43rd International Conference of the Linguistic Society of India (ICOLSI-43), 2021 Dec 21

# Language description and language universals

MARTIN HASPELMATH
*Max Planck Institute for Evolutionary Anthropology*

## 1. Two ways of approaching the study of Human Language

– the description-comparison approach

we describe many languages worldwide and compare them

– the "deep reality" approach

we look for some deep "underlying" structure
that is not immediately apparent

The description-comparison approach may look a bit boring –
like collecting butterflies and displaying them in a museum.

The "deep reality" approach may look more exciting –
like chemistry and physics, where it has been shown
that the stuff that surrounds us consists of elements of atoms.

But I will argue that the description-comparison approach is better suited to helping us understand the nature of Human Language.

## 2. The study of Human Language, and particular languages

I assume that
– we are interested in theoretical linguistics (not just in applied linguistics)
– we are interested in Human Language, not just in particular languages (*p-languages*)
– in other words, we want to do ***general linguistics*** *(g-linguistics)*

but how do we solve the **general linguistics paradox**?
(Haspelmath 2021b)

We want to explore and understand the nature of Human Language,
but what we can observe directly is particular languages.

answer: – we study **universals**
[– alternatively: we study nonconventional aspects of Human Language]

When we observe a particular language, we can describe it, but it does not tell us anything about Human Language directly.

Consider words such as

| | | |
|---|---|---|
| *house* | дом | গৃহ |
| *tree* | дерево | গাছ |
| *red* | красный | লাল |

they differ across languages in **random, accidental** ways – they do not tell us anything about the nature of Human Language.

The situation is basically the same with morphosyntax – languages differ in **seemingly random** ways, e.g.

*I have a red house.*  
У меня красный дом  
আমার একটি লাল ঘর আছে

English has a transitive verb that shows agreement with the possessor, while Russian has no verb and a spatial preposition (*u menja* 'near me'), and Bengali has a genitive possessor and a copula ('of me is a red house').

Thus, particular languages (*p-languages*) are not an immediate window to human cognition – they differ across populations, and they are learned like other aspects of human culture.

so we are confronted with the **general linguistics paradox**:  
(Haspelmath 2021b)

> We want to explore and understand the nature of Human Language, but what we can observe directly is particular languages.

Cf. religious studies:

> What we can observe directly is particular religions.  
> But how do we learn from them about Human Religion in general?

The solution: **comparing cultures:**

| | |
|---|---|
| – comparing religions | (comparative religious studies) |
| – comparing languages | (comparative grammar studies) |

## 3. The "deep reality" approach

Some linguists seem to think that there is a shortcut – that we can get deep insights into the human language faculty by studying a few languages.

The stuctural descriptions of sentences are often very intricate, as for example in Baker's (2015: 125-126) theory of "dependent case":

Differential object marking in Sakha (a Turkic language of Siberia)
(see further discussion in Haspelmath 2019b)

(22) a. Masha salamaat-y sie-te.
Masha porridge-ACC eat-PAST.3sS
'Masha ate the porridge.'

b. Masha salamaat sie-te.
Masha porridge eat-PAST.3sS
'Masha ate porridge.'

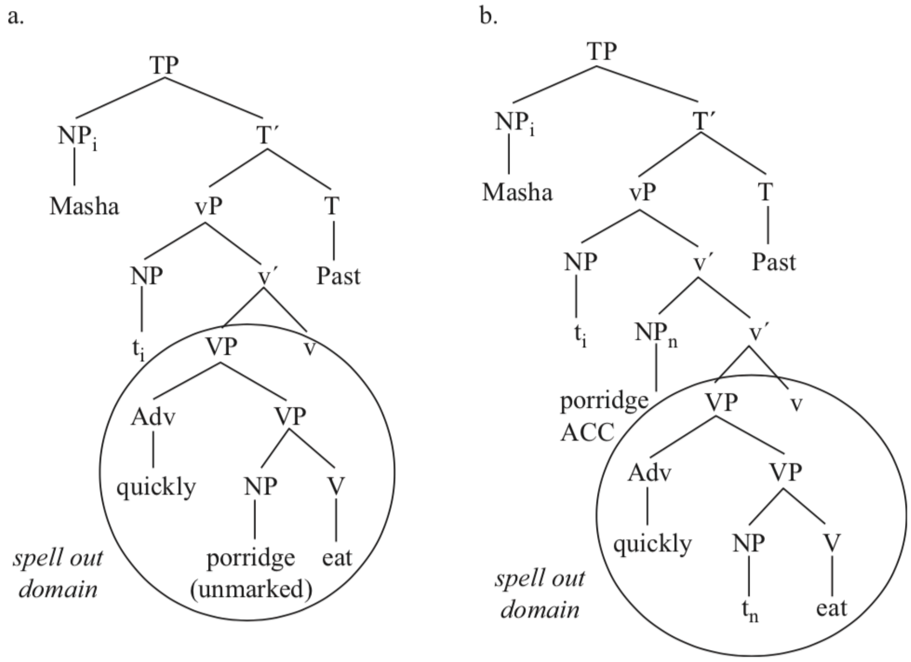
(23) a. Masha salamaat-y türgennik sie-te.
Masha porridge-ACC quickly eat-PAST.3sS
'Masha ate the porridge quickly.'

*(Object is * without ACC)*

b. Masha türgennik salamaat sie-te.
Masha quickly porridge eat-PAST.3sS
'Masha ate porridge quickly.'

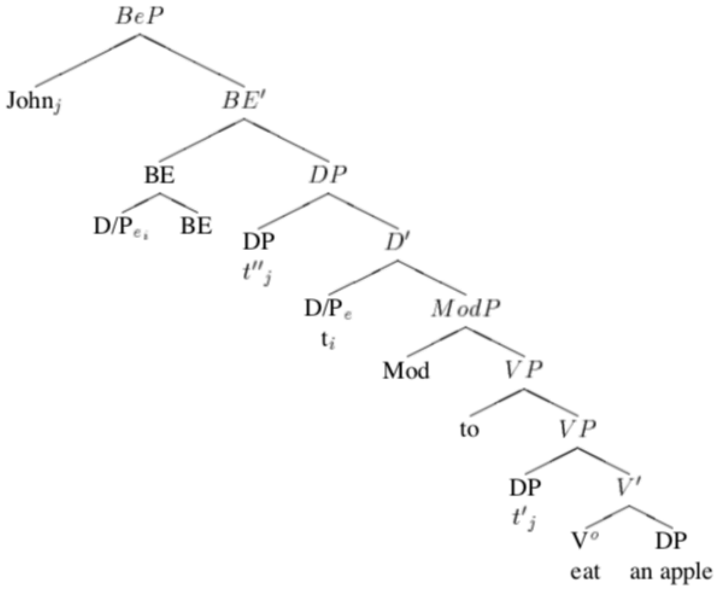
*(ACC would imply that the object is focused)*

(24)



Or take Bhatt's (1998) analysis of the English sentence *John has to eat an apple*.

(39) John has to eat an apple.



It would indeed be fascinating if it could be shown that simple sentences like *John has to eat an apple* have a rich underlying structure, made up of **a small set of innate building blocks**.

Finding these innate building blocks would be like finding the elements of chemistry (cf. Baker 2001). This would be a major success.

BUT:

– We **do not know** what the innate building blocks of syntax are.

– We have **no research programme** for finding out.

– Even quite a few linguists who work in the Chomskyan tradition now think that there is no rich Universal Grammar

even Chomsky himself!

Many generative linguists agree with Chomsky, e.g. Julie Anne Legate:

> "I have now convinced myself of a framework whereby merge is innate, but any (other) **language-specific innate properties are highly suspect** and require significant evidence. (This is due to Noam's writings on evolution finally sinking in, and due to my accumulating knowledge about the extent of language variation.) Case, both the distribution of noun phrases and the case morphology, is not universal, varies considerably across languages, and so must be learned." (2018; see https://dlc.hypotheses.org/1392)

## 4. Description (and comparison)

The good thing about language description is: **We know how to do it.**

– we know how to design orthographies
– we know how to write dictionaries
– we know how to describe syntactic constructions
  – we even write entire grammars

This is solid work that future generations will be able to use, even if it does not give us direct insight into Human Language.

An insight of the early 20th century: Different languages not only have different labels, but also **different categories**.

– different semantic categories

| | | |
|---|---|---|
| cf. Hindi-Urdu kinship terms | दादा/ دادا | grandfather |
| | दादी, دادی | |
| | नाना, نانا | grandmother |
| | नानी, نانی | |

– different morphosyntactic categories

| | | |
|---|---|---|
| e.g. | *laṛk-ā* | ‘boy’ |
| | *laṛk-e* | ‘boy’ |
| | *laṛk-e-kā* | ‘boy’s’ |
| | *laṛk-e-kī* | ‘boy’s’ |

“Each language has its own categories – we should describe each language in its own terms.” (Boas 1911)

If we assume that all languages make use of the same categories, then we are tempted to impose the categories of our languages on other languages.

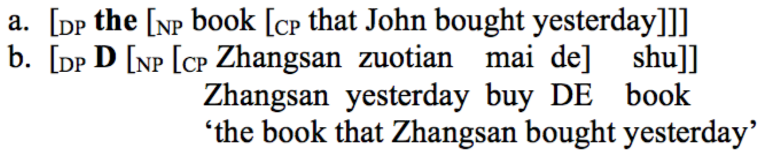
> Carl Velten (1901), *Grammatik des Kinyamuesi.* Göttingen.
>
> “The nominative and the accusative have the same form. They can be recognized from the position in the clause.”

So do all languages have a “nominative” and an “accusative”?

> (6) **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)

With such a principle, it is easy to justify “nominative” and “accusative” for the Bantu language Kinyamwezi.

We might also justify **zero "determiners"** for Chinese:



a. [$_{DP}$ **the** [$_{NP}$ book [$_{CP}$ that John bought yesterday]]]
b. [$_{DP}$ **D** [$_{NP}$ [$_{CP}$ Zhangsan zuotian mai de] shu]]
Zhangsan yesterday buy DE book
'the book that Zhangsan bought yesterday'

(from Cheng 2011)

Or we might justify **"pro-drop"** in Italian:

(1) *Cant-a*[3SG]. 'She sings.' (**affixal subject expression**)
*Cant-ano*[3PL]. 'They sing.'
*Cant-o*[1SG]. 'I sing.'

(2) *La ragazza*[SG] *canta*[3SG]. 'The girl sings.' (**double subject expression**)
*Le ragazze*[PL] *cantano*[3PL]. 'The girls sing.'

From the perspective of English, one may posit **"zero pronouns"** ("pro-drop"):

(3) PRONOUN[SG] *canta*[SG]. 'She sings.'
PRONOUN[PL] *cantano*[PL]. 'They sing.'

(cf. Haspelmath 2013)

Thus:

– We know how to describe languages (each language has its own categories).
– We **don't know** what the innate building blocks are (if any).

If we want to avoid ethnocentrism and do justice to each language,
we must **create new categories for each language**.

(see Haspelmath 2020)

## 5. Comparison based on measurement uniformity (not on building block uniformity)

To compare languages, we need **uniform yardsticks for comparison**.

For example, to test the claim in (1), we need to determine the order of **adpossessor** and **noun**, and the order of **object** and **verb** in a representative set of the world's languages.

> (1) If the adnominal possessor precedes the noun, the object tends to precede the verb; and if the adnominal possessor follows the noun, the object tends to follow the verb.

– how do we measure "order"?
(Dryer 2005: dominant order = more than 67% of occurrences in texts)

– how do we "measure" "possessor", "noun", "object" and "verb"?
(Greenberg 1963: semantically)

NOTE: Comparison is not based on ***the rules*** of the languages
– because the rules do not make reference to text frequencies
and to semantic notions

Compare economics:

We measure economic indicators (like inflation) by uniform yardsticks, while ignoring culture-specific rules about money and buying (let alone mental representations of money).

Comparison of phonological systems:
by means of **phonetic** properties, not phonological values

In general: comparison requires **comparative concepts**
not **descriptive categories** (Haspelmath 2010)

Comparative concepts are like units of measurement in other sciences.
They are somewhat arbitrary, but are applied in the same way to all languages.

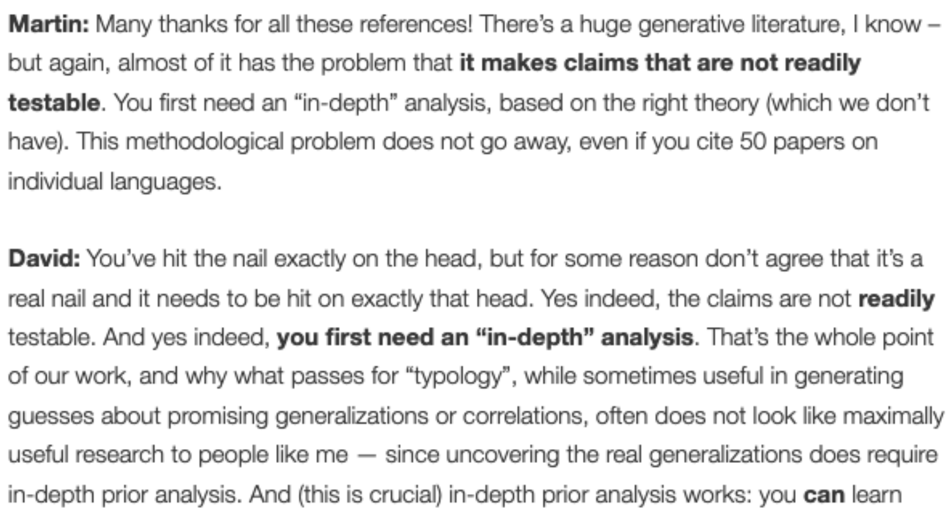
(cf. blogpost: https://dlc.hypotheses.org/2305)

Measurement uniformity allows large-scale quantitative testing of universals.

Hypothetical building-block uniformity does not allow large-scale quantitative testing of universals.

→ Universal-testing must proceed in a slow and piecemeal fashion...

cf. a recent interaction on Facebook with David Pesetsky (https://dlc.hypotheses.org/2235):



**Martin:** Many thanks for all these references! There's a huge generative literature, I know – but again, almost of it has the problem that **it makes claims that are not readily testable**. You first need an "in-depth" analysis, based on the right theory (which we don't have). This methodological problem does not go away, even if you cite 50 papers on individual languages.

**David:** You've hit the nail exactly on the head, but for some reason don't agree that it's a real nail and it needs to be hit on exactly that head. Yes indeed, the claims are not **readily** testable. And yes indeed, **you first need an "in-depth" analysis**. That's the whole point of our work, and why what passes for "typology", while sometimes useful in generating guesses about promising generalizations or correlations, often does not look like maximally useful research to people like me — since uncovering the real generalizations does require in-depth prior analysis. And (this is crucial) in-depth prior analysis works: you **can** learn

But this slow process has not really led to any lasting findings that one can build on.

There are many ideas floating around, and many papers and dissertations, but **no theories that have been tested** and found to be solid.

## 6. Universals (and how to explain them)

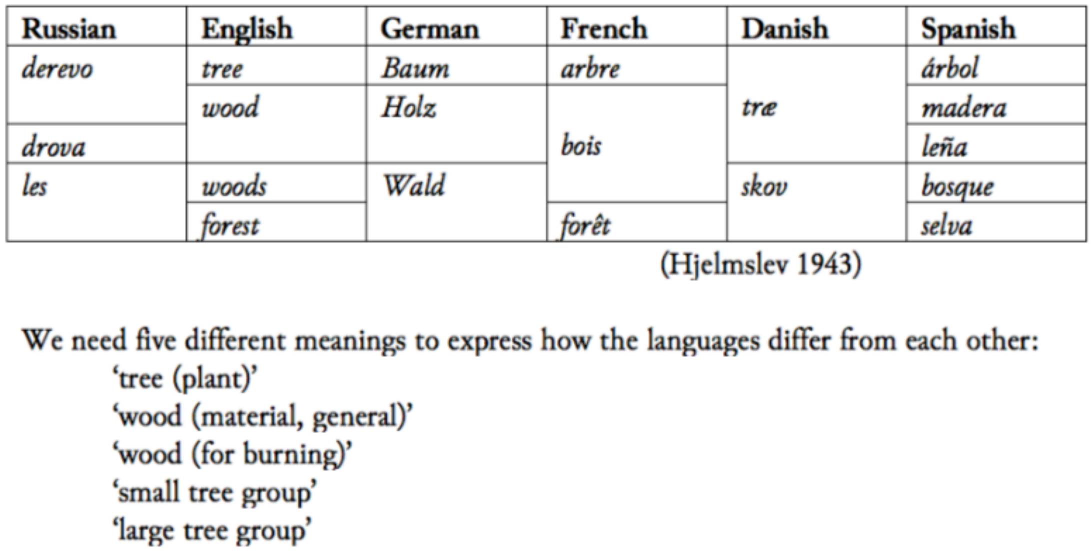
Once we have found universals, we can **look for explanations** of universals.

The two most important types of universals:

– coexpression universals

– asymmetric coding universals

### 6.1. Coexpression universals: Semantic maps



| Russian | English | German | French | Danish | Spanish |
|---|---|---|---|---|---|
| *derevo* | *tree* | *Baum* | *arbre* | *træ* | *árbol* |
| | *wood* | *Holz* | *bois* | | *madera* |
| *drova* | | | | | *leña* |
| *les* | *woods* | *Wald* | | *skov* | *bosque* |
| | *forest* | | *forêt* | | *selva* |

(Hjelmslev 1943)

We need five different meanings to express how the languages differ from each other:
'tree (plant)'
'wood (material, general)'
'wood (for burning)'
'small tree group'
'large tree group'

Different languages **coexpress** the comparison meanings in different ways, but there is a universal: In the implicational sequence

tree – wood (material) – wood (for burning) – small woods – large forest

languages **may only coexpress adjacent meanings**.

It's the same for grammatical meanings, e.g. beneficiary, recipient, goal, patient

English: recipient & goal

She gave money **to** her brother.
She went **to** town.

Russian: recipient & beneficiary

*Ona dala den'gi brat-**u**.*
she gave money brother-DAT

*Ona ispekla tort brat-**u**.*
she baked cake brother-DAT

Imbabura Quechua: beneficiary & adnominal possessor

*wasi-ta rura-rka-ni ñuka churi-**paj***
house-ACC make-PST-1SG my son-GEN
'I made a house for my son.' (Cole 1982: 113)

*Juzi-**paj** warþi*
Jose-GEN wife
'Jose's wife' (Cole 1982: 115)

Hindi-Urdu: recipient & patient

a. *Us-ne Anuu-**ko** kitaab dii.*
she-ERG Anu-DAT book gave.PST.1SG.F
'She gave Anu a book.'

b. *Admii-ne lạrke-**ko** dekh-aa.*
man-ERG boy.obl-ACC see-3M.SG.PST
'The man saw the boy.'

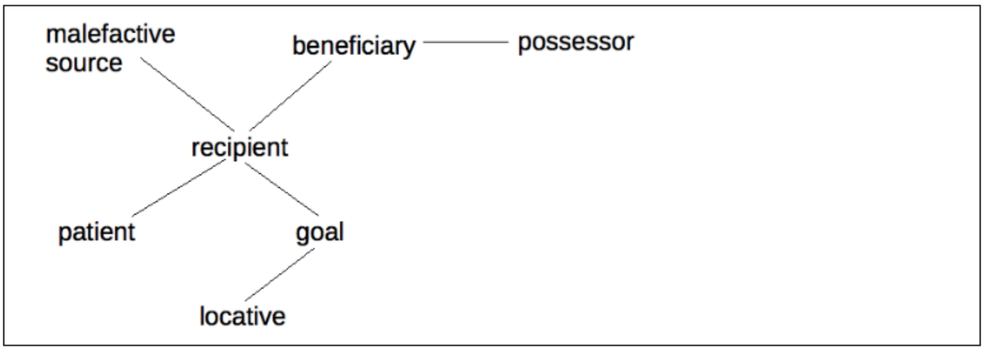
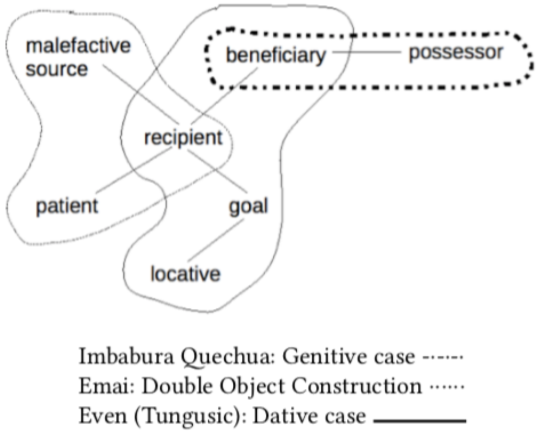


**Figure 11.3. A semantic map for recipient and related functions**

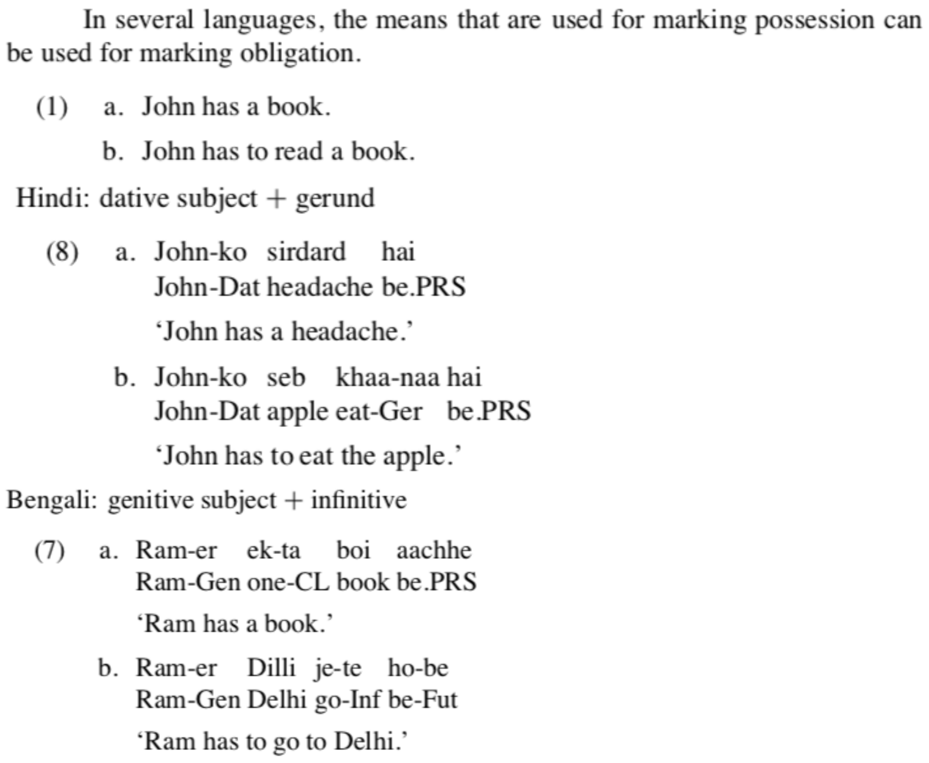
(see Haspelmath 2003)

And this is how the distribution of markers in different languages is displayed :


malefactive
source
beneficiary
possessor
recipient
patient
goal
locative
Imbabura Quechua: Genitive case -·-·-·
Emai: Double Object Construction ······
Even (Tungusic): Dative case ———

The coexpression perspective allows us to state universals – to say what is general across languages – without claiming that we have detected some "deep reality".

Contrast this with Bhatt (1998), who notes the coexpression of obligation and possession :



In several languages, the means that are used for marking possession can be used for marking obligation.

(1) a. John has a book.
b. John has to read a book.

Hindi: dative subject + gerund

(8) a. John-ko sirdard hai
John-Dat headache be.PRS
'John has a headache.'

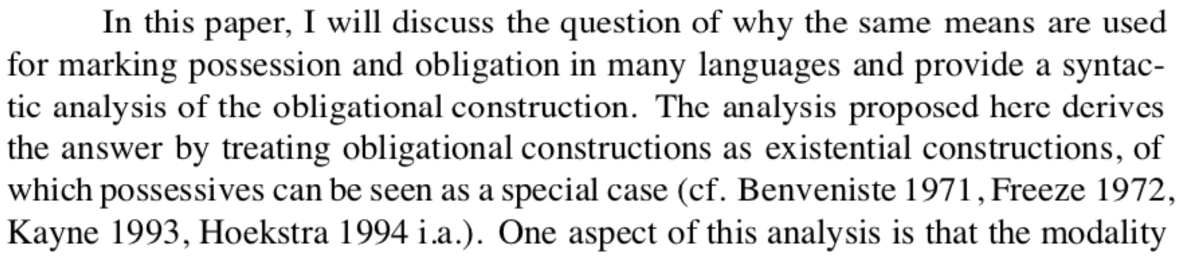
b. John-ko seb khaa-naa hai
John-Dat apple eat-Ger be.PRS
'John has to eat the apple.'

Bengali: genitive subject + infinitive

(7) a. Ram-er ek-ta boi aachhe
Ram-Gen one-CL book be.PRS
'Ram has a book.'

b. Ram-er Dilli je-te ho-be
Ram-Gen Delhi go-Inf be-Fut
'Ram has to go to Delhi.'

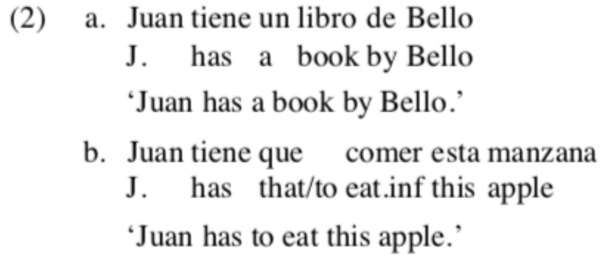
Bhatt (1998) claims that obligational constructions and possessive constructions are really existential constructions, at some level of "deep reality":



In this paper, I will discuss the question of why the same means are used for marking possession and obligation in many languages and provide a syntactic analysis of the obligational construction. The analysis proposed here derives the answer by treating obligational constructions as existential constructions, of which possessives can be seen as a special case (cf. Benveniste 1971, Freeze 1972, Kayne 1993, Hoekstra 1994 i.a.). One aspect of this analysis is that the modality

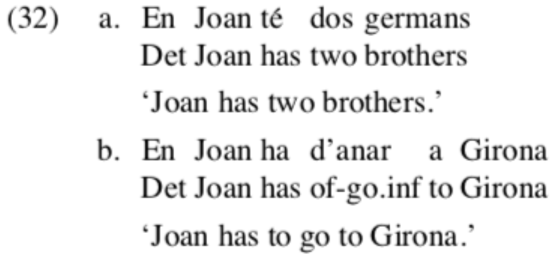
But this is very difficult to demonstrate, so much of the work in this tradition remains at a speculative level. It is unclear how it can be extended to other languages.

Bhatt himself admits that it works for English, German, and Spanish, but not for Catalan



(2) a. Juan tiene un libro de Bello  
J. has a book by Bello  
'Juan has a book by Bello.'

b. Juan tiene que comer esta manzana  
J. has that/to eat.inf this apple  
'Juan has to eat this apple.'

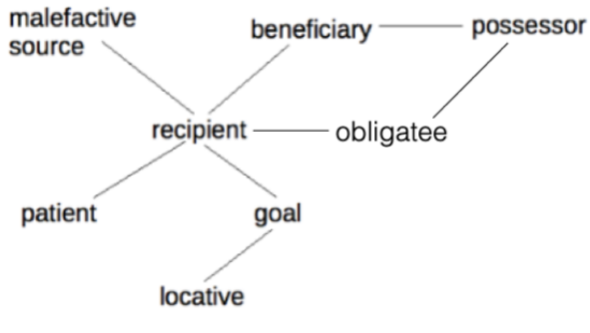


(32) a. En Joan té dos germans  
Det Joan has two brothers  
'Joan has two brothers.'

b. En Joan ha d'anar a Girona  
Det Joan has of-go.inf to Girona  
'Joan has to go to Girona.'

In the semantic-map approach, **there is no need to determine some underlying, deep reality**. We can simply "measure" the differences between languages by mapping their forms onto our comparison meanings.

We can add another semantic role, let's call it **obligatee**, to the semantic map, and then see how languages behave :



## 6.2. Asymmetric coding universals

I have written many papers about asymmetric coding universals over the last few years, so here is just a very short overview:

Many phenomena exhibit asymmetric coding, i.e. zero vs. short, or short vs. long, in a cross-linguistically systematic way, as in this table:

**Table 1: Examples of universal grammatical coding asymmetries**

| | | |
|---|---|---|
| singular | plural | (*book* – *book-**s***) |
| nominative (A/S) | accusative (P) | (*he* – *hi-**m***) |
| allative | ablative | (*to* – ***from***) |
| positive | comparative | (*small* – *small-**er***) |
| present | future | (*go* – ***will*** *go*) |
| affirmative | negative | (*go* – ***don't*** *go*) |
| inanimate patient | animate patient | (Spanish *Ø la casa* – ***a*** *la mujer*) |
| 3rd person | 2nd person | (Spanish *canta*$_{3SG}$ / *canta-**s***$_{2SG}$ 'sing(s)') |
| 2nd person imperative | 3rd person imperative | (*praise!* – **let** *her praise!)* |
| attributive adjective | attributive verb | (*small* – *play-**ing***) |

These phenomena have typically been treated under the heading of "markedness" or "differential marking".

I have argued that they can all be treated together and that they follow a single generalization:

**frequently expressed grammatical distinctions get short coding**
(form-frequency correspondences)

Again, as in the case of semantic maps, there is no need to establish any kind of "deep reality" in order to test these universals (see Haspelmath 2021a).

### 6.3. How to explain universals

I have proposed that there are three main types of explanatory factors (or constraints):

– biocognitive constraints (on possible representations)
– functional-adaptive constraints
– mutational constraints (on possible changes) (see Haspelmath 2019a)

I have claimed that the asymmetric coding universals are due to functional-adaptive constraints (Haspelmath 2021a), and I think that the coexpression universals are due to mutational constraints.

Biocognitive constraints seem to be far less important than is often thought.

## 7. Conclusion

– to understand Human Language, we need to adopt a comparative approach

– to describe languages, we should use language-particular categories, because we have not found an innate set of universal building blocks yet

– to compare languages, we need comparative concepts as uniform yardsticks of measurement

– language universals are of two main types:
coexpression universals (expressed via semantic maps)
asymmetric coding universals (Haspelmath 2021a)

– the description-comparison approach is more promising than an approach that posits some "deep reality" and tries to go directly from particular languages to human cognition, without worldwide comparison

# References

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

Baker, Mark C. 2015. *Case: Its principles and parameters.* Cambridge: Cambridge University Press.

Bhatt, Rajesh. 1998. Obligation and possession. *Papers from the UPenn/MIT roundtable on argument structure and aspect, MITWPL* 32. 21–40.

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

Dryer, Matthew S. 2005. Determining dominant word order. In Haspelmath, Martin & Dryer, Matthew S & Gil, David & Comrie, Bernard (eds.), *The world atlas of language structures*, 371. Oxford: Oxford University Press. ((http://wals.info/chapter/s6))

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

Haspelmath, Martin. 2003. The geometry of grammatical meaning: Semantic maps and crosslinguistic comparison. In Tomasello, Michael (ed.), *The New Psychology of Language*, vol. 2, 211–243. New York: Lawrence Erlbaum. (10.5281/zenodo.831410)

Haspelmath, Martin. 2010. Comparative concepts and descriptive categories in crosslinguistic studies. *Language* 86(3). 663–687. (doi:10.1353/lan.2010.0021)

Haspelmath, Martin. 2013. Argument indexing: A conceptual framework for the syntax of bound person forms. In Dik Bakker & Martin Haspelmath (eds.), *Languages across boundaries: Studies in memory of Anna Siewierska*, 197–226. Berlin: De Gruyter Mouton.

Haspelmath, Martin. 2019a. Can cross-linguistic regularities be explained by constraints on change? In Karsten Schmidtke-Bode, Natalia Levshina, Susanne Maria Michaelis & Ilja A. Seržant (eds.), *Competing explanations in linguistic typology*, 1–23. Berlin: Language Science Press. http://langsci-press.org/catalog/book/220.

Haspelmath, Martin. 2019b. Ergativity and depth of analysis. *Rhema* 2019(4). 108–130. (doi:10.31862/2500-2953-2019-4-108-130)

Haspelmath, Martin. 2020. The structural uniqueness of languages and the value of comparison for description. *Asian Languages and Linguistics* 1(2). 346–366.

Haspelmath, Martin. 2021a. Explaining grammatical coding asymmetries: Form-frequency correspondences and predictability. *Journal of Linguistics* 57(3). 605–633. (doi:10.1017/S0022226720000535)

Haspelmath, Martin. 2021b. General linguistics must be based on universals (or nonconventional aspects of language). *Theoretical Linguistics* 47(1–2). 1–31. (doi:10.1515/tl-2021-2002)

Hjelmslev, Louis. 1943. *Omkring sprogteoriens grundlæggelse*. København: Ejnar Munksgaard.