The diachrony of parataxis and hypotaxis

http://walkden.space/zurich.pdf

George Walkden
DCLS Colloquium, Zurich
November 2021
The context

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

parataxis > hypotaxis

(where “>” is to be read as “precedes”)

November 2021
The diachrony of parataxis and hypotaxis
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 Greek (opposed to **syntaxis** there; **hypotaxis** only in later works)
- Very prevalent in historical linguistics before the advent of structuralism (e.g. Gildersleeve 1883; Delbrück 1900: 411; Small 1924: 125)
- Reiterated in more recent works with a functionalist orientation (e.g. Jucker 1991: 203; Deutscher 2001: ch. 11; Dąbrowska 2015: 230)
- Influential in a different form in the generative tradition (O’Neil 1977, Kiparsky 1995)
Roadmap

Part 1: establish what the content of the PPHH is

- Spoiler alert: there are several different versions and much unclarity

Part 2: empirically evaluate a particular version of the PPHH:
    parataxis > hypotaxis as change in frequency

- Precondition for this kind of research: parsed diachronic corpora of various languages
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.
Flavours of parataxis > hypotaxis
What is parataxis? What is hypotaxis?

Harris & Campbell (1995: 284): “in approaching the question of whether hypotaxis develops out of parataxis we encounter the problem that different linguists have in mind different ideas of parataxis, and that at least some of them are vague”

**Basic intuition: parataxis is a ‘looser’ relation than hypotaxis.**

Unconnected utterances are paratactic. A complement clause is hypotactic.

But there’s a lot in between. What about:
- Coordination?
- Non-finite clauses?
- Embedded clauses other than complementation? (e.g. relativization)
- The non-clausal domain? (e.g. possessor recursion)

Some authors – e.g. Hopper & Traugott (2003: 177–178) – distinguish between hypotaxis and subordination.

**Before a claim about parataxis and hypotaxis can be evaluated, we need to know what these notions mean.**
A cognitive claim?

In the twentieth and twenty-first centuries, there have been no explicit claims that parataxis > hypotaxis in the historical record is a direct result of change in human cognitive capacities.

Working assumption (uniformitarianism; see Walkden 2019):
the human faculty of language (in whatever sense we understand this notion)
has not varied or changed during the period in which languages are historically attested.

The claim that parataxis > hypotaxis is not a cognitive claim.

- “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)
Do such languages exist? It depends who you believe.

- Futrell et al. (2016), for Pirahã, develop a regular 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 Pirahã as an infinite regular language” (2016: 20).
- Kornai (2014): “time and again we come across languages where only a finite presentation seems to make sense”.
  - But his list includes Akkadian, for which the case has only been made that it lacks finite complement clauses (Deutscher 2001), and dubious cases such as Proto-Uralic (Ravila 1960).
- “Any finite corpus or set of examples can be given a description as a finite language in principle” (Futrell et al. 2016: 3) – hence, historically attested languages are not a good testing ground.
- Widmer et al. (2017): no language in their sample of 55 Indo-European languages (present or historically attested) lacks NP-within-NP embedding.

No truly compelling case has yet been uncovered.
Variant: Early human languages lacked finite subordinate clauses
(Givón 1979; Karlsson 2009; O’Neil 1977: 207, tentatively)

Givón (1979: 306):
“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 Assyrian Empire as a non-literate society of intimates, but Akkadian is one of the languages where the case has been best made 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!
  - The history of English per se can’t tell us much about this version of the claim.

Working definition of (clausal) subordination:

a CP is subordinated if it is dominated by another CP.
Variant: Early human languages lacked (finite) subordinate clauses
(Givón 1979; Karlsson 2009; O’Neil 1977: 207, tentatively)

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 clausal subordination of the familiar type” (2007: 174).

- Delbrück (1900) claims that Proto-Indo-European lacked finite subordination, on the grounds that finite subordinators are not reconstructable.
  - This is bad reasoning; cf. negation, basic vocabulary
  - Languages like Mandarin have subordination but no (overt) subordinator
Some cautionary notes

For this variant of parataxis > hypotaxis to be correct, there has to be an asymmetry between early and more recently spoken languages in having/lacking (finite) subordination.

− King & Cookson (1890: 204): “We cannot … suppose that hypotaxis 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.”
− Gildersleeve (1893: xxv): “we have to be on our guard. Hypotaxis is older than our record, and we cannot argue safely as to prehistoric processes”
− Even Karlsson (2009): “Evidence from many language families indicates that non-finite clausal subordination and initial stages of finite clausal subordination existed already in preliterate languages.”

As a categorical claim, it is certainly false that early (attested) human languages lacked (finite) subordinate clauses.

Does it hold statistically? I’m not aware of any studies addressing the question.
Different claim: Diachronically, embedded structures replace adjoined structures


- 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 Portuguese. (Cf. also O’Neil 1977, Suárez-Gómez 2006)
- Whether Kiparsky’s version of parataxis > hypotaxis extends beyond Indo-European hasn’t ever really been addressed, to my knowledge.

Today’s results won’t bear directly on this claim.
Claim: Diachronically, hypotactic structures become more common

This is possibly the most frequently-encountered version of parataxis > hypotaxis in the literature. Usually interpreted with respect to (finite) clausal subordination in particular.

Dąbrowska (2015: 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 gradual”

Karlsson (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”

This is a quantitative claim. It can only be assessed using quantitative data from historical corpora.
The causal argumentation for this claim 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?
- **Orality > literacy**
  - Chafe (1982) and Biber (1995) show that finite subordinate clauses are more common in written than in spoken texts
  - But very difficult to disentangle “real change” from genre effects; cf. present day
- **Cultural complexity & communicative needs** (e.g. Givón 1979; Deutscher 2001)
  - Difficult to find a robust proxy; different notions of “complexity”/“needs”
  - Important to avoid discredited notions of “primitiveness” and lower intelligence
  - Causal chain rarely made explicit (though see Deutscher 2001: 166–186)
Does (finite) clausal subordination become more common over time?
Methods: investigating parataxis > hypotaxis

Crucially relies on availability of parsed diachronic corpora.

**Hypotaxis level: proportion of all clauses that are subordinate/embedded, including all non-finite clauses.**

- Finite **unembedded** clauses: IP-MAT* in Penn-style parsed corpora (includes e.g. imperatives, exclamatives, coordinated clauses)
- Finite **subordinate**/embedded clauses: basically IP-SUB* (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)
- **Non-finite** clauses: IP-INF*

Languages investigated: English, Icelandic, French, Portuguese, Irish, Chinese
YCOE (Taylor et al. 2003)
PPCME2 (Kroch & Taylor 2000)
PPCEME (Kroch et al. 2005)
PPCMBE2 (Kroch et al. 2016)
Trend line is locally estimated scatterplot smoothing (LOESS) with confidence intervals
On eyeballing, no sign of an overall upward trend.

Pre-1500 not easy to interpret.

High point in 16th century (not 17th, *pace* Karlsson 2009).

Apparent decline after 16th century.

BUT: genre seems to explain much more of the variation than time does.
Non-finite clauses are the dark dots in the centre.
English: distribution

- Gentle increase in non-finite clauses between OE and Modern English.
- Window: 50 years
The diachrony of parataxis and hypotaxis

November 2021
Clause types: generalizations

- Window: 50 years
- Gentle increase in non-finite clauses between ca. 1150 and 1500
- Change is less striking than continuity here
- Some developments – such as the apparent general decrease in that-clauses, and in adverbial clauses – don’t have an obvious explanation
- (The final time point should probably be ignored)
Acte of Apparell.

Forasmuche as the great and costly arraye and apparell used within this Realme cont=arie to good Statut~ thereof made hath ben the occasion of great impov~ysshyng of div~se of the King~ Subgiett~ and p~voked of theym to robbe and to do extorcion and other unlaufall ded~ to mayntayn therby their costely arraye; In eschewyng wherof be it ordeyned by the auctoritie of this p~sent p~liament that …

And that …

And that …

And that …

And that …
- IcePaHC (Wallenberg et al. 2011)
- Sagas typically have less than average hypotaxis
Icelandic: distribution

- Gentle increase in non-finite clauses between 1500 and 1900.
- Window: 100 years
- MCVF (Martineau et al. 2010)
- Apparent early rise is exclusively due to dominance of verse texts in this period
- Again, gentle rise of non-finite clauses
- Window: 100 years
- Tycho Brahe Corpus (Galves, Andrade & Faria 2017)
- News texts & dramas typically low
Portuguese: distribution

- Only clear trend is reduction in finite subordinate clauses
- Window: 50 years
Old and Middle Irish

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

The diachrony of parataxis and hypotaxis

November 2021
Old and Middle Irish: distribution

- No clear trends
- Window: 100 years
- ChiParHC (Li 2017)
- Again, hard to generalize about genre
No clear trends
Window: 500 years
Evaluation

Mixed-effects linear regression using R and lme4 package
- Dependent variable: proportion of unembedded vs. (finite or non-finite) subordinate clauses in each text
- Fixed effect: date
- Random intercept: genre

Positive 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.
- Marginal $R^2$: only fixed effects (date)
- Conditional $R^2$: fixed and random effects (date and genre)

<table>
<thead>
<tr>
<th></th>
<th>English</th>
<th>Icelandic</th>
<th>French</th>
<th>Portuguese</th>
<th>Irish</th>
<th>Chinese</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effect of date</td>
<td>-0.00006</td>
<td>0.00001</td>
<td>0.00008</td>
<td>-0.00064</td>
<td>-0.00030</td>
<td>0.00002</td>
</tr>
<tr>
<td>Marginal $R^2$</td>
<td>1.2%</td>
<td>0.1%</td>
<td>1.3%</td>
<td>44.4%</td>
<td>11.3%</td>
<td>12.5%</td>
</tr>
<tr>
<td>Conditional $R^2$</td>
<td>43.2%</td>
<td>39.9%</td>
<td>54.6%</td>
<td>49.0%</td>
<td>44.6%</td>
<td>12.5%</td>
</tr>
</tbody>
</table>
Constellations

I don't trust linear regressions when it's harder to guess the direction of the correlation from the scatter plot than to find new constellations on it.

xkcd #1725, “Linear Regression” (Randall Munroe, CC-BY-NC 2.5)
Evaluation (with genre)

- Marginal $R^2$: only fixed effects (date)
- Conditional $R^2$: fixed and random effects (date and genre)
- $p$-value of date effect calculated using package lmerTest

Effect of date explains little of the data, with the exception of Portuguese.
- Portuguese, Irish: effect is in the wrong direction.
- English, Icelandic, French: effect explains almost nothing.
- Irish, Chinese: probably not enough data to be hugely confident.

Genre explains much, much more of the data, except in Chinese.

Potential problem: each text treated equally as single data point. Logistic regression more appropriate?
Overview

No robust support for parataxis > hypotaxis.

- English, Icelandic, Irish, Chinese: no consistent direction of change.
- French: apparent increase in hypotaxis 1100–1200, but early texts are in verse.
- Portuguese: gentle but steady decrease in hypotaxis over the timespan of the corpus.
- Gentle upward trend for non-finite clauses in English, Icelandic and French.

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

- The most hypotactic texts in English are legal texts.
- A consistent role for genre is exactly what we’d predict given Chafe’s (1982) and Biber’s (1995) results, if performance effects are constant.
- So unless the corpora are unbalanced and genre effects are counteracting a real diachronic trend, the result basically stands.
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. Correlation for Icelandic shown (incl. nonfinite).
- PROIEL (Haug & Jøhndal 2008)
- Again, hard to generalize about genre
Slavic/Russian

- PROIEL (Haug & Jøhndal 2008)
- Bible texts are Old Church Slavonic; narrative texts are Russian
- Too little here to say anything meaningful at all
Georgian National Corpus (Gippert & Tandashvili 2015)

- Philosophical and legal texts most hypotactic
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 diachronic corpora of English, Icelandic, French, Portuguese, Irish, or Chinese.
  − Maybe the corpus annotation is wrong.
  − Maybe the choice of languages is wrong.
  − But insofar 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.
  − Suggestions welcomed!

Thank you for your attention!
References (1)

References (2)

References (3)

References (4)

- Lash, Elliott. 2014. The Parsed Old and Middle Irish Corpus (POMIC). Version 0.1.
- Li, Man. 2017. Chinese Parsed Historical Corpus (ChiParHC).
References (5)

References (6)