UNIVERSITÀ
DI PAVIA

LINGUISTICS
IN PAVIA

ASSESSING MORPHOLOGICAL PRODUCTIVITY IN A CORPUS LANGUAGE: A DIACHRONIC STUDY OF ANCIENT GREEK DEVERBAL NOMINAL SUFFIXES

Silvia Zampetta - *University of Pavia*

Fifth International Workshop on Resources and Tools for Derivational Morphology
September 4 and 5, 2025, University of Fribourg, Switzerland

Roadmap

1. Introduction
2. Methodological Framework
3. Measuring the Productivity of Ancient Greek Deverbal Nominal Suffixes
   - Distribution and Relative Frequency Across Time
   - *P* Measure
   - *P** Measure
   - LNRE Models
   - Suffix Interaction and Resolution of Rivalry
4. Conclusion

UNIVERSITÀ DI PAVIA

1. Introduction

Background & Research Gap

- **Ancient Greek Deverbal Nominal Domain**
- **Well-studied from an Indo-European perspective**
  - *e.g.*, Debrunner 1916, Chantraine 1933, Benveniste 1948, Risch 1974
  - ➢ Focus: morphophonology & cross-linguistic comparison

- **Recent developments**
  - *-mo- in diachronic/typological framework (Napoli 2009)
  - Synchronic nominalizations (Civilleri 2010)

Background & Research Gap

- **Ancient Greek Deverbal Nominal Domain**
- **Well-studied from an Indo-European perspective**
  - E.g., Debrunner 1916, Chantraine 1933, Benveniste 1948, Risch 1974
  - ➢ Focus: morpho-phonology & cross-linguistic comparison
- **Recent developments**
  - -mo- in diachronic/typological framework (Napoli 2009)
  - Synchronic nominalizations (Civilleri 2010)

- **No quantitative & diachronic analysis of morphological productivity**

UNIVERSITÀ DI PAVIA

Background & Research Gap

- Most empirical research on productivity in derivational morphology has focused on modern languages, mainly due to the availability of large electronic corpora and computational tools

  - ✓ English (Baayen 1989, 1992, 1993, 2009)
  - ✓ German (Evert and Lüdeling 2001)
  - ✓ Italian (Gaeta and Ricca 2003, 2005, 2006, Varvara 2019, 2020)
  - ✓ **Old Italian** (Štichauer 2006), which introduced a diachronic dimension

Aims, Corpus & Data Extraction

- **Research Aims**

1. **Measure productivity** of six AG deverbal nominal suffixes in diachrony:
   - ✓ -*eía*, -*mos*/-*mós*, -*sia*, -*sis*, -*tis*, -*tus* + their allomorphes

   (Chantraine 1953: only suffixes whose function of creating abstract names from verbs is already recognized, and whose phonetic substance is clear)

2. Using **corpus-based statistical methods** (Baayen 1989 et seq.)
3. **Evaluate applicability** of modern productivity measures to **Ancient Greek**

Aims, Corpus & Data Extraction

- **Corpus (< *Thesaurus Linguae Graecae*)**

1. ~4 million tokens from 8th c. BC to 6th c. AD
2. Divided into 4 sub-corpora: Archaic, Classical, Hellenistic, Imperial
3. Balanced by token count and genres
4. Philological consistency: only texts with available critical editions, commentaries, and translations

Historical Period	*Archaic*	*Classic*	*Hellenistic*	*Imperial*
Token	277.876	1.231.944	1.158.453	1.288.522
Literary genres	3	5	6	5

UNIVERSITÀ DI PAVIA

Aims, Corpus & Data Extraction

- **Data extraction**

  - ✓ Based on Liddell-Scott-Jones lexicon (Perseus)
  - ✓ Manual checking for relevant deverbal nouns, excluding:
    - POS ≠ Noun
    - Non-deverbal derived nouns
    - Compounds
    - Proper nouns
    - Borrowings
    - Baseless formations
  - ✓ **Final dataset**: 1905 types and 50,637 tokens

2. Methodological Framework

Define productivity

- **Theoretical definition**

  - Plag 2006: The productivity of a given affix refers to its **potential** to form new words and the **extent** to which this potential is actually realized in language use

Define productivity

- **Theoretical definition**

  - Plag 2006: The productivity of a given affix refers to its **potential** to form new words and the **extent** to which this potential is actually realized in language use

  ✓ Potential = qualitative feature

Define productivity

- **Theoretical definition**

  - Plag 2006: The productivity of a given affix refers to its **potential** to form new words and the **extent** to which this potential is actually realized in language use

  ✓ Potential = qualitative feature

  ✓ Extent = quantitative feature

Define productivity

- Theoretical definition
  - Plag 2006: The productivity of a given affix refers to its **potential** to form new words and the **extent** to which this potential is actually realized in language use

✓ Potential = qualitative feature

✓ **Extent = quantitative feature**

OPERATIONALIZATION

Define productivity

- **Operative definition** (< corpus-based statistical methods)
  - Productivity is:

✓ Synchronic

✓ Linked to the number of **hapax legomena**, i.e., words with a frequency of 1 in a given corpus

Define productivity

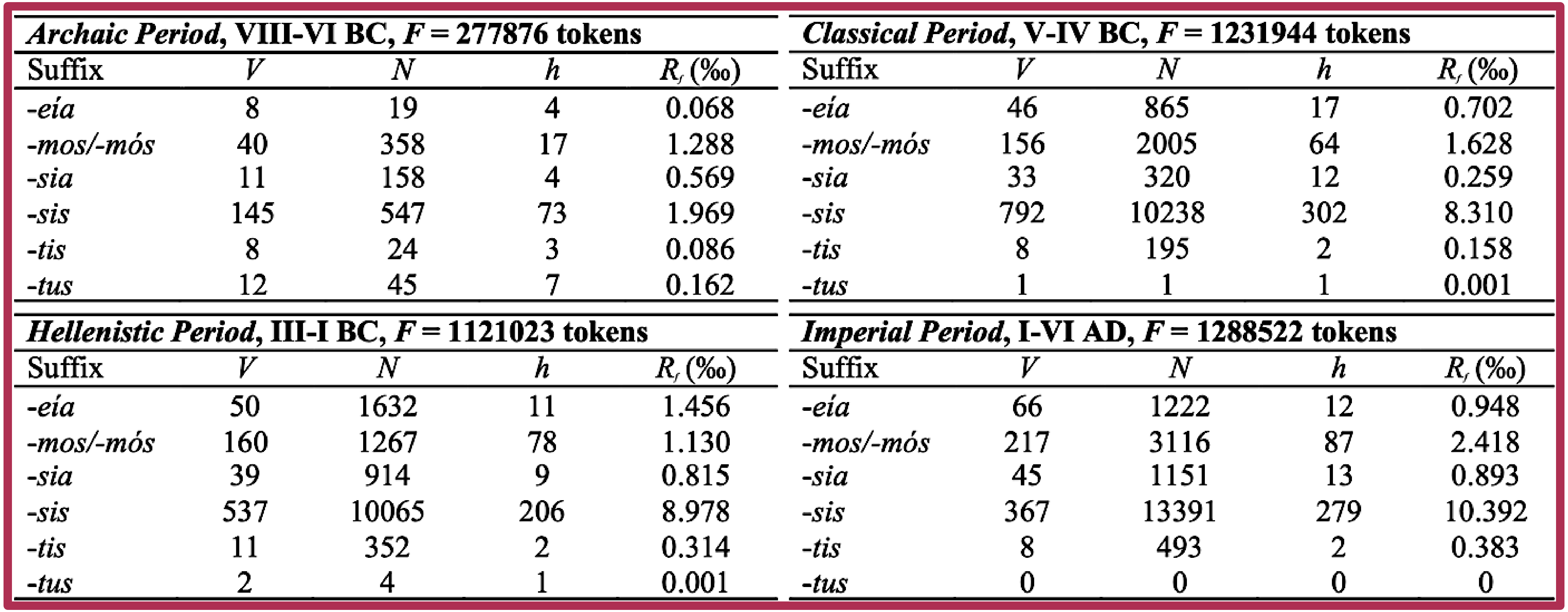
- **Operative definition** (< corpus-based statistical methods)
  - Productivity is:
  - ✓ Synchronic
  - ✓ Linked to the number of hapax legomena, i.e., words with a frequency of 1 in a given corpus

- **Hapax legomenon** = approximation of neologism
  - o In large corpora: unfamiliar words indicate an ongoing word formation process
  - o Psycholinguistics view:
  - ➤ Speakers decompose rare words into known morphemes
  - o **Productive rules** → many rare/new forms
  - o **Unproductive rules** → few high-frequency, well-established words

UNIVERSITÀ DI PAVIA

3. Measuring the Productivity of Ancient Greek Deverbal Nominal Suffixes

Distribution and Relative Frequency Across Time



Archaic Period, VIII-VI BC, *F* = 277876 tokens

Suffix	*V*	*N*	*h*	R_f (‰)
-eía	8	19	4	0.068
-mos/-mós	40	358	17	1.288
-sia	11	158	4	0.569
-sis	145	547	73	1.969
-tis	8	24	3	0.086
-tus	12	45	7	0.162

Classical Period, V-IV BC, *F* = 1231944 tokens

Suffix	*V*	*N*	*h*	R_f (‰)
-eía	46	865	17	0.702
-mos/-mós	156	2005	64	1.628
-sia	33	320	12	0.259
-sis	792	10238	302	8.310
-tis	8	195	2	0.158
-tus	1	1	1	0.001

Hellenistic Period, III-I BC, *F* = 1121023 tokens

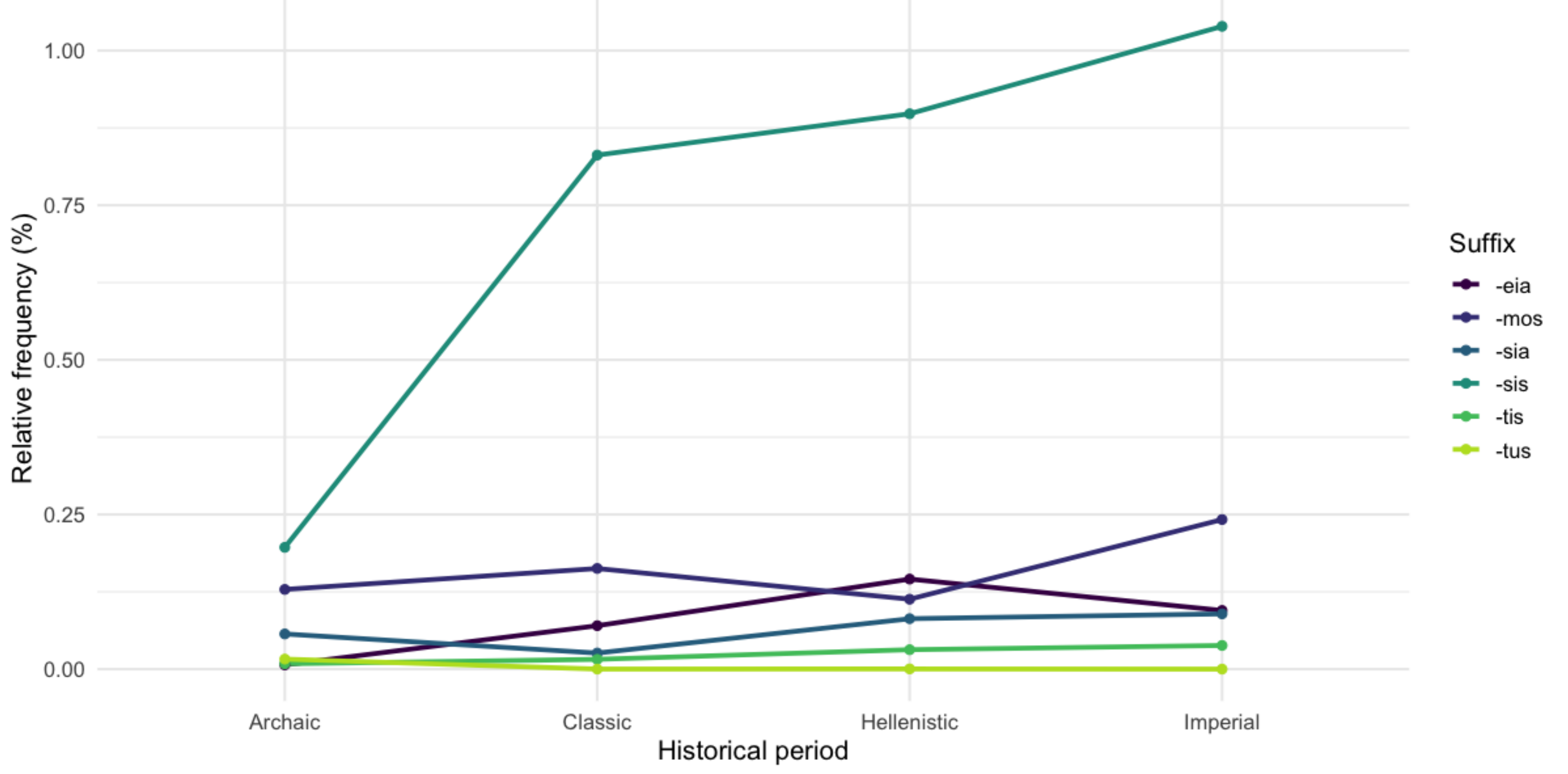
Suffix	*V*	*N*	*h*	R_f (‰)
-eía	50	1632	11	1.456
-mos/-mós	160	1267	78	1.130
-sia	39	914	9	0.815
-sis	537	10065	206	8.978
-tis	11	352	2	0.314
-tus	2	4	1	0.001

Imperial Period, I-VI AD, *F* = 1288522 tokens

Suffix	*V*	*N*	*h*	R_f (‰)
-eía	66	1222	12	0.948
-mos/-mós	217	3116	87	2.418
-sia	45	1151	13	0.893
-sis	367	13391	279	10.392
-tis	8	493	2	0.383
-tus	0	0	0	0

Distribution and Relative Frequency Across Time

- **Relative frequency trends** of suffixes across historical periods



UNIVERSITÀ DI PAVIA

Distribution and Relative Frequency Across Time

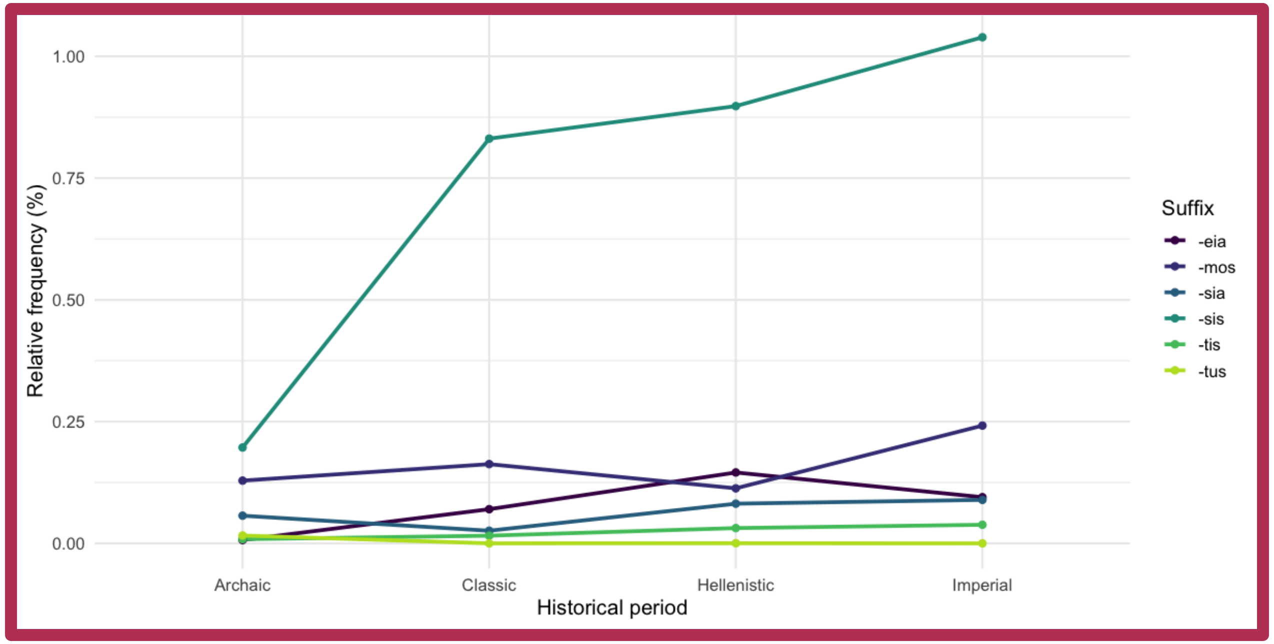
Chi-squared with simulated p-values (10,000 replicates)

➤ Result: $\boldsymbol{\chi^2}$ **= 3236.7, $\boldsymbol{p}$ = 9.999e-05**

➤ **Significant** association between **suffix type** and **period**

But weak effect size

→ Cramér's V = 0.149



P Measure

- ***P* (Potential Productivity, Baayen 2009)**
  - Formula: $P = h / N$
  - h = hapaxes with a given affix
  - N = total tokens with that affix

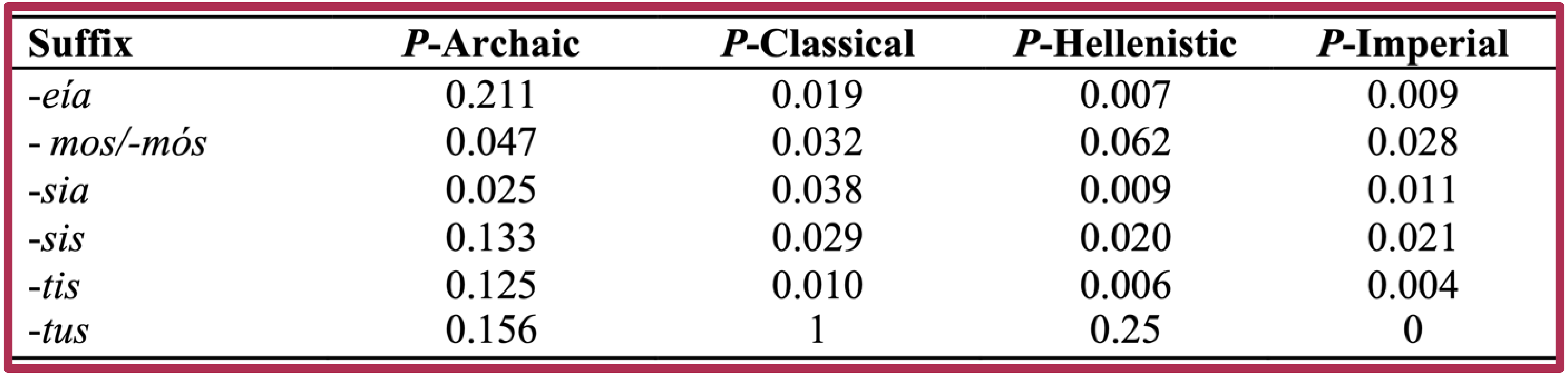
➤ Estimates the probability of encountering a new type after sampling N tokens with an affix

➤ Reflects the affix’s speed and capacity to expand its lexical inventory

- P is a decreasing function
- Approaches zero as N increases
- Overestimates rare suffixes
- Produces counterintuitive results when suffixes with very different token frequency are compared

P Measure in AG

- **Archaic phase → inflated *P* values due to small corpus size**
- -*tus* (rarest suffix) appears highly productive
- -*sis* (most frequent suffix) scores very low *P*, esp. in Classical & Hellenistic



Suffix	***P*-Archaic**	***P*-Classical**	***P*-Hellenistic**	***P*-Imperial**
-*eía*	0.211	0.019	0.007	0.009
- *mos*/-*mós*	0.047	0.032	0.062	0.028
-*sia*	0.025	0.038	0.009	0.011
-*sis*	0.133	0.029	0.020	0.021
-*tis*	0.125	0.010	0.006	0.004
-*tus*	0.156	1	0.25	0

*P** Measure

- ***P**** **(Expanding Productivity, Baayen 2009)**
  - Formula: *P* = h / H*
  - *h* = hapaxes with a given affix
  - *H* = total hapaxes in a corpus

➤ Enables comparisons across affixes

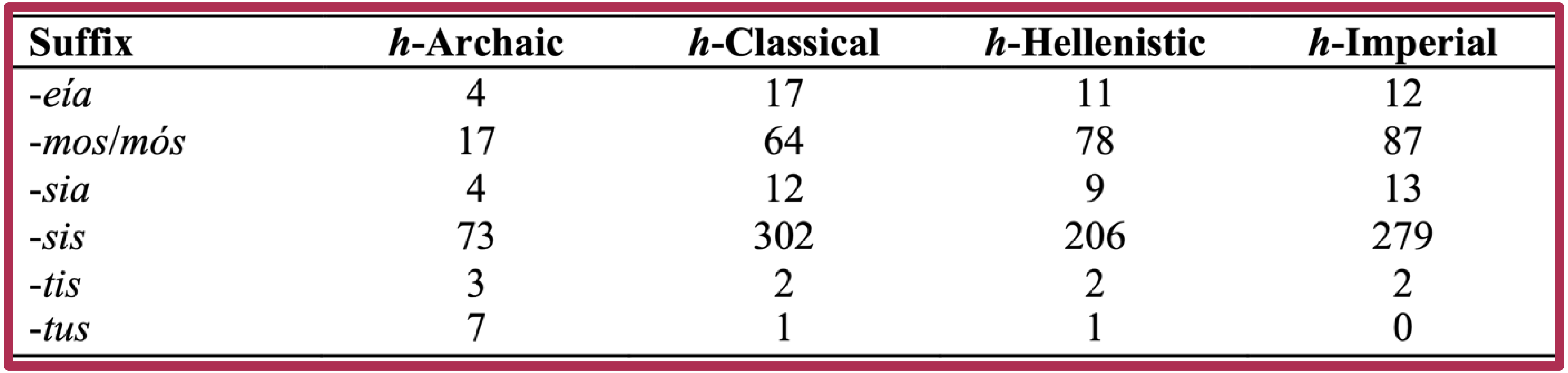
➤ Since *H* is constant, comparing *P** for the six suffixes is equivalent to directly compare the number of their hapaxes, regardless their total respective frequency

➤ **Conceptual critique**: reflects affix share of new words, not true productivity rate

I introduced the label P* for practical reasons

*P** Measure in AG

- -*sis* = most productive suffix across all periods → **core role in deverbal nominalization**
- -*mos/-mós* = **2nd most productive**, peaks in Hellenistic period (stylistic influence?)
- -*tus* = high in Archaic, then rapid decline, absent in Imperial era → **genre-specific use?**
- -*eía* & -*sia* = **low productivity overall**; -*sia* surpasses -*eía* only in Imperial phase (minor fluctuation not statistycally significant)
- -*tis* = **not productive** in any period



Suffix	*h*-Archaic	*h*-Classical	*h*-Hellenistic	*h*-Imperial
-*eía*	4	17	11	12
-*mos*/*mós*	17	64	78	87
-*sia*	4	12	9	13
-*sis*	73	302	206	279
-*tis*	3	2	2	2
-*tus*	7	1	1	0

LNRE Models

- P is negatively sensitive to token frequency variation across affixes
- P^* is less informative

- **Solution: LNRE Models (Large Number of Rare Events)**

➤ Predict hapax distribution beyond observed corpus size

➤ Estimate P for any N, even larger than observed

LNRE Models

Popular Models:

- **GIGP** (Generalized Inverse Gauss-Poisson)
- **fZM & ZM** (finite Zipf-Mandelbrot, Zipf-Mandelbrot)

  → (Implemented in zipfR - *R* package)

➤ Allow balanced comparisons across affixes with different frequencies

➤ Useful for ancient language corpora with uneven data distributions

LNRE models in AG

Model Used:

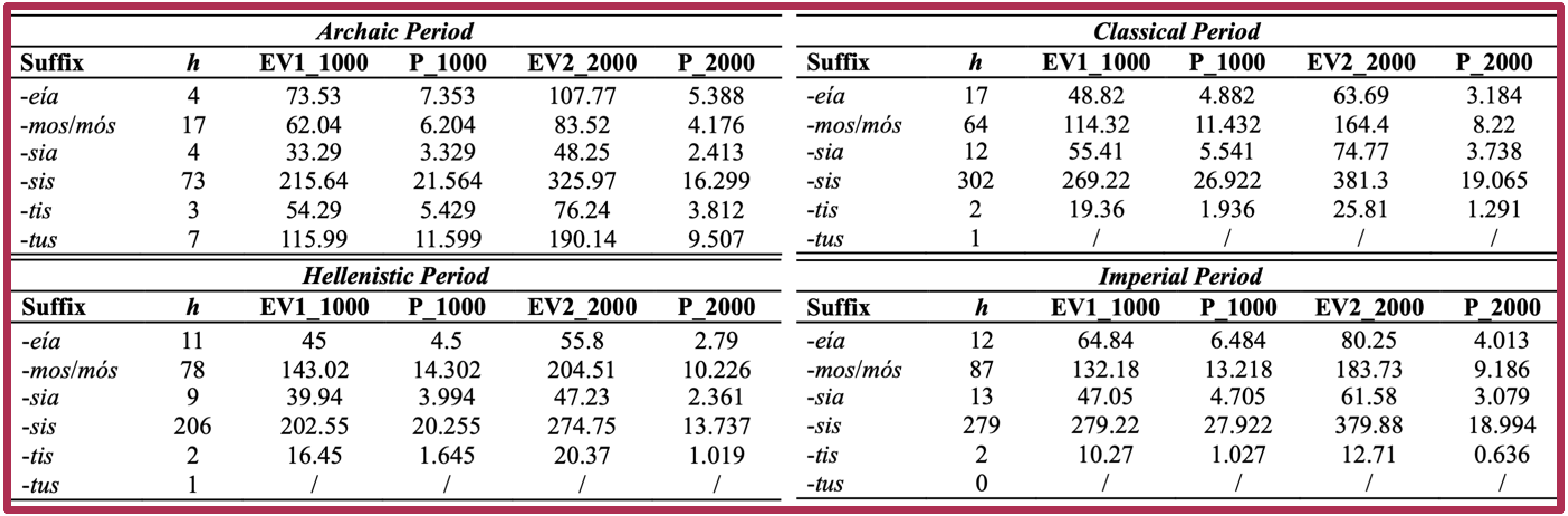
- **Zipf-Mandelbrot (ZM)**

➤ More reliable with small samples than fZM or GIGP (Evert & Baroni 2006)

➤ Based on observed frequency distributions estimates:

- Expected hapaxes for N = 1000 and N = 2000
- Corresponding productivity values: P_{1000}, P_{2000}

LNRE models in AG

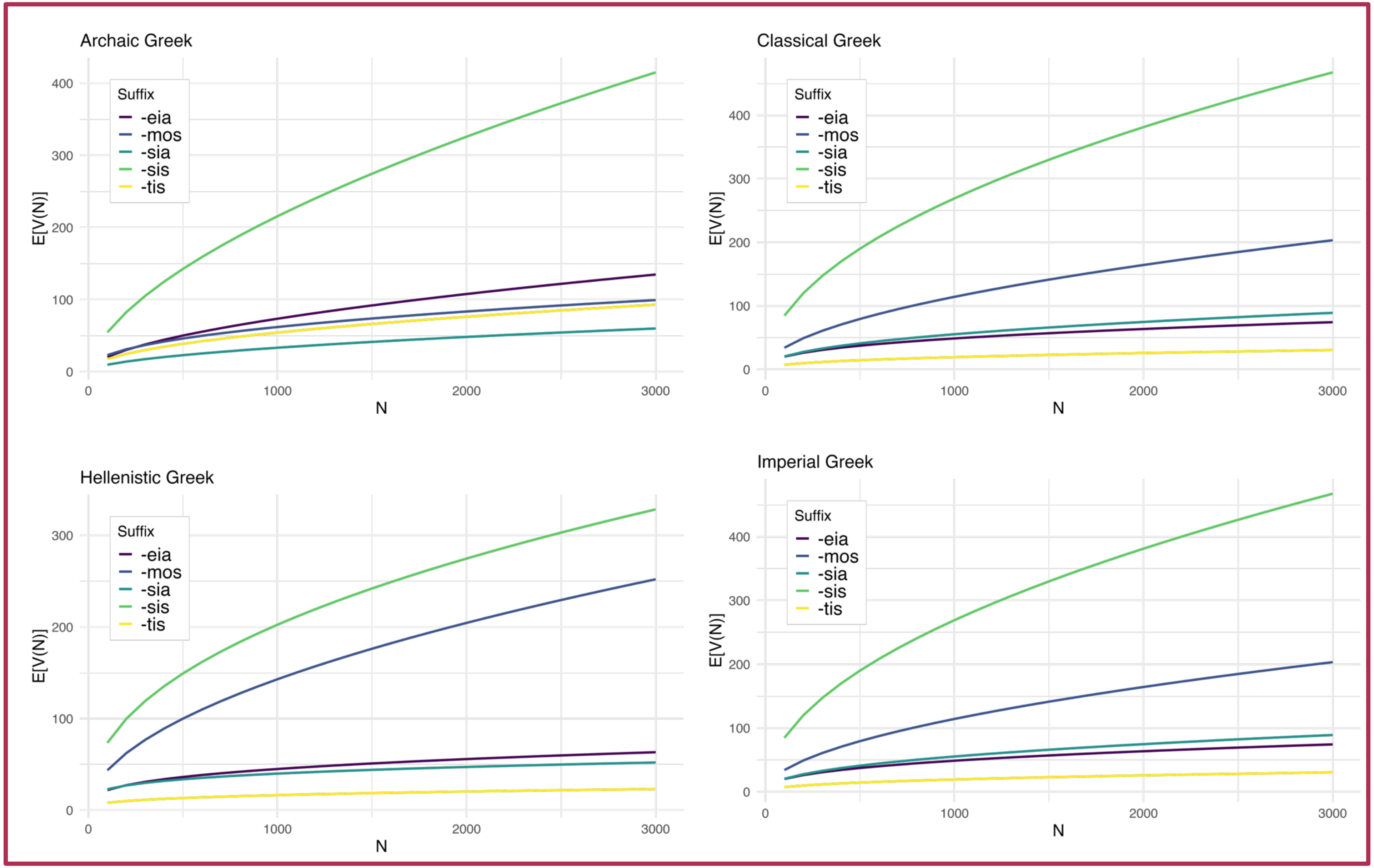


		Archaic Period			
Suffix	***h***	**EV1_1000**	**P_1000**	**EV2_2000**	**P_2000**
-eía	4	73.53	7.353	107.77	5.388
-mos/mós	17	62.04	6.204	83.52	4.176
-sia	4	33.29	3.329	48.25	2.413
-sis	73	215.64	21.564	325.97	16.299
-tis	3	54.29	5.429	76.24	3.812
-tus	7	115.99	11.599	190.14	9.507

		Classical Period			
Suffix	***h***	**EV1_1000**	**P_1000**	**EV2_2000**	**P_2000**
-eía	17	48.82	4.882	63.69	3.184
-mos/mós	64	114.32	11.432	164.4	8.22
-sia	12	55.41	5.541	74.77	3.738
-sis	302	269.22	26.922	381.3	19.065
-tis	2	19.36	1.936	25.81	1.291
-tus	1	/	/	/	/

		Hellenistic Period			
Suffix	***h***	**EV1_1000**	**P_1000**	**EV2_2000**	**P_2000**
-eía	11	45	4.5	55.8	2.79
-mos/mós	78	143.02	14.302	204.51	10.226
-sia	9	39.94	3.994	47.23	2.361
-sis	206	202.55	20.255	274.75	13.737
-tis	2	16.45	1.645	20.37	1.019
-tus	1	/	/	/	/

		Imperial Period			
Suffix	***h***	**EV1_1000**	**P_1000**	**EV2_2000**	**P_2000**
-eía	12	64.84	6.484	80.25	4.013
-mos/mós	87	132.18	13.218	183.73	9.186
-sia	13	47.05	4.705	61.58	3.079
-sis	279	279.22	27.922	379.88	18.994
-tis	2	10.27	1.027	12.71	0.636
-tus	0	/	/	/	/



Archaic Greek

Classical Greek

Hellenistic Greek

Imperial Greek

Suffix Interaction and Resolution of Rivalry

Goal: Explore possible competition among AG deverbal nominal suffixes

Motivation:

➤ Hypotheses in literature suggest morphological rivalry

➤ Notably: Chantraine (1933) proposes a competitive link between:

-sis ↔ *-mos/-mós*

-sis ↔ *-sia*

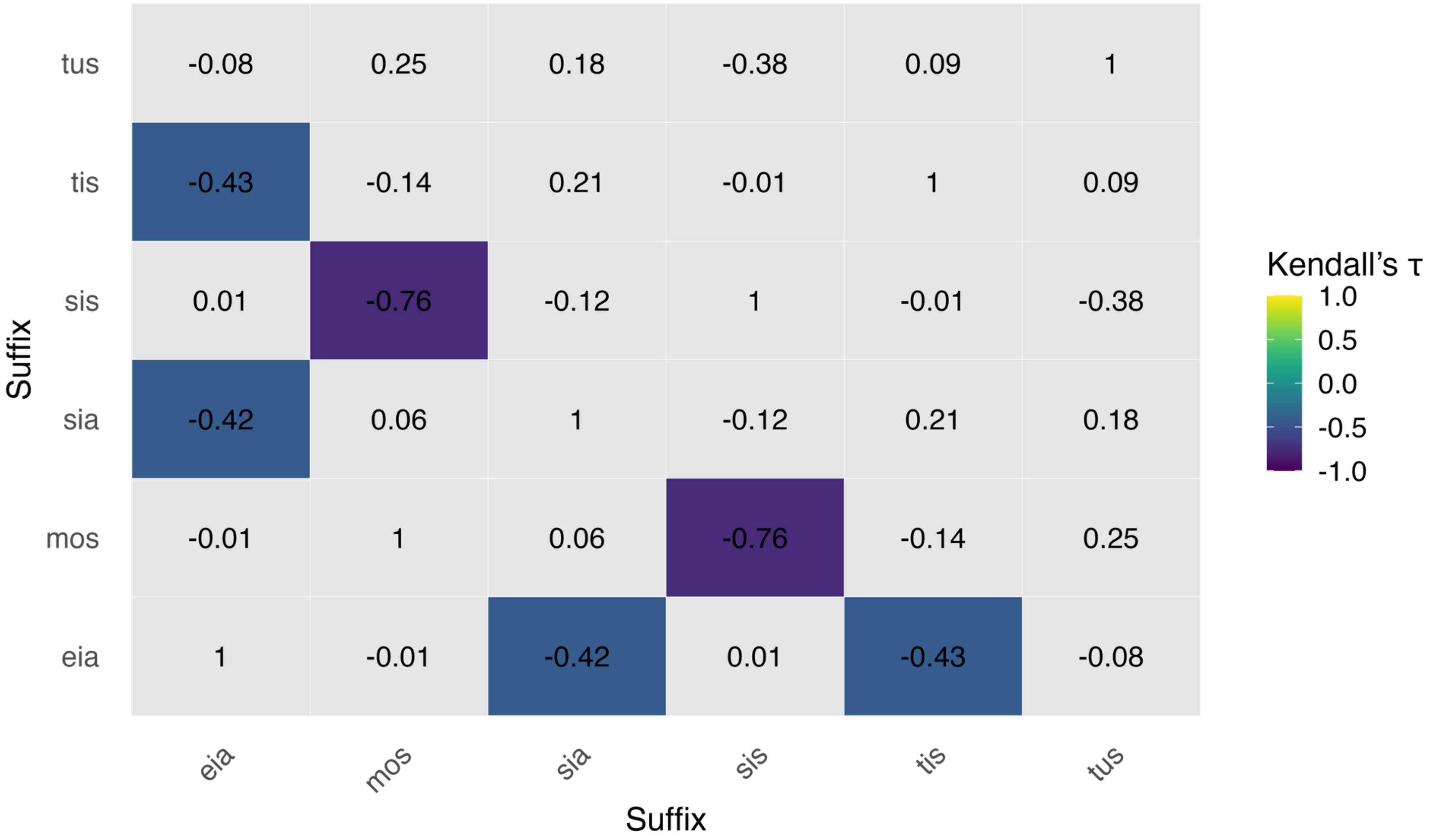
➤ Use quantitative data to test these claims and uncover new patterns of competition within the suffix system

UNIVERSITÀ DI PAVIA

Suffix Interaction and Resolution of Rivalry (Kendall’s Tau correlation)

Suffix Interaction and Resolution of Rivalry

-sis* vs *-mos/-mós
→ strong negative correlation (*p = 5.34e-06*)



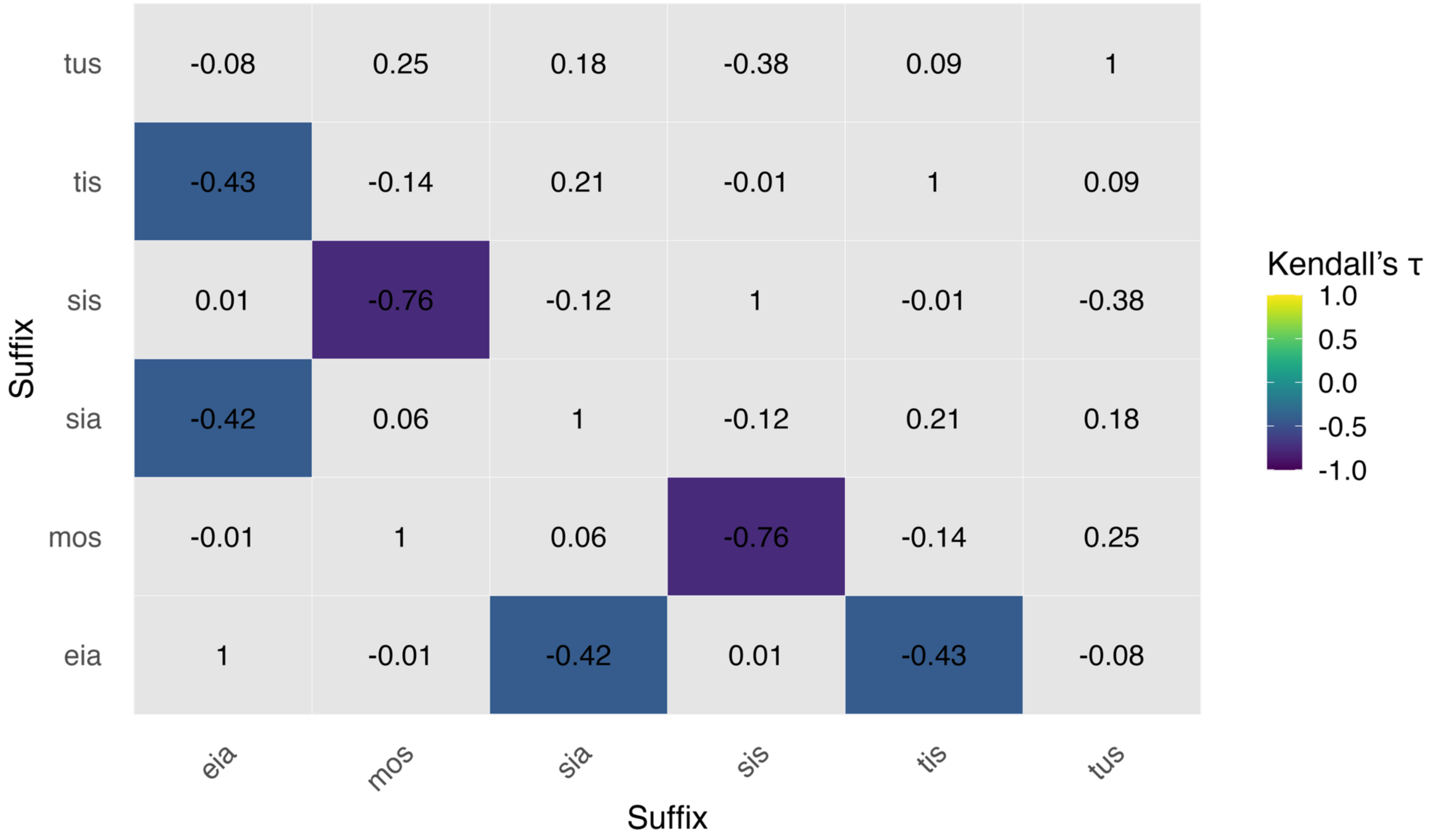
Suffix	eia	mos	sia	sis	tis	tus
tus	-0.08	0.25	0.18	-0.38	0.09	1
tis	-0.43	-0.14	0.21	-0.01	1	0.09
sis	0.01	-0.76	-0.12	1	-0.01	-0.38
sia	-0.42	0.06	1	-0.12	0.21	0.18
mos	-0.01	1	0.06	-0.76	-0.14	0.25
eia	1	-0.01	-0.42	0.01	-0.43	-0.08

Suffix Interaction and Resolution of Rivalry

×

-sia* vs *-sis

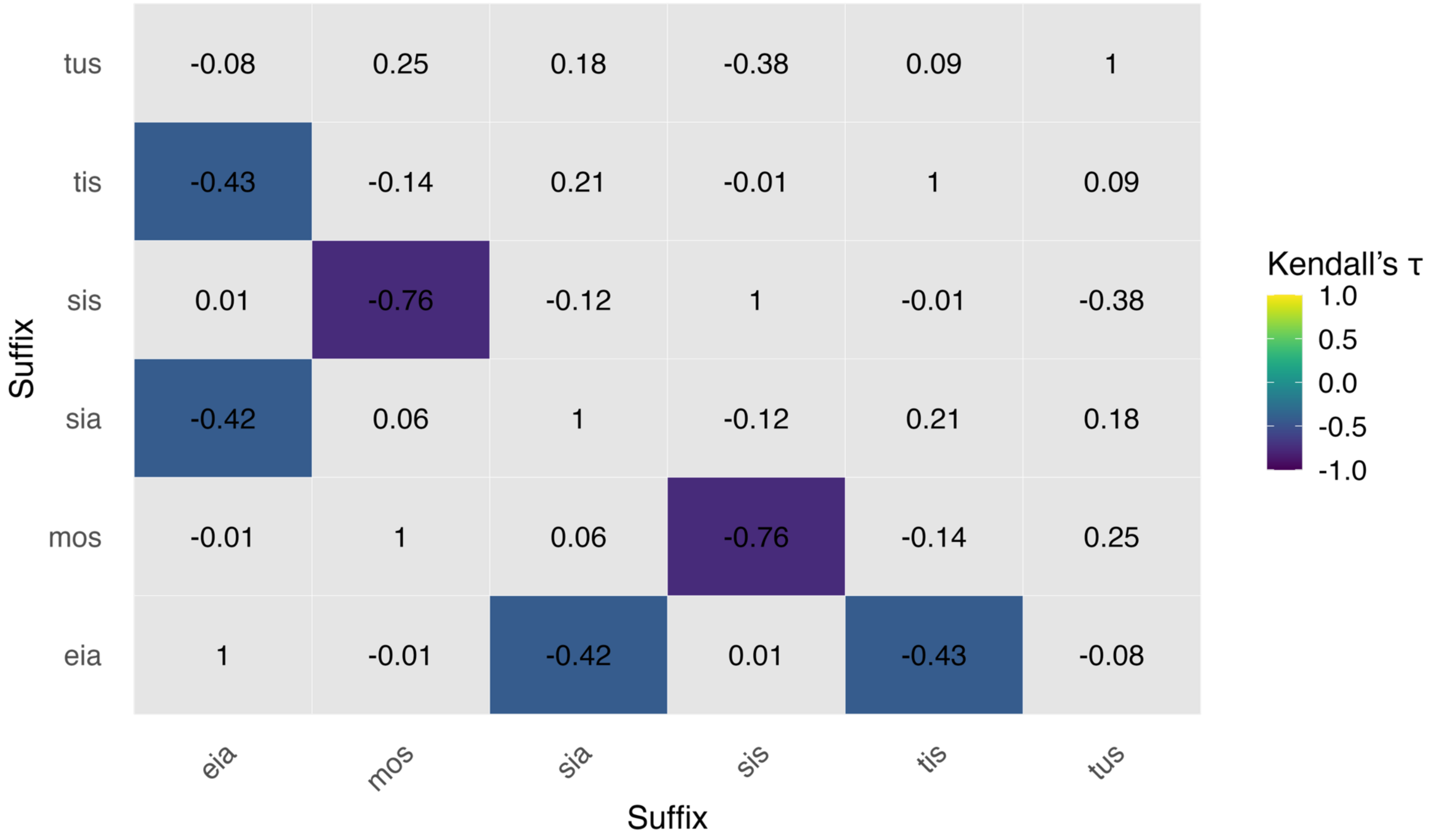
p = 4.84e-01



Suffix	eia	mos	sia	sis	tis	tus
tus	-0.08	0.25	0.18	-0.38	0.09	1
tis	-0.43	-0.14	0.21	-0.01	1	0.09
sis	0.01	-0.76	-0.12	1	-0.01	-0.38
sia	-0.42	0.06	1	-0.12	0.21	0.18
mos	-0.01	1	0.06	-0.76	-0.14	0.25
eia	1	-0.01	-0.42	0.01	-0.43	-0.08

Suffix Interaction and Resolution of Rivalry

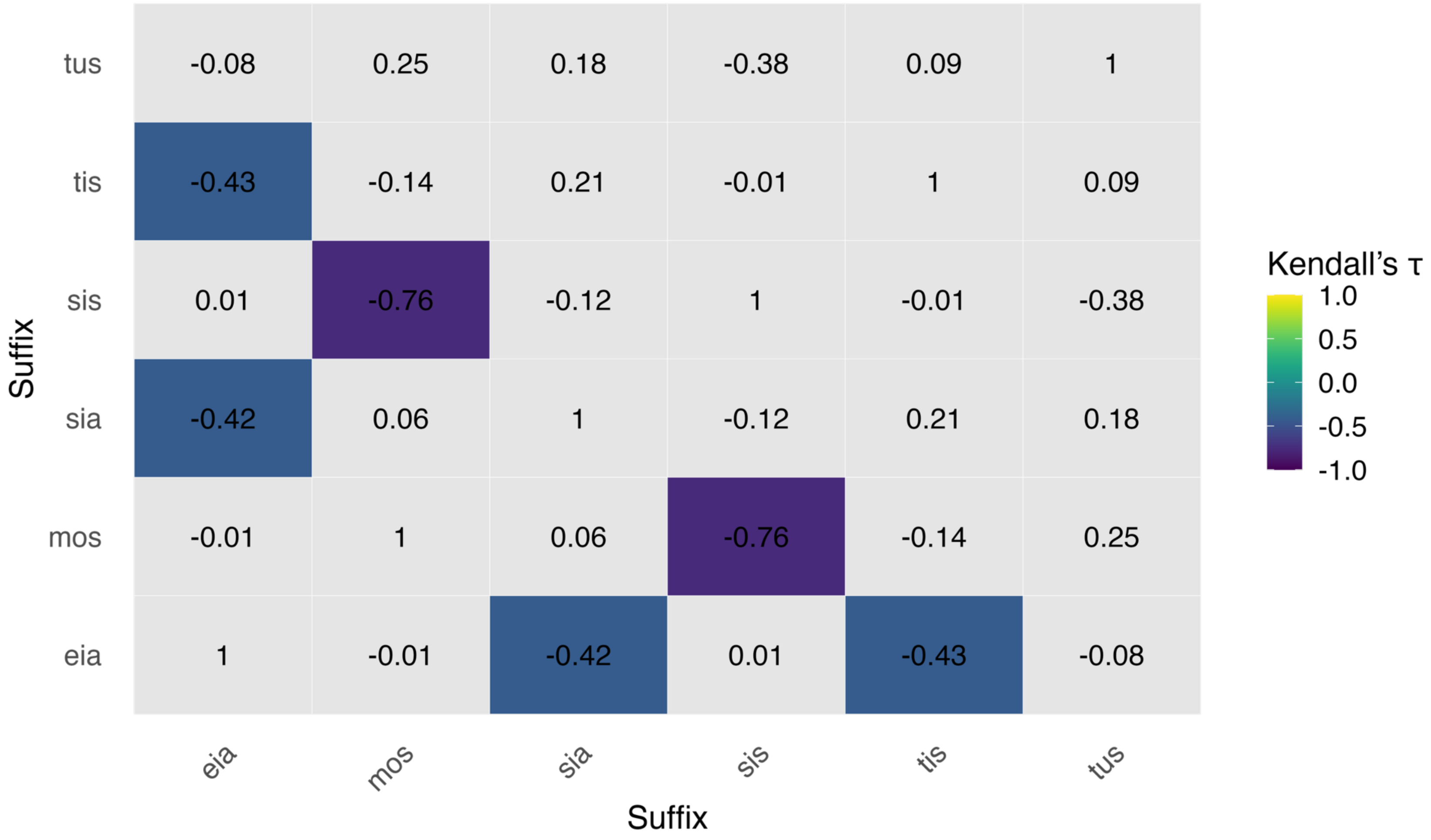
-sia* vs *-eía
→ moderate negative correlation (p = 0.0116)



Suffix	eia	mos	sia	sis	tis	tus
tus	-0.08	0.25	0.18	-0.38	0.09	1
tis	-0.43	-0.14	0.21	-0.01	1	0.09
sis	0.01	-0.76	-0.12	1	-0.01	-0.38
sia	-0.42	0.06	1	-0.12	0.21	0.18
mos	-0.01	1	0.06	-0.76	-0.14	0.25
eia	1	-0.01	-0.42	0.01	-0.43	-0.08

Suffix Interaction and Resolution of Rivalry

-eía **vs** ***-tis***
→ weak negative correlation (*p* = 0.0138)



Suffix	eia	mos	sia	sis	tis	tus
tus	-0.08	0.25	0.18	-0.38	0.09	1
tis	-0.43	-0.14	0.21	-0.01	1	0.09
sis	0.01	-0.76	-0.12	1	-0.01	-0.38
sia	-0.42	0.06	1	-0.12	0.21	0.18
mos	-0.01	1	0.06	-0.76	-0.14	0.25
eia	1	-0.01	-0.42	0.01	-0.43	-0.08

Suffix Interaction and Resolution of Rivalry

- …–*sia* and –*sis*: Chantraine was wrong?
- Negative correlations suggest **resolved** past competition
  - Functional specialization
  - Genre-specific preferences
  - Suffix decline
- **Correlations reflect outcomes**, not active competition
- Further qualitative analysis needed to explore:
  - Genre-specific uses
  - Functional overlap or specialization
  - Possible **overabundance** patterns

4. Conclusions

Conclusions: Results

1. Methodological Contribution:

- First quantitative and diachronic study of the deverbal nominal domain in AG
- I applied:
  - ➤ *P* (Potential Productivity)
  - ➤ *P** (Expanding Productivity)
  - ➤ ZM Model (LNRE)

Conclusions: Results

2. Key Findings:

- -*sis* = most productive suffix across all periods
- -*mos*/*mós* = productive, esp. in Hellenistic period
- -*eía* and -*sia* = limited, unstable productivity
- -*tis* and -*tus* = non-productive (esp. -*tus*, limited to archaic epic)

Conclusions: Results

3. Suffix Competition:

- Significant negative correlations suggest resolved rivalry
  - ✓ *-sis* vs *-mos/mós*
  - ✓ *-sia* vs *-eía*
  - ✓ *-eía* vs *-tis*

No correlation ≠ no rivalry

Conclusions: Challenges

- **Challenges in Quantifying Morphological Productivity in Ancient Greek**

1. Data Sparsity and Imbalance
2. Structural Inhomogeneity of Diachronic Corpora
3. Limitations of automatic POS-tagging in Ancient Greek

Conclusions: Challenges

1. **Data Sparsity and Imbalance**

- **The corpus size negatively influences the metrics**
  - Ancient Greek corpora are limited in size, especially in early periods like the Archaic era
  - This leads to inflated productivity estimates for rare suffixes and underrepresentation of more common ones
  - Uneven suffix frequency across periods can distort quantitative results (e.g., high *P* for rare -*tus*, low *P* for frequent -*sis*)

Conclusions: Challenges

2. **Structural Inhomogeneity of Diachronic Corpora (cf. Štichauer 2006)**

- The corpus includes texts of diverse genres and authorship, unevenly distributed over time

- Some genres are absent in certain periods (e.g., historiography in the Archaic phase), which biases affix visibility

- Repetition effects from single authors can skew data – e.g., a coined form may appear multiple times within one work but not elsewhere
  → **LOSING OF A NEW COINAGE**

Conclusions: Challenges

3. Errors in Automatic PoS-tagging

- Automatic annotation can misclassify homographic forms (e.g., *amúxeis* as a noun or verb)

- This introduces noise into suffix frequency counts

- **Solution**: manual review of a representative sample to estimate and minimize error rate
  → **CURRENTLY IN PROGRESS**

Future Directions and Methodological Considerations

1. Integrated Approach to Productivity

- No single measure (*P*, *P**, or LNRE) is sufficient alone
- Combined use of multiple metrics offers a more reliable view, especially when results converge (e.g., ZM and P*)
- Requires **critical interpretation** informed by frequency distribution and corpus structure

Future Directions and Methodological Considerations

2. Suffix Usage by Literary Genre

- Currently analyzing suffix productivity across literary genres

- **Goal**: determine whether shifts in productivity reflect genuine morphological trends or stylistic preferences

UNIVERSITÀ DI PAVIA

Future Directions and Methodological Considerations

3. Qualitative Exploration of Morphological Rivalry

- **Beyond correlation**

✓ Assess functional overlap, semantic nuances, and genre constraints

✓ Investigate cases of overabundance (multiple suffixes coexisting for the same function) and polyfunctionality

→ **CURRENTLY IN PROGRESS**

THANK YOU FOR YOUR ATTENTION ☺

(And special thanks to Richard Huyghe, who taught me most of this)

Silvia Zampetta
Contact: silvia.zampetta01@universitadipavia.it

References

- Aronoff, Mark. 1976. Word formation in generative grammar. Cambridge,
- Mass.: MIT Press.
- Aronoff, Mark & Schvaneveldt, Peter. 1978. Testing Morphological Productivity. Annals of the New York Academy of Sciences: Papers in Anthropology and Linguistics 318. 106-114.
- Baayen, R. Harald. 1992. Quantitative aspects of morphological productivi- ty. In Booij, Geert & Van Marle, Jaap (a cura di), Yearbook of Morphology 1991, 109-149. Dordrecht: Kluwer Academic Publishers.
- Baayen, R. Harald. 1993. On frequency, transparency, and productivity. In Booij, Geert & van Marle, Jaap (a cura di), Yearbook of Morphology 1992, 181-208. Dordrecht: Kluwer Academic Publishers.
- Baayen, R. Harald. 2001. Word Frequency Distributions. Dordrecht: Kluwer.
- Baayen, R. Harald. 2009. Corpus linguistics in morphology: Morphological productivity. In Lüdeling, Anke & Kytö, Merja (a cura di), Corpus Linguistics. An International Handbook. Vol. 2, 899-919. Berlin: Mouton de Gruyter.
- Baayen, R. Harald & Renouf, Antoinette. 1996. Chronicling The Times: Productive lexical innovations in an English newspaper. Language 72. 69-96.
- Bauer, Laurie. 2001. Morphological Productivity. Cambridge: Cambridge University Press.
- Bauer, Laurie. 2005. Productivity: theories. In Štekauer, Pavol & Lieber, Rochelle (a cura di), Handbook of word-formation, 315-334. Dordrecht: Springer.
- Corbin, Danielle. 1987. Morphologie dérivationnelle et structuration du lexique, vol. 1. Tübingen: Niemeyer.

References

- Evert, Stephanie. 2004. A simple LNRE model for random character sequences. In Proceedings of the 7èmes Journées Internationales d'Analyse Statistique des Données Textuelles (JADT 2004), Louvain-la-Neuve, Belgium, 411- 422.
- Evert, Stephanie & Baroni, Marco. 2007. zipfR: Word frequency distributions in R. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Posters and Demonstrations Session, Prague, 29-32.
- Gaeta, Livio & Ricca, Davide. 2002. Corpora testuali e produttività morfolo- gica: i nomi d'azione italiani in due annate della Stampa (1996-1997). In Bauer, Roland & Goebl, Hans (a cura di), Parallela IX. Testo-variazione- informatica/Text-Variation-Informatik. Atti del IX Incontro italoaustria- co dei linguisti, Salzburg, 1-4 novembre 2000, 223-249. Wilhelmsfeld: Egert.
- Gaeta, Livio & Ricca, Davide. 2006. Productivity in Italian word formation: A variable-corpus approach. Linguistics 44(1). 57-89.
- Plag, Ingo. 1999. Morphological Productivity. Structural Constraints in English Derivation. Berlin/New York: Mouton de Gruyter.
- Plag, Ingo. 2003. Word-formation in English. Cambridge: Cambridge University Press.
- Plag, Ingo. 2006. Productivity. In Aarts, Bas & McMahon, April M.S. (a cura di), The Handbook of English Linguistics, 537-556. Malden, MA: Blackwell.
- Rainer, Franz. 2005. Constraints on productivity. In Štekauer, Pavol & Lieber, Rochelle (a cura di), Handbook of word-formation, 335-352. Dordrecht: Springer.
- Štichauer, Pavel. 2009. Morphological productivity in diachrony: The case of the deverbal nouns in -mento, -zione and -gione in Old Italian from the 13th to the 16th century. In Montermini, Fabio & Boyé, Gilles & Tseng, Jesse (a cura di), Selected Proceedings of the 6th Décembrettes, 138-147. Somerville, MA: Cascadilla Proceedings Project.
- Thornton, Anna Maria. 1988. Sui nomina actionis in italiano, Pisa: Università di Pisa (Tesi di dottorato).
- Thornton, Anna Maria. 2005. Morfologia. Roma: Carocci.
- Zampetta, Silvia. *Accepted*. Assessing Morphological Productivity in a Corpus Language: A Diachronic Study of Ancient Greek Deverbal Nominal Suffixes. Description and application in a productivity study. Proceedings of the 5th Int. Workshop on Resources and Tools for Derivational Morphology (DeriMo 2025).