Advances in Complex Systems, Vol. 11, No. 3 (2008) 415–420
© World Scientific Publishing Company

LINEAR ORDER AS A PREDICTOR OF WORD ORDER REGULARITIES

MICHAEL CYSOUW
Max Planck Institute for Evolutionary Anthropology,
Deutscher Platz 6,
04103 Leipzig, Germany
cysouw@eva.mpg.de

Received 20 March 2008

This is a reply to Ramon Ferrer-*I*-Cancho's paper in this issue "Some Word Order Biases from Limited Brain Resources: A Mathematical Approach." In this reply, I challenge the Euclidean distance model proposed in that paper by proposing a simple alternative model based on linear ordering.

Keywords: Linear order; word order; language universals; human language.

1. The Linear Order Model

The paper by Ramon Ferrer-i-Cancho (henceforth RF) takes a novel and interesting approach to the study of word order regularities. In particular, the principle of looking for the minimal linear arrangement as a constraint on the coding of the nonlinear structure of thought is an interesting proposal. However, I think that the argumentation put forward in the paper is not convincing. The problem is that the model presented by RF is needlessly complex, and much better predictions can be made with much simpler models. This does not mean that RF's model does not have its merits. I find the notion of minimizing Euclidean distance between linguistic entities highly interesting, and I can see many interesting applications for it. However, it does not seem to be very well suited for modeling word order preferences among the world's languages.

Let me sketch an alternative model, which has actually (more or less explicitly) been the standard model for explaining word order preferences in the field of linguistic typology ever since the original papers of Greenberg in the 1960s (see Refs. 2, 10–12, 14 and 15, just to cite the classic works on this topic). Consider the following model, which states that the *linear order* of constituents is important, and not the *Euclidean distance* between constituents (as RF suggests). To be precise, the model would say something like the following (ignoring free word order):

- SV is very strongly preferred to VS;
- SO is strongly preferred to OS;
- There is no preference for OV or VO.

This very simple model predicts the following relative frequencies of word order types (from most frequent to least frequent):

$$SVO, SOV > VSO > OSV > VOS, OVS. \tag{1}$$

In this model one can even try to match the attested frequencies [6] by fitting the following probabilities:

- SV order is preferred in 96.3% of the languages to VS;
- SO order is preferred in 88.8% of the languages to OS.

This model predicts the relative frequencies of the word order types, as shown in Table 1. For example, the predicted frequency of the combination of VS and SO order (resulting in VSO order) is calculated by multiplying the respective probabilities $(1 - 0.888) \cdot 0.963 \cdot 100\ \% = 10.8\ \%$. However, there is one slight complication. Note that the two pairwise choices SV/VS and SO/OS only specify four possible combinations, so this model is underdetermined for the six possible word orders. For example, consider a language that chooses the pairwise orders SV and SO. Such a language could be either SVO or SOV. The model predicts a probability of $0.888 \cdot 0.963 = 0.855$ for this combination, which is then simply equally distributed over both SVO and SOV (i.e. each get a predicted frequency of 42.8 %). Although this is not completely correct (SOV is actually slightly more frequent than SVO), the result is a very accurate prediction of the attested frequencies from Ref. 6 (ignoring the free word order languages). The correlation between the attested and the predicted frequencies is almost perfect ($r^2 = 0.98$; $p < 0.0001$).

When this is contrasted with the basic prediction of RF's model (viz. SVO and OVS are more economical than the other orders), I fail to see the rationale for invoking the complex mathematical machinery that is used. Even by adding a completely ad-hoc "broken symmetry" condition in the final section of the paper, the predicted worldwide frequencies (2) still only vaguely match the empirical situation

Table 1. Predictions of relative word order frequencies.

Word order	Predicted	Attested
SOV	42.8%	47.1%
SVO	42.8%	41.2%
VSO	10.8%	8.0%
VOS	0.2%	2.4%
OVS	0.2%	0.8%
OSV	3.3%	0.4%

(see the top-down order in Table 1):

$$SVO > SOV, VSO, OSV, VOS > OVS. \tag{2}$$

2. The Order of Other Elements

Let me add two notes on other details from RFs paper. The first note concerns the following quotation from the end of Sec. 4: "In recent studies of the ordering of S, V and O, the dominant word order of a language is not defined as the most frequent word order globally but as the most frequent word order of declarative sentences (not interrogatives or exclamatives), with the further constraint that S and O cannot be a pronoun (Our approach to the ordering of S, V and O that we will introduce in this article is not limited to these particular cases)."

Actually, two of the topics addressed here have received a considerable amount of attention in cross-linguistic studies. As for the order of elements in questions, this involves various other factors, including a widespread tendency for interrogative words to be fronted, independent of their argument status in the sentence [8]. Further, there are many different ways to mark sentences as polar questions, but only a vanishing minority of the world's languages use a change in word order (about 1%, mainly European languages) [7]. In both cases, I do not immediately see any profit in using the distance between constituents as an explanatory factor for the cross-linguistic preferences.

As for the usage of pronouns, the problem here is that many languages regularly "drop" them, so it is often difficult to make any claim about their order. The more common situation is to use inflectional cross-referencing on the verb as the main marker, at least for S [4]. Many languages even have inflectional markers for both S and O (almost 51% of the world's languages, according to the data in Ref. 13). For these languages, it is interesting to look at the relative order of the inflectionally marked S and O, relative to the verb root (for about 40% of the world's languages it is possible to establish this order unequivocally). The attested cross-linguistic frequencies are shown in the second columns of Table 2, adapted from Ref. 13. These numbers suggest that SVO order is by far the most frequent for affixes (3–4 times as frequent as the other possibilities), with all other orders being roughly equally frequent. Although the high frequency of SVO would fit in with the approach proposed by RF, the more striking observation is that the attested frequencies of

Table 2. Ordering frequencies of affixes.

Affix ordering	Attested	Predicted
SVO	63 (40.9%)	40.3%
OVS	21 (13.6%)	13.3%
SOV	19 (12.3%)	11.8%
VOS	19 (12.3%)	11.3%
OSV	17 (11.0%)	11.8%
VSO	15 (9.7%)	11.3%

the ordering of affixes do not correlate with the attested frequencies of the ordering of constituents (cf. the third column of Table 1 with the second column of Table 2: $r^2 = 0.31$; $p = 0.25$). This indicates that one single overarching approach is not sufficient to explain cross-linguistic ordering preferences (as suggested by RF in the quoted sentence from Sec. 4).

The fact that the ordering preferences for cross-referencing affixes are different from the order of major sentence constituents implies that the model that I proposed above for constituent order will not work for affixes. However, it is possible to make a slightly different model in the same spirit to fit the attested frequencies of affix ordering, e.g.:

- *SV* order is preferred in 64.3% of the languages over *VS*
- *VO* order is preferred in 63.0% of the languages over *OV*

This model nicely predicts the attested frequencies, as shown in the third column of Table 2 ($r^2 = 0.99$, $p < 0.0001$). It is enlightening to compare the two models for the different kinds of ordering. Apparently, to be able to model the order of cross-referencing affixes, it is important to specify the order of *S* and *O* relative to *V*. In contrast, for sentence constituent order it is important to specify the order of *O* and *V* relative to *S*. This might be interpreted as showing that the verb is the pivotal element for affix ordering and the subject is the pivotal element for sentence constituent ordering. Further, note that the asymmetry between the alternatives is much larger for constituent ordering than for affix ordering, which might be interpreted that the left-right ordering is more important for sentence constituents than for affixes.

3. Interactions Between Orders

My second note concerns the appendix, added to the paper to justify a purported language universal "with overwhelmingly more than chance frequency, languages with dominant order *SVO* have the adjective after the noun and languages with dominant order *SOV* have the adjective before the noun." To claim such a correlation is rather audacious, given the repeated rebuttals of closely related claims by Matthew Dryer [1–3, 9]. At face value, the numbers presented by RF, taken from Refs. 5 and 6, seem convincing. I have repeated the numbers in a slightly different presentation in Table 3. The numbers shown indicate the number of languages, and

Table 3. *SVO*/*SOV* vs. *NA*/*AN* counting languages.

	NA	*AN*	other
SOV	223 (−40.6)	166 (+50.4)	21 (−9.8)
SVO	303 (+56.8)	56 (−52.0)	24 (−4.8)
other	142 (−16.2)	81 (+1.6)	33 (+14.4)

Table 4. *SVO/SOV* vs. *NA/AN* counting genera.

	NA	*AN*	other
SOV	113 (+2.6)	65 (+5.9)	17 (−8.5)
SVO	76 (+8.6)	25 (−11.0)	18 (+2.4)
other	84 (−11.2)	56 (+5.1)	28 (+6.1)

the differences from the statistical expectation are added in brackets. These differences are impressive, and so is the statistical significance ($\chi^2 = 72.9$; $p < 10^{-15}$ against the null hypothesis that the distribution of the cell counts in the table is the product of the row and column marginals). So, why has Dryer argued so fervently against such correlations? The main problem with these numbers is that they represent counts of languages, and there are very many languages in this particular data set that are closely related. Doing the same counts, but now counting the number of genera (a genus is a low level genetic group roughly of the time depth of linguistic families like Germanic or Romance), results in the numbers in Table 4. This time the differences from the statistical expectation do not look very impressive, and neither does the statistical significance ($\chi^2 = 11.8$; $p < 0.05$).

Acknowledgments

I thank the editors of *ACS* and Ramon Ferrer-i-Cancho for giving me the possibility to write this reply. The points raised in this reply originated in the reviewing process of the paper by Ferrer-i-Cancho. The discussion that arose in that closed discussion was deemed to be interesting enough to be made available in this form to all readers of this journal.

References

[1] Dryer, M. S., SVO languages and the OV:VO typology, *J. Ling.* **27** (1991) 443–482.

[2] Dryer, M. S., The Greenbergian word order correlations, *Language: J. Ling. Soc. Am.* **68** (1992) 80–138.

[3] Dryer, M. S., Why statistical universals are better than absolute universals, *Chicago Ling. Soc.* **33** (1997) 123–145.

[4] Dryer, M. S., Expression of pronominal subjects, in *World Atlas of Language Structures*, eds. Haspelmath, M., Dryer, M. S., Gil, D. and Comrie, B. (Oxford University Press, Oxford, 2005), pp. 410–413.

[5] Dryer, M. S., Order of adjective and noun, in *World Atlas of Language Structures*, eds. Haspelmath, M., Dryer, M. S., Gil, D. and Comrie, B (Oxford University Press, Oxford, 2005), pp. 354–357.

[6] Dryer, M. S., Order of subject, object and verb, in *World Atlas of Language Structures*, eds. Haspelmath, M., Dryer, M. S., Gil, D. and Comrie, B. (Oxford University Press, Oxford, 2005), pp. 330–333.

[7] Dryer, M. S., Polar questions, in Haspelmath, M., Dryer, M. S., Gil, D. and Comrie, B. (eds.), *World Atlas of Language Structures* (Oxford University Press, Oxford, 2005), pp. 470–473.

[8] Dryer, M. S., Position of interrogative phrases in content questions, in *World Atlas of Language Structures*, eds. Haspelmath, M., Dryer, M. S., Gil, D. and Comrie, B. (Oxford University Press, Oxford, 2005), pp. 378–381.

[9] Dryer, M. S., Relationship between the order of object and verb and the order of adjective and noun, in *World Atlas of Language Structures*, eds. Haspelmath, M., Dryer, M. S., Gil, D. and Comrie, B. (Oxford University Press, Oxford, 2005), pp. 394–397.

[10] Greenberg, J. H., Some universals of grammar with particular reference to the order of meaningful elements, in *Universals of Language*, ed. Greenberg, J. H. (MIT Press, Cambridge, Mass, 1963), pp. 73–113.

[11] Greenberg, J. H., *Language Universals: With Special Reference To Feature Hierarchies* (Mouton, The Hague, 1966).

[12] Hawkins, J. A., *Word Order Universals* (Academic Press, New York, 1983).

[13] Siewierska, A., Order of person markers on the verb, in *World Atlas of Language Structures*, eds. Haspelmath, M., Dryer, M. S., Gil, D. and Comrie, B. (Oxford University Press, Oxford, 2005), pp. 422–425.

[14] Tomlin, R. S. *Basic Word Order: Functional Principles* (Croom Helm, London, 1986).

[15] Vennemann, T. *Language Type and Word Order* (LAUT, Trier, 1973).