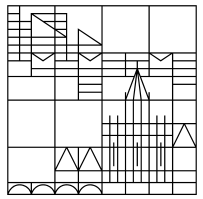
Universität Konstanz



Detecting syntactic change and stability

George Walkden

Statistics in Historical Corpus Linguistics

Maynooth University

October 2019

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")

Focus today:

- **How do we evaluate claims like this?**
- **What types of model/approach are appropriate?**
- **What do the models need to tell us in order for us to take them seriously?**

I'm a linguist, not a statistician, and this is work in progress – feedback welcomed!

parataxis > 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)

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)

Content warning

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.

Versions of the PPHH

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"

Version that's relevant today:

diachronically, hypotactic structures become more common.

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"

This is a quantitative claim.

It can only be assessed using quantitative data from historical corpora.

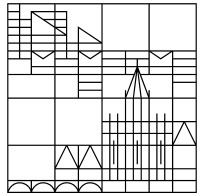
Implications

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

- If it 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's a "real change", and if the causal argument works, it indicates that sociocultural 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** (Kroch 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!

Universität
Konstanz



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.

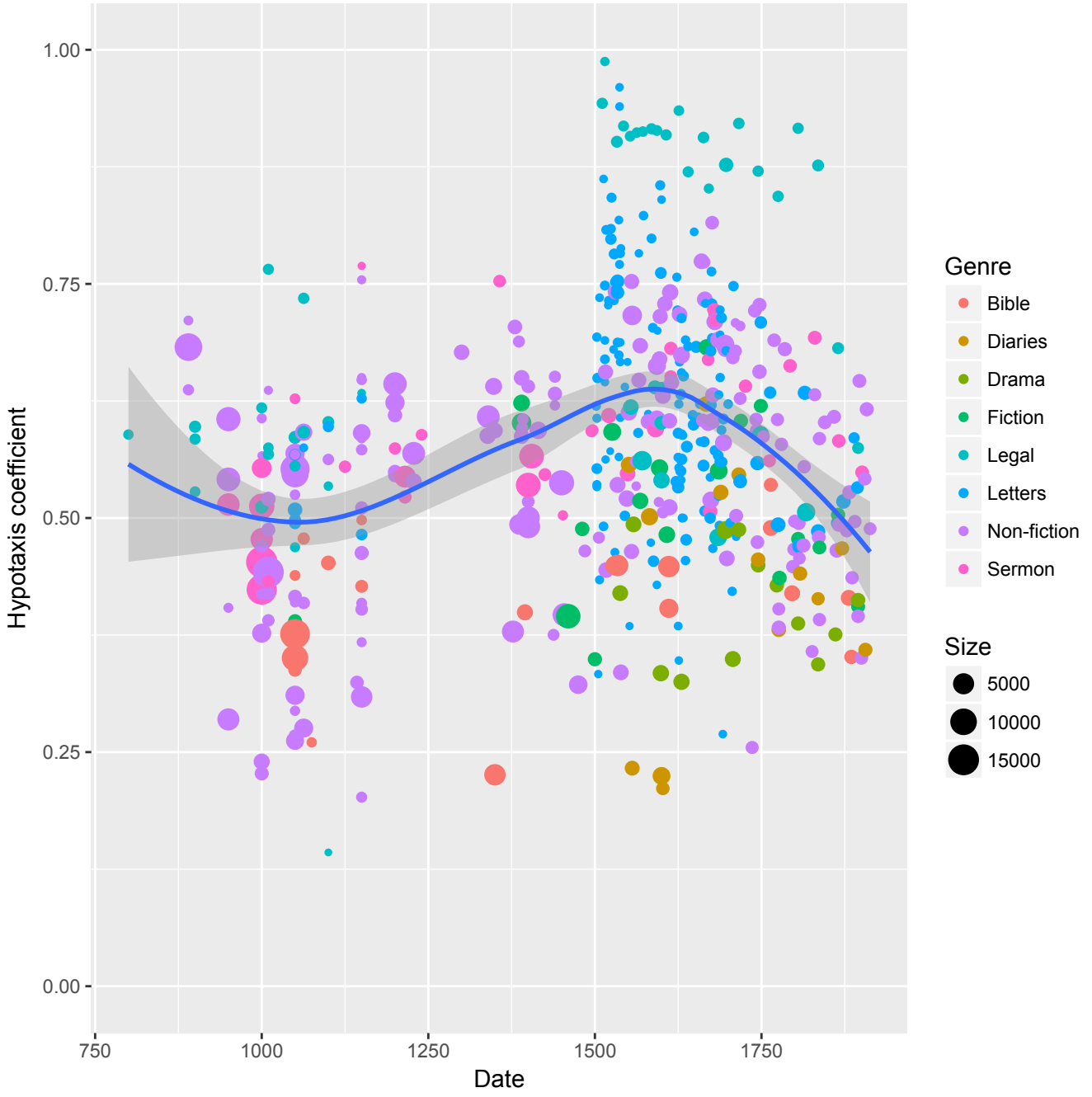
- Finite **unembedded** clauses: IP-MAT* in Penn-style parsed corpora
  (includes e.g. imperatives, exclamatives, coordinated clauses)
- Finite **subordinate**/embedded clauses: basically IP-SUB*
  (includes e.g. relatives, complement clauses, adverbial clauses)
  - Some variation in how interrogatives 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

English

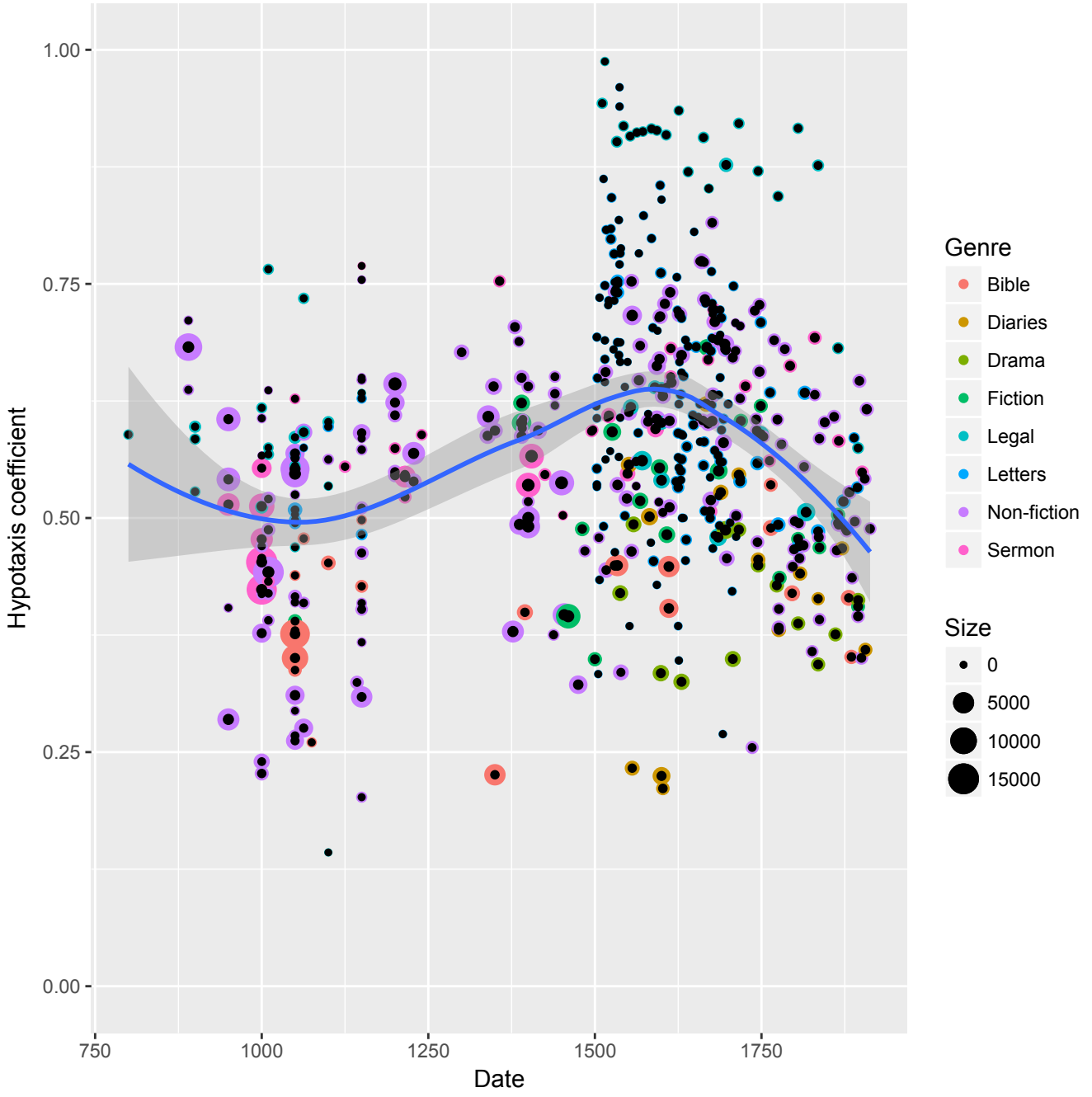
- YCOE
  (Taylor et al. 2003)
- PPCME2
  (Kroch & Taylor 2000)
- PPCEME
  (Kroch et al. 2005)
- PPCMBE
  (Kroch et al. 2010)

- "Non-fiction" (purple) is something of a dustbin category.
- Legal texts high; diaries and bibles low



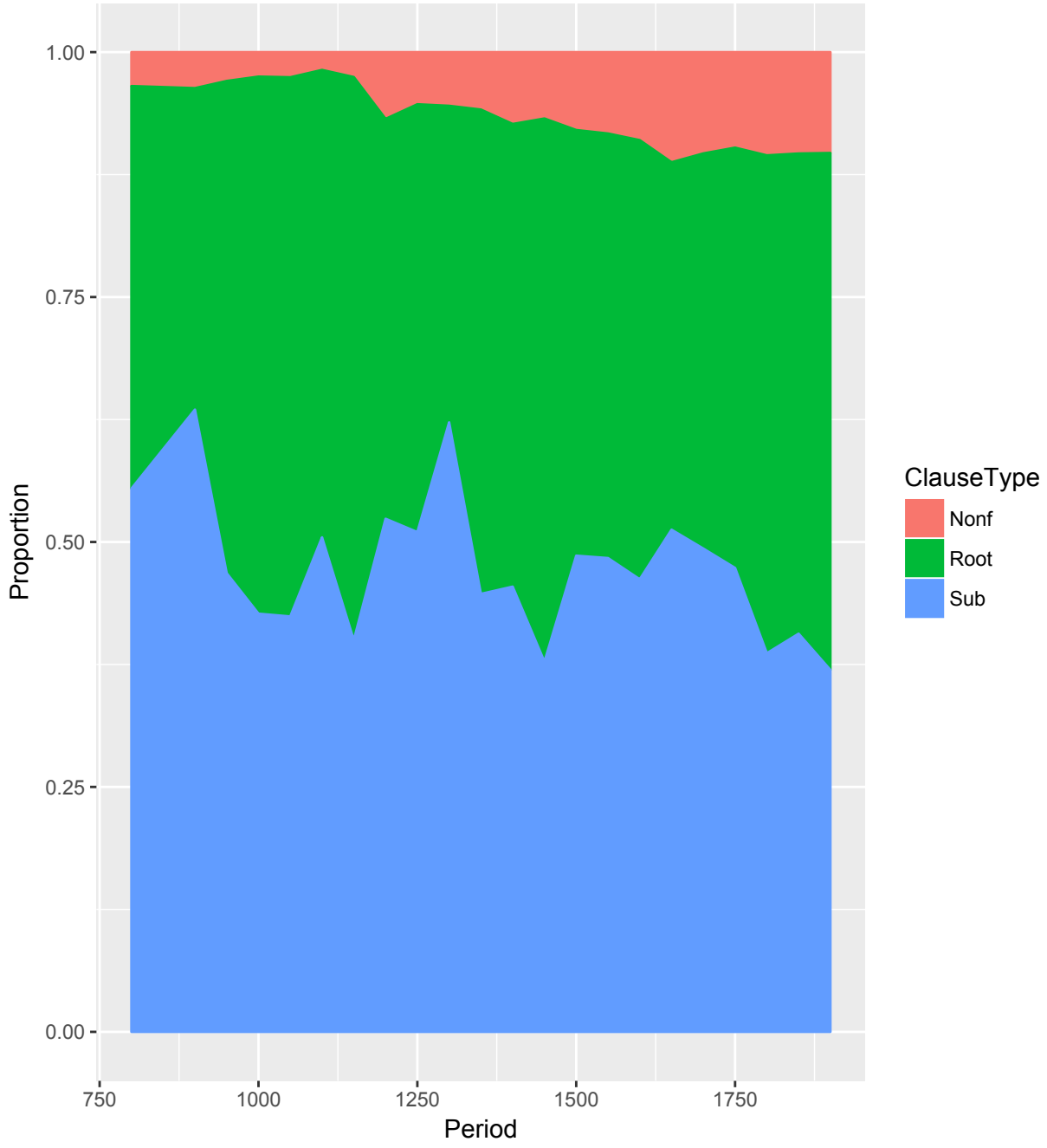
English with non-finite

- Non-finite clauses are the dark dots in the centre.



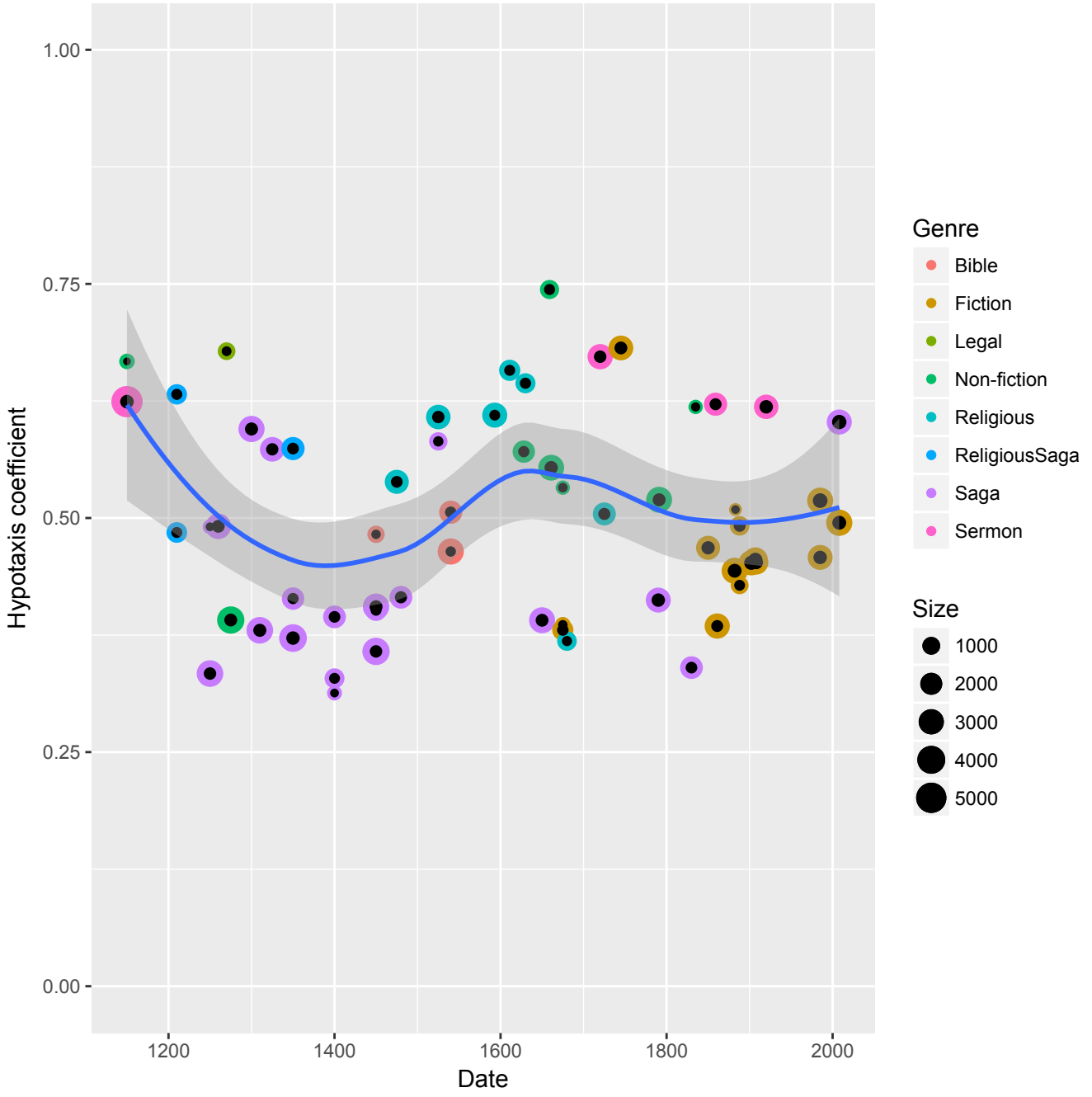
English: distribution

- Gentle increase in non-finite clauses between OE and Modern English.
- Window: 50 years



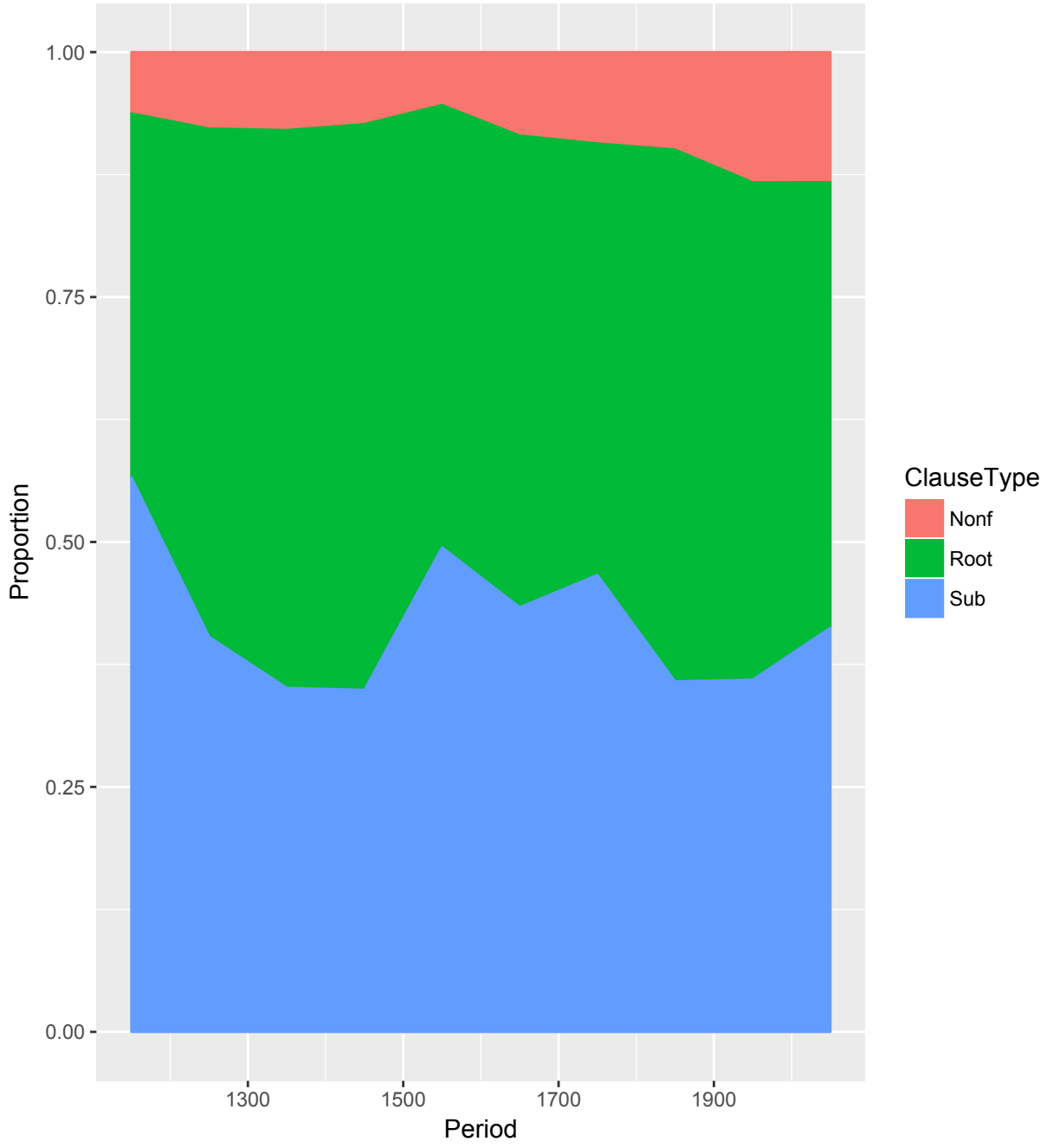
Icelandic

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



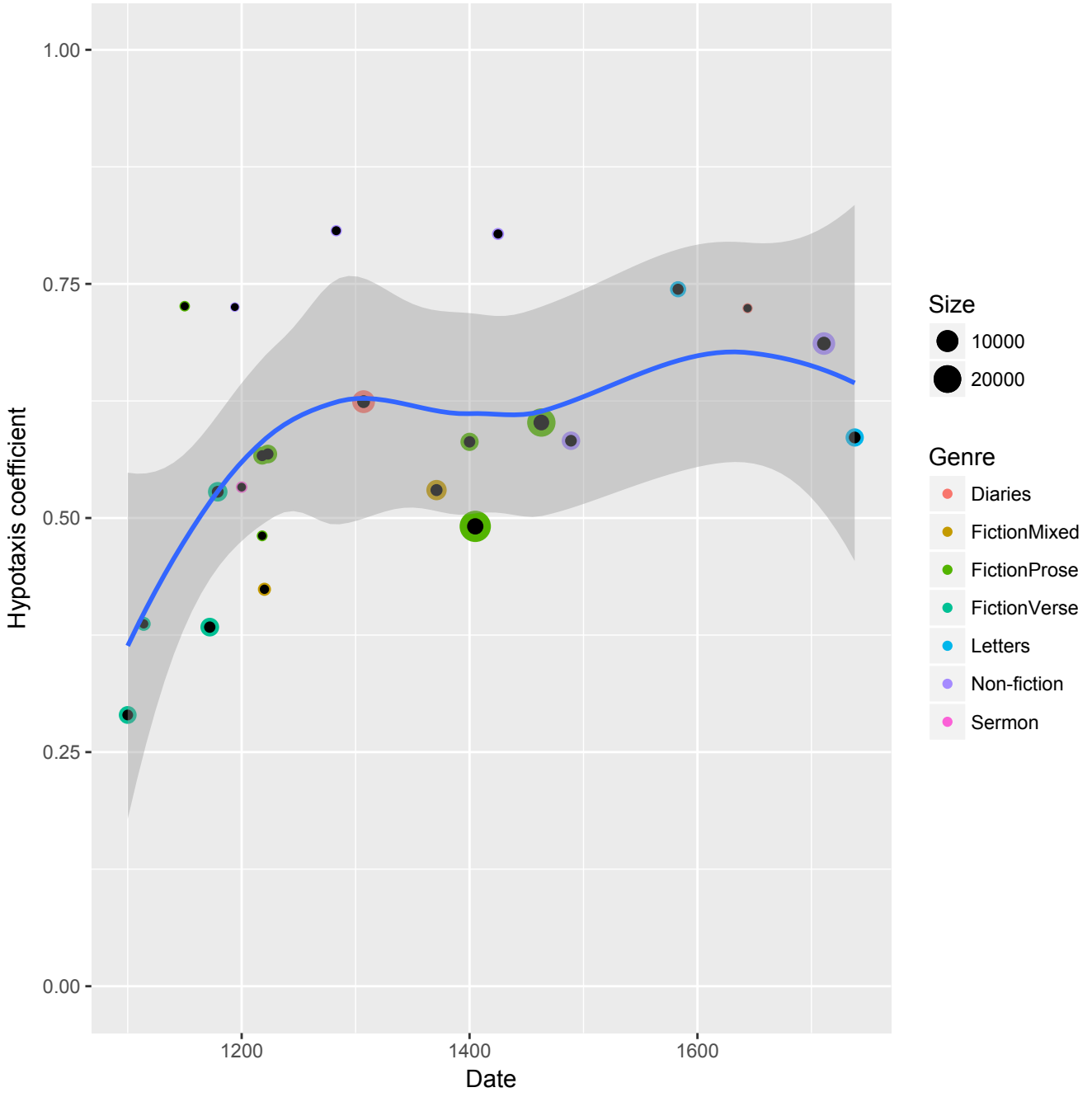
Icelandic: distribution

- Gentle increase in non-finite clauses between 1500 and 1900.
- Window: 100 years



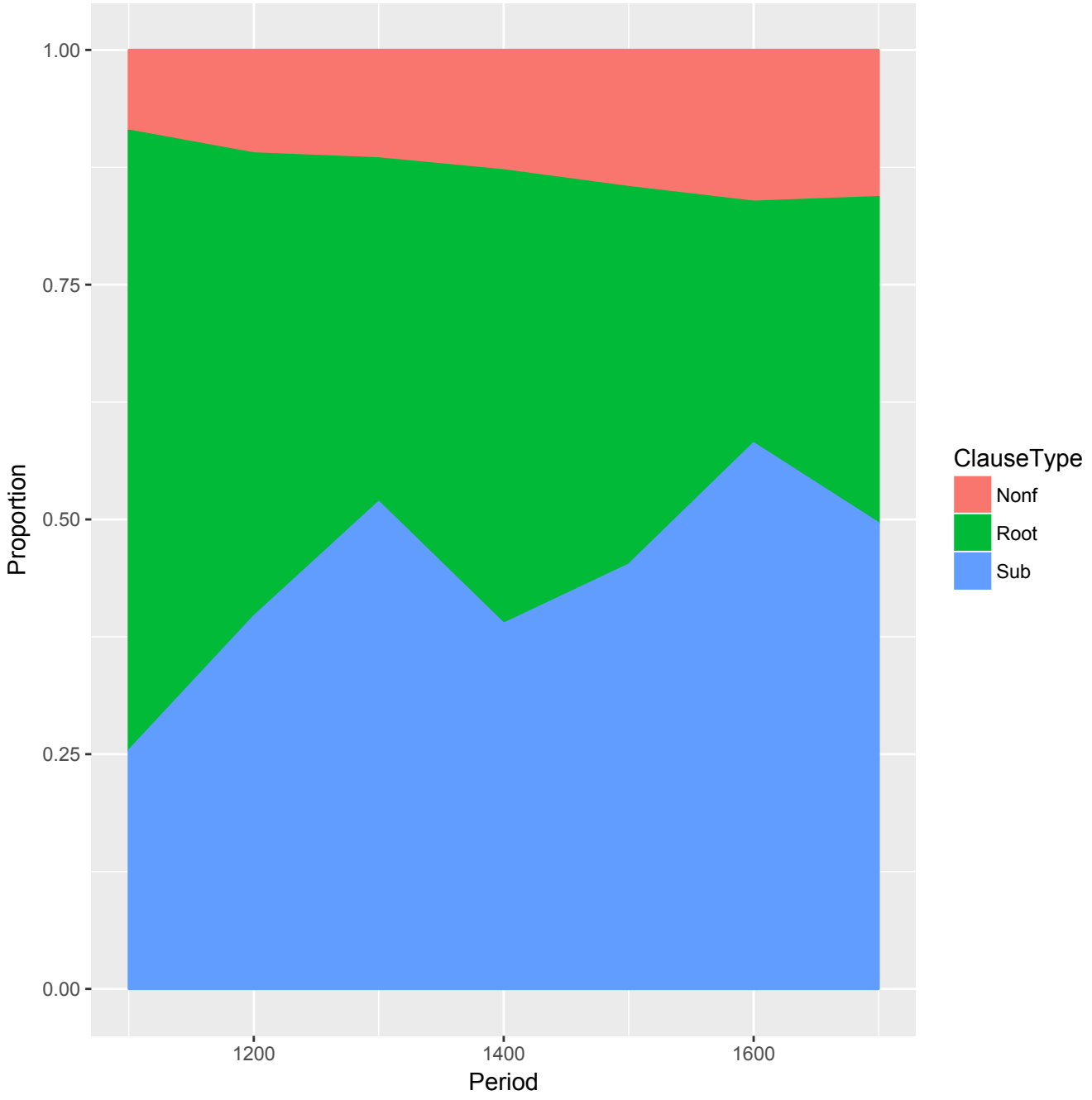
French

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



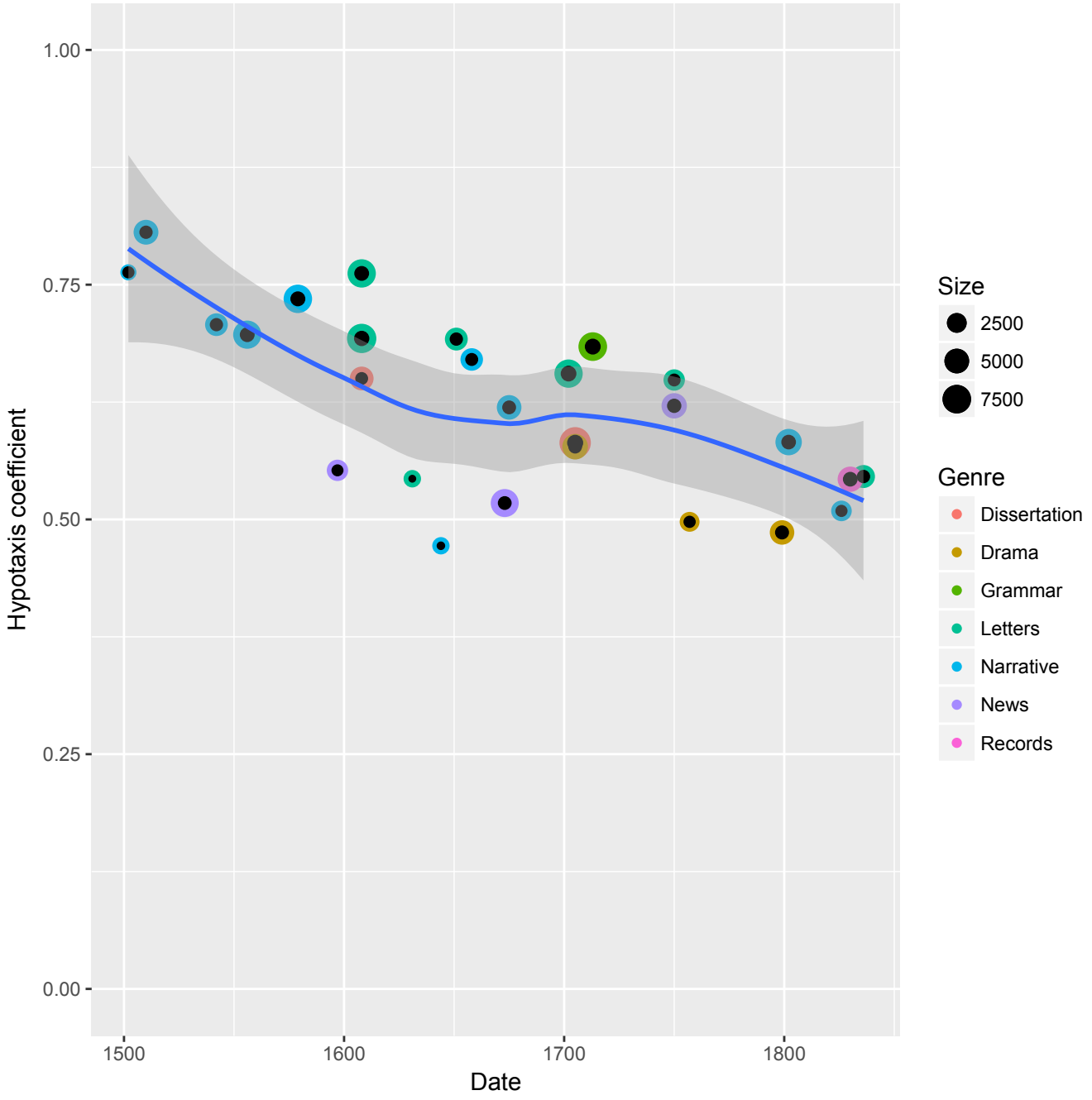
French: distribution

- Again, gentle rise of non-finite clauses
- Window: 100 years



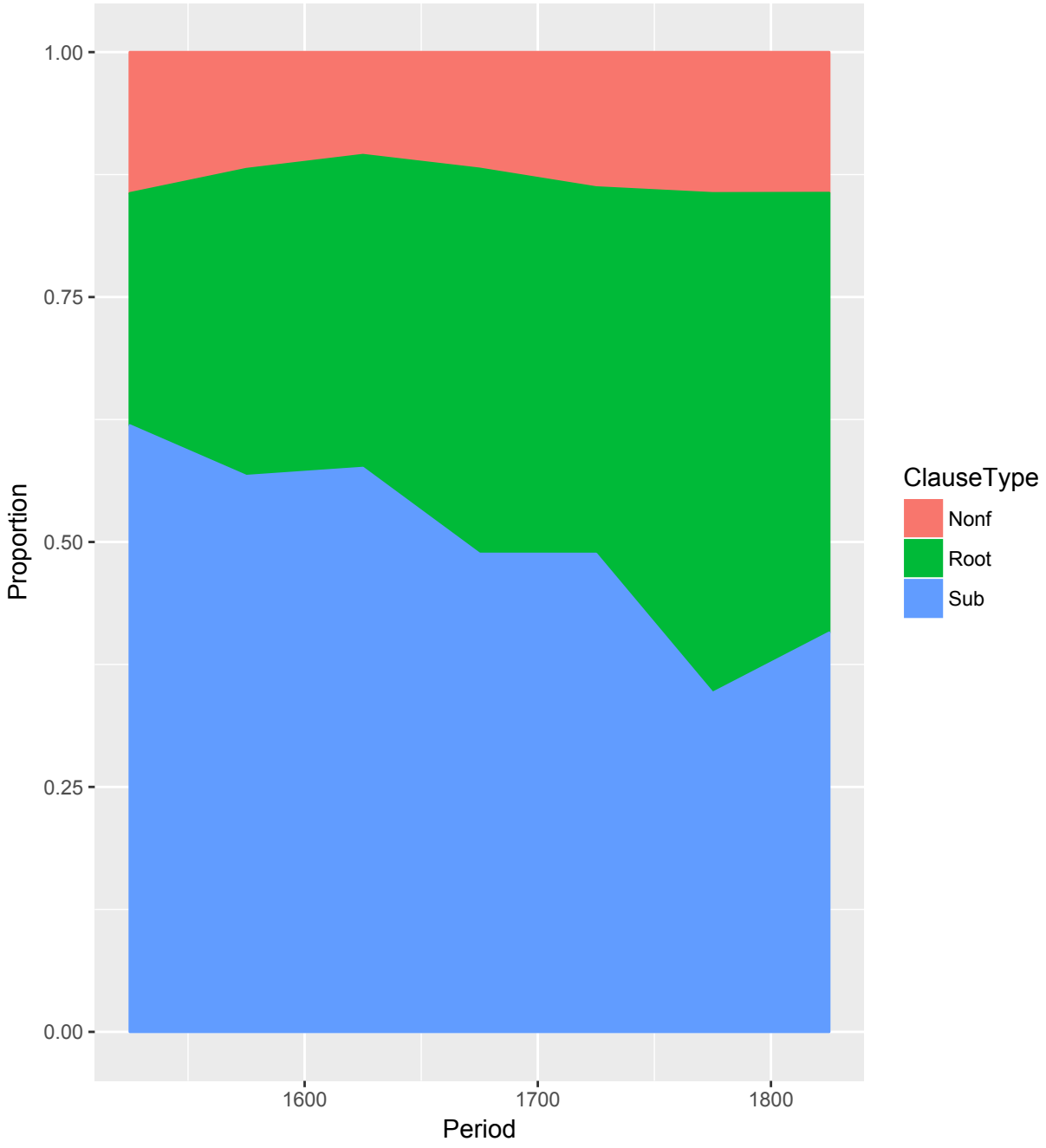
Portuguese

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



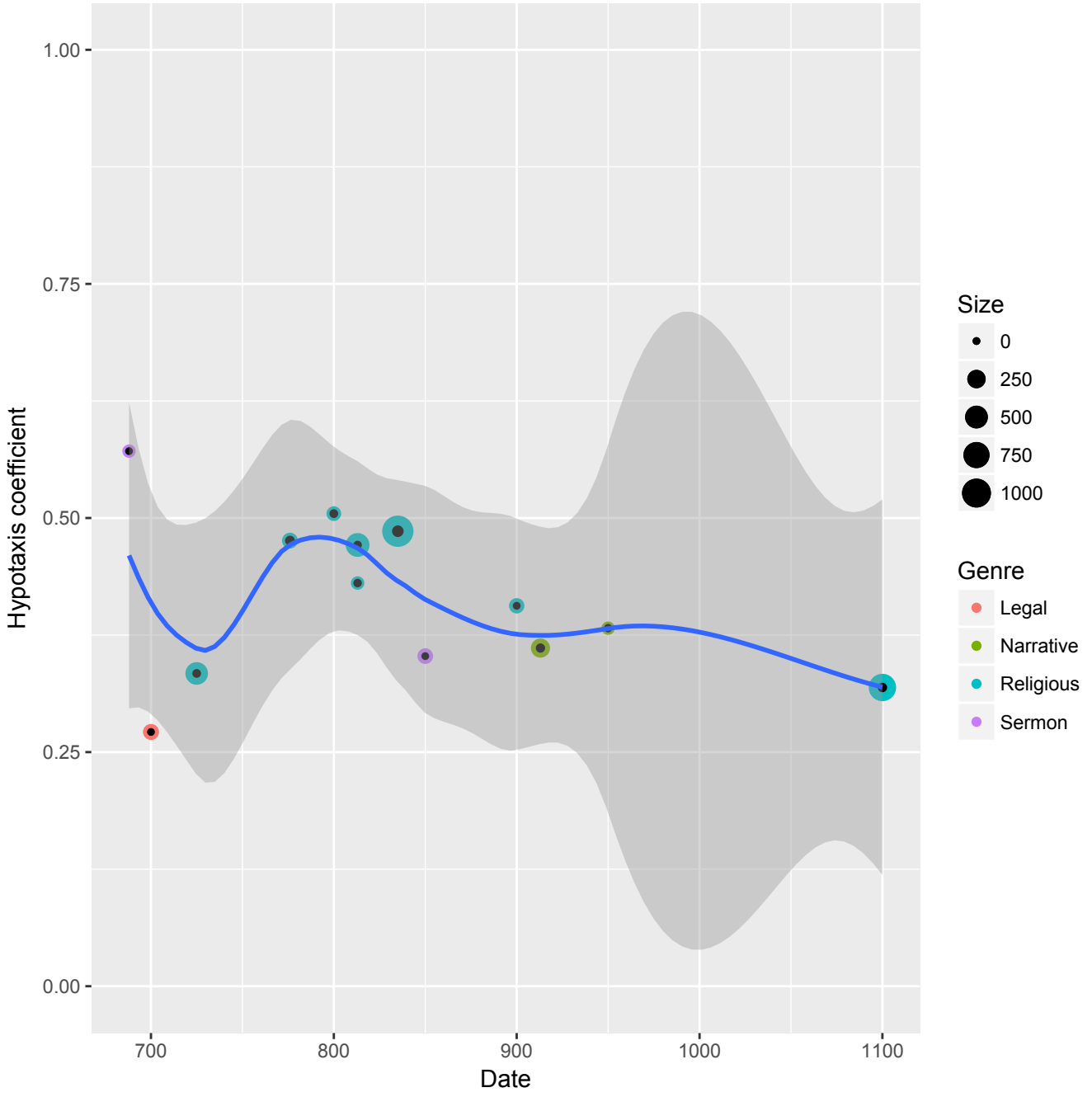
Portuguese: distribution

- Only clear trend is reduction in finite subordinate clauses
- Window: 50 years



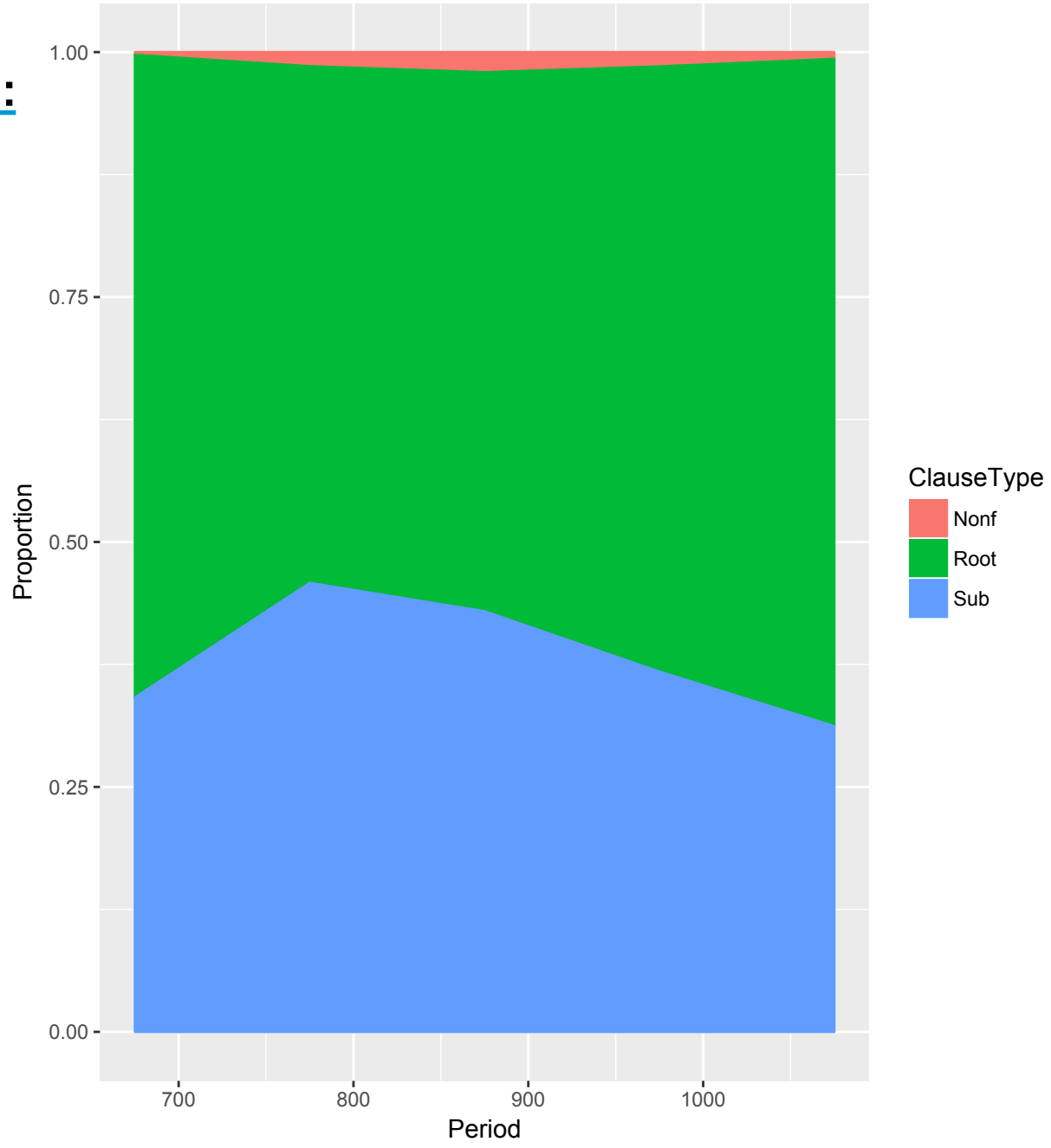
Old and Middle Irish

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



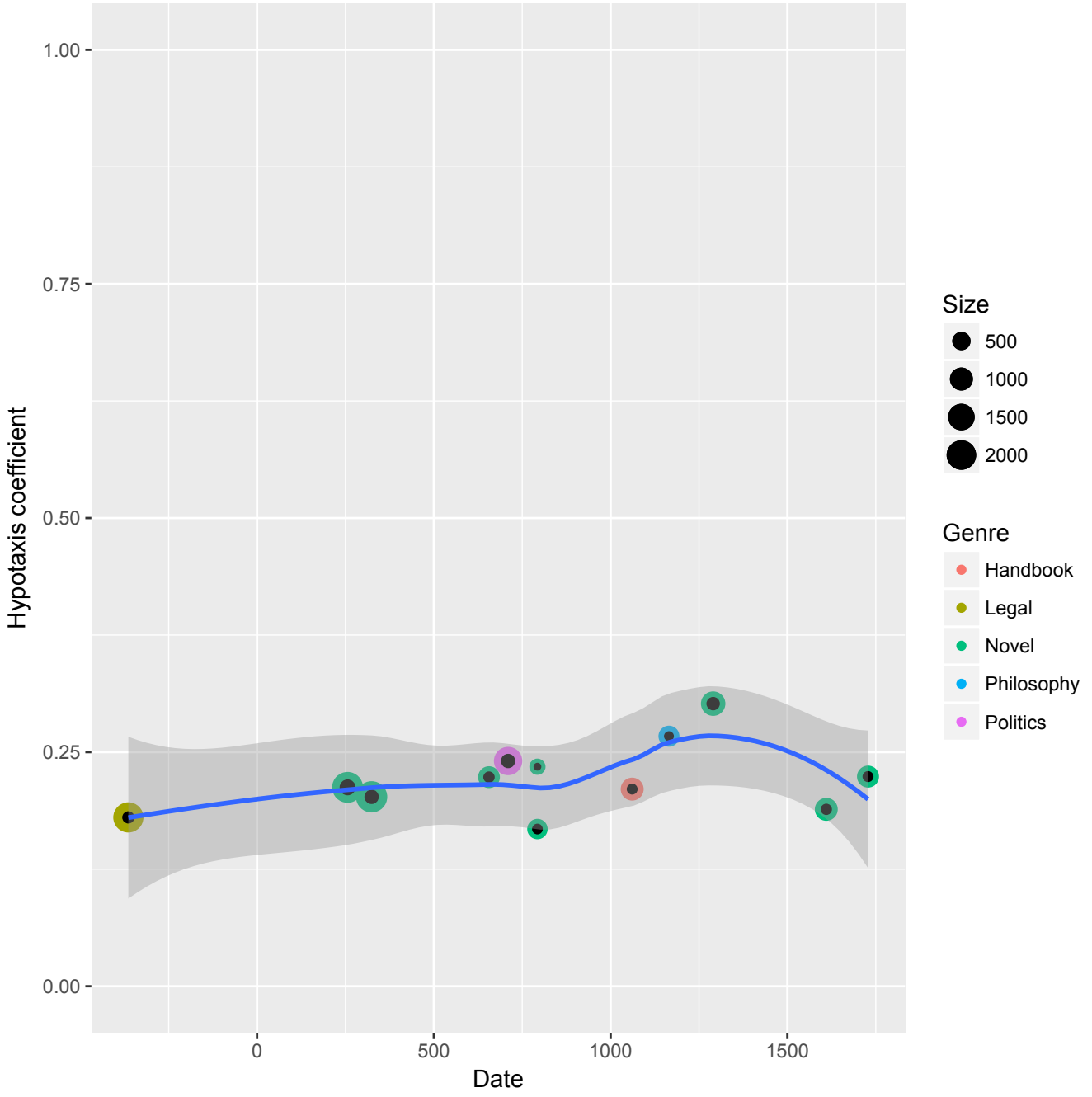
Old and Middle Irish: distribution

- No clear trends
- Window: 100 years



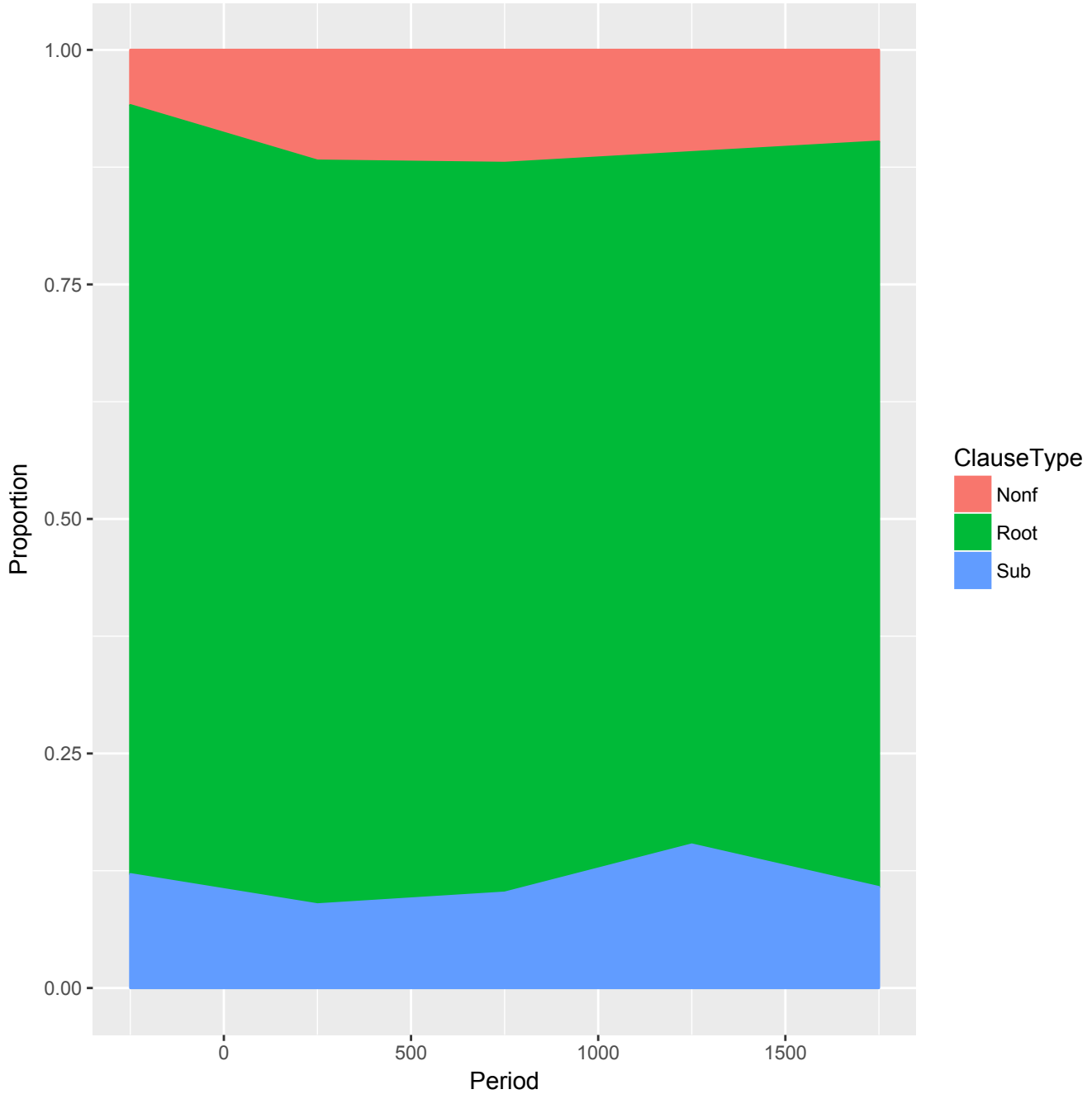
Chinese

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



Chinese: distribution

- 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

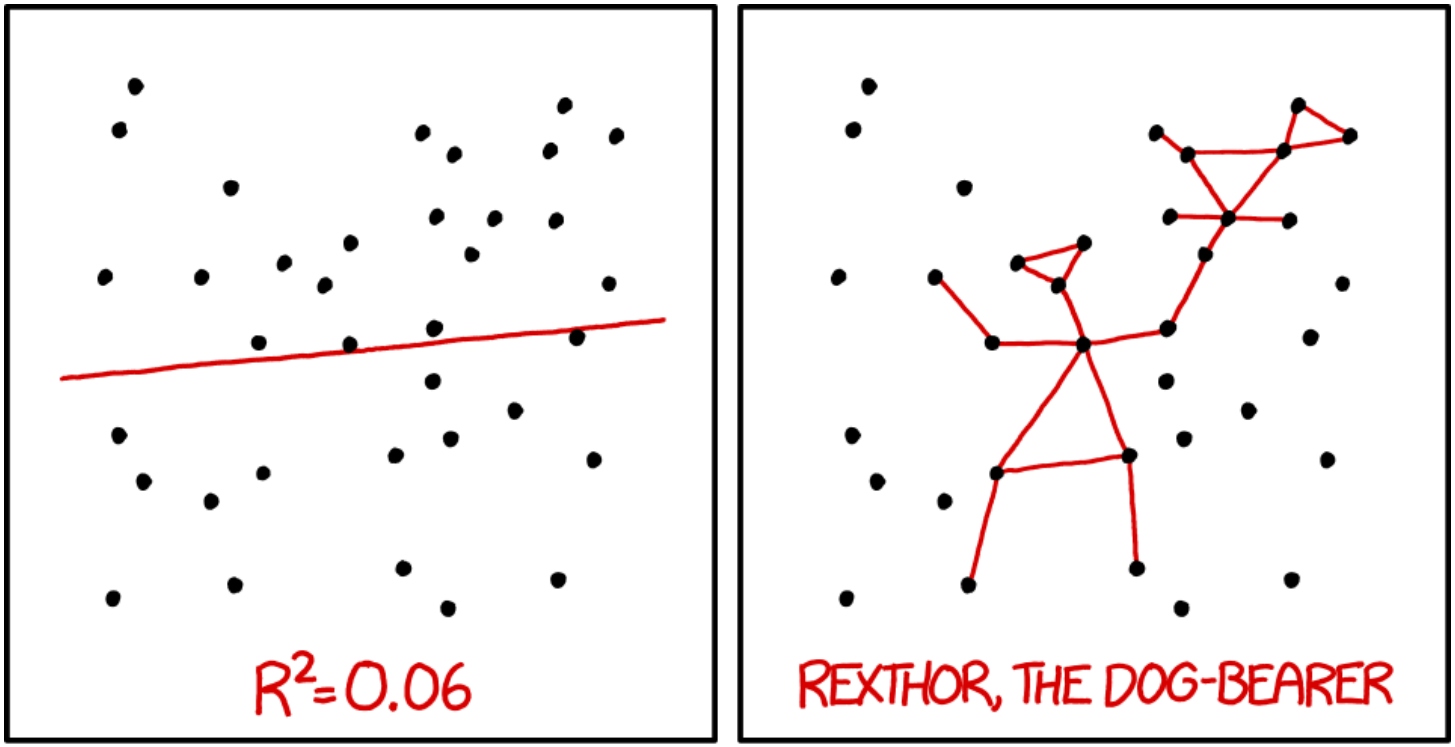
Positive linear effect of time should at least be detectable if the hypothesis is correct!

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)

	English	Icelandic	French	Portuguese	Irish	Chinese
Effect of date	0.00011	0.00001	0.00008	-0.00064	-0.00030	0.00002
Marginal R^2	3.2%	0.1%	1.3%	44.4%	11.3%	12.5%
Conditional R^2	47.8%	39.9%	54.6%	49.0%	44.6%	12.5%

Constellations



xkcd #1725, "Linear Regression" (Randall Munroe, CC-BY-NC 2.5)

Evaluation (with genre)

	English	Icelandic	French	Portuguese	Irish	Chinese
Effect of date	0.00011	0.00001	0.00008	-0.00064	-0.00030	0.00002
Marginal R^2	3.2%	0.1%	1.3%	44.4%	11.3%	12.5%
Conditional R^2	47.8%	39.9%	54.6%	49.0%	44.6%	12.5%
p	<0.001	0.829	0.567	<0.001	0.182	0.238

- 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.

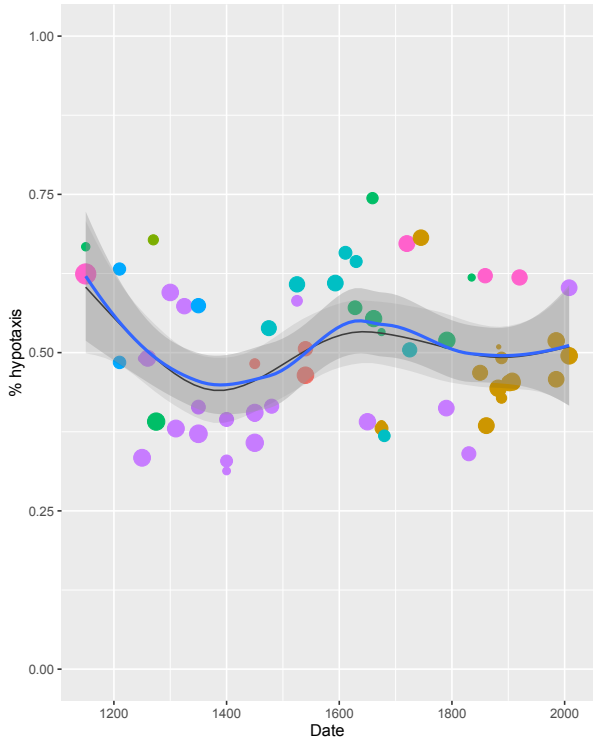
Genre explains much, much more of the data, except in Chinese.

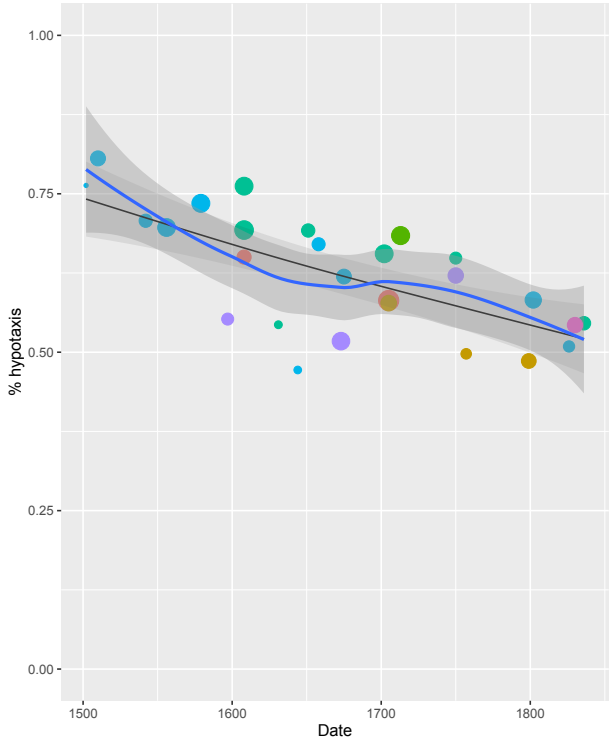
Potential problem: each text treated equally as single data point.
Logistic regression more appropriate?

GA(M)M vs LOESS

Since generalized additive (mixed) models (GA(M)Ms) are trendy, I decided to try them.

- Packages: mgcv (for GA(M)Ms), tidymv (for plotting). k=10





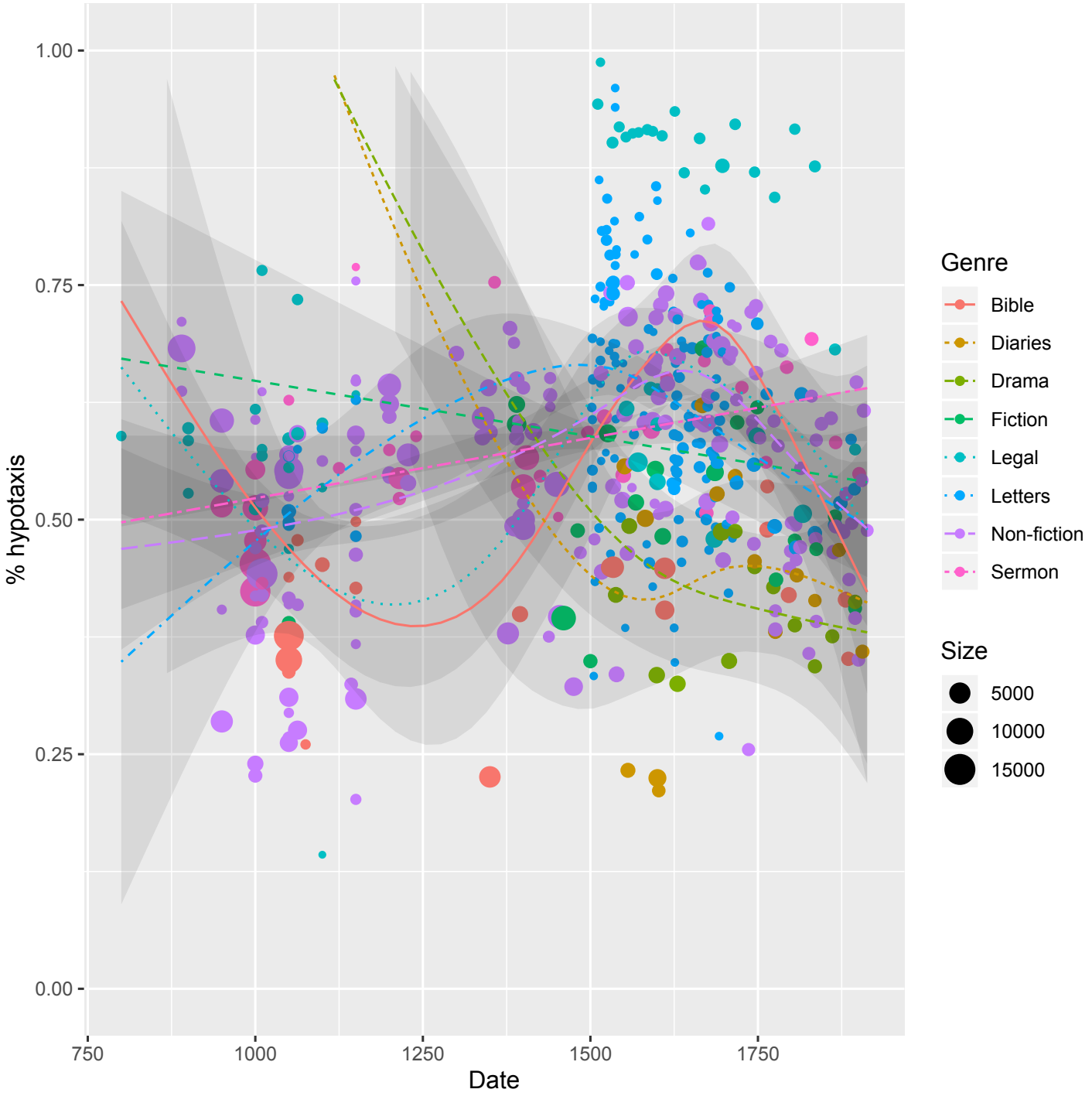
The results are very similar to the LOESS smooth provided out-of-the-box by ggplot2.

- Above: English (left), Icelandic (centre), Portuguese (right)
- LOESS (blue line) is jerkier

Genre effects

Obvious, and major, effects of genre (in English)

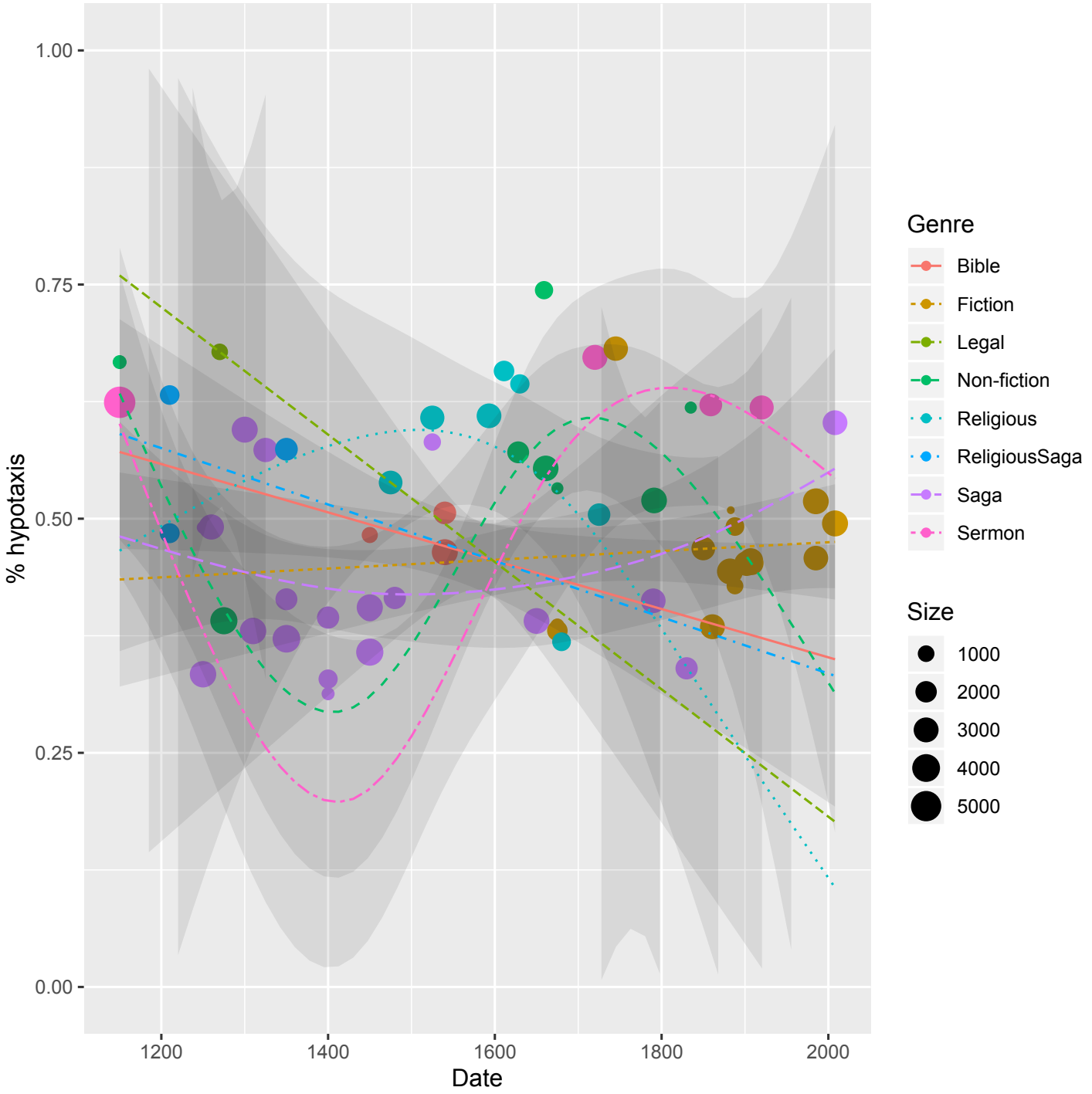
- k=10
- Genres not equally distributed over time.



Genre effects

Obvious, and major, effects of genre (in Icelandic)

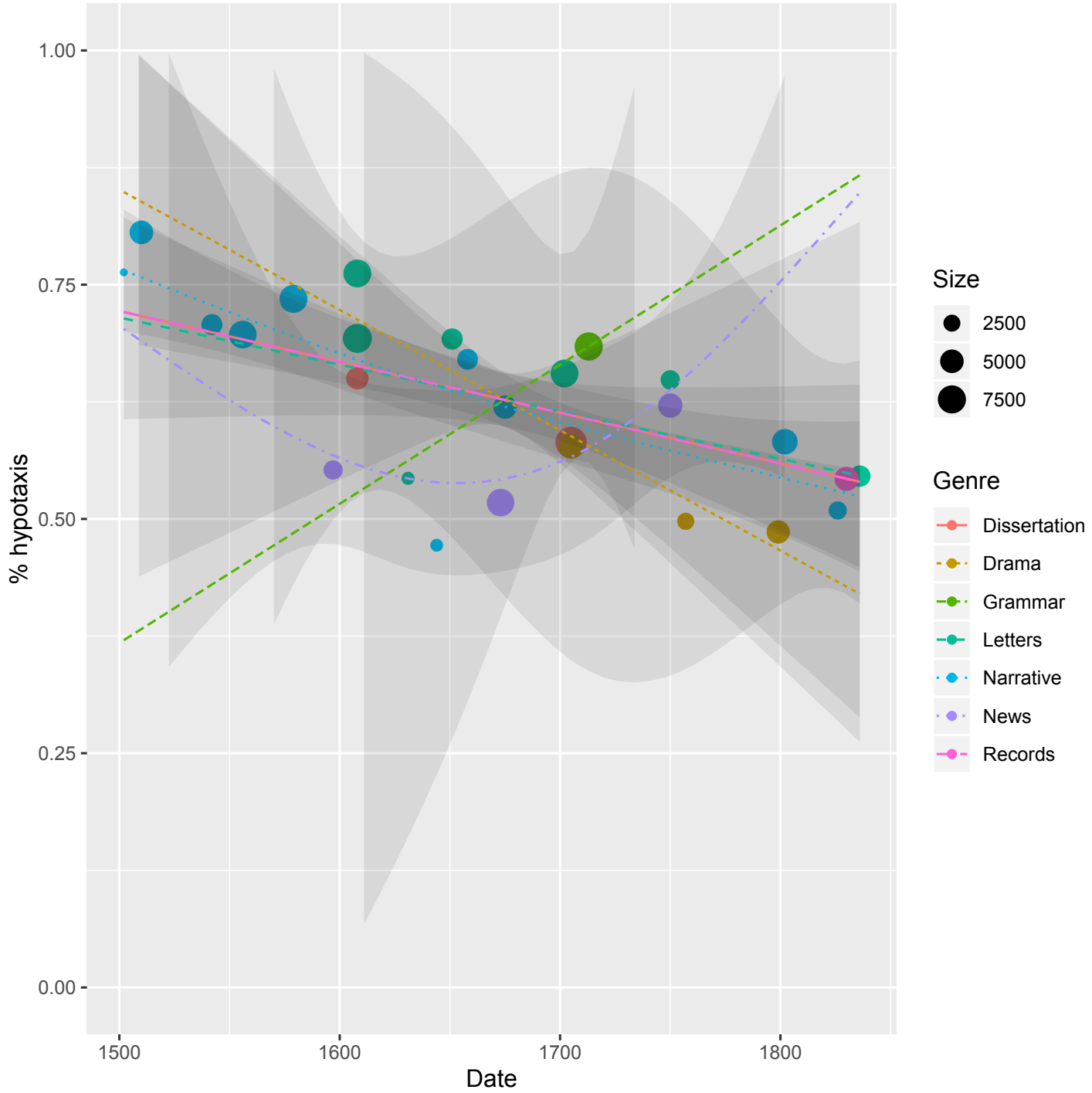
- k=5
- Genres not equally distributed over time.



Genre effects

Confusing (Portuguese)

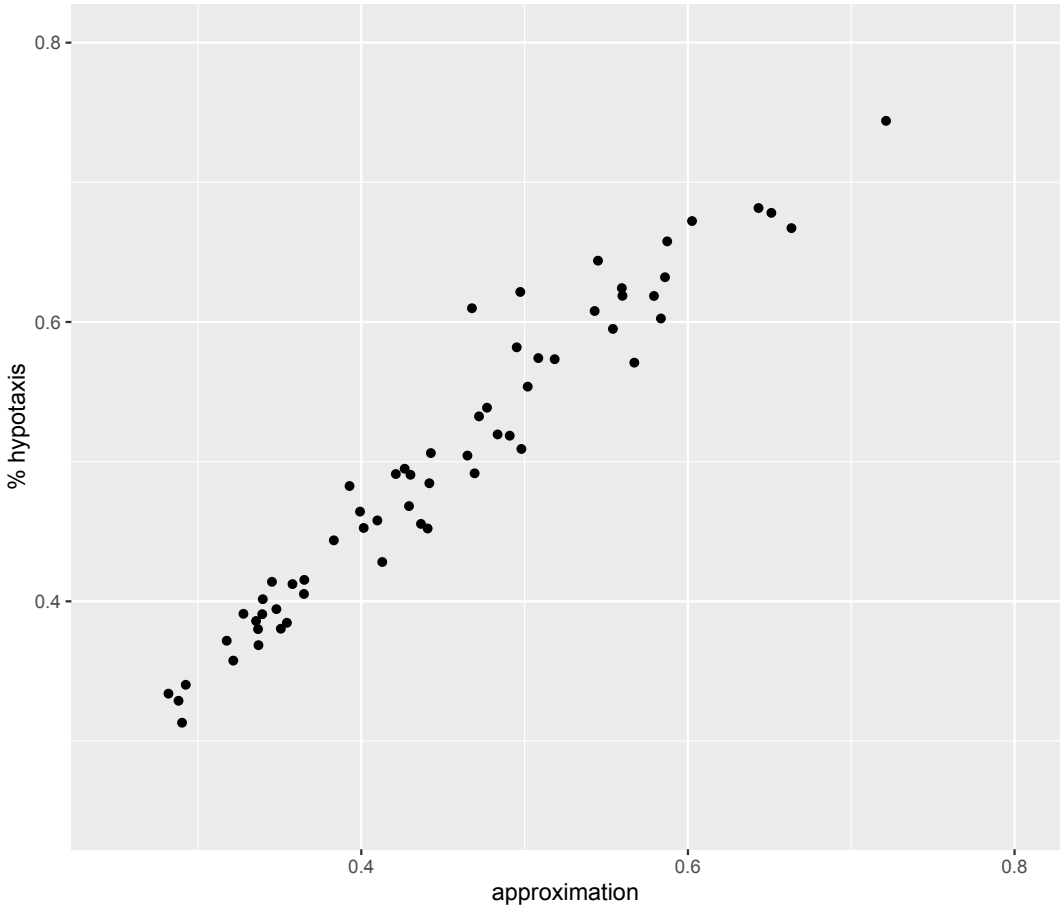
- k=4



Bonus languages: Latin, Slavic/Russian, Georgian

These corpora don't have constituency parsing.

- Latin: PROIEL
- Slavic/Russian: PROIEL
- Georgian: Georgian National Corpus

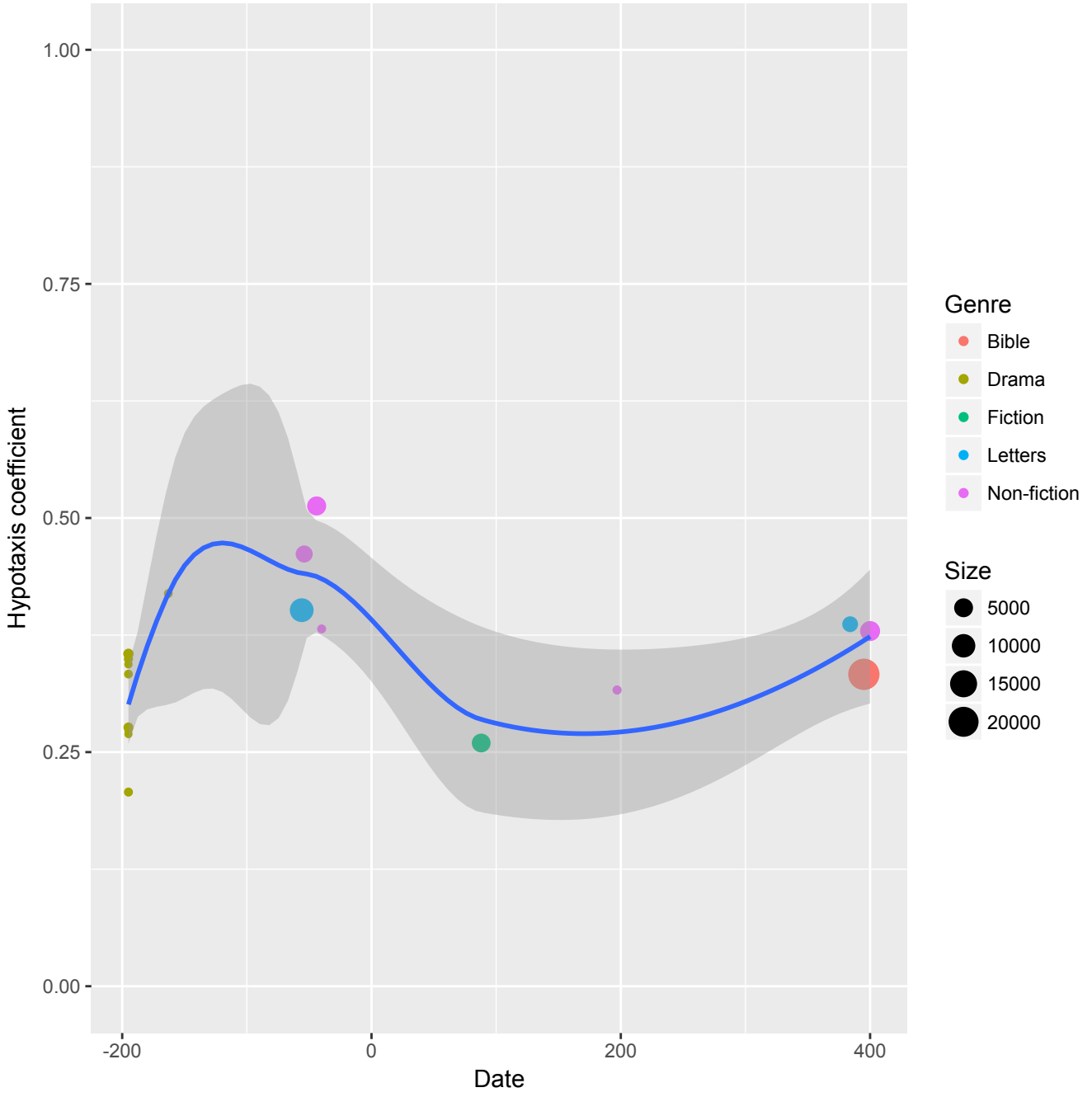


Approximation to the hypotaxis coefficient: number of overt subordinators divided by the number of finite verbs.

This seems to work reasonably well. Correlation for Icelandic shown (incl. nonfinite).

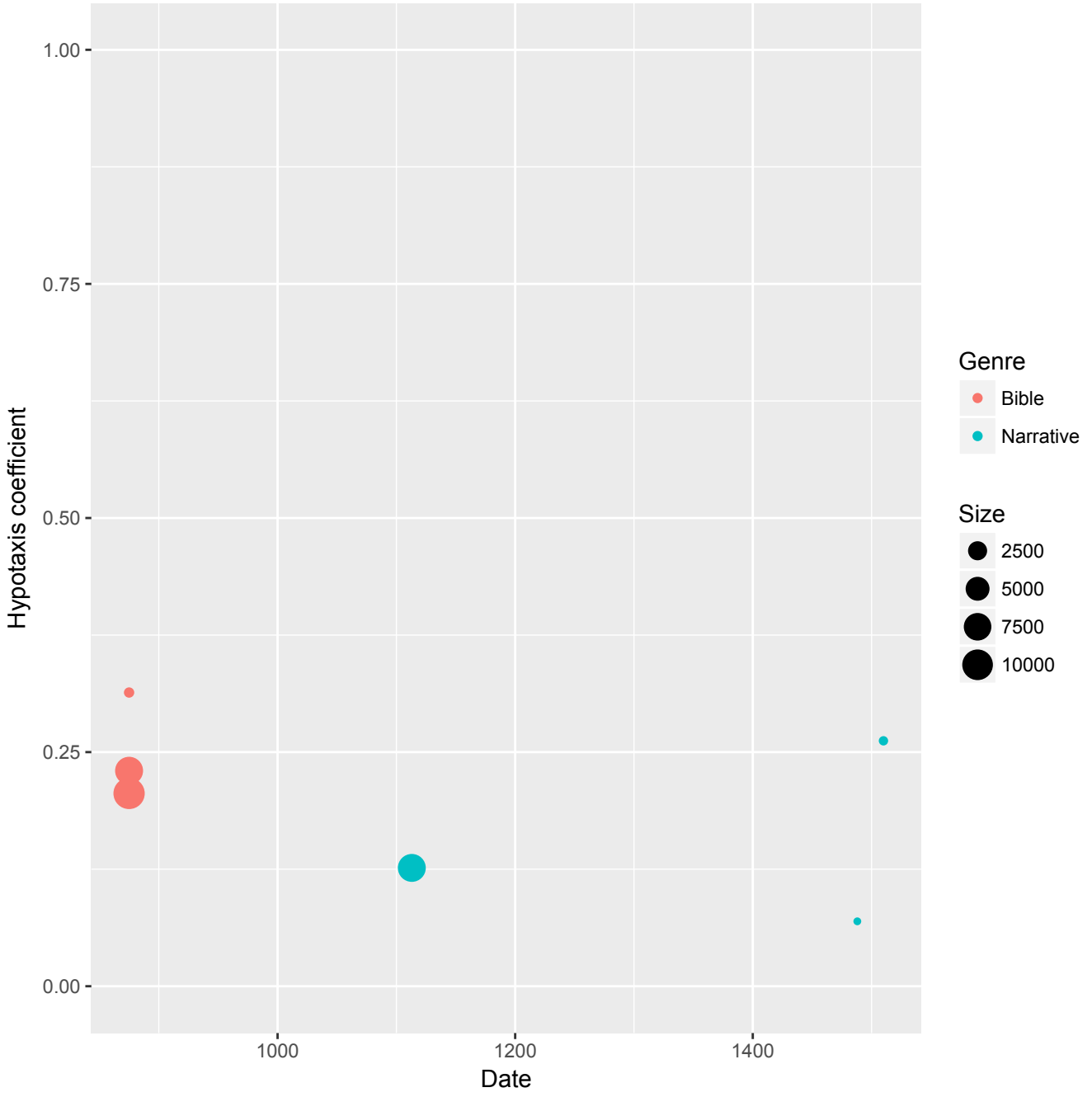
Latin

- PROIEL (Haug & Jøhndal 2008)
- Again, hard to generalize about genre



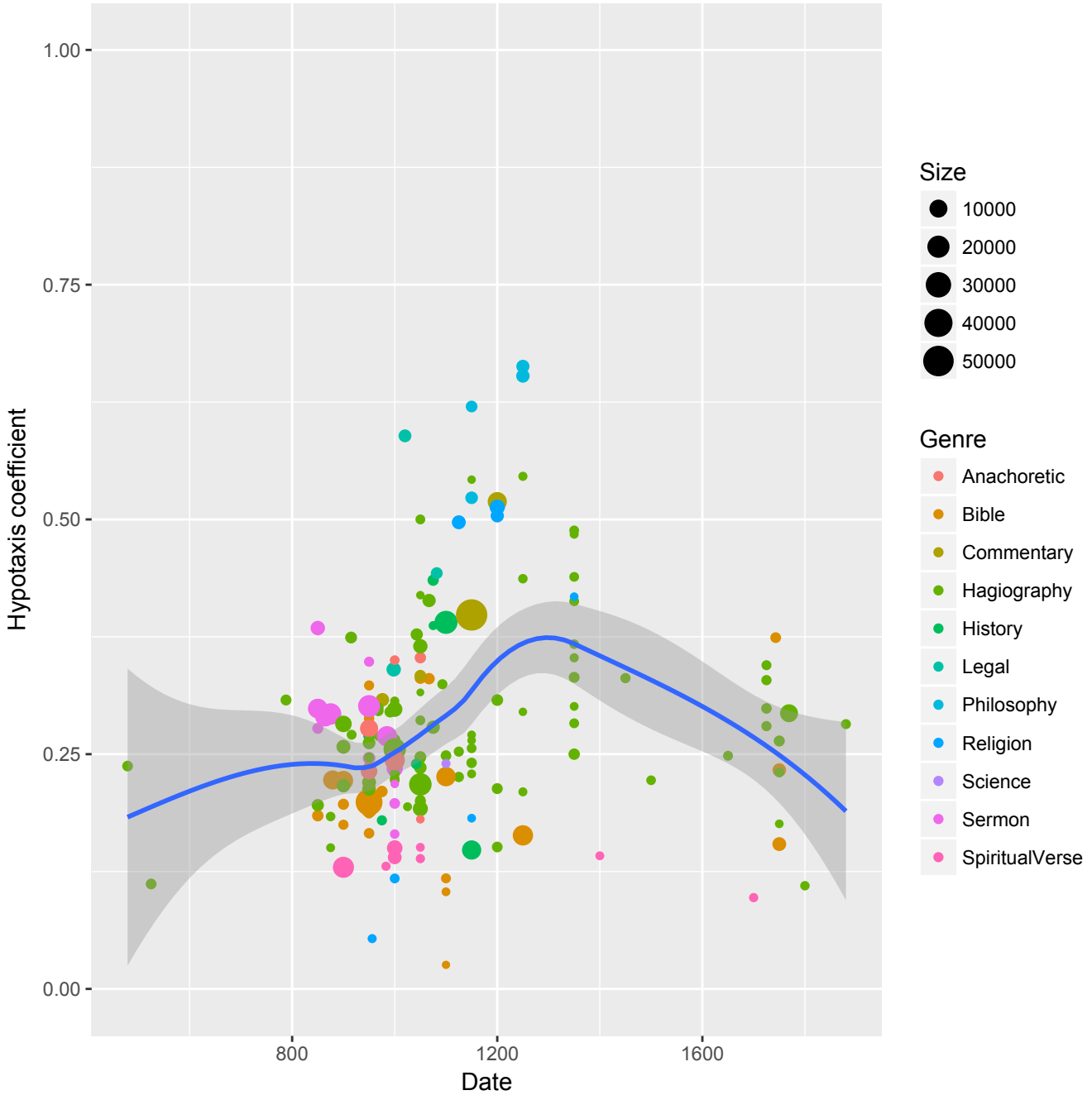
Slavic/Russian

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

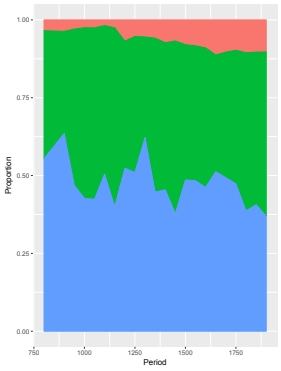


Georgian

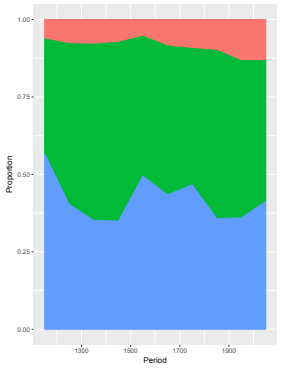
- Georgian National Corpus (Gippert & Tandashvili 2015)
- Philosophical and legal texts most hypotactic



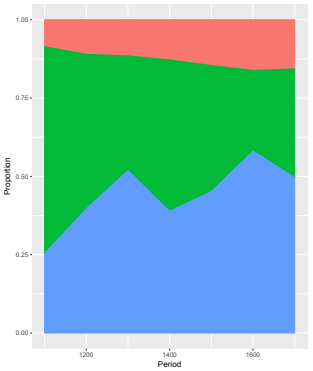
Overview



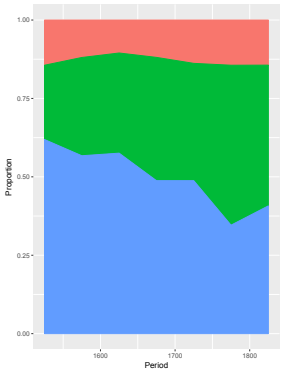
English



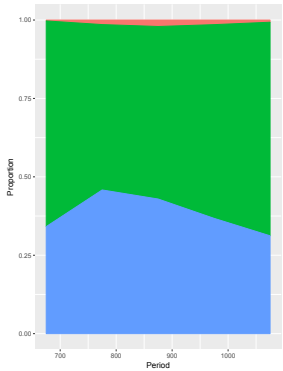
Icelandic



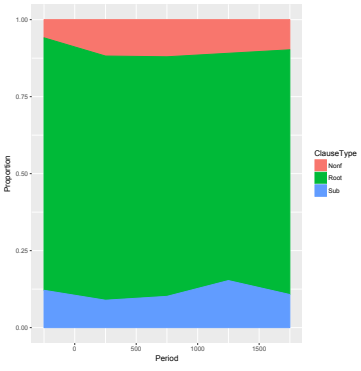
French



Portuguese



Irish



Chinese

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.

Conclusion

- **It's widely agreed that parataxis > hypotaxis.**
  Much less widely agreed what this actually means.

- **Focusing on the idea that 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.
  - Maybe my use of statistics 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)

- **Andrew**, S. O. 1940. *Syntax and style in Old English*. Cambridge: Cambridge University Press.
- **Axel-Tober**, Katrin. 2017. The development of the declarative complementizer in German. *Language: Historical Syntax* 93, e29–e65.
- **Biber**, Douglas. 1995. *Dimensions of register variation: a cross-linguistic comparison*. Cambridge: Cambridge University Press.
- **Chafe**, Wallace. 1982. Integration and involvement in speaking, writing, and oral literature. In Deborah Tannen (ed.), *Spoken and written language: Exploring orality and literacy*, 35–53. Norwood, NJ: Ablex.
- **Chomsky**, Noam. 1995. *The Minimalist Program*. Cambridge, MA: MIT Press.
- **Chomsky**, Noam. 2001. Beyond explanatory adequacy. *MIT Working Papers in Linguistics* 20, 1–28.
- **Chomsky**, Noam. 2013. Problems of projection. *Lingua* 130, 33–49.
- **Collins**, Chris, & Edward **Stabler**. 2016. A formalization of Minimalist syntax. *Syntax* 19, 43–78.
- **Dąbrowska**, Ewa. 2015. Language in the mind and in the community. In Jocelyne Daems, Eline Zenner, Kris Heylen, Dirk Speelman & Hubert Cuyckens (eds.), *Change of paradigms – new paradoxes: recontextualizing language and linguistics*, 221–236. Berlin: de Gruyter.
- **Delbrück**, Berthold. 1900. *Vergleichende Syntax der indogermanischen Sprachen*. Vol. 3. Strasbourg: Karl J. Trübner.

References (2)

- **Deutscher**, Guy. 2001. *Syntactic change in Akkadian: the evolution of sentential complementation*. Oxford: Oxford University Press.
- **Enfield**, Nicholas J. (ed.) 2002. *Ethnosyntax: explorations in grammar and culture*. Oxford: Oxford University Press.
- **Futrell**, Richard, Laura **Stearns**, Daniel L. **Everett**, Steven T. **Piantadosi** & Edward **Gibson**. 2016. A corpus investigation of syntactic embedding in Pirahã. *PLoS ONE* 11, 1–20.
- **Galves**, Charlotte, Aroldo L. de **Andrade**, & Pablo **Faria**. 2017. *Tycho Brahe Parsed Corpus of Historical Portuguese*.
- **Gildersleeve**, Basil L. 1883. On the final sentence in Greek. *The American Journal of Philology* 4, 416–444.
- **Gildersleeve**, Basil L. 1893. Some problems in Greek syntax. *Transactions of the American Philological Association* 24, xxiv–xxvii.
- **Gippert**, Jost, & Manana **Tandashvili**. 2015. Structuring a diachronic corpus: the Georgian National Corpus project. In Jost, Gippert & Ralf Gehrke (eds.), *Historical corpora: challenges and perspectives*, 305–322. Tübingen: Narr.
- **Givón**, Talmy. 1979. From discourse to syntax: Grammar as a processing strategy. In Talmy Givón (ed.), *Syntax and semantics*, vol. 12: *Discourse and syntax*, 81–112. New York, NY: Academic Press.
- **Harris**, Alice C., & Lyle **Campbell**. 1995. *Historical syntax in cross-linguistic perspective*. Cambridge: Cambridge University Press.

References (3)

- **Haug**, Dag T. T., & Marius L. **Jøhndal**. 2008. Creating a parallel treebank of the old Indo-European bible translations. In Caroline Sporleder & Kiril Ribarov (eds.). *Proceedings of LaTeCH 2008*, 27–34.
- **Hornstein**, Norbert, & Jairo **Nunes**. 2008. Adjunction, labeling, and Bare Phrase Structure. *Biolinguistics* 2, 57–86.
- **Jucker**, Andreas H. 1991. Between hypotaxis and parataxis: clauses of reason in *Ancrene Wisse*. In Dieter Kastovsky (ed.), *Historical English syntax*, 203–219. Berlin: de Gruyter.
- **Karlsson**, Fred. 2009. Origin and maintenance of clausal embedding complexity. In Geoffrey Sampson, David Gil & Peter Trudgill (eds.), *Language complexity as an evolving variable*, 192–202. Oxford: Oxford University Press.
- **Kayne**, Richard S. 1994. *The antisymmetry of syntax*. Cambridge, MA: MIT Press.
- **King**, John E., & Christopher **Cookson**. 1890. *An introduction to the comparative grammar of Greek and Latin*. Oxford: Clarendon.
- **Kiparsky**, Paul. 1995. Indo-European origins of Germanic syntax. In Adrian Battye & Ian Roberts (eds.), *Clause structure and language change*, 140–169. Oxford: Oxford University Press.
- **Kornai**, András. 2014. Resolving the infinitude controversy. *Journal of Logic, Language and Information* 23, 481–492.
- **Kroch**, Anthony. 1989. Reflexes of grammar in patterns of language change. *Language Variation & Change* 1, 199–244.

References (4)

- **Kroch**, Anthony, & Ann **Taylor**. 2000. *The Penn-Helsinki Parsed Corpus of Middle English (PPCME2)*.
- **Kroch**, Anthony, Beatrice **Santorini**, & Lauren **Delfs**. 2004. *The Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME)*.
- **Kroch**, Anthony, Beatrice **Santorini**, & Ariel **Diertani**. 2010. The Penn Parsed Corpus of Modern British English (PPCMBE).
- **Lash**, Elliott. 2014. *The Parsed Old and Middle Irish Corpus (POMIC)*. Version 0.1.
- **Li**, Man. 2017. *Chinese Parsed Historical Corpus (ChiParHC)*.
- **Martineau**, France, Paul **Hirschbühler**, Anthony **Kroch**, & Yves Charles **Morin**. 2010. *Modéliser le changement: Les voies du français*.
- **Mitchell**, Bruce. 1985. *Old English syntax*. 2 vols. Oxford: Clarendon.
- **Nakagawa**, Shinichi, & Holger **Schielzeth**. 2013. A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution* 4, 133–142.
- **Nevins**, Andrew, David **Pesetsky**, & Cilene **Rodrigues**. 2009. Pirahã exceptionality: a reassessment. *Language* 85, 355–404.
- **Nomura**, Masashi. 2017. Pair-merge and feature-valuation via minimal search: evidence from Icelandic. In Aaron Kaplan, Abby Kaplan, Miranda K. McCarvel, & Edward J. Rubin (eds.), Proceedings of the 34th West Coast Conference on Formal Linguistics, 395–403. Somerville, MA: Cascadilla Proceedings Project.

References (5)

- **O'Neil**, Wayne. 1977. Clause adjunction in Old English. *General Linguistics* 17, 199–211.
- **Pintzuk**, Susan. 2003. Variationist approaches to syntactic change. In Brian Joseph & Richard Janda (eds.), *The handbook of historical linguistics*, 509–528. Oxford: Blackwell.
- **Pullum**, Geoffrey K., & Barbara **Scholz**. 2010. Recursion and the infinitude claim. In Harry van der Hulst (ed.), *Recursion and human language*, 113–137. Berlin: de Gruyter.
- **Ravila**, Paavo. 1960. Proto-Uralic. In B. Collinder (ed.), *Comparative grammar of the Uralic languages*, 250–251. Stockholm: Almqvist & Wiksell.
- **Richards**, Marc. 2009. Internal pair-merge: the missing mode of movement. *Catalan Journal of Linguistics* 8, 55–73.
- **Roberts**, Ian. 2007. *Diachronic syntax*. Oxford: Oxford University Press.
- **Small**, George W. 1924. *The comparison of inequality*. Baltimore: University Press.
- **Suárez-Gómez**, Cristina. 2006. *Relativization in early English (950–1250): the position of relative clauses*. Berlin: Peter Lang.
- **Taylor**, Ann, Anthony **Warner**; Susan **Pintzuk**, & Frank **Beths**. 2003. The YorkToronto-Helsinki Parsed Corpus of Old English Prose.
- **Thiersch**, Friedrich. 1826. *Griechische Grammatik, vorzüglich des homerischen Dialekts*. 3rd edition. Leipzig: Fleischer.

References (6)

- **Wallenberg**, Joel C. 2016. Extraposition is disappearing. *Language: Historical Syntax* 92, e237–e256.
- **Wallenberg**, Joel C., Anton K. **Ingason**, Einar F. **Sigurðsson**, & Eiríkur **Rögnvaldsson**. 2011. Icelandic Parsed Historical Corpus (IcePaHC). Version 0.9.
- **Whitman**, John. 2000. Relabelling. In Susan Pintzuk, George Tsoulas & Anthony Warner (eds.), *Diachronic syntax: models and mechanisms*, 220–238. Oxford: Oxford University Press.
- **Widmer**, Manuel, Sandra **Auderset**, Johanna **Nichols**, Paul **Widmer** & Balthasar **Bickel**. 2017. NP recursion over time: evidence from Indo-European. *Language* 93, 799–826.
- **Yang**, Charles D. 2002. *Knowledge and learning in natural language*. Oxford: Oxford University Press.