Morphological complexity of Catalan: a diachronic and sociolinguistic perspective

Table of contents

- Background
- Methods
- Some preliminary results
- Caveats

Background

Language complexity

- We know that languages encode information in the morpho-syntactic realm in very different ways
- For example, some languages have more **elaborate morphological systems**, while others mark relations with **stricter word order**
- Vietnamese: *Anh ãý sẽ ngôi* (Badosa Roldós, personal communication)
    - 3SG.M DEM.MED FUT sit - “He will sit”
- Catalan: *seurà*
    - sit.3SG.FUT.IND - “He/she will sit”

Why are languages different?

- The language internal hypothesis: languages change from one configuration to another through internal processes only, such as grammaticalization and lexicalization
    - Sapir (1933): all languages are essentially perfect
        - Does this mean that they are equally complex?
    - Bickerton (2007), Bickel et al. (2024): language is given as it is from human genetics, and language change and variation is merely “recycling” through this space

Why are languages different?

- The cultural evolution hypothesis: language is a cultural tool which adapts to a combination of cognitive, cultural, demographic, and ecological factors
- Evans & Levinson (2009): biological and cultural evolution interact with one another. Languages change because of cultural factors, constrained by biology
- As languages are spoken by people in different (cultural, social, ecological) situations, one can consider it a *complex adaptive system* (Bentz, 2018)

The Language Niche Hypothesis (LNH)

- Out of the ideas of language as an adaptive system comes the **Language Niche Hypothesis:** languages adapt their morphosyntactic complexity to their social niche (Dale & Lupyan, 2012; Wray & Grace, 2007)
  - Languages are ordered on an *esotericity* continuum
    - *Exoteric* languages are spoken by large groups of people, over large areas
      - These languages tend to have more L2 speakers
        - They will tend to, then, use **simpler morphology** to accommodate for adult learners, and favor lexical (syntactic) strategies
  - *Esoteric* languages are spoken by smaller, more tightly connected groups of people
    - These languages are pressured towards informational redundancy in order to favor child acquisition
      - They will tend to acquire, then more **complex morphology**

Developments in the LNH

- The validity of the LNH has been debated over the last 15 years. Some recent developments:
    - **In favor:** Chen et al. (2024)
        - “[S]ociopolitical esotericity tends to correlate with morphological complexity, in the sense of more explicit markings and distinctions” (p. 9)
        - “[S]ociopolitical exotericity tends to correlate with more complex syntax, including more syntactic layering and more obligatory syntactic categories and distinctions” (p. 9)
    - **Against:** Shcherbakova et al. (2023)
        - “Both [proxy measures for complexity] scores are best predicted by the combination of phylogenetic and spatial effects” (p. 4)
        - “None of these relationships are negative as predicted by prior studies” (p. 6)

How to measure language complexity

- We can see two main ways to measure morpho-syntactic complexity

1. The **descriptive** (grammar-based) approach: use descriptions of language (usually as compiled in databases such as WALS (Dryer & Haspelmath, 2013) or Grambank (Skirgård et al., 2023))
    - **Number of categories** in features such as grammatical case, number distinctions [e.g. Chen et al. (2024)
        - A language with 5 nominal cases is more complex in this metric than a language with 3 cases, which is in turn more complex than a language with no case markings
    - **Fusion** and **obligatoriness** of markers (e.g. Shcherbakova et al., 2023)
        - A language which fuses markers of different morphological information together is more complex than one in which form and function approach a 1:1 ratio

How to measure language complexity

2. The **information-theoretical** (corpus-based) approach

    - Type-to-token ratio (TTR)
        - Divide the number of unique “words” (types) by the total number of words (tokens)
    - Shannon Entropy (Shannon, 1948)
        - How much information does a message contain in the context of other messages?
    - Kolmogorov complexity (Kolmogorov, 1963)

Why Catalan?

- Catalan is an interesting language to study as (potentially) an adaptive system because
    - It has a multifaceted nature and history
        - In the Middle Ages, an international language under the Crown of Aragon
        - Despite receiving variable state support and being considered a minoritized language in all four states where it is spoken (Baylac-Ferrer & Ferrerós-Pagès, in press), it has maintained high vitality
    - It is very well documented

Research questions

- Can we get a sense of what the **complexity of one language** is over time using corpus-based methods?
    - Will this match descriptive accounts of diachronic changes in grammar?
- Can we see a morphology-syntax tradeoff in the complexity of a language over time? (Nijs et al., 2025)
- Can we **correlate** changes in the complexity of a language using extra-linguistic (demographic, historical) factors?
- Will changes in the complexity of the language match with changes in the social structure of the societies that speak it?

Methods

Kolmogorov complexity

- Kolmogorov complexity is understood as the **shortest possible description length** of a string (example from Ehret & Szmrecsanyi, 2016)
    1. *cdcdcdcdcd* (10 characters) → *5×cd* (4 characters): less complex
    2. *cdgh39aby7* (10 characters) → *cdgh39aby7* (10 characters): more complex
- In language terms, this means that languages with more regularities will tend to reflect a lower Kolmogorov complexity
    1. *Fish fish fish fish fish fish fish* has a lower Kolmogorov complexity score than
    2. *Els peixos que els peixos pesquen pesquen peixos que els peixos pesquen*
- Kolmogorov complexity has been used to measure language complexity since Juola (1998). See Juola (2008), Ehret & Szmrecsanyi (2016), Nijs et al. (2025)
- The way one approaches Kolmogorov complexity in language studies is by way of **consumer-grade compression programs** (`gzip`, `xz`, `bzip2`) on plain text files.

Morphological and syntactic distortion

- The way that morphological and syntactic complexity is teased apart in Kolmogorov complexity techniques is usually through **distortion at different levels**
  - Morphological distortion is achieved through random deletion, substitution or permutation of characters **within a word**
    - This is supposed to disturb less complex texts (which have fewer overall word forms) more than more complex texts (which have more overall word forms)

$$\text{morphological complexity} = -\frac{\text{compressed file size after distortion}}{\text{compressed file size before distortion}}$$

Morphological and syntactic distortion

- Syntactic distortion is achieved through random deletion, substitution or permutation of **entire words**
  - This supposedly disturbs more syntactically complex (rigid) languages more than more syntactically "simple" (free) languages, as the former have fewer overall syntactic combinations
    - This is also known as **Inter Word Information** (IWI) (see Oh & Pellegrino, 2023)

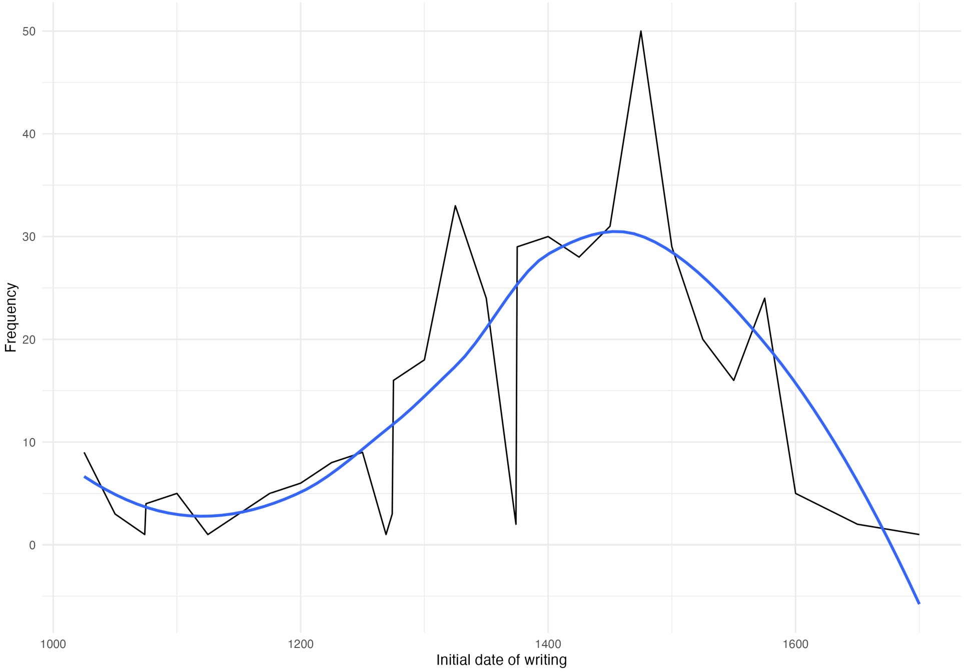
$$\text{syntactic complexity} = \frac{\text{compressed file size after distortion}}{\text{compressed file size before distortion}}$$

Statistical techniques

- Granger causality (Granger, 1969): a series of statistical tests which compares two **time series** and looks at its correlations
- Bayesian statistics with `brms`
    - Beta regressions
        - Probably non linear (polynomial) models

What is the corpus like?

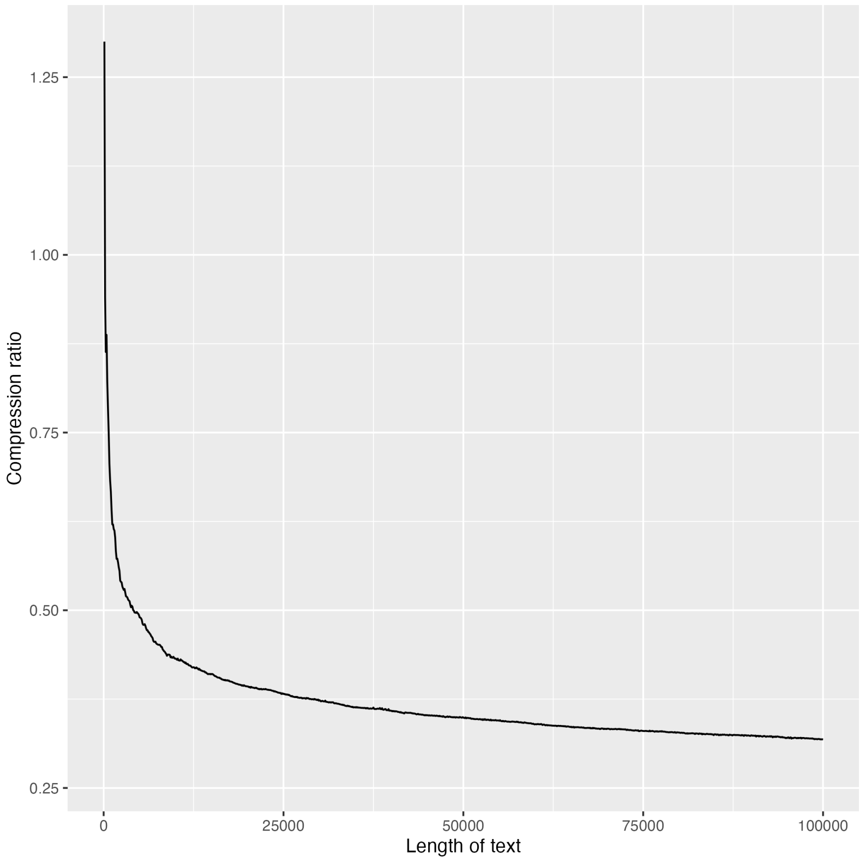
- I am using the Computerized Corpus of Old Catalan (Torruella et al., 2010). It contains 414 texts
- Ranging from the 11th to the 18th century
- Covering a variety of genres (legal, religious, poetic…)
- Tagged by dialect (Eastern, Western, Balearic, Valencian…)



Distribution of texts across the centuries

What is the corpus like?

- We have a lot of small texts that might be distorting the results
  - From my benchmarking, compressibility scales dramatically with the length of a text



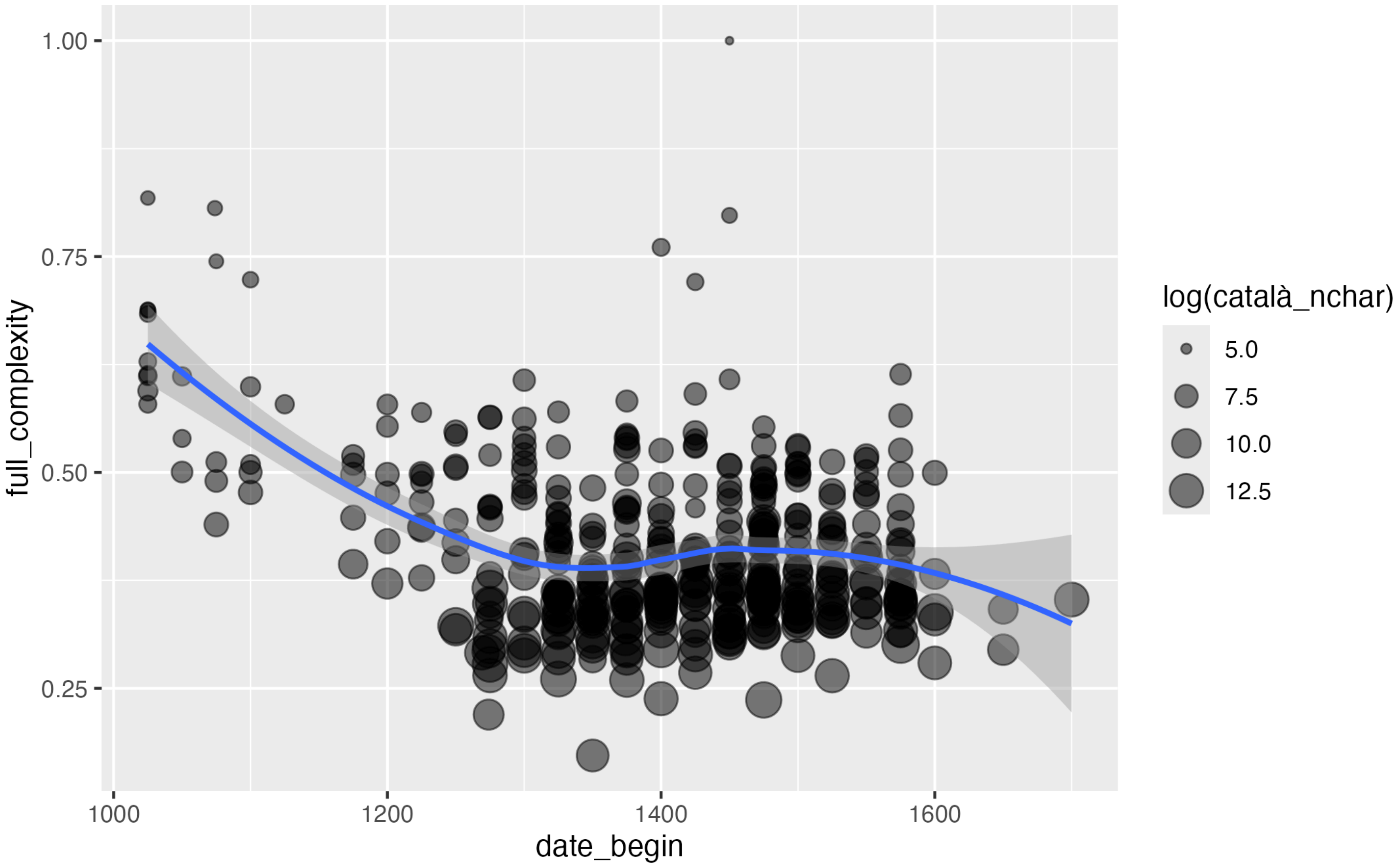
Testing the compressibility of Joanot Martorell’s *Tirant lo Blanc* as we slice it into bigger chunks

Variables taken into account

- Dependent variables: complexity of the text
  - Overall complexity
  - Morphological vs syntactic complexity
- Independent variables: possible predictors
  - Population size at time of writing
  - Historical events
  - Genre of text
  - Dialect
  - Author
  - Date
  - Multilingualism of the text
    - Most of the texts in the corpus contain passages in Latin, Occitan, Spanish, French, Italian… Which are tagged and accounted for

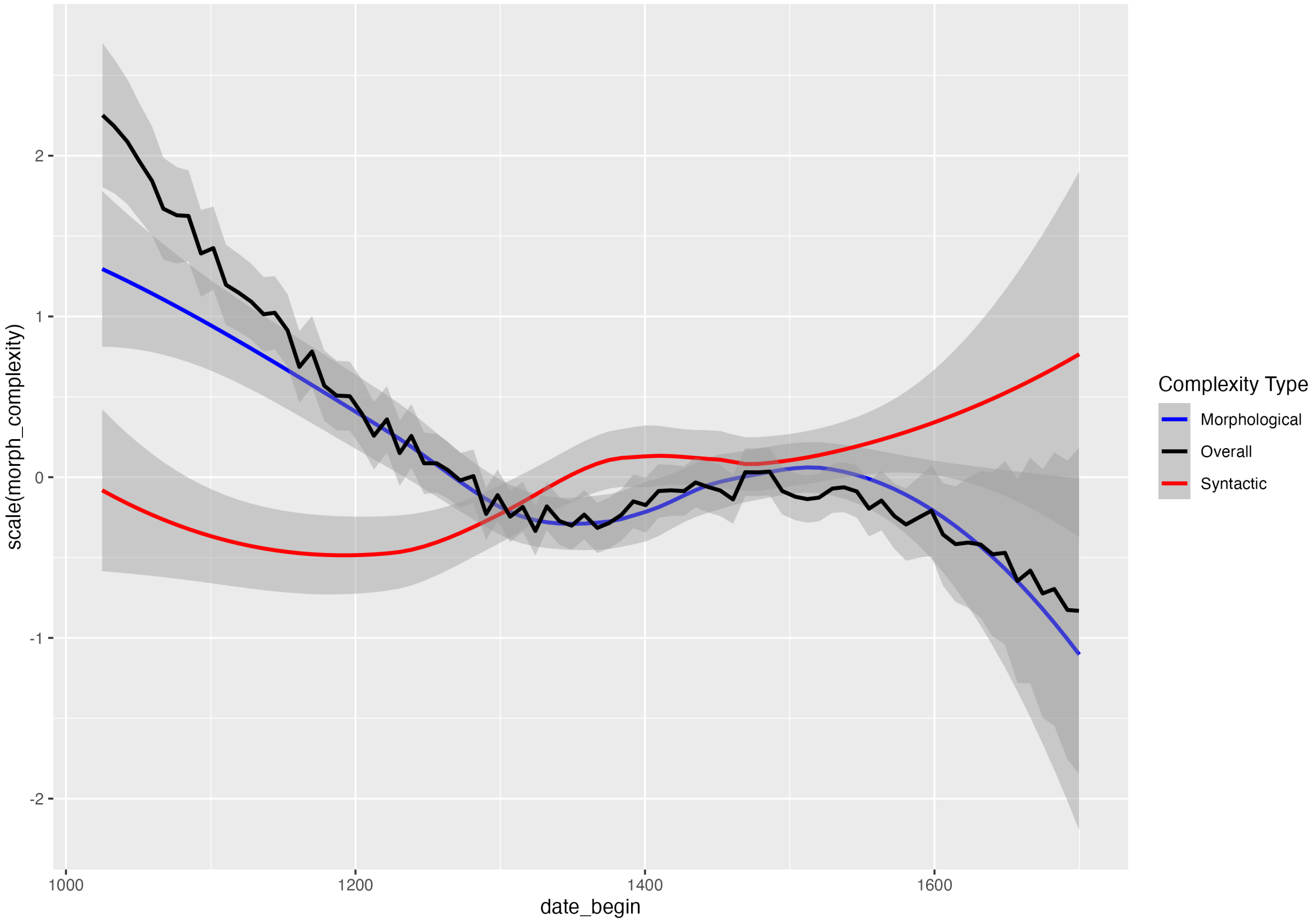
Some preliminary results

Full complexity versus time



Full complexity across time

Morphology-syntax trade-off



Morphological complexity vs syntactic complexity

Caveats

Character encodings

- We have a possible conundrum with encoding fromats
  - UTF-8 (the standard) is a variable-width encoding
    - This means that more common characters, such as <a>, take up less space than less “common” characters such as <à>
    - The text then, is already “compressed” to some extent.
    - Other formats such as UTF-32, which use fixed-width encoding, might (or might not) be more representative of “natural” language

Compression: what does it mean to compress?

- The standard in Kolmogorov complexity measuring of language is using the `gzip` program
- However, in terms of compression power, `gzip` is worse for text than other programs such as `xz` or (especially) `bzip2`

Type of compression	Average size of text	Minimum size of text	Maximum size of text
gzip	36715.86	129	815893
bzip2	28535.66	153	590667
xz	31024.44	188	625680
none	109876.86	141	2396447

Compression: what does it mean to compress?

- But is the program that’s best at *compressing text* also the best at *reflecting language complexity*?
    - Still an open question, which probably needs investigating at each step in the compression pipelinne

Morphological and syntactic distortion

- What does distortion capture?
- What is the best way to distort?
  - The most widely used technique is **deletion** of characters Nijs et al. (2024)
    - *Molt excel·lent, virtuós e gloriós príncep → Mot excel·let, virtós e glriós pícep*
- But this, naturally, affects the size of the text, which we know changes its overall compressibility.
- Some alternatives:
  - Change random characters by one specific character (e.g.)
    - *Molt excel·lent, virtuós e gloriós príncep → Mzlt excel·lenz, virtuós e gzoriós príncep*
  - Permutate characters (switch their positions)
    - *Molt excel·lent, virtuós e gloriós príncep → Melt oxceg·lent, virtuós e lloriós príncep*

Future steps

- Control better for length of text
- Compare results to descriptive accounts of grammatical change in Catalan
- Find historical events that might have triggered changes in the language
- Include texts from more modern corpora (e.g. Corpus Textual Informatitzat del Català)

References

Baylac-Ferrer, A., & Ferrerós-Pagès, C. (in press). Catalan across three states: A comparison of language minoritisation features. In M. Gandarillas (Ed.), *Fighting Stigma and Discrimination in Minoritised Languages*. Cambridge Scholars Publishing.

Bentz, C. (2018). *Adaptive languages: An information-theoretic account of linguistic diversity*. De Gruyter Mouton.

Bickel, B., Giraud, A.-L., Zuberbühler, K., & van Schaik, C. P. (2024). Language follows a distinct mode of extra-genomic evolution. *Physics of Life Reviews*, *50*, 211–225. https://doi.org/10.1016/j.plrev.2024.08.003

Bickerton, D. (2007). Language evolution: A brief guide for linguists. *Lingua*, *117*(3), 510–526. https://doi.org/10.1016/j.lingua.2005.02.006

Chen, S., Gil, D., Gaponov, S., Reifegerste, J., Yuditha, T., Tatarinova, T., Progovac, L., & Benítez-Burraco, A. (2024). Linguistic correlates of societal variation: A quantitative analysis. *PLOS ONE*, *19*(4), e0300838. https://doi.org/10.1371/journal.pone.0300838

Dale, R., & Lupyan, G. (2012). Understanding the origins of morphological diversity: The linguistic niche hypothesis. *Advances in Complex Systems*, *15*, 1150017. https://doi.org/10.1142/S0219525911500172

Dryer, M. S., & Haspelmath, M. (Eds.). (2013). *WALS online (v2020.4)* [Data set]. Zenodo. https://doi.org/10.5281/zenodo.13950591

Ehret, K., & Szmrecsanyi, B. (2016). An information-theoretic approach to assess linguistic complexity. In R. Baechler & G. Seiler (Eds.), *Complexity, Isolation, and Variation* (pp. 71–94). De Gruyter. https://doi.org/10.1515/9783110348965-004

Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, *32*(5), 429–448. https://doi.org/10.1017/S0140525X0999094X

Granger, C. W. J. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*, *37*(3), 424–438. https://doi.org/10.2307/1912791

Juola, P. (1998). Measuring linguistic complexity: The morphological tier. *Journal of Quantitative Linguistics*, *5*(3), 206–213. https://doi.org/10.1080/09296179808590128

Juola, P. (2008). Assessing linguistic complexity. In M. Miestamo, K. Sinnemäki, & F. Karlsson (Eds.), *Language Complexity: Typology, contact, change* (pp. 89–108). John Benjamins Publishing Company. https://doi.org/10.1075/slcs.94.07juo

Kolmogorov, A. N. (1963). On Tables of Random Numbers. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, *25*(4), 369–376. https://www.jstor.org/stable/25049284

Nijs, J., Van de Velde, F., & Cuyckens, H. (2024). An Information-Theoretic Approach to Morphosyntactic Complexity in English, Dutch and German. *Journal of Quantitative Linguistics*, *31*(4), 275–297. https://doi.org/10.1080/09296174.2024.2374613

Nijs, J., Van de Velde, F., & Cuyckens, H. (2025). Is Word Order Responsive to Morphology? Disentangling Cause and Effect in Morphosyntactic Change in Five Western European Languages. *Entropy*, *27*(1, 1), 53. https://doi.org/10.3390/e27010053

Oh, Y. M., & Pellegrino, F. (2023). Towards robust complexity indices in linguistic typology: A corpus-based assessment. *Studies in Language. International Journal Sponsored by the Foundation "Foundations of Language"*, *47*(4), 789–829. https://doi.org/10.1075/sl.22034.oh

Sapir, E. (1933). Language. In *Encyclopaedia of the Social Sciences* (Vol. 9, pp. 155–169).

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Shcherbakova, O., Michaelis, S. M., Haynie, H. J., Passmore, S., Gast, V., Gray, R. D., Greenhill, S. J., Blasi, D. E., & Skirgård, H. (2023). Societies of strangers do not speak less complex languages. *Science Advances*, *9*(33), eadf7704. https://doi.org/10.1126/sciadv.adf7704

Skirgård, H., Haynie, H. J., Hammarström, H., Blasi, D. E., Collins, J., Latarche, J., Lesage, J., Weber, T., Witzlack-Makarevich, A., Dunn, M., Reesink, G., Singer, R., Bowern, C., Epps, P., Hill, J., Vesakoski, O., Abbas, N. K., Ananth, S., Auer, D., … Gray, R. D.