

Author: Harald Hammarström
University/Affiliation: Chalmers University
Email address: harald2@cs.chalmers.se

Basic Word Order Frequencies and Transition Probabilities in the Languages of the World

Traditionally typologists look at frequencies of various types of languages of the world to gain insight about possible human languages. At least potentially, this reflection might be skewed by “historical accidents” that happened to surface as large-scale areal relationships. Whether or not this is an actual problem, one solution to it has already been suggested (i.e., a method to estimate the natural incidence of various types of languages that is [meant to be] immune to historical accidents). Originally proposed by Maslova (2000) and taken up by Cysouw (2007), the idea is to change from estimating probabilities of occurrence to estimating probabilities of transition. At the center of this approach lies the assumption that there is a *constant probability of change* inherent in every linguistic parameter, henceforth CPCH (“constant probability of change hypothesis”). This further allows the interpretation of frequent types as *stable*, i.e., the constant probability distribution favours changes to the type and disfavors changes from it, versus infrequent types as *less stable*, i.e., the constant probability distribution disfavors changes to the type and favours changes from it (Maslova and Nikitina tted).

In addition to CPCH, the Maslova/Cysouw model also allows birth- and death effects, henceforth BDE (“birth-death effects”). That is, languages, in addition to transitioning in features, can also die and/or fork into to more languages. Thus, languages we find today are not only the result of independent feature transitions from earlier versions of the same languages – they are the surviving members of isolate languages or languages which inherited features from an ancestor language. The specific rates of birth- and death are kept open, but we may assume that birth- and death processes are independent of features. For example, a language is no more (or less) likely to die (or fork) if it has SVO rather than some other value.

We do not question BDE, but we will attempt to show that CPCH is not valid.

We have put together three databases on basic word order:

1. **Ethnologue:** This database contains 1097 data points (Gordon 2005). Sources for the data points are not indicated. It is not clear how the data points/languages were selected.
2. **WALS:** This database contains 1203 data points (Dryer 2005). Sources for the data points are indicated. It is not clear how the data points/languages were selected, but it may be guessed that it is some kind of convenience sample.
3. **Hammarström:** This database contains 338 data points (Hammarström 2007a). Sources for the data points are indicated. The languages were sampled *at random*, one for *every* attested language family in the world.

These three databases put together, without overlap, amount to 2086 languages – possibly the biggest database of a syntactic parameter so far assembled in linguistic typology. Using the classification of Hammarström (2007b), these 2086 languages are fall into 338 distinct families.¹ 198 of the families have only one [language with a] data point (henceforth ‘isolates’), and 140 of them have more than one. Intuitively, the word order distribution in the Hammarström sample, the isolates, and the majority word order for the non-isolates, should agree. This property is

¹According to this classification, a family is a set of languages which have been shown, in publication, using orthodox comparative methodology to be genetically related. This classification, in general, is ignorant of subgrouping matters.

Table 1: Incidence of word order types across samples (see text)

	All 2086		Hammarström		Isolates		Majority	
SOV	977	46.8%	208	61.5%	121	61.1%	86	61.4%
SVO	659	31.5%	49	14.4%	28	14.1%	25	17.8%
NODOM	166	7.9%	30	8.8%	17	8.5%	11	7.8%
VSO	181	8.6%	21	6.2%	12	6.0%	9	6.4%
VOS	46	2.2%	9	2.6%	6	3.0%	3	2.1%
OVS	14	0.6%	6	1.7%	3	1.5%	2	1.4%
VSO/VOS	9	0.4%	7	2.0%	6	3.0%	0	0.0%
OSV	13	0.6%	1	0.2%	1	0.5%	2	1.4%
SVO/VSO	6	0.2%	2	0.5%	1	0.5%	1	0.7%
SOV/OVS	4	0.1%	2	0.5%	2	1.0%	0	0.0%
SVO/VOS	6	0.2%	0	0.0%	0	0.0%	0	0.0%
SOV/OSV	2	0.0%	2	0.5%	0	0.0%	1	0.7%
SOV/SVO	2	0.0%	1	0.2%	1	0.5%	0	0.0%
SOV/VOS	1	0.0%	0	0.0%	0	0.0%	0	0.0%
	2086		338		198		140	

fully satisfied, as shown in Table 1. The discrepancy to the full 2086-language database is readily understood as distorting effects of a certain few large SVO-prominent families.

The validity of the CPCH may then be assessed by looking at intra-family divergence.

A rigorous statistical test cannot be built because 1) the CPCH is not sufficiently precisely formulated; for example, there are question marks for how much divergence from the constant transition probability is acceptable, and it is not obvious how to quantify time-depth/family-heterogeneity/family-size (or any other transition unit), and 2) the data is not uniformly sampled within families. However, one prediction of the CPCH in any variant, is that the estimates of the constant probabilities, if they exist, should become better the larger the family/the more data points we have for the family. For example, if the CPCH gives rise to a stationary distribution of 61% SOV, then we could look at (say) SOV-original families and see how many of its synchronic languages are no longer SOV. If the number of data points for the family is 2, then we expect to find 0.5 or 1.0, but as the number of data points for a family grows, we expect to find incidences that gravitate towards the assumed stationary ratio, in this example 0.61%. More precisely, the logic is as follows:

1. We assume that CPCH is true.
2. Given that PCH is true, it should give rise to a *stationary distribution*.
3. This stationary distribution should be the distribution evidenced above (in the isolates and Hammarström sample).
4. Given the stationary distribution, for each family, we can calculate the maximum likelihood hypothesis of the word order of its ancestor language (we may also note that the ML, MAP and majority vote on a family turns out to give essentially the same results for this data set).
5. Given the ancestor word order and the samples of synchronic word orders attested, we can compare families with the same ancestor word order.
6. Under the assumption that CPCH is true, we expect that families with the same ancestor word order should show similar synchronic distributions. In particular, we expect that the

larger the family/the more data points we have, the synchronic distribution should approach the stationary distribution.

7. We find that the data do not show a converging behaviour.

For space reasons, the full data cannot be given but Table 2, shows synchronic distributions for the biggest SOV- and SVO-original families respectively. For whatever reason, different language families display very different transition patterns, and there is no observable tendency towards oscillation towards a constant as data points increase. Lexically very diverse families as well as lexically very tight-knit families show divergent rates of word order change.

Intuitively, the presence of BDE introduce some perturbation, to the effect that different families should show diverging behaviour even if CPCH is true. However, we can cope with this, given reasonable assumptions on BDE, mathematical state-of-the-art and that CPCH should be falsifiable at all within practical limits of world's attested languages. We will pay special attention to argue that, for the dataset of this size, the steps outlined above all remain robust in the wake of BDE.

It follows that the CPCH hypothesis, at least for the basic word order parameter, must be rejected or reformulated, though a profitable reformulation appears hard to attain. An introduction of subgrouping distinctions will remain infeasible for a long time ahead, as detailed evidence of subgrouping is much less developed (than mere relatedness demonstration) for most of the world's language families. The same can be said for attempts at a better guess (rather than majority vote) at the diachronic original of a family.

References

- Cysouw, M. (2007). Investigating transition probabilities in the world atlas of language structures (wals). Paper Presented at The seventh International Conference of the Association for Linguistic Typology (ALT VII), CNRS, Paris, September 25-28, 2007.
- Dryer, M. S. (2005). Order of subject, object, and verb. In B. Comrie, M. S. Dryer, D. Gil, and M. Haspelmath (Eds.), *World Atlas of Language Structures*, pp. 330–333. Oxford University Press.
- Gordon, Jr., R. G. (Ed.) (2005). *Ethnologue: Languages of the World* (15 ed.). SIL International, Dallas.
- Hammarström, H. (2007a). A genetically stratified language sample for basic word order typology. Paper Presented at The seventh International Conference of the Association for Linguistic Typology (ALT VII), CNRS, Paris, September 25-28, 2007.
- Hammarström, H. (2007b). The language families of world: A critical synopsis. Manuscript available at http://www.cs.chalmers.se/~harald2/language_families_full.pdf accessed 25 Sept 2007.
- Maslova, E. (2000). A dynamic approach to the verification of distributional universals. *Linguistic Typology* 4(3), 307–333.
- Maslova, E. and T. Nikitina (Submitted). Stochastic universals and dynamics of cross-linguistic distributions: the case of alignment types. MS available online at <http://www.stanford.edu/~emaslova/Publications/ProbabilityPubl.html>, accessed 11 Feb 2008.

Table 2: Transition probabilities for the biggest SOV-original families (top) and the biggest SVO-original families (bottom).

SOV-Family	n	SOV	SVO	NODOM	VSO	VOS	Other
Sino-Tibetan	172	91.2%	8.7%	0.0%	0.0%	0.0%	0.0%
Indo-European	106	52.8%	33.9%	7.5%	5.6%	0.0%	0.0%
Trans New Guinea	94	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Afro-Asiatic	85	44.7%	40.0%	5.8%	9.4%	0.0%	0.0%
Pama-Nyungan	57	47.3%	12.2%	31.5%	0.0%	3.5%	5.2%
Quechuan	42	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Uto-Aztecan	31	41.9%	22.5%	19.3%	12.9%	0.0%	3.2%
Omotic	23	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Turkic	20	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Tupi	18	50.0%	33.3%	0.0%	5.5%	0.0%	11.1%
Uralic	16	50.0%	37.5%	12.5%	0.0%	0.0%	0.0%
Mande	16	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Sepik	14	92.8%	0.0%	0.0%	0.0%	0.0%	7.1%
Chibchan	14	92.8%	0.0%	7.1%	0.0%	0.0%	0.0%
Tucanoan	12	66.6%	0.0%	8.3%	8.3%	0.0%	16.6%
Panoan	12	83.3%	0.0%	16.6%	0.0%	0.0%	0.0%
Dravidian	12	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Siouan	10	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Nakh-Dagestanian	10	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Mongolian	10	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%
...							
SVO-Family	n	SOV	SVO	NODOM	VSO	VOS	Other
Austronesian	240	7.9%	64.1%	6.2%	12.0%	5.8%	3.7%
Atlantic-Congo	201	4.4%	91.5%	3.9%	0.0%	0.0%	0.0%
Mayan	50	0.0%	44.0%	4.0%	26.0%	16.0%	10.0%
Austroasiatic	32	6.2%	84.3%	6.2%	0.0%	3.1%	0.0%
Tai-Kadai	21	4.7%	95.2%	0.0%	0.0%	0.0%	0.0%
Central Sudanic	21	0.0%	71.4%	28.5%	0.0%	0.0%	0.0%
Torricelli	9	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%
Miao-Yao	9	11.1%	88.8%	0.0%	0.0%	0.0%	0.0%
Totonacan	5	0.0%	60.0%	0.0%	40.0%	0.0%	0.0%
Songhay	5	40.0%	40.0%	20.0%	0.0%	0.0%	0.0%
North Halmahera	5	40.0%	60.0%	0.0%	0.0%	0.0%	0.0%
Guaicuruan	5	0.0%	80.0%	20.0%	0.0%	0.0%	0.0%
Zaparoan	3	33.3%	66.6%	0.0%	0.0%	0.0%	0.0%
Koman	3	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%
Iwaidjan Proper	3	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%
...							