Proto-Indo-European: a language without Merge?

George Walkden
Workshop on Syntax and Reconstruction, York, 18.06.2018

The context

Comparative and diachronic work with a generative orientation has unquestionably led to a great many discoveries and insights. However, it’s also part of our duty as researchers to consider prominent claims made by scholars of very different theoretical persuasions – especially if those claims seem poorly evidenced, poorly motivated, or simply false.

Today’s talk is the early stage of a project to assess the following claim(s):

Proto-Indo-European: a language without Merge?

Comparative and diachronic work with a generative orientation has unquestionably led to a great many discoveries and insights. However, it’s also part of our duty as researchers to consider prominent claims made by scholars of very different theoretical persuasions – especially if those claims seem poorly evidenced, poorly motivated, or simply false. Today’s talk is the early stage of a project to assess the following claim(s):

Proto-Indo-European: a language without Merge?

The Parataxis-Precedes-Hypotaxis Hypothesis (PPHH) has a long history:

− The term parataxis in its modern sense was introduced by Thiersch (1826) in the context of historical German (opposed to syntaxis there; hypotaxis only in later works)
− Very prevalent in historical linguistics before the advent of structuralism (e.g. Götzenwerke 1883, Débdub 1907: 411; Small 1924: 125)
− Reiterated in more recent works with a functionalist orientation (e.g. Jucker 1991: 203; Deutscher 2001: ch. 11; Hajnawska 2015: 235)

But almost never explicitly addressed in the generative literature:

− Its influence can be seen in O’Neil (1977) and Kiparsky (1995)
− Rejected summarily in Roberts (2007: 174–175)
− Rejected summarily in Roberts (2007: 174–175)

Roadmap

Part 1: establish what the content of the PPHH is

− Spoiler alert: there are several different versions of the PPHH
− Only one in principle threatens standard Minimalist assumptions about the architecture of grammar
− That version is very obviously wrong (as far as we can tell)

Part 2: empirically evaluate a particular version of the PPHH

− Not one that is inherently problematic for generative linguists if correct
− But one that is interesting nonetheless
− Precondition for this kind of research: parsed diachronic corpora of various languages

The Parataxis-Precedes-Hypotaxis Hypothesis (PPHH) has a long history:

− The term parataxis in its modern sense was introduced by Thiersch (1826) in the context of historical German (opposed to syntaxis there; hypotaxis only in later works)
− Very prevalent in historical linguistics before the advent of structuralism (e.g. Götzenwerke 1883, Débdub 1907: 411; Small 1924: 125)
− Reiterated in more recent works with a functionalist orientation (e.g. Jucker 1991: 203; Deutscher 2001: ch. 11; Hajnawska 2015: 235)

But almost never explicitly addressed in the generative literature:

− Its influence can be seen in O’Neil (1977) and Kiparsky (1995)
− Rejected summarily in Roberts (2007: 174–175)
− Rejected summarily in Roberts (2007: 174–175)

Roadmap

Part 1: establish what the content of the PPHH is

− Spoiler alert: there are several different versions of the PPHH
− Only one in principle threatens standard Minimalist assumptions about the architecture of grammar
− That version is very obviously wrong (as far as we can tell)

Part 2: empirically evaluate a particular version of the PPHH

− Not one that is inherently problematic for generative linguists if correct
− But one that is interesting nonetheless
− Precondition for this kind of research: parsed diachronic corpora of various languages

Flavours of parataxis > hypotaxis

Ideas don’t arise in a vacuum. Some of the ways in which the PPHH is stated (and motivated) in earlier literature make for uncomfortable reading today.

− Mitchell (1985) approvingly quotes Small (1924: 125): “It may be laid down as a general principle that in the progress of language parataxis precedes hypotaxis.”
− Small’s following sentence: “The former is associated with the uncultivated mind; the latter, with the cultivated mind of civilized peoples.”
− Andrew (1940: 87): early Old English was characterized by “simply a lack of grammatical subordination such as we find in the language of children and some primitive people.”

This doesn’t mean that (every version of) the PPHH is wrong, of course. But claims (in science as elsewhere) may persist because of ideology rather than merit.
Version of the PPHH

Harris & Campbell (1995: 284): “In approaching the question of whether hypotheses develop out of parallelism we encounter the problem that different languages have in mind different ideas of parallelism, and that at least some of them are vague”

A non-exhaustive list:
1. Early human languages lacked Merge.
2. Early human languages lacked self-similar embedding.
3. Early human languages lacked finite subordinate clauses (specifically).
4. Diachronically, hypothetic structures develop out of paraletic structures.
5. Diachronically, adjacent structures replace ‘embedded’ structures.
6. Diachronically, hypothetic structures become more common. (possible causal factors: complex culture/ society; Latin, literacy from orally)

There may be entailment relations between these hypotheses. But each needs to be considered separately.

PPHH 1: Early human languages lacked Merge

(a.k.a. “The horse, the wheel, and Merge”)

• Merge, the operation hypothesized, in Minimalist syntactic theorizing (Chomsky 1985), to be fundamental to structure building.

• “Given any two distinct syntactic objects A, B Merge(A,B) = {A,B}.”

(Poll & Stabler 2016: 5)

PPHH 1 can quickly be put to rest:

• In a language without an operation such as category-neutral Merge (defined such that it can be applied to its own output):

  “No sentence could contain more than two words” (Nevins, Pesce, & Rodrigues 2009: 386)

  Neither Proto-Indo-European nor any other human language has ever been argued to display this property.

  “The cognitive ability to handle finite complementation must have already been a feature of the human brain in the more distant past” (Deutscher 2001: 184–185)

PPHH 2: Early human languages lacked self-similar embedding

As an empirical claim, this ought to be taken more seriously than PPHH 1.

- although Merge may not minimize combine any two lexical items or phrases an unbounded number of times, not every imaginable instance of Merge is acceptable in actual languages (Kruemel, Pesce, & Rodrigues 2009: 366)

- “It is theoretically possible, though unlikely, that some language might be so impoverished in lexical and other resources that only a finite number of non-deviant sentences could be generated by its GP (generative procedure) GPE. If so, it would be a minor curiosity, with no bearing on UG, acquisition, or other significant issues, contrary to much media confusion.” (Chomsky 2013: 36)

- Cf. also Pullum & Scholz (2010): “recursion does not guarantee infinite.

No grammatical theory I know of predicts that self-similar embedding is a necessary property of human languages.

- In the Triggered Merge formalization of Collins & Stabler (2016: section 7), it’s easy to construct a language where nothing triggers Merge of a constituent of the same type (or of a constituent that could contain a constituent of the same type).

PPHH 3: Early human languages lacked finite subordinate clauses


Assuming that all clauses are CPs, PPHH 2 entails PPHH 3.

Givón (1976: 368): “certain types of languages—those which have only coordination (‘clause chaining’) but no subordination—are found only in preliterate ‘societies of intimates’.

- Givón’s claim is actually stronger: not restricted to finite clauses, uses word “only”.

- Hard to view the Akkadian or Old Egyptian Empire as a non-literate society of intimates, but Akkadian is one of the languages where the case has been made best for lack of finite subordination (Deutscher 2001).

- This claim needs to be assessed with reference to present-day languages; obviously, we have no historical records of non-literate societies pre-20th century.

- Coordination is a classic instance of self-similar embedding, of course. PPHH 3 could be true without PPHH 2 being true in general.

Working definition of (clause) subordination:

- a CP is subordinated if it is dominated by another CP.

Are there languages without self-similar embedding?

It depends who you believe.

- Fulink et al. (2016), for Piraha, develop a grammar yielding a finite output.

  But this stipulates that “up to 3 instances” of certain elements are allowed.

  If instead unbounded repetition is allowed, then they “analyze Piraha as an infinite regular language” (Fulink et al. 2016: 30) – hence, historically attested languages are not a good testing ground.

- Wimler et al. (2017): no language in their sample of 55 Indo-European languages (present or historically attested) lacks NP- within-RP embedding.

No truly compelling case has yet been uncovered.

Are there languages without finite subordinate clauses?

- Roberts (2007: 174): “the claim that earlier stages of certain languages may have lacked subordination altogether violates the uniformitarian hypothesis, the idea that all languages at all times reflect the same basic UG, and so cannot be taken seriously in the approach adopted here.”

  This of course depends on what we think UG contains.

- “On the other hand, it is quite plausible that a language may lack finite clause subordination of the familiar type” (2007: 174).

- Deubrich (1990) claims that Proto-Indo-European lacked finite subordination, on the grounds that finite subordinators are not reconstructable.

  This is bad reasoning: no rejection, basic vocabulary

  Languages like Mandari have subordination but no (overt) subordinator

PPHH 3: some cautionary notes

For PPHH 3 to be correct, there has to be an asymmetry between early and more recently spoken languages in having/facilitating finite subordination.

- King & Cockburn (1980: 204): “We cannot … suppose that hypothesis is of recent origin in language; for as far as we can go back in the history of human speech, we find the degradation of sentences to a completely subordinate position fully established.”
- Giderstein (1983: xxx): “we have to be on our guard. Hypothesis is older than our record, and we cannot argue safely as to prehistoric processes”
- Even Karlsson (2003): “Evidence from many language families indicates that non-finite clause subordination and initial stages of finite clause subordination existed already in prehistoric languages.”

As a categorical claim, PPHH 3 is certainly false. Does it hold statistically? I’m not aware of any studies addressing the question.

PPPH 4: Diachronically, hypotactic structures develop out of paratactic structures

This is a very different beast. Classic case: reanalysis having two independent clauses.

- [I think that John is here] > [I think that John is here]
- Varied (PPHH 4a); actually very different grammatical operations of [adjacent] clause as [embedded]

This kind of reanalysis necessarily involves violation of Whitman’s (2000) “conservation of structure” constraint: c-command relations change.

Most famous case has been powerfully challenged (Ansel-Tober 2017). See Harris & Campbell (1985: 283–310) for sceptical discussion.

PPHH 4 is not the focus of this talk.

Interlude: adjunction in current theories of syntax


Chomsky (2001: 18): “For structure building, we have so far assumed only the free symmetrical operation Merge, yielding syntactic operations that are ale, all binary: call them simple. … But it is an empirical fact that there is also an asymmetric operation of adjunction, which takes two objects P and a and forms the ordered pair <P, a>. A merged is p. Roll-merge and pair-merge are descendants of substitution and adjunction in earlier theories.”

Pair Merge is prime face a departure from the Strong Minimalist Thesis, and is motivated by the interface with the C-I system, where it can yield predicate combination.

In cartographic and Kaynean approaches (Hayne 1994), adjunction is not different from specifier formation, and hence PPHH 4 is not formulable.

“It is fair to say that what adjunctions are and how they function grammatically is not well understood” (Hofmann & Nunes 2006)

PPHH 5: Diachronically, adjoined structures replace embedded structures

Since adjunction is formally a case of self-similar embedding par excellence, and involves (Pair) Merge, PPHH 5 is orthogonal to PPHH 1 and 2.


- Proto-Indo-European has only adjoined S; CP is innovated in the history of the subfamilies (e.g. Germanic)
- Wallenberg (2016) presents a supporting quantitative tendency: relative clause “extraposition” has been getting rarer for centuries in English, Icelandic, French and (e.g. Germanic).
- Whether Kiparsky’s version of PPHH 5 extends beyond Indo-European hasn’t ever really been addressed, to my knowledge.

Today’s results won’t bear directly on PPHH 5.

PPHH 6: Diachronically, hypotactic structures become more common

This is possibly the most frequently-encountered version of PPHH in the literature. Usually interpreted with respect to finite clause subordination in particular (PPHH 6a).

Dignanovska (2013: 230): “Further telling evidence can be gleaned from historical data. The earliest written texts in a language are usually highly paratactic; … while later texts typically show more use of subordination. The historical increase in the frequency of subordination is greatest”

Karnison (2009): “It is a well-known fact that, mainly due to Latin influences. German and English were syntactically most complex in the 17th century and Swedish in the 19th century”

PPHH 6 is a quantitative claim. It can only be assessed using quantitative data from historical corpora.

Implications of PPHH 6

PPHH 6 has no bearing on questions of grammatical architecture. But it is interesting nonetheless for a variety of reasons.

- If PPHH 1-6 is correct as far as the corpora are concerned, is it a “real change” in the sense of differences in knowledge of language between generations?
- Could in principle be an artefact of the texts available to us from different periods (poetry, literacy)
- Could in principle be a real, but non-linguistic, change
- If it is a “real-change”, and if the causal argument works (see next slide), it indicates that socio-cultural factors have an impact on language change (cf. ethnosyntax, Enfield 2002)
- If it’s not a “real change”, it has important implications for the variationist approach to syntactic change (Knoch 1989, Yang 2002, Pintzuk 2003, etc.) how much change in corpus frequency involves change in the weightings of different grammatical options?

But let’s assess the hypothesis first before speculating further!
Causal factors

The causal argumentation for all versions of the PPHH has varying levels of quality. Three broad groups of possible factors:

- Latin influence (e.g. Karlsson 2009)
  - Only really applicable to the Early Modern European written context
  - Non-finite clauses leading to an increase in finite subordination?
  - How likely is this to lead to a "real change" in principle?

- Quality > literacy
  - Chafe (1982) and Biber (1995) show that finite subordinate clauses are more common in spoken than in written texts.
  - But very difficult to disentangle "real change" from genre effects; cf. present day
  - Cultural complexity & communicative needs (e.g. Givón 1979; Deutscher 2001)
  - Important to avoid discredited notions of "primitiveness" and lower intelligence
  - Causal chain rarely made explicit (though see Deutscher 2001: 166–186)

- Orality > literacy
  - Financial unembedded clauses: IP-MAT* in Penn-style parsed corpora (includes e.g. imperatives, exclamatives, coordinated clauses)
  - Financial subordinate/embedded clauses: basically IP-SUB* in Penn-style parsed corpora (includes e.g. relatives, complement clauses, adverbial clauses)

Overview of hypotheses

<table>
<thead>
<tr>
<th>Categorical</th>
<th>Tendency</th>
<th>Structures</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Merge &gt; Merge</td>
<td>PPHH 1</td>
<td>?</td>
</tr>
<tr>
<td>No self-embedding &gt; self-embedding</td>
<td>PPHH 2</td>
<td>PPHH 6</td>
</tr>
<tr>
<td>No subordination &gt; finite subordination</td>
<td>PPHH 3</td>
<td>PPHH 6A</td>
</tr>
<tr>
<td>Causal adjunction &gt; embedding</td>
<td>PPHH 5</td>
<td>?</td>
</tr>
</tbody>
</table>

In the rest of today’s talk: PPHH 6a will be the focus.

Methods: investigating PPHH 6a

Crucially relies on availability of parsed diachronic corpora.

"Hypotaxis coefficient": proportion of finite clauses that are subordinate/embedded.

- Financial unembedded clauses: IP-MAT* in Penn-style parsed corpora
  - Includes e.g. imperatives, exclamatives, coordinated clauses
- Financial subordinate/embedded clauses: basically IP-SUB* in Penn-style parsed corpora
  - Includes e.g. relatives, complement clauses, adverbial clauses
- Some variation in how interrogatives are treated – ask me if interested (shouldn’t affect the overall results much)

Languages investigated: English, Icelandic, French, Portuguese, Irish, Chinese

Non-finite clauses left out of consideration entirely.

Does finite clausal subordination become more common over time?

<table>
<thead>
<tr>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>YCDE</td>
</tr>
<tr>
<td>PPMCME2</td>
</tr>
<tr>
<td>PPMCME</td>
</tr>
<tr>
<td>PPCMBE</td>
</tr>
</tbody>
</table>

English with genre

<table>
<thead>
<tr>
<th>&quot;Non-fiction&quot; (purple)</th>
</tr>
</thead>
<tbody>
<tr>
<td>something of a dustbin category.</td>
</tr>
</tbody>
</table>
| Legal texts high, genres and titles low
Icelandic

- IcePaHC (Weidenberg et al. 2011)
- Sagas typically have less than average hypotaxis

French

- MCVF (Martineau et al. 2010)
- Apparent early rise is exclusively due to dominance of verse texts in this period

Portuguese

- Tycho Brahe Corpus (Galves, Andrade & Faria 2017)
- News texts & dramas typically low

Old and Middle Irish

- Parsed Corpus of Old and Middle Irish (Lash 2014)
- Hard to generalize about genre

Chinese

- ChiParHC (Li 2017)
- Again, hard to generalize about genre

Evaluation

Mixed-effects binomial logistic regression using R and lme4 package
- Fixed effect: date
- Random intercept: text

Linear effect of time should at least be detectable if the hypothesis is correct!

Nagelkerke $R^2$, a measure of goodness of fit, calculated using Nakagawa & Schielzeth (2013) method and MuMIn R package. Gives percentage of variance explained by the model.

<table>
<thead>
<tr>
<th>Language</th>
<th>English</th>
<th>Icelandic</th>
<th>French</th>
<th>Portuguese</th>
<th>Irish</th>
<th>Chinese</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log odds</td>
<td>0.00022</td>
<td>0.00014</td>
<td>-0.00141</td>
<td>0.00325</td>
<td>0.00052</td>
<td>-0.00018</td>
</tr>
<tr>
<td>Marginal $R^2$</td>
<td>0.1%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>2.3%</td>
<td>6.4%</td>
<td>0.3%</td>
</tr>
<tr>
<td>Conditional $R^2$</td>
<td>12.4%</td>
<td>8.7%</td>
<td>9.6%</td>
<td>4.7%</td>
<td>2.9%</td>
<td>3.0%</td>
</tr>
</tbody>
</table>
Overview

No robust support for any version of the PPHH.
- English, Icelandic, Irish, Chinese: no consistent direction of change.
- French: increase in hypotaxis 1500–1200, but early texts are in verse.
- Portuguese: gentle but steady decrease in hypotaxis over the timespan of the corpus.

Does genre play a role? Yes, but irrelevant to the hypothesis as far as we can tell.

No robust support for any version of the PPHH.

Bonus languages: Latin, Slavic/Russian, Georgian

These corpora don’t have constituency parsing.
- Latin: PROIEL
- Slavic/Russian: PROIEL
- Georgian: Georgian National Corpus

Approximation to the hypotaxis coefficient: number of overt subordinators divided by the number of finite verbs.

This seems to work reasonably well (more testing needed).

Overview

Bonuses:
- Latin
- Slavic
- Russian
- Georgian

Overview

Conclusion

- It’s widely agreed that parataxis > hypotaxis.
  Much less widely agreed what this actually means.
- Focusing on the idea that finite clausal subordination becomes more prevalent over time, I have found no support for this in parsed electronic corpora of English, Icelandic, French, Portuguese, Irish, or Chinese.
  - Maybe the corpus annotation is wrong.
  - Maybe the choice of languages is wrong.
  - But more as parataxis = hypotaxis is an empirical question, the burden of proof should be shifting at least somewhat.
- Much future work suggests itself:
  - More languages. More consideration of genre. What about non-finite clauses?
  - Suggestions welcomed!
References (1)


References (2)


References (3)


References (4)