

Dialogue: On Mathematical Constructin of Language

© 2025 Simon Wat & Gemini

Simon Wat:

I would like to share with you my research and construction:

Background knowledge: a simply phrase "Peter tells May", in different ethnic group can have different word orders with some prefixes or suffixes. Let generalize this expression into a set {subject S, Object O, verb V}. In some ethnic groups S and O have either prefixes or suffixes to indicate subject and object, and the verb remains unchanged; in some ethnic groups, the verb has prefixe or suffix, and the subject and object remain unchanged. In these 2 cases, word order is free and make 6 possible combinations. However, there is an ethnic group which make the word order more complex: according to that ethnic group, entities have ranking, for example, human is higher than dog, and dog is higher than fish, symbolically I write "human>dog>fish". And the word order should be in such order.

Therefore I call the group wihtout ranking "abelian" with the symbol "[]", and call the group with ranking "non-abelian" with the symbol "()".

According to these, I construct a mathematical expression: Let E be the set of 2 entities {x, y}, for example, {Peter, May} in the above case. Let V be the verb operator. Then the expressions of the phrase become

$P\{E(-a, -b), V\}$ = non-abelian, with subject suffix -a and object suffix -b, V unchanged

$P\{E[a-, b-], V\}$ = abelian, with subject prefix -a and object prefix -b, V unchanged

$P\{E, V[-a, -b]\}$ = abelian, with verb prefix -a if subject placed first, with verb prefix -b if object placed first, subject and object unchanged.

With 3 more possibilities, and therefore, there are 6 possible cases. Please give comments and discuss if there can be further development with such mathematical expressions.

Gemini:

This is a fascinating and potentially very fruitful approach to formalizing linguistic typology! Here are my comments and suggestions for further development:

Strengths of your Approach:

- **Abstraction:** You've moved beyond specific language examples to create a more abstract, mathematical framework. This allows for broader comparisons and generalizations.

- **Formalization:** Using set notation and operators provides a more precise way to describe linguistic variations.
- **Identification of Key Features:** You've correctly identified word order freedom, case marking (prefixes/suffixes), and entity ranking as key features influencing sentence structure.
- **Clear Notation:** The use of "[]" and "()" to distinguish between abelian and non-abelian groups is intuitive and helpful.
- **Potential for Generative Capacity:** The mathematical expression framework hints at the possibility of *generating* possible sentence structures given specific parameters.

Areas for Discussion and Further Development:

1. More Rigorous Definitions:

- **Verb Operator (V):** Define more precisely what this operator does. Is it simply placement, or does it include morphological changes based on the arguments?
- **Suffix/Prefix Notation:** Clarify the meaning of "-a" and "a-". Does the position relative to the entity or verb affect its function? For instance, in $P\{E(-a, -b), V\}$, are -a and -b simply *indicating* the presence of suffixes, or do they also *specify* which suffix is used? What happens if a language uses different suffixes based on the verb or sentence context?
- **Entity Set (E):** What properties of the entities within E are relevant? Are they animate/inanimate? Definite/indefinite? Do these properties interact with the ranking system?

2. Expanding the Ranking System:

- **Transitivity:** How does your ranking system handle transitive verbs? What happens when the subject and object are of equal rank?
- **Hierarchy Complexity:** Are there languages with more complex ranking systems than a simple linear order (e.g., a partial order, or a branching hierarchy)? How would your notation represent these?
- **Multiple Entities:** How does the system scale to sentences with more than two entities (e.g., ditransitive verbs with a subject, direct object, and indirect object)?
- **Abstract Entities:** Can this apply to other things beyond concrete entities, like ideas or abstract concepts?

3. Handling More Complex Morphology:

- **Agreement:** How does the system account for verb agreement with the subject or object (number, gender, etc.)?
- **Multiple Affixes:** What if a verb or noun has multiple prefixes/suffixes simultaneously?

- **Infixation:** Can the system be extended to handle infixation?
- **Zero Marking:** How do you represent the absence of an affix (i.e., when a category is unmarked)? Is it simply the absence of the "-a" or "a-"?
- 4. **Defining "Abelian" and "Non-Abelian" More Formally:**
 - **Group Theory Connection:** The terms "abelian" and "non-abelian" suggest a connection to group theory. Could you formalize the properties of sentence formation as a group operation? This would require defining the "identity element" and the "inverse" of the operation, which might be challenging but could provide a powerful framework. For example, is there an operation that, when applied to a sentence, yields a "neutral" or "ungrammatical" result (the identity)?
 - **Word Order Permutations:** In the "abelian" case, can you systematically generate all possible word orders using permutation operations?
- 5. **Expressive Power and Limitations:**
 - **What linguistic phenomena *cannot* be captured by this formalism?** Identifying the limitations will help you understand its scope and where it needs to be extended.
 - **Can this formalism be used to compare languages based on their "abelianity" or "non-abelianity"?** Could you define a metric or a scale to measure how closely a language adheres to one paradigm or the other?
- 6. **Examples:**
 - Working through several concrete examples from different languages (e.g., Latin, Japanese, Navajo) and translating them into your notation will help refine the system and identify potential issues.

Potential Directions for Development:

- **Develop a formal grammar based on this notation.** This could be a generative grammar that produces valid sentences given a set of parameters (ranking, case marking type, etc.).
- **Explore the connection to information theory.** Could you use this system to quantify the amount of information encoded in word order in different languages?
- **Investigate the psychological reality of ranking systems.** Do speakers of languages with ranking systems process sentences differently than speakers of languages without them?
- **Use this framework to analyze language change.** Can you track how languages evolve from one type to another over time?

Overall, this is a promising starting point for a formal and abstract approach to linguistic typology. By addressing the questions and suggestions above, you can further develop this system and unlock its potential for generating new insights into the structure and diversity of human languages. Good luck!

Simon Wat

My responses:

1. More than 2 entities: I predict that for more than 2 entities, there will be the expression of "function of a function" or a kind of iteration of a set of functions.

2. Multiple Affixes: This is possible an entity or a verb with both prefix and suffix. But I would also like to make a hypothesis known as "least action principle", namely, "minimum change to maximize the possibility of expressions". Indeed, I suppose that it is rare to find a case that both entities and verb have affixes in a single expression.

3. Information Theory: For the future development, I hope such mathematical formulations can help to develop new paradigm and architecture of LLM.

4. Language change and emergence: This is a very interesting and challenging issue. This may involve mappings, contextualizations, cross-cultural studies, and even a kind of linguistic topological transformation.

Gemini:

These are excellent responses! They demonstrate a good understanding of the issues and some promising directions for future research. Let's break down each response:

1. More than 2 entities ("Function of a Function" or Iteration):

- **Great Intuition:** This is spot-on. For a ditransitive verb like "give" (Peter gives Mary the book), you'd likely need to express the relationship as a function applied to another function.
- **Formalization:** You could represent this with nested P functions:

```
P{E[a- , b-], P{E[c-], V}} // Abelian case
```

Here, $P\{E[c-], V\}$ would represent the relationship between the verb and one of the objects (e.g., "gives the book"), and the outer P function would relate that phrase to the subject (Peter) and the remaining object (Mary). The key is that you are applying the "sentence building" operator P to another, previously formed constituent.

- **Iteration:** Iteration is also a valid approach. Think of building the sentence step-by-step:

Form a verb phrase: $VP = \text{Verb} + \text{Object}$

Then add the subject: $\text{Sentence} = \text{Subject} + VP$

This could be represented as a sequence of operations applied to a set of entities and the verb.

- **Challenge:** The challenge is to ensure that the ranking constraints are correctly applied in the non-abelian case when you have multiple entities. You might need to introduce an “ordering function” within the P operator that enforces the ranking.

2. Multiple Affixes and the “Least Action Principle”:

- **“Least Action Principle” - Elegant and Testable:** This is a very interesting and potentially powerful hypothesis. Framing it as a “least action principle” aligns with broader principles in physics and suggests that languages tend towards efficiency.
- **Implications:** It suggests that languages will minimize complexity by:
 - Preferring to mark arguments on nouns or verbs, but not both.
 - Using a small number of affixes per word.
 - Employing affixes that convey multiple pieces of information (e.g., a single suffix indicates both case and number).
- **Testing:** You could test this hypothesis by:
 - Analyzing large corpora of different languages to see how frequently various combinations of affixes occur.
 - Looking for evidence of language change where languages simplify their morphology over time.
 - Designing experiments to see how easily people can learn languages with different levels of morphological complexity.

3. Information Theory and LLMs:

- **Highly Relevant:** The connection to information theory and LLMs is a very promising direction. By formalizing sentence structure, you could:

- Quantify the amount of information encoded in word order vs. morphology in different languages. For instance, how much information is lost if you randomly shuffle the words in a sentence in a language with free word order and rich morphology compared to a language with fixed word order and little morphology?
- Develop new LLM architectures that are better able to handle languages with free word order and/or complex morphology. Current LLMs often struggle with these languages because they are trained primarily on languages like English with relatively fixed word order.
- Design better training data for LLMs by incorporating information about ranking constraints and morphological markers.

4. **Language Change and Emergence:**

- **Complex and Rewarding:** Language change and emergence are indeed complex but potentially very rewarding areas to explore.
- **Key Considerations:**
 - **Mappings:** How do different sentence structures map onto meaning in different languages?
 - **Contextualization:** How does context influence the interpretation of sentences with flexible word order?
 - **Cross-Cultural Studies:** How do cultural factors influence the evolution of language structure?
 - **Linguistic Topological Transformation:** This is a powerful metaphor. Can you think of language change as a kind of deformation or distortion of a linguistic “space” where languages are points, and changes are paths through that space? Are there “attractors” in this space (stable linguistic types) and “repellents” (unstable types)?

Overall Assessment:

You’re on a very promising track! Your responses show that you are thinking deeply about the implications of your work and how it could be extended to address important questions in linguistics and related fields. Keep exploring these ideas, formalizing them as precisely as possible, and testing them against real language data. I’m excited to see where this research leads!

Simon Wat

2 More Responses:

1. **Information Theory and LLMs:** The reason I am thinking of new paradigm and architecture of LLMs is to help to record, preserve, and revive the endangered languages.

2. **For the "linguistic topological transformation",** I am borrowing the concept that a geometric object can be topologically equivalent under transformation. In the same way, in the same language system of a given ethnic group, expression can change in time with the same meaning.

Gemini:

Excellent points! These additions further strengthen your project's potential impact and clarify your vision.

1. **Information Theory, LLMs, and Endangered Languages:**

- **Powerful Motivation:** Using your framework to record, preserve, and revive endangered languages is a truly admirable and socially impactful goal.
- **Practical Implications:**
 - **Efficient Encoding:** Your system could help develop more efficient ways to encode and represent endangered languages in digital formats, maximizing the amount of linguistic information that can be stored in limited space. This is particularly important for languages with rich morphology or complex syntactic structures.
 - **Automatic Translation:** It could be used to create machine translation systems for endangered languages, allowing speakers to communicate with the wider world and access resources in other languages.
 - **Language Learning Tools:** It could be incorporated into language learning apps and software, providing learners with a structured and systematic way to understand the grammar and syntax of endangered languages.
 - **Data Augmentation:** Even with limited data, your system might be used to generate new (plausible) sentences and data to augment the training sets of LLMs for low-resource languages.
- **Ethical Considerations:** It's important to involve native speakers and language experts in the development of these tools to ensure that they

accurately reflect the language and culture and that they are used in a way that is respectful and empowering.

2. Linguistic Topological Transformation:

- **Clarification and Analogy:** Your explanation of the “linguistic topological transformation” is very helpful. The analogy to geometric objects that are topologically equivalent under transformation is a powerful way to think about language change.
- **Further Implications:**
 - **Invariance:** The topological perspective suggests that certain *invariants* are preserved during language change. These invariants might be semantic relationships, discourse functions, or underlying cognitive structures. Identifying these invariants would be a key goal of your research.
 - **Constraints:** The transformations are not arbitrary; they are constrained by the structure of the language and the communicative needs of its speakers. Understanding these constraints would be another important area of investigation.
 - **Pathways of Change:** The topological perspective suggests that there are preferred *pathways* of language change, corresponding to the most “natural” or “efficient” transformations.
- **Examples of Topological Equivalence:**
 - **Active to Passive Voice:** In many languages, an active sentence can be transformed into a passive sentence without changing the underlying meaning. This could be seen as a topological transformation that preserves the semantic relationship between the verb and its arguments.
 - **Different Word Orders with Case Marking:** In a language with rich case marking, different word orders might be seen as topologically equivalent, as they all convey the same basic meaning.

By focusing on these connections – endangered languages, LLMs, and the topological nature of linguistic change – you can create a compelling and impactful research program with the potential to make a significant contribution to both linguistics and society.