Distinguishing features of natural languages

Peter M. Hines

York Cross-Disciplinary Center for Systems Analysis

University of York

YCCSA – York – Jan. 2019

Mathematical models of natural language

"We logicians (and non-logicians) are traditionally reluctant to think of **language** *as a* **material phenomenon embedded in space and time**.

We have learned along the years to treat reasoning as a purely dis-incarnated and formal activity living in the æther of symbolic logic. This formalist inclination of the field is probably temporary..."

– Paul-André Mellièrs (Abramsky Festschrift - 2013)

Origins and References

This is loosely based on a talk given at:

"Compositional Approaches to Physics, Natural Language, and Social Science"

Nice, France (Sept. 2018)

and resulting paper:

"Information Flow in Pregroup Models of Natural Language"

E.P.T.C.S. (Nov. 2018)

The general area of the talk:

A subtle distinction

This talk is not just about mathematical linguistics; it is about linguists:

- What they do
- How & why they do it
- What assumptions / preconceived ideas are involved.
- Whether these assumptions are warranted
- Whether they have any concrete implications

Mathematical models of Language?

Why do we use *Mathematical* models of *Natural Language*?

— *we do not refer to* **Chemistry** *as*

Mathematical Models of Molecules.

Language: the most complex subject we will ever master?

- Learning & relating words and concepts.
- Manipulating exceedingly complex grammatical rules.

We have an (inaccurate) tendency to believe our native language is straightforward, and other languages are unreasonably awkward!

Surprisingly complex structures!

An illustration from **"A Computational Approach to Biblical Hebrew Conjugation"** – *N. Yonofsky, J. Lambek*

We distinguish 140 possible finite verb forms $C_{i,j,k}(V)$ for every verb

⋮

The 140 verb forms are calculated by the formula $C_{i,j,k}(V) \mapsto P_{i,j}S_{j,k}(V)Q_{i,j}$
where $P_{i,j}$ and $Q_{i,j}$ are given by the following table

⋮

This table encapsulates a number of rewrite rules

⋮

The patterns $S_{j,k}(V)$ have the following shades of meaning

⋮

Before embarking on any actual calculations, we must state a number of phonological rewrite rules

⋮

[23 pages later ...] We have shown how the *finite* forms of *regular* verbs can be calculated

⋮

We are not satisfied with our efforts for the following reasons: **[List of simplifications made / gaps in account]**

Complexity of grammars in novel languages

The study of Pidgin & Creole languages (– Muysken & Smith)

Pidgin Language "Speech-forms which do not have native speakers and are used as a means of communication between people who do not share a common language."

Creole Language What a Pidgin becomes when it acquires native speakers: "The children of mixed marriages frequently grow up speaking the pidgin as their native language".

"Creole languages frequently develop as the result of linguistic (and often social) violence."

Introducing complexity where there was none

Very different grammatical structures:

- Pidgins have (unsurprisingly) highly simplified grammars.
- The same does not appear to be true for Creole languages!

> *"Creole languages are not in the slightest distinguishable from other languages. To claim a language as a Creole, we need to know something about its history"*
>
> — *Muysken & Smith*

A remarkable claim!

Are languages that arise in a single generation

structurally indistinguishable from

Languages that have evolved over hundreds of years?

Muysken & Smith again ...

"...this inevitably means that there may be many unrecognized creole languages around the world."

Whenever a high level of grammatical complexity is absent,

it is rapidly introduced to a language!

Why do we ‘need’ such complex structures?

- Necessary for understanding?

  — We can understand highly ungrammatical sentences!

- To communicate nuances of meaning?

  — This question has some cultural baggage!

  “Labov (1969) demonstrated that working class black youth were fully capable of abstract syllogistic reasoning, and that their non-standard vernacular dialect was not ’a basically nonlogical mode of expressive behavior’ as some psychologists had alleged.”

  “Adequacy, Expressiveness, and the Creole Speaker” – J. Rickford, J. Linguistics (1986)

- Is it something we simply enjoy?

  “Mastering verbs in Spanish requires many years of practice. In Esperanto, you can do the same in 5 to 10 minutes . . . There are no exceptions [irregular verbs], not even the verb ‘to be’. ”

  — Verbs in Esperanto, Jakub Marian (2017)

- Is it a ‘peacock’s tail’, inevitable, or even illusory?

How to get a handle on this complexity?

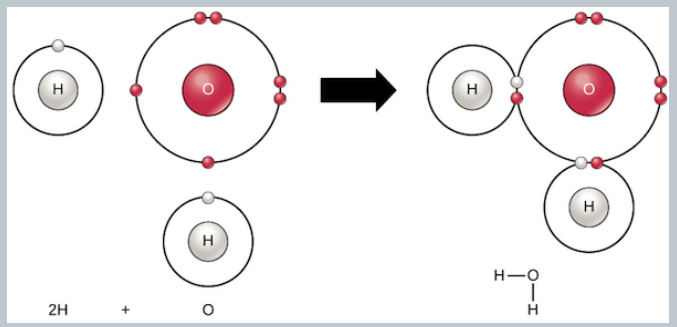
Claim: The mathematics is there to give *as simple as possible* an account of *highly complex* structures.

Question: Does it actually do this?

Some motivating ideas

Charles Sanders Peirce

"A transitive verb behaves like a molecule with two unsaturated bonds, one looking for a subject, and one looking for an object"



This views grammar as a 'typing system' that controls how linguistic entities may be 'plugged together'.

A system for grammar, from J. Lambek

Pierce is quoted by J. Lambek, as motivation for his **pregroup models** of natural language grammar.

Pregroup grammars are:

- popular among mathematically inclined linguists.
- currently fashionable in (some branches of) Natural Language Processing.
- good illustrations of how & why linguistic models are made and used.

Two very different questions!

~~What makes pregroups particularly suitable for linguistics~~?

Why do (some) linguists use pregroups?

Motivation for working with pregroups

1. Undoubted mathematical elegance.
2. A convenient & powerful graphical formalism.
3. Deep connections to many other areas of mathematics.
   - Linear logic
   - Category theory
   - Topology
4. They seem to model languages (quite) well.

1. “Undoubted Mathematical Elegance”

Simply as abstract algebra:

A **pregroup** is a monoid P with a partial order $\leq$, satisfying:

Compatibility of order & composition

$$p \leq q \text{ and } r \leq s \Rightarrow pr \leq qs$$

Existence of left- and right- adjoints

For every $p \in P$, there exists p^l, p^r satisfying

$$p^l p \leq 1 \leq p p^l \quad \text{and} \quad p p^r \leq 1 \leq p^r p$$

1. “Undoubted Mathematical Elegance” (cont.)

Simple consequences of the axioms:

- Adjoints reverse order: $(uv)^r = v^r u^r$ and $(uv)^l = v^l u^l$
- Left and right adjoints are dual: $(a^r)^l = a = (a^l)^r$
- Adjoints are unique: $qp \leq 1 \leq pq \;\Rightarrow\; q = p^l$
- The identity is its own adjoint: $1^r = 1 = 1^l$.

Such *derived properties* are frequently used

as *axioms* for other (lesser?) algebraic systems.

2. “Powerful graphical formalism”

A convenient notation

1. The **contractions** $aa^r \leq 1 \quad , \quad a^l a \leq 1$ are represented by (nested) underscores:

$$\underline{a^l \; \underline{a \; a^r} \; a}$$

2. The **expansions** $1 \leq a^r a \quad , \quad 1 \leq a a^l$ are represented by (nested) overscores

$$\overline{a \; \overline{a^r \; a} \; a^l}$$

2. “Powerful graphical formalism” (cont.)

We may combine and overlap, under- / over- scores:

$$\overline{n^r \; \overline{n \; \underline{n^l} \; \underline{n}} \; \underline{n^l} \; \underline{n}}$$

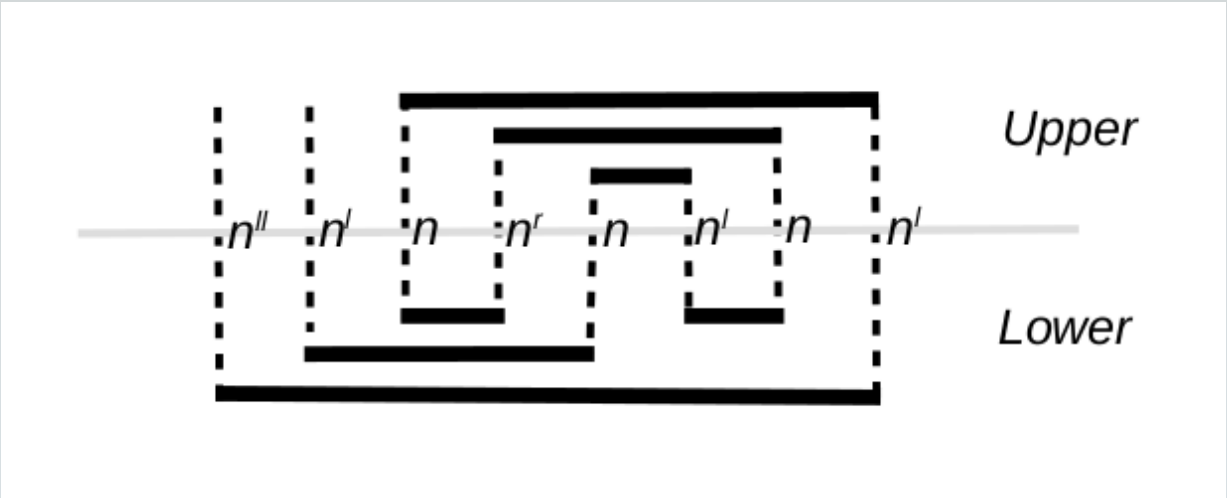
We may understand **algebra** (language?)
entirely in terms of **pictures**.

2. “Powerful graphical formalism” (cont.)

A straightforward result

The resulting diagrams are always *planar*

i.e. there are no crossings.



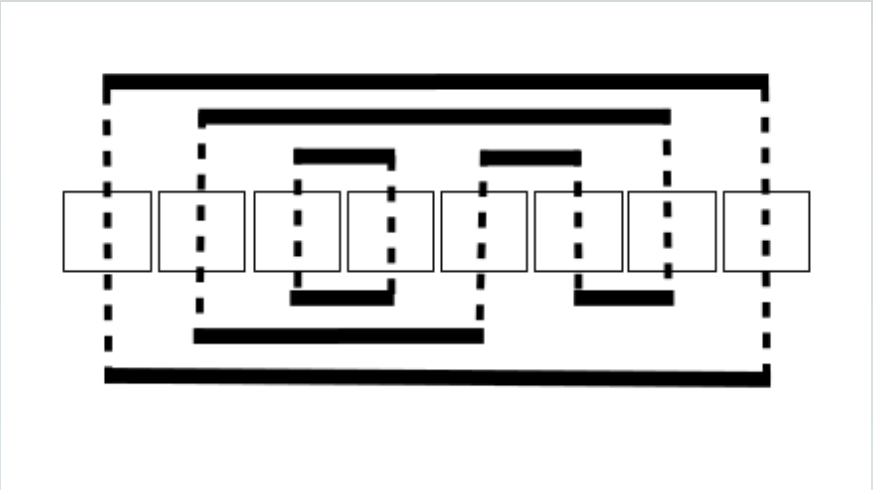
2. “Powerful graphical formalism” (cont.)

A slightly more subtle result

The types of pregroups used in linguistics (free pregroups) are characterised by:

“It is not possible to form closed loops”

— “Information flow in pregroup models of language” (EPTCS 2018)



3. “Deep connections to other areas of maths”

Possibly only of interest to mathematicians!

The graphical results are equally:

Algebraically	Normal form results.
Logically	Cut-elimination properties.
Categorically	Coherence theorems.

4. “Model languages (quite) well”

Let’s have a look ...

how do linguists use pregroups?

The key to grammatical interpretations

Think of the partial ordering as an *information ordering*:

$$\textit{Specific} \leq \textit{General}$$

states that *Specific* is a **special case** of *General*.

The game we play: We fix some distinguished **sentence** type in a pregroup, and look for elements beneath this type.

The toy models we teach ...

We have two special elements in our pregroup:

The **sentence type** s, and the **noun type** n.

Following C.S. Pierce ...

A **transitive verb** might have type $n^r s n^l$.

It 'seeks' a noun to its left, and its right.

Mathematically:

$$n.n^r s n^l.n \leq 1.s.1 = s$$

$(Noun)(TransitiveVerb)(Noun)$ is a special case of $(Sentence)$.

Entirely graphically:

- **Sentences** have type s
- **Nouns** have type n
- **Transitive verbs** have type $n^r s n^l$.

The pattern of underscores:

$$\underline{n\ n^r}\ s\ \underline{n^l\ n}$$

demonstrates that

$$(Noun)(Transitive\ Verb)(Noun)$$

is a special case of *Sentence*.

From toy examples to the real world

In actual languages

$$(\textit{Noun})(\textit{Transitive Verb})(\textit{Noun})$$

is a very coarse approximation of a grammatical structure:

"Man bite Dog."

"Tourist buy Hat"

"Student hate Mathematics"

A real-world example (Cymraeg):

A Modern Welsh sentence:

Dyma fy nghath i

A word-for-word translation			
Dyma	fy	nghath	i
Here is	**my**	**cat**	**??**

Dyma Defective verb

fy Possessive adjective / weak pronoun

nghath Noun *(mutated form)*

i Confirming pronoun

Welsh mutations

In Modern Welsh: The start of nouns, adjectives, and verbs varies, according to the grammatical context:

Cat	Cath	My cat	Fy nghath
Father	Tad	Your Father	Dy Dad
Bally	Bali	to Bally	i Fali
Bread	Bara	Bakery	Siop fara
I heard	Clywais i	I didn't hear	Chlywais i ddim

Confirming pronouns:

No English analogue of confirming pronouns ...

... part of spoken (rather than written) Welsh.

... differ significantly between N. and S. Wales.

Similar constructions in (rather informal) French:

Dyma	**fy**	**nghath**	**i**
Voici	**mon**	**chat**	**à moi**

A pregroup formalisation

We use the pregroup ‘freely generated by’:

- s [sentence]
- n [noun phrase]
- n_p [noun (pos.)]
- c_1 [conf. pronoun (1st person)]

Dyma	**“here is”**	sn^l
fy	**“my”**	$nc_1^l n_p^l$
nghath	**“cat”** [pos.]	n_p
i	[confirming pronoun]	c_1

A demonstration of grammaticality:

Pregroup typing for Welsh grammar

Dyma	**fy**	**nghath**	**i**
sn^l	$nc_1^l n_p^l$	n_p	c_1

The pattern of underscores

$$s\ \underline{n^l\quad n}\ \underline{c_1^l\ \underline{n_p^l\quad n_p}\quad c_1}$$

demonstrates that

$$sn^l nc_1^l n_p^l n_p c_1 \ \leq\ s$$

i.e. we have special case of the sentence type.

Some notable points

In this particular example:

1. We have used (single) left- adjoints only.
2. There are no possible expansions.

In general:

1. Most grammatical types are modeled by composites.
2. The contractions take place between symbols in distinct (natural language) words.
3. Expansions / overscores play no role in any linguistic applications.

Entia non sunt multiplicanda praeter necessitatem?

Linguistic applications have no use for overscores / expansions.

Occam's Razor alert

Lambek's pregroups appear to have precisely twice the amount of structure that is actually needed.

- This was noted by J. Lambek in his original paper(!)
- Spurious justifications have been proposed:

> *"The extra structure, although not linguistically necessary, is needed to determine the algebra" – J. L.*

What about protogroups ??

A philosophical confrontation

Mathematical Platonism

"Our choice of structures for modeling natural language is determined by their mathematical elegance & æsthetic appeal"

Pragmatism & Occam's razor

"Language is already very complicated. Why should we use structures that are twice as complex as necessary?"

Is any reconciliation possible between these viewpoints?

An alternative perspective

Information Content vs. Information Flow

Inspiration from Natural Language Processing

Compositional Distributional Semantics

- Introduced by S. Clarke, B. Coeke, M. Sadrzadeh
- A very practical branch of N.L.P.
- Uses a highly degenerate form of pregroups
- *Claims an interpretation for overscores / expansions.*

A scientific(?) hypothesis

The Categorical Hypothesis

The pattern of underscores & overscores in a (grammatically correct) sentence models the ‘flow of information’ or ‘causal connections’ between the distinct words of the sentence.

The obvious question: (W.T.F.) Where does That come From??

We also need to ask:

- Is this testable?
- What assumptions are made?
- Are there any concrete implications?

An argument by analogy

This comes from:

The identification of Lambek pregroups within a particular branch of mathematics

We see the same structures in a wide range of fields:

- Linear logic, Geometry of Interaction, & game semantics [Abramsky 96, PMH 97, Haghverdi 2000]
- Turing machines [PMH 03,08], lambda calculus [Abramsky, Haghverdi, Scott 03] and its models [PMH 01]
- QM teleportation [Abramsky, Coecke 04, PMH, Braunstein 09]
- Program semantics [Lutz, Derby (implicitly) 84, PMH 08]

The interpretation of under / over scores is similar in each case.

Back to linguistic considerations

What does the Categorical Hypothesis

assume, or imply,

for pregroup models of natural languages?

Information flow in sentences

The interpretation

"Underscores & overscores model the information flow between words in a sentence"

Trivially, replies upon:

"There is information flow between words in a sentence."

Algebraically: all words in the model of a (grammatically correct) sentence are directly or indirectly connected by under- / over- scores.

A linguistically important property of pregroups?

How could we formalise this?

We need to:

- Rigorously describe the process of making a pregroup grammar.
- Axiomatise the intuitive notion of ‘direct or indirect connection’.
- State & prove a theorem about *pregroups generally*.

A relevant example

Assume the following grammatical types:

$$\{SENTENCE, FOO, BAR, DOG, DUCK\}$$

and a model in the free pregroup over $\{s, a, b, c\}$

SENTENCE	s
FOO	sac^l
BAR	ca^r
DOG	$a^r b^l$
DUCK	ba^{rr}

A sentence with the type $FOO.BAR.DOG.DUCK$ is grammatically correct:

$$\underline{s\ a\ \underline{c^l\ c}\ a^r}\ \ \underline{a^r\ \underline{b^l\ b}\ a^{rr}}$$

Grammatical, but content-free

Let us compare the underscores

$$s\ \underline{a\ \underline{c^l}\ c\ a^r}\ \underline{a^r\ \underline{b^l}\ b\ a^{rr}}$$

with the typing of individual words:

FOO	BAR	DOG	DUCK
(sac^l)	(ca^r)	$(a^r b^l)$	(ba^{rr})

Similarly to the Welsh example, there are no possible overscores at all.

In contrast to the Welsh example, there is no information flow between the two halves of this sentence.

What is, and should never be

Towards a conjecture ...

We do not expect *this kind of behaviour*

in (pregroup models of) natural languages

This, of course, needs formalising.

Creating a pregroup model for a natural language

1. We have some a priori set G of **grammatical types**

$$G = \{SENTENCE,\ ARTICLE,\ TRANS\_VERB,\ NOUN,\ PRONOUN,\ CONF\_PRONOUN, \ldots\}$$

   together with *some mapping*[1] from **natural language words** to grammatical types.

2. A **pregroup model** is a function $model : G \to P$, for some pregroup P.

3. This extends to strings of members of G in the obvious way.

[1] This may be multi-valued, context-dependent, probabilistic, &c.

A relevant definition:

Given a string of types $w = T_1 T_2 T_3 \ldots T_n$

The causal graph $\mathfrak{C}_w$ is defined by:

Nodes These are $model(T_1), model(T_2), \ldots$

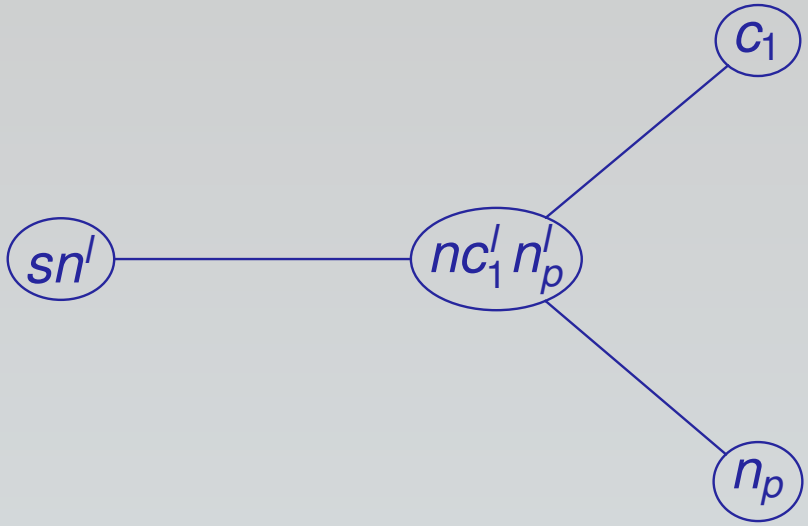
Edges Two nodes are linked by the underscores / overscores connecting them in

$$model(T_1 T_2 T_3 \ldots T_n)$$

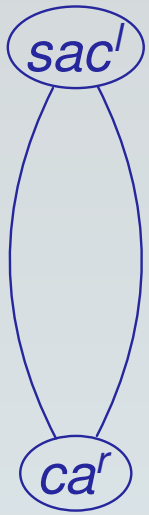
Best illustrated by example!

An explanation in 2000 words

Dyma fy nghath i



FOO BAR DOG DUCK



Formalising a conjecture

Given:

- A pregroup model $model : G \to P$,
- A distinguished grammatical type $S \in G$,

we say that *model* is S**-connected** when, for all strings w,

$$model(w) \leq model(S) \;\Rightarrow\; \mathfrak{C}_w \text{ is a connected graph.}$$

The connectedness hypothesis

Pregroup models of natural languages

are Sentence-connected.

Caution: It is much harder to prove connectedness for a *model* than for a particular *example*.

The status of the hypothesis

Concrete predictions about *models of Natural Languages*.

Provable? No — as difficult as proving all swans are white!

Disprovable? Yes – by a single convincing counterexample!

What about potential counterexamples?

- These would need to be *generally accepted*.
- It is easy to claim that, *'the pregroup typing is incorrect'*.
- Failure of connectivity could be taken as *evidence* of this!

Is there anything special about *Sentence*?

Games we can play with a hypothesis (I)

Extending the hypothesis:

Various stronger versions ...

Do we also expect a pregroup model $model : G \rightarrow P$ to be:

- $NOUN\_PHRASE$ - connected?
- $VERB\_PHRASE$ - connected?
- T-connected, for any grammatical type $T \in G$?

Games we can play with a hypothesis (II)

Look for 'reasonable exceptions':

Where would failure of connectivity be 'reasonable'?

How about for connectives?

In something like "$SENTENCE_1$ *and* $SENTENCE_2$", do we expect 'causal connection' between the two sentences?

Caution: pregroup models of connectives are rather complex.

"Types and forgetfulness in categorical linguistics" (PMH 2013)

Games we can play with a hypothesis (III)

Look for algebraic characterisations::

Can we characterise pregroup models $model : G \to P$ that are:

- S-connected, for some distinguished type S?
- T-connected, for *all* $T \in G$

... some results on this in EPTCS paper

Looking for meaning where there is none?

Games we can play with a hypothesis (IV)

Treat it as a ‘working assumption’.

Somewhat speculatively(!)

Consider a document with:

- a well-analyzed grammar,
- no known / generally accepted meaning,
- doubts as to whether it represents a natural language!

Is it reasonable to draw any conclusions from the presence or absence of T-connectivity?

What is a type, anyway?

Games we can play with a hypothesis (V)

Consider it as (part of) a definition.

The definition of types

We have treated grammatical types as though they have been 'handed down from on high'.

In practice, they may be thought of as:

Semantic constructs *Structural Grammar*

Syntactic constructs *Generative Grammar*

What do we -not- accept as a type?

Consider languages with a **V**erb-**S**ubject-**O**bject word order:

Biblical Hebrew

Bara	Elohim	et	ha-shamayim
Created	*God(s)*	[object-marker]	*the Heavens*

Modern Welsh

Gwellodd	Anwen	ddefaid
Saw	*Anwen*	*(a) sheep*

What is not a type?

In a **V-S-O** language

Both semantically and syntactically, we *could* treat

$$(Noun - phrase)\ [object - marker]\ (Noun - phrase)$$

as a single grammatical type that 'forms a double bond' with the $(Transitive - Verb)$ type . . . but we don't.

Conjecture

We could produce a *consistent* pregroup grammar by doing this, but not a *connected* one.