The DNA and evolution of language structures

Martin Haspelmath

April 2017, statement for Crecchio workshop on "Variation and universals in language" (June 2017)

1. Understanding limits on diversity

The world around us presents us with diversity in quite different domains

– many different chemical compounds and elements
– many different biological species
– many different geological formations
– many different languages and grammatical units
– many different religious beliefs
– many different recipes

None of this diversity is limitless, and it is the task of science to find the causal factors that put limits on the observed variation.

In biocultural systems, there are generally two kinds of constraints that explain gaps in the attested variation:

– **representational constraints**: limits imposed by the underlying building blocks

– **functional-adaptive constraints**: limits imposed by the likelihood to survive in a process of variation and selection (evolutionary adaptation)

For example, wooden houses strongly tend to have pitched (nonflat) roofs because wood is not stable enough (a **representational constraint** on a cross-cultural generalization about house construction).

Houses also strongly tend to have doors and windows, across cultures, because otherwise they would not be usable by people. This generalization can be explained by a **functional-adaptive constraint**: People make houses with doors and windows to adapt them to their purposes.

Representational constraints on language structures

The "material" that knowledge of language is made of is **mental representation** – hence, constraints imposed by our cognitive endowment for language are here called **representational constraint**.

For example, grammatical systems do not contain numerical conditions (e.g. "prenominal possessor phrases cannot be longer than three words/five syllables") – this can plausibly be attributed to a representational constraint (GRAMMARS DON'T COUNT).

I take UG ("universal grammar") to be a synonym of the set of representational constraints on human languages.

Functional-adaptive constraints on language structures

Some aspects of human language are unquestionably due to functional adaptation, e.g. that rarer words strongly tend to be longer across languages (Zipf 1935), e.g.

(1)	'in' (inessive)	'through' (perlative)
French	*dans*	*à travers*
Russian	*v*	*čerez*
Swedish	*i*	*genom*
Mandarin	*zài*	*tōngguò*
	(frequent)	(rare)

There is little doubt that this tendency has an **efficiency explanation**: Shorter words for more frequent and hence more expected words help speakers save production energy, and they help both speakers and hearers save time.

In chemistry, where evolutionary explanation has been irrelevant (or in the background), the burden of explanation is on the representational constraints, i.e. the underlying building blocks (atoms and their properties).

Linguistics is like biology in that both the **underlying building blocks** ("the DNA", or "UG") and the **evolutionary processes** play a role.

But how do we find out which of the two kinds of constraints are responsible for a given gap in the observed variation?

A beginning of an answer is: **By considering both possibilities.**

2. A first example: Disjoint reference with anaphoric pronouns

In English, a personal pronoun in object position cannot be bound by the subject.

(2) **Gianni$_1$ saw him$_1$.* (OK: **Gianni$_1$ saw him$_2$.*)

For a binding relation to obtain, one needs a reflexive pronoun (*Gianni saw himself*).

What are the causal factors that limit diversity here?

A possible **representational answer**:

Principle B of the binding theory (Reinhart 1983; Chomsky 1981) is part of universal grammar (UG) and rules out (2) and similar examples in other languages.

A possible **functional-adaptive answer**:

Objects are normally referentially disjoint from subjects (in over 95% of the cases), so hearers expect disjoint reference. Some languages such as English therefore require special marking for the less expected case (i.e. the reflexive pronoun *himself*) (see Haspelmath 2008).

It could also be that both causal stories play a role. But how can we tell?

It seems to me that Occam's Razor demands that we first consider functional-adaptive explanations before hyothesizing representational causes. In biology, wings and eyes arising independently in different lineages are widely thought to be adaptive responses, and therefore nobody proposes DNA-based explanations for them.

Moreover, the Principle B account presupposes that learners somehow know that *him* is a pronoun while *himself* is not, and it is quite unclear how they can learn this. (Moreover, there are languages which do not have a rule like this, i.e. where sentences like (1) are perfectly grammatical. Principle B could at most be a preference constraint, and traditional UG-based accounts have not considered preference constraints.)

In addition, the functional-adaptive account also explains the following observations, which play no role in the Principle B account (Haspelmath 2008):

– reflexives are always longer than (or as long as) ordinary anaphoric pronouns
– languages show a much weaker tendency to require special (or longer) reflexives in possessor position, in locational expressions (*He saw a snake behind him*), and in subordinate clauses (long-distance reflexives, Pica 1987)
– languages show a much weaker tendency to require special (or longer) reflexives when the verb has a self-directed meaning (e.g. 'wash', 'shave')

The functional-adaptive explanation **does not include a mechanism for elegant language-particular description** ("analysis"), but recall that the initial question was not how to describe/analyze languages, but how to understand limits on observed diversity.

3. A second example: Limits on variation in flagging patterns

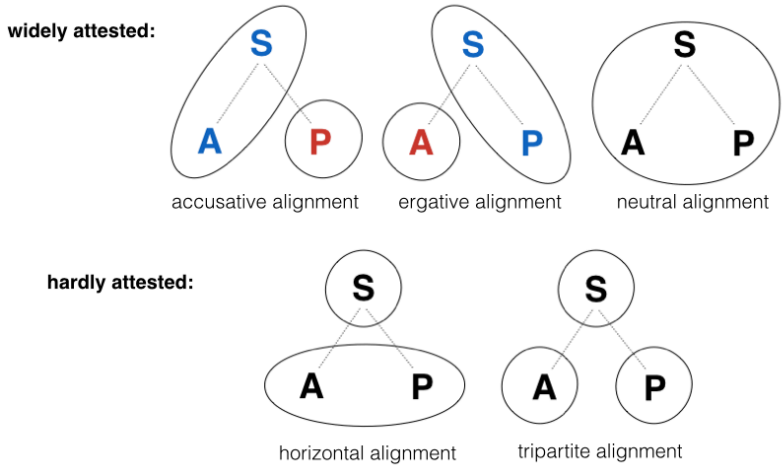
3.1. Some well-known Greenbergian and Greenberg-style universals

> **(A)** With overwhelmingly greater than chance frequency, languages with normal SOV order are **postpositional**. (Greenberg 1963: Universal 4)

> **(B)** There is no language in which the **plural** does not have some **nonzero** allomorphs, whereas there are languages in which the singular is expressed only by zero. (Greenberg 1963: Universal 35)

(C) Where there is a **case system**, the only case which ever has only zero allomorphs is the one which includes among its meanings that of the **subject** of the intransitive verb. (Greenberg 1963: Universal 38)

(D) With overwhelmingly greater than chance frequency, languages do not use a special form for the S argument. (I.e., the other two conceivable alignment patterns, horizontal and tripartite alignment, do not occur.)



Functional-adaptive explanation:

communication is efficient if grammatical patterns do not contain superfluous structure, i.e. if they **minimize domains, distinctions and forms**

3.2. Domain minimization: universal (A)

The word orders OV & NPostp, as well as VO & PrepN, provide optimally efficient lengths of syntactic processing domains (Hawkins 2004; 2014) – hence, VO languages also tend to have prepositions, and OV languages postpositions.

3.3. Distinction minimization: universal (D)

Tripartite alignment needs an additional form distinction that is not needed by the bipartite alignments, and horizontal alignment uses a form distinction wastefully.

3.4. Form minimization: universals (B), (C)

Frequently occurring grammatical meanings are predictable/expected and hence need not be coded by long forms, and can often be left uncoded. This applies to lexical items (as seen above), but also to all asymmetrical grammatical oppositions, and is supported by massive evidence:

"typological markedness" patterns (Greenberg 1966; Croft 2003), showing oppositions of short/frequent/expected categories and long/rare/unexpected categories:

(3) short and frequent	**long and rare**		
singular	plural	(*book – book-s*)	(Universal B)
present tense	future tense	(*go – will go*)	
3rd person	2nd person	(Spanish *canta – canta-s*)	
nominative	accusative	(Hungarian *ember – ember-t*)	(Universal C)
absolutive	ergative	(Lezgian *gada – gada-di*)	(Universal C)
active	passive	(Latin *cantat – cantat-ur*)	
affirmative	negative	(*go – don't go*)	
allative	ablative	(*to – from*)	
declarative	interrogative	(Polish *widzi – czy widzi?*)	
positive	comparative	(*small – small-er*)	
action word	agent noun	(*bake – bak-er*)	
property	change of state	(*red – redd-en*)	
cardinal	ordinal	(*four – four-th*)	
male	female	(*poet – poet-ess*)	

Special coding for accusative and ergative is just a special case of this extremely general pattern.

And longer forms for coreference/binding ("Principle B", see §2 above) are another special case of this pattern.

4. Baker (2015): Representational constraints on flagging patterns?

Baker (2015): some highly general case-assignment rules:

(4) a. High case in the clause is ergative.
b. Low case in the clause is accusative.
c. High case in VP is dative.
d. Low case in VP is secundative
e. High case in NP is genitive (there is no low case in NP).
f. Unmarked case is nominative-absolutive.

But these statements only rule out the horizontal alignment pattern (= one half of universal (D)); they explicitly allow the tripartite alignment pattern (and Baker says that it is probably for functional reasons that tripartite alignment is rare).

The rules do not explain universal (C), which Baker leaves to "morphology".

In addition, the rules would allow ditransitive patterns of the type "ERG-ERG-ABS", and Baker has to go to great lengths to exclude them.

In the end, the "theory" has enormous complexity, but it explains few universals (cf. Haspelmath 2017).

The primary reason is that Baker wants to achieve **elegant language-particular description** at the same time ("principles of case assignment that are as unified as possible").

Baker's principles are extended in a complex way so that they also allow him to subsume definiteness-based differential object marking in Sakha:

(5) a. *Masha salamaat-y türgennik sie-te.*
Masha porridge-ACC quickly eat-PST.3SG.SBJ
'Masha ate the porridge quickly.'

b. *Masha türgennik salamaat sie-te.*
Masha quickly porridge eat-PST.3SG.SBJ
'Masha ate porridge quickly.'

Baker claims that this contrast is due to dependent case assignment as well, because the object is in the same domain as the subject only when it raises from the VP as in (5a). In (5b), where the object stays in its base VP position, it does not count as a case competitor in this kind of language.

But this is not sufficiently general to explain differential object marking, because this is not always associated with different word orders, as it happens to be in Sakha.
Baker is led to make his theory vastly more complicated than a universal case-marking theory would have to be because **language-particular elegant description** is an important goal, in addition to explaining universals of case marking.

(In fact, it is probably a more important goal, because Baker does not highlight the universal predictions his theory makes, and perhaps they do not make any universal predictions, because Baker does not claim that dependent case assignment is the only mode of case assignment; there might be many other modes. Moreover, Baker accepts that the rarity or nonexistence of some kinds of case-marking patterns is functionally motivated, i.e. outside the purview of his system.)

5. A plea for reproducibility and cumulativity of research results

– a successful science produces not only ideas, hypotheses and theories, but real discoveries

– success (i.e. the distinction between ideas and discoveries) is measured by reproducibility and cumulativity:

(i) findings of one research team can be **reproduced** by another research team

(ii) findings of one generation serve as the **foundation** of more and better findings of a subsequent generation

– since the 1870s (Leipzig-based neogrammarian revolution), revolutions have been popular in linguistics, but it is not easy to say that subsequent generations have generally been building on earlier generations (except in historical linguistics)

A plea for the comparative study of language structures

– linguists tend to focus most of their efforts on systems of particular languages

– but language-particular systems are **social conventions** that are very largely the result of historical contingencies, so they do not tell us very much about human language

– if general linguistics wants to make claims about conventional aspects of language (rather than, say, study language processing), it needs to study a broad range of languages from around the world

– it may be that some aspects of language structure are not learned, but present in humans from the very beginning; but in practice, it is very difficult to show that a structural trait is not learned, and as a result, **linguists rarely argue for specific claims about unlearned structural traits**

– the comparative study of a broad range of languages from around the world is easier than ever before, but it is still very difficult, because the points of comparison are not obvious at all

– perhaps more importantly, very few linguists are engaged in this endeavor – probably not more than two dozen; so the world-wide study of language structure is still in its infancy

– hence, to make progress, we should encourge more of our students to compare languages world-wide

6. Can the study of individual languages bring us closer to the "DNA of language"?

– for practical reasons, and for reasons of cultural attachment, many linguists study one language, or at most a few related languages

– clearly, if the study of one language could bring us closer to the DNA of language, this would be wonderful, because we would be able to combine our cultural preferences with deeper insights into human languages

– but we have access only to phenotypical traits of languages, i.e. speakers' utterances and their conventions; how might it be possible to infer the DNA from the phenotype?

– in practice, linguists often formulate hypotheses about the "DNA", i.e. the unlearned aspects of language structure, on the basis of a single language (or a group of closely related languages, e.g. the Scandinavian languages, Roberts & Holmberg 2005; or Romance varieties, e.g. Manzini & Savoia 2011)

– in a second step, they ask whether the same proposal would also work in another language; this often seems to be the case, but it is usually **impossible to rule out the alternative hypothesis that the similarities between languages are due to the same functional-adaptive constraints that apply to all languages**

– unless we **consider both possibilities**, we will hardly make progress

– in particular, when we find clear evidence for adaptation (e.g. hiatus resolution), as is widely recognized by phonologists ("It is well known that phonological operations apply in ways such that **optimal outputs** emerge either language-internally ... or cross-linguistically", Newell), there seems to be no good reason to attribute the adaptive patterns to representational constraints.

7. The natural-kind presumption

Generative grammarians often adopt an explicitly biological point of view, a trend that has become stronger in recent decades ("biolinguistics").

It may thus seem natural to regard the categories of languages as biological objects, like biological species, or like the biomolecules that are the foundation of living beings.

And indeed, languages show recurring similarities across continents, just as biomolecules recur in different forms of life, and biological species occur over wide geographical areas (some, like the red fox (*Vulpes vulpes*), occur in both hemispheres).

More abstractly, linguistic categories on this view are **natural kinds**. An even better example of natural kinds is chemical elements, which occur on different planets and even in different galaxies (presumably throughout the universe, as true universals of physics). And just as it makes sense to ask which molecules occur on Mars (e.g. whether there is water, and if so how much), it may well make sense to ask whether a language on a different continent such as Quechua has "gerunds", or "determiners", or "VP", or "coronal consonants", or any other category that was originally found to exist in English (see Baker 2001 for an explicit analogy between elements/molecules and linguistic categories).

If linguistic categories are natural kinds, then it makes perfect sense to hypothesize that a category identified for one language (say, "anaphor" in English) will exist in another language.

(It even makes sense to **assume** that a similar category in a different language is the same category, until there is evidence to the contrary, cf. Chomsky's Uniformity Principle: "In the absence of compelling evidence to the contrary, assume languages to be uniform, with variety restricted to easily detectable properties of utterances." (Chomsky 2001: 2)

But again, there is an alternative possibility that needs to be considered:

In biocultural systems, we often see **culture-specific categories** arising.

In human social systems, there are different **kinship categories** (younger sister, uncle, different-sex sibling, moieties, etc.), different **governance categories** (chief, king, mayor, president, consul, chairman, general secretary, etc.), different **religious categories** (angel, gospel, surah, stupa, shaman, etc.), different **food categories** (soup, dumpling, noodle, sandwich, dessert, confectionery), different **visual art categories** (painting, sculpture, photograph, movie, woodblock print, fashion, etc.), and so on and so forth.

Few people would claim that similarities in social categories across cultures should be explained by our uniform genetic endowment, and nobody treats such categories are natural kinds. They are seen as culture-specific.

Comparison across cultures is possible through comparative concepts, e.g. when authors such as Watts et al. (2015) carry out large-scale comoartive religious research with a comparative concept "big god".

Likewise, comparative linguistics can be carried out with **comparative concepts** (Haspelmath 2010), a set of concepts that are distinct from the categories used in analyzing particular languages.

It could of course be that the natural-kinds approach is better (at least for some linguistic categories), but one needs to consider both possibilities.

So far, it seems that the natural-kinds approach has had a mixed record of success:

> "this volume shows, first and foremost, that an actual theory of parameters and indeed their general format is still a distant prospect; the contributions, while uniformly optimistic in their outlook, hardly try to conceal this fact. Consequently, details of variation in the surface data by far outweigh the discussion of parameters with any predictive power in this volume, with a few notable exceptions such as Rizzi's (1982, 1986) seminal null-subject parameter. Proposals of this kind, which make a genuine attempt at cutting through the complexity of phenomenology rather than merely restating observations (e.g. in the guise of arbitrary morphosyntactic features or otherwise), remain few and far between; consequently, the empirical investigation of linguistic variation proceeds with little theoretical guidance. The volume under review is highly recommended as a sobering reminder of how much work still lies ahead." (Dennis Ott 2016, review of Fábregas et al. 2015, LINGUIST List 27.4380)

8. Description and comparison are distinct enterprises

Probably 98% of all linguists are almost exclusively concerned with describing/analyzing particular languages.

Perhaps the confusions surrounding our concepts (cf. Haspelmath 2018) and the slow progress in understanding the limits on linguistic diversity have to do with the fact that few linguists engage in comparison at the appropriate scale, and that many linguists who mostly care about individual languages nevertheless expect to take their concepts from other languages or from typology.

It seems that in other disciplines dealing with sociocultural diversity, this problem of confusing description and comparison arises to a lesser degree, for reasons that are not entirely clear to me.

References

Baker, Mark C. 2001. *The atoms of language*. New York: Basic Books.

Baker, Mark C. 2015. *Case*. Cambridge: Cambridge University Press.

Chomsky, Noam A. 2001. Derivation by phase. In Michael Kenstowicz (ed.), *Ken Hale: A life in language*, 1–52. Cambridge, Mass.: MIT Press.

Croft, William. 2003. *Typology and universals*. 2nd edition. Cambridge: Cambridge University Press.

Fábregas, Antonio, Jaume Mateu i Giral & Michael T. Putnam (eds.). 2015. *Contemporary linguistic parameters*. London: Bloomsbury.

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

Greenberg, Joseph H. 1966. *Language universals, with special reference to feature hierarchies*. The Hague: Mouton.

Haspelmath, Martin. 2008. A frequentist explanation of some universals of reflexive marking. *Linguistic Discovery* 6(1). 40–63. (Open access: http://journals.dartmouth.edu/cgi-bin/WebObjects/Journals.woa/1/xmlpage/1/article/331)

Haspelmath, Martin. 2010. Comparative concepts and descriptive categories in crosslinguistic studies. *Language* 86(3). 663–687.

Haspelmath, Martin. 2017. Review of "Baker, Mark. 2015. Case. Cambridge: Cambridge University Press." *Studies in Language* 41 (to appear)

Haspelmath, Martin. 2018. How comparative concepts and descriptive linguistic categories are different (draft). DOI: http://doi.org/10.5281/zenodo.570000.

Hawkins, John A. 2004. *Efficiency and complexity in grammars*. Oxford: Oxford University Press.

Hawkins, John A. 2014. *Cross-linguistic variation and efficiency*. New York: Oxford University Press.

Manzini, Maria Rita & Leonardo Maria Savoia. 2011. *Grammatical categories: Variation in Romance languages*. Cambridge: Cambridge University Press.

Pica, Pierre. 1987. On the nature of the reflexivization cycle. *North-Eastern Linguistics Society* 17. 483–499.

Reinhart, Tanya. 1983. *Anaphora and semantic interpretation*. London: Routledge.

Roberts, Ian F. & Anders Holmberg. 2005. On the role of parameters in universal grammar: a reply to Newmeyer. In Hans Broekhuis, N. Corver, Riny Huybregts, Ursula Kleinhenz & Jan Koster (eds.), *Organizing grammar: Linguistic studies in honor of Henk van Riemsdijk*, 538–553. (Studies in Generative Grammar 86). Berlin: Mouton de Gruyter.

Watts, Joseph, Simon J. Greenhill, Quentin D. Atkinson, Thomas E. Currie, Joseph Bulbulia & Russell D. Gray. 2015. Broad supernatural punishment but not moralizing high gods precede the evolution of political complexity in Austronesia. *Proceedings Royal Society B* 282(1804). 20142556. doi:10.1098/rspb.2014.2556.

Zipf, George Kingsley. 1935. *The psycho-biology of language: An introduction to dynamic philology*. Cambridge, Mass.: M. I. T. Press.