table 17: Movements needed for each order with [Dem [Adj [Num [N]]]] as base order.

	Num merge below	No pied piping	[XP[NP]] moves	[NP[XP]] moves	Partial move	Split move	NPless move
a′. DEM NUM ADJ N	1	0	0	1	1	0	1
b′. DEM NUM N ADJ	1	0	1	0	1	0	0
c′. DEM N NUM ADJ	1	1	0	1	1	0	0
d′. N DEM NUM ADJ	1	1	0	1	0	1	0
e′. NUM DEM ADJ N	1	0	0	1	0	0	1
f′. NUM DEM N ADJ	1	1	0	0	0	0	1
g′. NUM N DEM ADJ	1	0	1	0	0	0	0
h′. N NUM DEM ADJ	1	1	0	1	0	0	0
i′. ADJ DEM NUM N	1	0	1	0	0	0	1
j′. ADJ DEM N NUM	1	0	0	1	0	0	1
k′. ADJ N DEM NUM	1	1	0	0	0	0	1
l′. N ADJ DEM NUM	1	1	0	1	0	0	1
m′. DEM ADJ NUM N	1	0	0	0	0	0	0
n′. DEM ADJ N NUM	1	0	0	1	1	0	0
o′. DEM N ADJ NUM	1	1	0	0	1	0	0
p′. N DEM ADJ NUM	1	1	0	0	0	0	0
q′. NUM ADJ DEM N	1	0	0	1	0	1	1
r′. NUM ADJ N DEM	1	0	1	1	0	0	1
s′. NUM N ADJ DEM	1	0	0	1	0	0	0
t′. N NUM ADJ DEM	1	0	0	1	0	0	0
u′. ADJ NUM DEM N	1	1	0	0	0	0	1
v′. ADJ NUM N DEM	1	0	1	0	0	0	0
w′. ADJ N NUM DEM	1	0	1	1	0	0	0
x′. N ADJ NUM DEM	1	1	0	1	0	0	0

language. For this reason, we partition the number of languages having a given observed word order in two different proportions assigned to the two derivations that result in that word order. Since we don't know which are the proportions that best predicts the data, we explore a representative sample of the space of proportions.

We sample all the proportions of derivations where all the observed word orders can be generated in the same proportions. We test all the proportions in increments (or decrements) of 10%. This gives eleven combinations (derivation1 100%, derivation2 0%; derivation1 90%, derivation2 10%, …, derivation1 0%, derivation2 100%). For example, the DEM NUM ADJ N order is generated 100% of the time by derivation $<0, 0, 0, 0, 0, 0, 0>$, and 0% of the time by derivation $<1, 0, 0, 1, 1, 0, 1>$. Then, it is generated 90% of the time by derivation $<0, 0, 0, 0, 0, 0, 0>$, and 10% of the time by derivation $<1, 0, 0, 1, 1, 0, 1>$, and so forth. All the other orders are generated by these same proportions.

Table 18: Sampling of the space of possibilities when treating the merge position of the numeral as a parametric choice.

	Assumed Base Order		
	DEM NUM ADJ N (**High NUM**)	**DEM ADJ NUM N** (**Low NUM**)	**Word orders**
Distr 0:	100% of languages	0% of languages	all word orders
Distr 0–1:	100% of languages	0% of languages	half word orders
	90% of languages	10% of languages	half word orders
Distr 1:	90% of languages	10% of languages	all word orders
Distr 1–2:	90% of languages	10% of languages	half word orders
	80% of languages	20% of languages	half word orders
Distr 2:	80% of languages	20% of languages	all word orders
…			
Distr 9:	10% of languages	90% of languages	all word orders
Distr 9–10:	10% of languages	90% of languages	half word orders
	0% of languages	100% of languages	half word orders
Distr 10:	0% of languages	100% of languages	all word orders

Table 19: Linear regression models generated by the assumption that numerals are adjectives, different proportions (Int. = intercept; Corr. = correlation).

	No pied piping	**[XP[NP]] moves**	**[NP[XP]] moves**	**No move**	**Split move**	**NP-less move**	**Int.**	**Corr.**
19–0	−120.8	−115.9	−33.5	−78.5	−92.2	−133.9	236.9	0.76
19–01	−100.9	−104.6	20.2	−70.9	−85.7	−126.3	213.0	0.69
19–1	−87.9	−97.1	–	−58.3	−91.8	−118.6	192.6	0.68
19–12	−76.9	−91.8	–	−52.7	−83.8	−113.4	181.6	0.64
19–2	−75.8	−91.9	–	−42.9	−85.5	−107.7	178.7	0.61
19–23	−65.4	−86.7	–	−37.6	−70.0	−78.2	168.1	0.57
19–3	−64.4	−86.3	–	−28.1	−79.3	−96.8	164.9	0.53
19–34	−40.0	−66.7	+25.9	–	−73.7	−86.7	124.9	0.49
19–4	−38.6	−64.9	+29.1	–	−76.9	−81.8	121.0	0.49
19–45	−27.7	−59.1	+34.6	–	−73.9	−77.3	109.0	0.47
19–5	−26.6	−56.9	+37.6	–	−76.6	−72.5	105.0	0.45
19–56	–	−43.5	+49.1	–	−71.4	−66.5	80.1	0.43
19–6	–	−40.7	+51.4	–	−74.0	−61.9	77.2	0.41
19–67	–	−41.3	+51.0	–	−72.9	−59.1	75.2	0.40
19–7	–	−35.1	+52.8	+28.8	−68.5	−50.1	62.5	0.41
19–78	–	−35.9	+52.5	+32.4	−67.1	−46.8	60.0	0.40
19–8	–	−32.1	+54.4	+40.2	−66.0	−40.9	53.3	0.41
19–89	–	−33.4	+54.0	+43.9	−64.6	−37.5	51.1	0.38
19–9	+26.9	–	+67.9	+55.4	−59.1	−23.3	15.9	0.39
19–910	+33.1	–	+68.3	+58.8	−58.9	−24.3	10.8	0.42
19–10	+32.8	–	+68.9	+65.9	−57.6	−18.6	−5.6	0.41

We also sample some mixed combinations of proportions where different observed word orders can be generated by the two derivations in different proportions. The eleven samples with fixed proportions of derivations described above range from a 0–100% combination in favour of one base order (high NUM) to a 100–0% in favour of the other base order (low NUM). This means that in the space of values of 10% increments, we have sampled every 24th possible sample. For example, between a 0–100% proportion for all observed word orders to be derived by a high numeral base order and 10–90% proportion also in favour of high numeral base, there are 23 intermediate sample points, the first being a sample of 10–90% proportion for the first observed word order to be derived by high numeral and 0–100% for all other 23 observed word orders, and the last being a sample of 10–90% proportion for all observed word orders but the last one. We have therefore added 10 more data samples meant to represent these intermediate proportions, where the observed word orders are generated half by one base order (high NUM) and the other half by the other base order (low NUM), giving us mixed samples where the percentages of proportion of use of the base order are not the same across all word orders, in increasing proportions (half 100–0%, and half 90–10%; half 90–10% and half 80–20%, …, half 10–90% and half 0–100%). This method introduces variation while allowing us to know exactly the proportion of the underlying base order.

Twenty-one files containing the data with both derivations with the weights in Table 18 were loaded into WEKA (Hall et al. 2009) and the best linear regression model is automatically generated for each of them.

6.3 Results

As it turns out, treating numerals as high-merging gives rise to better predictions than any of the formulas allowing numerals to merge either higher or lower than adjectives in the dominant base structure of a language. That is to say that the assumption that numerals merge higher than adjectives in the dominant merge order of all languages makes better predictions than the assumption that numerals merge higher than adjectives in the dominant merge order of some languages, and lower than adjectives in the dominant merge order of others. The result are detailed in the following two paragraphs.

The linear regression model that was generated by assuming numerals merge high is the function in (21), with correlation coefficient of 0.751. Notice that while the features are the same as those illustrated in Table 7, we repeat these same features here ten times, producing the same number of data points as the datasets whose results are shown in Table 20, for a fairer comparison.

(21) Frequency = –129.0 × Uses NP movement without pied-piping
–115.6 × Uses NP movement, pied-piping [XP[NP]]
–37.8 × Uses NP movement, pied-piping [NP[XP]]
–65.6 × Partial Move
–91.6 × Uses NP-splitting movement
–135.9 × Requires moving a phrase not containing NP
+242.8

The linear regression models that are generated by the assumption that numerals are adjectives are shown in Table 19, depending on the proportions of languages given to each derivation, given a word order.

6.4 Discussion

The results of this experiment show a sharp degradation in data correlation as the proportions of high-adjectival word orders increases followed by fluctuation between 0.38 and 0.41 for the datasets that have a majority of high-adjectival word orders. This shows that the dominant merge order in any given language for a DP containing a demonstrative, a numeral, an adjective, and a noun must be DEM NUM ADJ N, and not DEM ADJ NUM N. In other words, going back to the question of whether numerals merge strictly higher than adjectives in the dominant order of all languages or not, the results of this experiment suggest, to a certain extent, they do.

It is important to note, however, that the results showing that including the second merge order deteriorates the typological predictions does not automatically entail that it is not a possible merge order. As alluded to in section 5.4, it is possible for various merge orders and movements to take place for semantic reasons. For example, it is possible to merge the numeral lower than some adjectives for scope reasons. That, however, will not affect the dominant order in the language, as it will only occur in the (infrequent) cases of a "high adjective", like *last*, or of a collective interpretation of the adjective, like *heavy* in (22-c).

(22) a. the last three boxes
b. the three heavy boxes (each is heavy)
c. the heavy three boxes (they are heavy as a whole)

Whether these are cases of numerals merging lower than their usual position (Ouwayda 2013; 2014), or of adjectives merging higher cannot be entirely determined using typological facts, and require an in-depth semantic analysis that goes beyond this paper. We note, however, that assuming only the high positions for the numeral, rather than two positions with the lower one never occurring in the dominant order of a language, performs slightly worse.

6.5 Experiment 4: The structural position of numerals: classification

The previous experiments attempt to predict the actual counts of languages per word order. Not all languages, however, have been documented in the typological literature and, for those that have, there is some debate on what is the dominant word order. This experiment addresses the possible criticism that results of experiment 3 are too dependent on the actual counts of different word orders, which have a tendency to change when new languages are attested and documented. For this reason, we group the exact frequencies in discrete equivalence classes, and we repeat the experiment as a classification task. The experiment concentrates on predicting the ranking of word order counts and confirms the results of experiment 3.

6.5.1 Materials

The same encoded data is used as the previous experiments. The goal attribute, the attribute we are trying to predict, is a given word order's frequency class. We can group the languages in different frequency groups, by discretising the frequencies in different ways: either as simply possible or impossible (two values), as was the original goal in Cinque's paper, or as having different levels of frequency.

We performed the classification task at two different levels of discretized granularity for the frequency. Given that the data is actually distributed according to a powerlaw, as shown in Figure 8, we binned the languages into classes according to the magnitude

of their frequency. Using averages or medians would not have properly represented the fundamental fact that the frequencies are distributed exponentially. For two levels of granularity, the cut-off point is whether the number of languages per word order was in the double or triple digits, or in the single digits or zero. So *Frequent* is assigned to word orders that occur in more than 10 languages (14 word orders), and *Infrequent* for word orders that occur in exactly or less than 10 languages or are unattested (10 word orders).

For three levels of granularity, the same magnitude-based mapping is used. So *Very Frequent* is assigned to word orders that occur in more than 100 languages (5 word orders), *Frequent* to word orders that occur betwen 99 and 10 languages (9 word orders), *Rare or zero* for word orders that occur for less than ten languages (10 word orders, and, according to our frequencies, they are all unattested).

6.5.2 Method

Among the many available learning algorithms, we use a simple probabilistic learning algorithms, Naive Bayes, and *n*-fold cross-validation as the training and testing protocol (Russel & Norvig 1995).

In the Naive Bayes algorithm, the objective of training is to learn the most probable word order type given the probability of each vector of features. This probability is decomposed, according to Bayes rule, into the probability of the attributes given the goal predicate and the prior probability of the goal predicate itself. In our setting, the attributes are the movement operations and the goal predicate is the typological frequency. This method is chosen because despite its simplicity it works well in practice. Results will be compared to a baseline which consists in assuming that all word orders belong to the most frequent class. The baseline tells us whether the model has learnt anything beyond class frequency. The baseline consists in always predicting that languages are not attested in the three-way classification, or that they are infrequent in the two-way classification. We used the WEKA Data Mining Software (Hall et al. 2009), to run a Naive Bayes Classifier on each of the encodings of the data.

6.5.3 Results and discussion

The results in Table 20 suggest that treating numerals as adjectives fares worse than assuming that numerals merge higher than adjectives. The latter structure predicts the frequency classes of different word orders relatively well beyond the broad strokes of probable and improbable, distinction it reaches without error. This confirms the results of the previous experiment, that treating numerals as high-merging in the dominant order in a given language may be closer to the right track than treating their merge position as parametric.

7 Related work

Linguistic universals have been discussed from very many different points of view. We concentrate here on those that are directly related to some aspects of our proposal.

Table 20: Results of the classification tasks on Naive Bayes. Correctly classified instances in parentheses.

	High NUM	Low NUM
3 classes	79.2% (38/48)	62.5% (30/48)
2 classes	100% (48/48)	79.2% (38/48)
Baseline	58.3% (28/48)	58.3% (28/48)

7.1 Related computational work on Universal 20

In this work, we have proposed using statistical models and classifiers to automatically model and evaluate some quantitative and qualitative aspects of proposals concerning Universal 20. Two previous pieces of work have also used linear regression modeling and classification to evaluate different explanations of linguistic proposals, specifically concerning Universal 20.

Cysouw (2010a) uses Dryer's (2006) typological data on the order of the four elements demonstrative, numeral, adjective, noun, in the DP. Cysouw's model is factorial, and not derivational. The features used are all observed, i.e. they can be directly observed in the data, and are based on the position of the elements with respect to each other and to the edge of the phrase. 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 also proposes a factorial explanation based on general principles of symmetry and harmony (Dryer 2006). Extending these factorial approaches to probabilistic modeling, Merlo (2015) uses Bayes Networks to determine the generalization ability of each of Cinque's, Dryer's, and Cysouw's formalizations. She shows that while all three proposals fare well, Dryer's formalization has the best generalization ability.

Cinque's (2005) proposal derives the grammatically possible and impossible word orders, as traditional in generative accounts, but also derives the exceptions, and the different degree of markedness of the various possible orders. The proposal is carefully calibrated to obtain exactly this distribution of possible and impossible word orders, and their frequencies. But Cinque's account faces an empirical problem: the frequency distribution of the attested orders is Zipfian, with a few common orders accounting for most of the languages, and a long tail of rarer attestations. Given this distribution, it is reasonable to wonder whether unattested orders are really grammatically impossible or if they simply have not yet been attested, but are, in fact, possible. On the latter hypothesis, Cinque is "overfitting" the data and the predictions are too strong. This observation is, indeed, confirmed by Cinque's (2013) database that we use here: some of the order that Cinque (2005) indicated as unattested are, in fact, attested in this larger sample.

To avoid developing too strong a theory, other methods set out to explain the frequency differentials between pairs of word orders. In contrast to Cinque's single universal base order, Abels and Neeleman (Abels & Neeleman 2009) propose a simplified movement theory. This alternative proposal makes much more use of base orders. Abels and Neeleman (2009) describe a system where eight word orders are base generated by linearising an unordered hierarchical structure demonstrably homomorphic to Cinque's (2005). The remaining six orders are generated by leftward movement. They argue that their account is more parsimonious and more empirically adequate than the LCA or Cinque's.

Since our account does not hinge on the LCA per se, which we do not explicitly represent, nor on any specific assumptions on the internal structure of the phrases, which differ in the two accounts, it is not clear if our method can say anything interesting about the difference of the two theories.

Abels and Neeleman's account however is very restricted and makes interesting quantitative predictions. Specifically, it predicts, first, that all eight base orders should be roughly equivalent in frequency (counted by genera or by languages) and, second, that all six derived orders should be less frequent or, at most, as frequent as the base order from which they are generated. Typological counts are shown in Table 21.

A strict interpretation of the first prediction appears not to be fulfilled, especially if we look at the counts for genera. We can however notice that the base orders are all frequent,

Table 21: Abels and Neeleman's predictions (DGen = Dryer's genera; C13Lg = Cinque's (2013) languages).

Base order	DGen	C13Lg	Derived order	DGen	C13Lg
DEM NUM ADJ N	44	300	DEM N NUM ADJ t N	3	37
			N DEM NUM ADJ t N	3	48
			(ADJ N) DEM NUM t ADJ N	2	14
NUM ADJ N DEM	3	40	N NUM (ADJ t N) DEM	7	35
DEM ADJ N NUM	6	35			
ADJ N NUM DEM	1	15			
DEM NUM N ADJ	17	114	(N ADJ) DEM NUM t N ADJ	11	69
NUM N ADJ DEM	21	180			
DEM N ADJ NUM	22	125	N DEM (t N ADJ) NUM	4	24
N ADJ NUM DEM	57	411			

and none is rare, which seems instead to support the theory. Prediction two is largely confirmed. In only one case, the derived order does not appear to be convincingly more costly than its corresponding base order (NUM ADJ N DEM, 3 genera and 40 languages, compared to the derived N NUM ADJ DEM, 7 genera and 35 languages).

Future work will develop a systematic comparison of these two approaches. Since the two approaches have a very considerable difference in complexity, with Cinque's allowing several types of movement, the comparison will have to include some measure of model complexity (for example by using Bayes factors) and is beyond the scope of the current article.

One of the clearest observations concerning Universal 20 above is that the two most frequently attested word orders are harmonic. This is a wide-spread observation in typology. While it is well-known that disharmonic patterns in the order of words exist, and, in fact most languages are not fully harmonic (Dryer 1992: fn12), typologists have long noted a preference for harmonic word orders, that is word orders in which the operand and the operator (i.e, complements and heads) are in the same sequence. The greater frequency of harmonic word orders follows in Cinque's and our model from the smaller costs of the operations that are used to derive them. Other explanations for the greater attestation of harmonic word orders are rooted in grammatical principles (FOFC, for example, proposed in Biberauer et al. 2008); general learning and regularisation biases (Culbertson & Smolensky 2012; Culbertson et al. 2012); parsing or processing principles (Hawkins 2004; Sheehan 2013), such as, recently proposed Dependency Length Minimisation effects in the Noun Phrase (Gulordava & Merlo 2015; Gulordava et al. 2015).

A fundamental assumption of this paper is that frequencies within a language and across languages are an aspect of language that is systematically related to its formal properties and that requires explanation, both in its numerical magnitude and its distribution. Like Yang (2011), we attract attention to the explanatory relevance of probabilistic accounts and expected frequencies of linguistic facts.

In this respect our work is also related to other computational probabilistic proposals for language universals. In a widely discussed, and controversial, paper, Dunn et al. (2011) investigate the correlations between word orders that have been used to argue in favour of the existence of language universals (Dryer 1992). They claim that to demonstrate that these correlations reflect underlying cognitive or system biases, the languages must be sampled in a way that controls for features linked only by direct inheritance from

a common ancestor. Correlation between two word orders may suggest that both are responding to some common evolutionary force or that one acts as a selective force for change in the other. To answer these questions, they develop a computational model that simultaneously accounts for phylogenetic uncertainty and estimates the posterior distributions of the parameters of the model of word order evolution.

This model is computationally sophisticated, but it aims to explain linguistic data that are relatively simple. Our approach shifts the focus of the investigation, and is computationally simpler, but linguistically more detailed. Although very simple, Naive Bayes belongs to the same class of models as it is a Bayesian model with latent variables, but our encoding of word orders is more distributed and encodes a derivational theory. To a large extent, our model confirms Cinque's approach and its differential cost of movement operations. These differential costs were set up to explain a pervasive universal asymmetry between prenominal and postnominal modifiers, the asymmetry between the very restricted choice of word orders among prenominal modifers and the much larger set of options for postnominal modifiers. This asymmetry is not captured by any of the other models that allow symmetric base orders.

7.2 Related theoretical work

The syntactic category of numerals has been in debate in recent years, with proposals claiming that numerals are quantifiers (Stavrou & Terzi 2007, 2008), adjectives (Corbett 1978; Landman 2000), or nouns (Hurford 1975; 1987; 2003; Ionin & Matushansky 2006). We have focused here on comparing the predictions of treating numerals as quantifiers, or treating them as adjectives. The claims that numerals are quantifiers or adjectives have been supported by intralinguistic facts, including agreement, case marking, and scope.

The idea that at least some numerals are adjectives is based on a number of properties of numerals that are typical of adjectives. Corbett (1978) notes that, like adjectives, low numerals in Russian, for example, agree with the following noun (23) in certain features including gender, number, case, and animacy.

(23) a. odin zurnal
one.MASC magazine.MASC
'one magazine'

b. odna gazeta
one.FEM newspaper.FEM
'one newspaper'

Stavrou & Terzi (2008), in contrast, propose that numerals are a subclass of weak quantifiers, endowed with a [card] feature and are located in the specifier of a number phrase. They base this on a number of similarities between numerals and weak quantifiers. They note that many of the resemblances between numerals and adjectives also apply to quantifiers. For example, like adjectives, numerals and quantifiers can both occur in existential *there*-sentences (24). They can both be preceded by determiners (25).

(24) There are many/three books on the table.

(25) The many/three books are on the table

Also, in negated contexts, both numerals and quantifiers have scope ambiguities, as in (26). The sentence can mean either 'There are three/many women I didn't see' or 'I didn't see many/three women, I saw few/five'.

(26) I didn't see many/three women.

In addition, they show that numerals and quantifiers share properties that adjectives do not have, namely that cardinals and quantifiers are both able to license bare subjects in Greek (27). They can both appear without a noun, while an adjective cannot (28). They can head a partitive construction which an adjective cannot (29). And they both allow for split topicalization in Greek which an adjective cannot (30).

(27) Tris/ligi fitites parusiasan to arthro.
three/few students presented the article
'Three/few students presented the article.'

(28) I met three/many/*tall.

(29) Many/three/*tall of the demonstrators caused trouble.

(30) Vivlia agorasa merika/lika/deka.
books bought.1s several/few/ten
'Books I bought several/few/ten.'

Many researchers view numerals as not having a uniform category, including Corbett (1978); Shlonsky (2004); Zweig (2005); Corver & Zwarts (2006); Danon (2012), but rather as being split into adjectives and nouns, or adjectives and quantifiers. Our reasoning in this paper assumes that numerals do form a uniform category. This is the same assumption that is made when they are referred to uniformly in the typological data.

There is much intra-linguistic evidence for more than one position for numerals. Researchers, including Zabbal (2005); Ouwayda (2013; 2014) propose that numerals do not have a single merge position in the DP, but rather can merge either higher or lower than adjectives. In Ouwayda (2013; 2014), for example, there are two functional projections that can host a numeral: A high one, which behaves much like quantifiers, and is associated mainly with a distributive interpretation, and another low position associated with a collective interpretation (like in sentence 20c)). It is important to note, however, that the two merge positions motivated in these works are semantically motivated. For Ouwayda (2013; 2014) for example, the difference is in distributivity. Similarly, in Pereltsvaig (2006); Ionin & Matushansky (2006), and Matushansky (2015), Russian numerals can have different positions depending on whether they are interpreted approximately or exactly. It is therefore crucial to distinguish between assuming multiple positions for numerals that are decided parametrically versus multiple positions for numerals that are semantically motivated within a given language.

Our work does not rule out the possibility of a lower merge position that is motivated semantically. It specifically rules out the possibility that different languages have different merge positions. The confirmation that the high merge position for numerals corresponds to the dominant order in all the languages, combined with intra-linguistic evidence of word orders that require a low merge position for numerals, lends support to the idea that the low merge position for numerals is semantically-motivated.

8 General discussion and conclusion

In this paper, we set out to compare and test different syntactic proposals concerning Universal 20 using vectorial representations and machine learning methods. Specifically, we set out to answer the following three questions:

- Can Cinque's ranking of the different kinds of movements be predicted automatically using Universal 20?
- Is movement always costly? (Is lack of movement always the less costly route?)
- Is the base structure proposed by Cinque the best predictor of the typological facts?

Since the syntactic proposals we are modelling are fairly complex, we kept the modelling as simple and as faithful as possible. We modelled the 24 possible permutations of DEM NUM ADJ N based on each of the different syntactic proposals, as a series of vectors of binary parameters. These binary parameters correspond to the different movements and merge positions that are assumed to lead to each of the 24 possible word orders given each of the syntactic proposals.

We then used linear regression in order to determine the weights of the different syntactic movements proposed in Cinque (2005). The results accurately predicted the key *costly* syntactic operations, but ranked them somewhat differently from Cinque surprisingly penalizing movement without pied-piping, and split movement, more than movement of a phrase not containing the NP. We used linear regression again in our second experiment to compare Cinque's (2005) proposal that penalizes partial movement to an identical analysis that penalizes lack of movement altogether, consistent with Shlonsky (2012) proposal that agreement-rich languages require movement. The results were similar, but fit the data slightly worse, suggesting that movement is not the less costly route. Finally, we used both linear regression and classification to compare the predictive power of assuming a high merge position for numerals as assumed by Cinque (2005), versus the predictive power of assuming a low merge position, as would be expected if numerals were adjectives. We also considered twenty-one intermediate points of view, where the numeral merges high in the dominant order of some languages and low in others. The high merge position of numerals had consistently better results. Given the limitations of the results for linear regression, a fourth experiment was conducted to replicate the third, but as a classification task. The results confirm experiment 3. They show that merging numerals lower does not fare as well in predicting the distinction between probable and improbable word orders as assuming that numerals are quantifiers. High merging of numerals, instead, correctly predicts finer-grained distinctions between the frequency classes among the more probable word orders, thus confirming the less confident result of experiment 3. Our experiments confirming that a high merge position for numerals is a better predictor of typological data lend support to the idea that the dominant merge order is semantically motivated.

Our large-scale automatic investigation allows us to discover some facts that would not have been accessible by more traditional methods. Determining the weights of the movement operations and establishing the preferred merge sequence requires computations that exhaustively explore the space of options and calculate the optimal solutions over the space of all languages, computations that are too costly to be done by hand and that would not be informative if done on a small scale.

Specifically, there has been evidence that numerals can merge lower than adjectives in some contexts, Ouwayda (2013; 2014); Ouwayda & Shlonsky (2015; 2017) among others. The conclusion that this low merger is semantically motivated rather than parametric could not have been reached without exploring the space of possibilities where different languages are given the option of having either a high or a low numeral in their base structure.

Abbreviations

DEM, Dem = demonstrative, NUM, Num = numeral, ADJ, Adj = adjective, N, N = noun, MASC = masculine, FEM = feminine, 1S = first person singular, [XP[NP]] = movement of a constituent containing the NP with pied-piping of the *picture of who* type, [NP[XP]] = NP movement with pied-piping of the *whose picture* type

Acknowledgements

We are very grateful to Guglielmo Cinque for giving us acces to his data and to Giuseppe Samo, for his very helpful attentive reading and comments. All remaining errors are our own.

Funding Information

The research described in this paper was partially funded by the Swiss NSF under grant 144362.

Competing Interests

The authors have no competing interests to declare.

Author Contributions

The order of authors is alphabetical to indicate equivalent contributions of the two authors. Most of the work by Sarah Ouwayda was performed while at the university of Geneva.

References

Abels, Klaus & Ad Neeleman. 2009. Universal 20 without the LCA. In José M. Brucart, Anna Gavarró & Jaume Solà (eds.), *Merging features: Computation, interpretation, and acquisition*, 60–79. Oxford: Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780199553266.003.0004

Biberauer, Theresa, Anders Holmberg & Ian Roberts. 2008. Structure and linearization in disharmonic word orders. In *26th West Coast Conference on Formal Linguistics*, 96–104. Berkeley, CA: Cascadilla Press.

Cinque, Guglielmo. 2005. Deriving Greenberg's universal 20 and its exceptions. *Linguistic Inquiry* 36(3). 315–332. DOI: https://doi.org/10.1162/0024389054396917

Cinque, Guglielmo. 2013. On the movement account of Greenberg's universal 20: Refinements and replies: Materials. Manuscript, University of Venice.

Corbett, Greville G. 1978. Universals in the syntax of cardinal numerals. *Lingua* 46(4). 355–368. DOI: https://doi.org/10.1016/0024-3841(78)90042-6

Corver, Norbert & Joost Zwarts. 2006. Prepositional numerals. *Lingua* 116(6). 811–835. DOI: https://doi.org/10.1016/j.lingua.2005.03.008

Culbertson, Jennifer & David Adger. 2014. Language learners privilege structured meaning over surface frequency. *Publications of the National Academy of Science* 111(16). 5842–5847. DOI: https://doi.org/10.1073/pnas.1320525111

Culbertson, Jennifer & 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. DOI: https://doi.org/10.1111/j.1551-6709.2012.01264.x

Culbertson, Jennifer, Paul Smolensky & Geraldine Legendre. 2012. Learning biases predict a word order universal. *Cognition* 122(3). 306–329. DOI: https://doi.org/10.1016/j.cognition.2011.10.017

Cysouw, Michael. 2010a. Dealing with diversity: Towards an explanation of NP word order frequencies. *Linguistic Typology* 14(2). 253–287. DOI: https://doi.org/10.1515/lity.2010.010

Cysouw, Michael. 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)*, 29–35. Berlin, Heidelberg: Springer-Verlag. DOI: https://doi.org/10.1007/978-3-642-14322-9_3

Danon, Gabi. 2012. Two structures for numeral-noun constructions. *Lingua* 122(12). 1282–1307. DOI: https://doi.org/10.1016/j.lingua.2012.07.003

Dryer, Matthew S. 1992. The Greenbergian word order correlations. *Language* 68. 81–138. DOI: https://doi.org/10.2307/416370

Dryer, Matthew S. 2006. The order demonstrative, numeral, adjective and noun: An alternative to Cinque. http://exadmin.matita.net/uploads/pagine/1898313034_cinqueH09.pdf.

Dunn, Michael, Simon J. Greenhill, Stephen C. Levinson & Russell D. Gray. 2011. Evolved structure of language shows lineage-specific trends in word-order universals. *Nature* 473. 79–82. DOI: https://doi.org/10.1038/nature09923

Greenberg, Joseph H. 1966. *Language universals*. The Hague, Paris: Mouton.

Gulordava, Kristina & Paola Merlo. 2015. Structural and lexical factors in adjective placement in complex noun phrases across Romance languages. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, 247–257. Beijing, China: Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/K15-1025

Gulordava, Kristina, Paola Merlo & Benoit Crabbé. 2015. Dependency length minimisation effects in short spans: A large-scale analysis of adjective placement in complex noun phrases. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL/CoNLL'15) (volume 2: Short papers)*, 477–482. Beijing, China: Association for Computational Linguistics. DOI: https://doi.org/10.3115/v1/P15-2078

Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann & Ian H. Witten. 2009. The Weka data mining software: An update. *SIGKDD Explorations Newsletter* 11(1). 10–18. DOI: https://doi.org/10.1145/1656274.1656278

Hawkins, John A. 2004. *Efficiency and complexity in grammars*. Oxford: Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780199252695.001.0001

Hurford, James R. 1975. *The linguistic theory of numerals*. Cambridge: Cambridge University Press.

Hurford, James R. 1987. *Language and number: The emergence of a cognitive system*. Oxford: Basil Blackwell.

Hurford, James R. 2003. The interaction between numerals and nouns. In Frans Plank (ed.), *Noun Phrase structure in the languages of Europe* 20. 561–620.

Ionin, Tania & Ora Matushansky. 2004. A singular plural. In *Proceedings of West Coast Conference on Formal Linguistics* 23. 399–412. Davis, CA: Cascadilla Press.

Ionin, Tania & Ora Matushansky. 2006. The composition of complex cardinals. *Journal of Semantics* 23(4). 315–360. DOI: https://doi.org/10.1093/jos/ffl006

Kayne, Richard. 1994. *The antisymmetry of syntax*. Cambridge, MA: MIT Press.

Landman, Fred. 2000. *Events and plurality: The Jerusalem lectures.* Deventer: Kluwer. DOI: https://doi.org/10.1007/978-94-011-4359-2

Matushansky, Ora. 2015. On Russian approximative inversion. In Gerhild Zybatow, Petr Biskup, Marcel Guhl, Claudia Hurtig, Olav Mueller-Reichau & Maria Yastrebova (eds.), *Slavic Grammar from a Formal Perspective. The 10th Anniversary FDSL Conference*, 303–316. Bern: Peter Lang.

Merlo, Paola. 2015. Predicting word order universals. *Journal of Language Modelling* 3(2). 317–344. DOI: https://doi.org/10.15398/jlm.v3i2.112

Ouwayda, Sarah. 2013. Where plurality is: Agreement and DP structure. In Stefan Keine & Shayne Sloggett (eds.), *Proceedings of the 42nd North Eastern Linguistics Society meeting*, 423–436. Toronto, Canada: GLSA.

Ouwayda, Sarah. 2014. *Where number lies: Plural marking, numerals, and the collective–distributive distinction*. Los Angeles, CA: University of Southern California dissertation.

Ouwayda, Sarah & Ur Shlonsky. 2015. Order in the DP! on word order and structure in the DP. *LSA Annual Meeting extended abstracts*. Portland, OR.

Ouwayda, Sarah & Ur Shlonsky. 2017. Word order variation in Lebanese Arabic DPs: In support of low numerals. *Linguistic Inquiry* 48(1). 181–193. DOI: https://doi.org/10.1162/LING_a_00240

Pereltsvaig, Asya. 2006. Passing by cardinals: In support of head movement in nominals. In James Lavine, Steven Franks, Mila Tasseva-Kurktchieva & Hana Filip (eds.), *Proceedings of FASL 14: The Princeton meeting*, 277–292. Ann Arbor, Michigan: Michigan Slavic Publications.

Russel, Stuart & Peter Norvig. 1995. *Artificial intelligence: A modern approach*. Upper Saddle River, NJ: Prentice Hall.

Scha, Remko. 1984. Distributive, collective and cumulative quantification. *Truth, Interpretation, and Information, GRASS* 2. 131–158.

Sheehan, Michelle. 2013. Explaining the Final-over-Final Constraint: formal and functional approaches. In Theresa Biberauer & Michelle Sheehan (eds.), *Theoretical Approaches to Disharmonic Word Order*, 407–444. Oxford: Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780199684359.003.0015

Shlonsky, Ur. 2004. The form of Semitic noun phrases. *Lingua* 114(12). 1465–1526. DOI: https://doi.org/10.1016/j.lingua.2003.09.019

Shlonsky, Ur. 2012. On some properties of nominals in Hebrew and Arabic, the construct state and the mechanisms of AGREE and MOVE. *Italian Journal of Linguistics* 24(2). 267–286.

Stavrou, Melita & Arhonto Terzi. 2008. Types of numerical nouns. In *Proceedings of the 26th West Coast Conference on Formal Linguistics*, 429–437. Berkeley, CA: Cascadilla Press.

Stavrou, Melita & Arhonto Terzi. 2009. Cardinal numerals and other numerical expressions. Manuscript, Aristotle University of Thessaloniki & Technological Educational Institute of Patras.

Steddy, Sam & Vieri Samek-Lodovici. 2011. On the ungrammaticality of remnant movement in the derivation of Greenberg's universal 20. *Linguistic Inquiry* 42(3). 445–469. DOI: https://doi.org/10.1162/LING_a_00053

Steedman, Mark. 2011. Greenberg's 20th: The view from the long tail. Manuscript, University of Edinburgh.

Yang, Charles. 2011. A statistical test for grammar. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, 30–38. Portland, OR: Association for Computational Linguistics.

Zabbal, Youri. 2005. *The syntax of numeral expressions*. Amherst, MA: University of Massachusetts generals paper.

Zweig, Eytan. 2005. Nouns and adjectives in numeral NPs. In *Proceedings of the North Eastern Linguistics Society Meeting* 35. 663–676. Storrs, CT: GLSA.

How to cite this article: Merlo, Paola and Sarah Ouwayda. 2018. Movement and structure effects on Universal 20 word order frequencies: A quantitative study. *Glossa: a journal of general linguistics* 3(1): 84. 1–35, DOI: https://doi.org/10.5334/gjgl.149

Submitted: 18 May 2016 **Accepted:** 10 January 2018 **Published:** 03 August 2018

]u[*Glossa: a journal of general linguistics* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS