Quantifying diachronic productivity

The example of Early New High German derivational morphology

Elizabeth Pankratz

June 12, 2019

Humboldt-Universität zu Berlin

Intuitions about derivational morphology

Illustration with Modern English *-ness* and *-ity*:

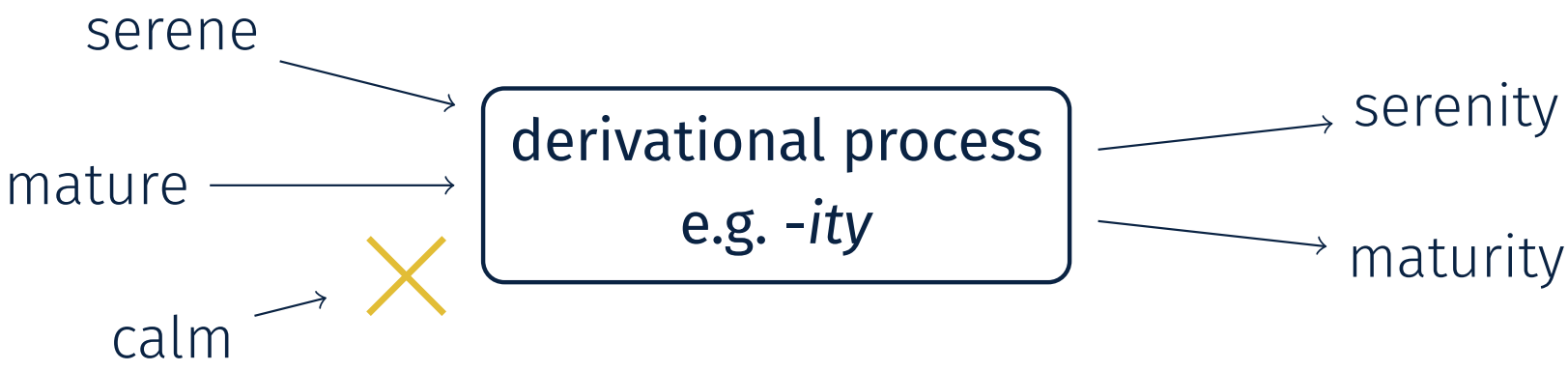
adjective	+	*-ness* *-ity*	→	noun

But, *-ness* and *-ity* aren't equally good for all adjectives:

calm	+	*-ness*	→	*calmness*
calm	+	*-ity*	→	**calmity*
serene	+	*-ness*	→	*?sereneness*
serene	+	*-ity*	→	*serenity*

"Though many things are possible in morphology, some are more possible than others" (Aronoff 1976: 35).

Describing derivational processes from two sides



What is possible as input?

→ selectional restrictions

→ qualitative

What exists as output?

→ actual usage

→ quantitative

(Plag 1999, Bauer 2001)

My project

My original goal: Using productivity measures from the literature, quantify how the productivity of three Early New High German (ENHG) nominalisers in the RIDGES Herbology corpus changed over time.

My project: The suffixes

Suffix	Examples	Tokens in RIDGES
-er	*Schreiber* ‘writer’ *Italiener* ‘person from Italy’ *Weggänger* ‘path-walker’	277
-heit/-keit	*Gesundheit* ‘health’ *Heimlichkeit* ‘secrecy’ *Reinigkeit* ‘cleanliness’	652
-ung	*Mischung* ‘mixture’ *Verstopfung* ‘blockage’ *Machung* ‘making’	1992

My project: Why do we care?

The ENHG era (approx. 1350–1650) was a period of transition from Latin to German in the sciences.

(Hartweg & Wegera 2005, Klein 2010, Odebrecht et al. 2017)

When a language needs to become functional in a new register, **new words need to be coined**: ideally, we can see productive use of derivational morphology in action.

My project: Why RIDGES?

We can track this register development especially well in botanical texts.

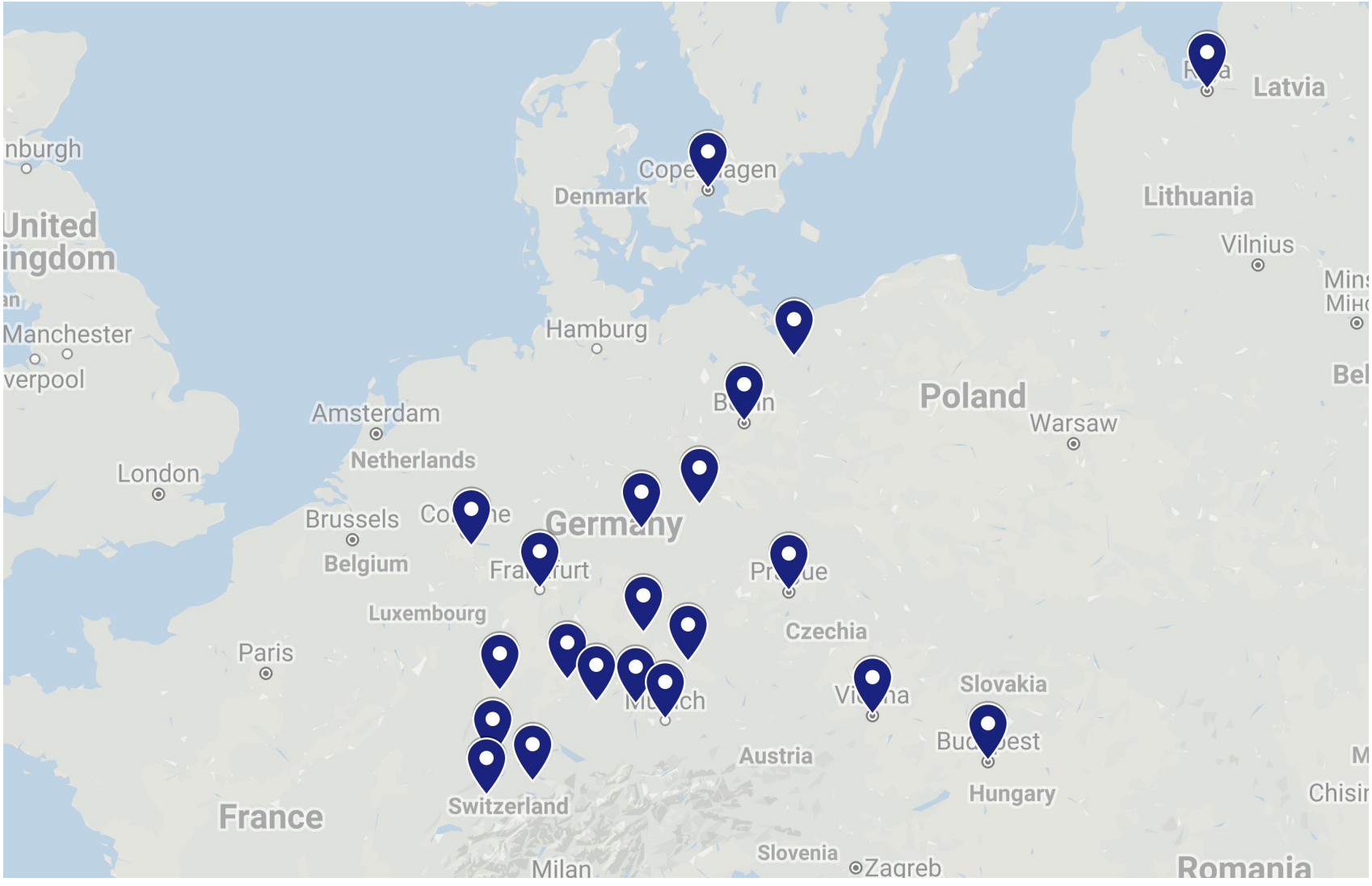
These were available in German earlier than many other scientific text genres.

(Klein 2010)

RIDGES (V.8) contains 257,537 tokens in 61 botanical texts from 1482–1914.

(Lüdeling, Odebrecht, et al. n.d., Odebrecht et al. 2017)

My project: Provenance of ENHG botanical texts



(Map accessible under `http://tiny.cc/ridges_map`)

My project: Methodology

Standard method for diachronic productivity studies:

1. Divide up overall time period into smaller subperiods,
2. calculate productivity measures for data in each subcorpus, and
3. compare the results.

(Scherer 2005, Schneider-Wiejowski 2011, Hartmann 2016, Kempf 2016, Cowie 1999)

This requires that the productivity measures **always produce comparable results** across subcorpora … which they don't do.

My project: The new plan

So, my goals now (and the plan for the rest of this talk):

- Explore the behaviour of the most widely-used measures to figure out where the problems are
  1. Type counts
  2. Hapax legomena and potential productivity
  3. LNRE models
- End up with several desiderata for what a reliable measure of diachronic productivity should do

Type counts

Background

Background: Type counts

Tokens, *N*: all instances of words in a corpus or category

Types, *V*: all unique words in that corpus or category

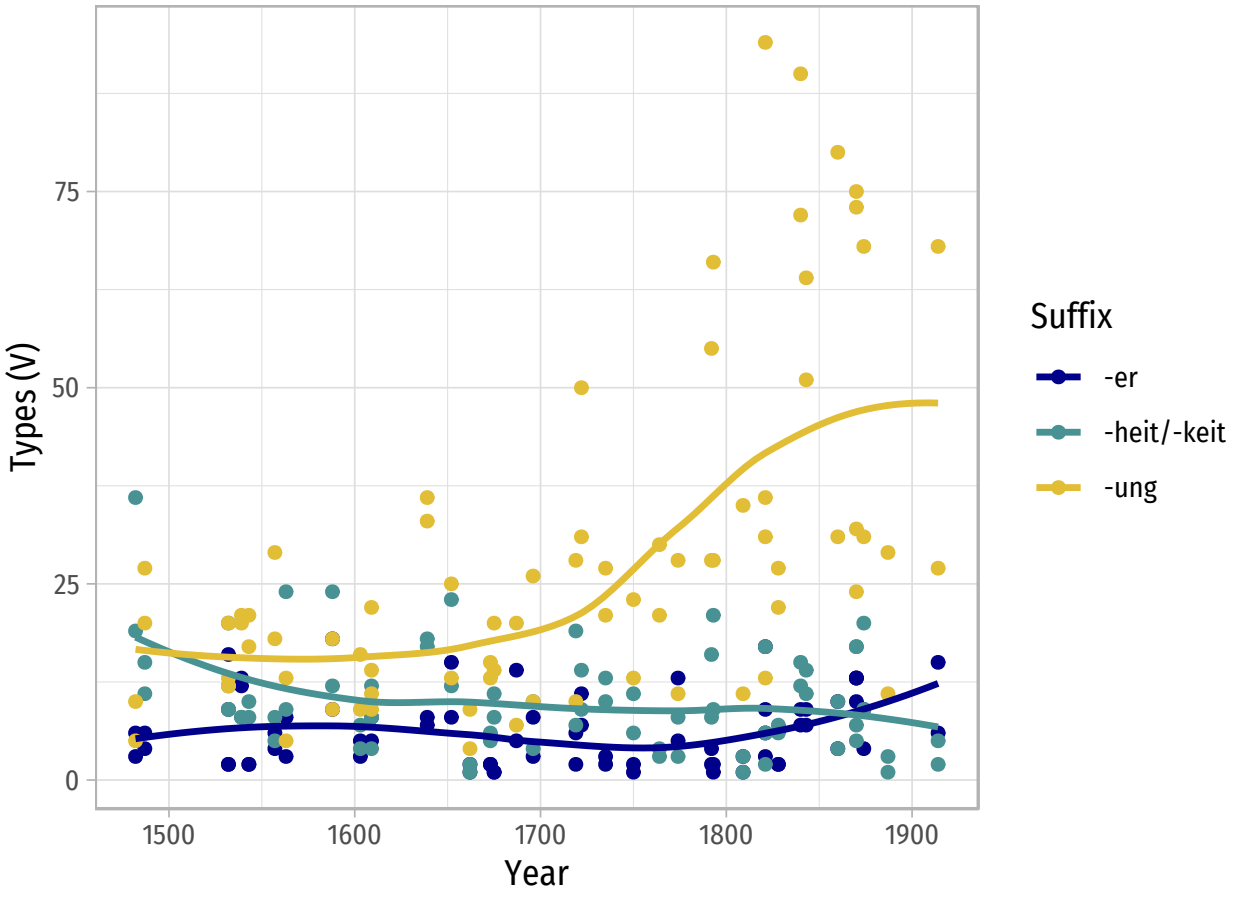
e.g.

	The	quick	brown	fox	jumps	over	the	lazy	dog
N	1	2	3	4	5	6	7	8	9
V	1	2	3	4	5	6	6	7	8

Background: Type counts

Type counts ≈ past productivity of a morphological process

(Baayen 2009, Bauer 2001, Cowie 1999)



Type counts

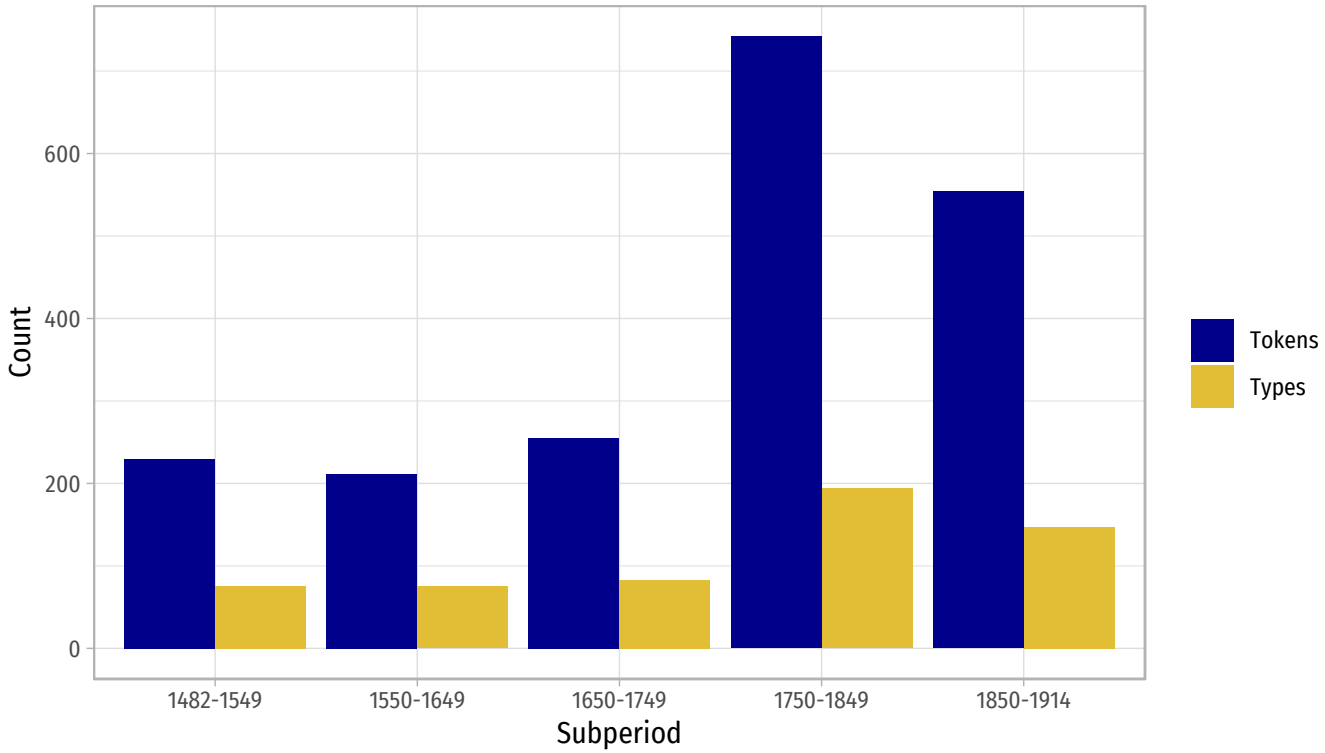
Validity

Validity: Type counts

Problem: An increase in type count could be due to an actual increase in usage, or just to a larger subcorpus, which contains more tokens and thus potentially more types.

Validity: Type counts (shown for *-ung*)

More **tokens** → more types.



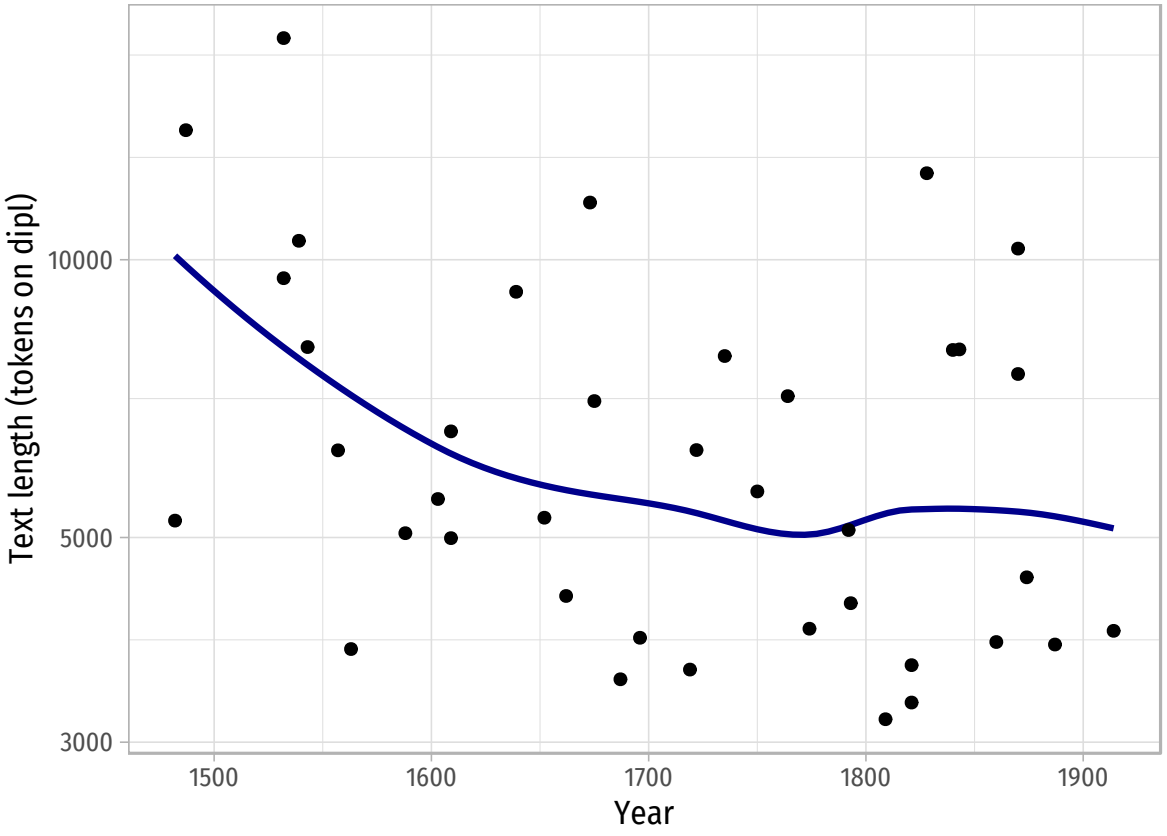
Type counts depend on the number of tokens they come out of.

(Zeldes 2012)

No way to be sure if type/token increase reflects increased usage without considering the size of the original (sub)corpus.

Validity: Type counts

In RIDGES, text length decreases over time.

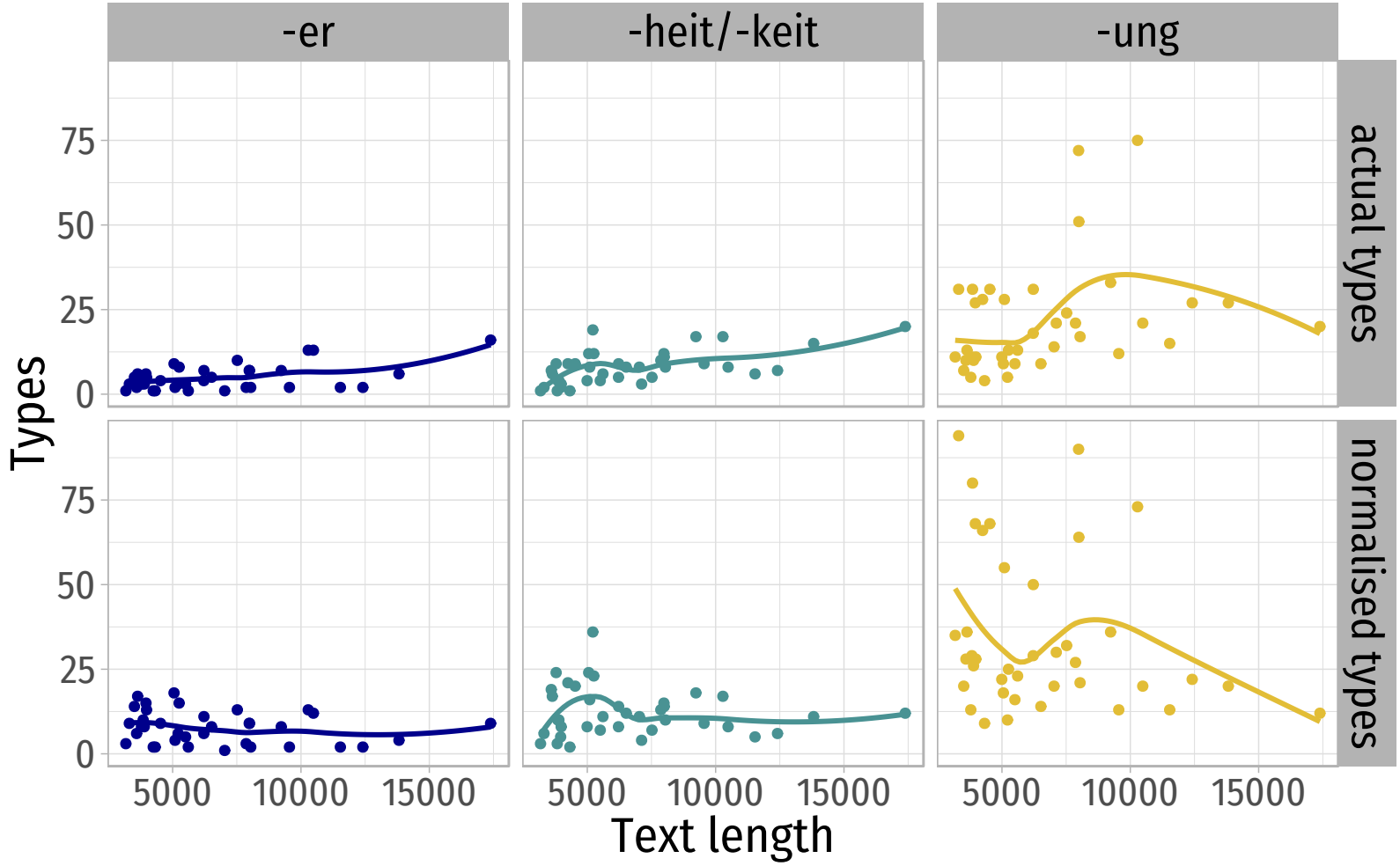


So, the higher frequency of -*ung* in later years is not an artefact of text size … but still can't quantify/compare it exactly.

Normalisation as a solution?

Possible solution: Normalisation to types per e.g. 10,000 words.

(Scherer 2005, Cowie & Dalton-Puffer 2002)



Normalisation as a solution?

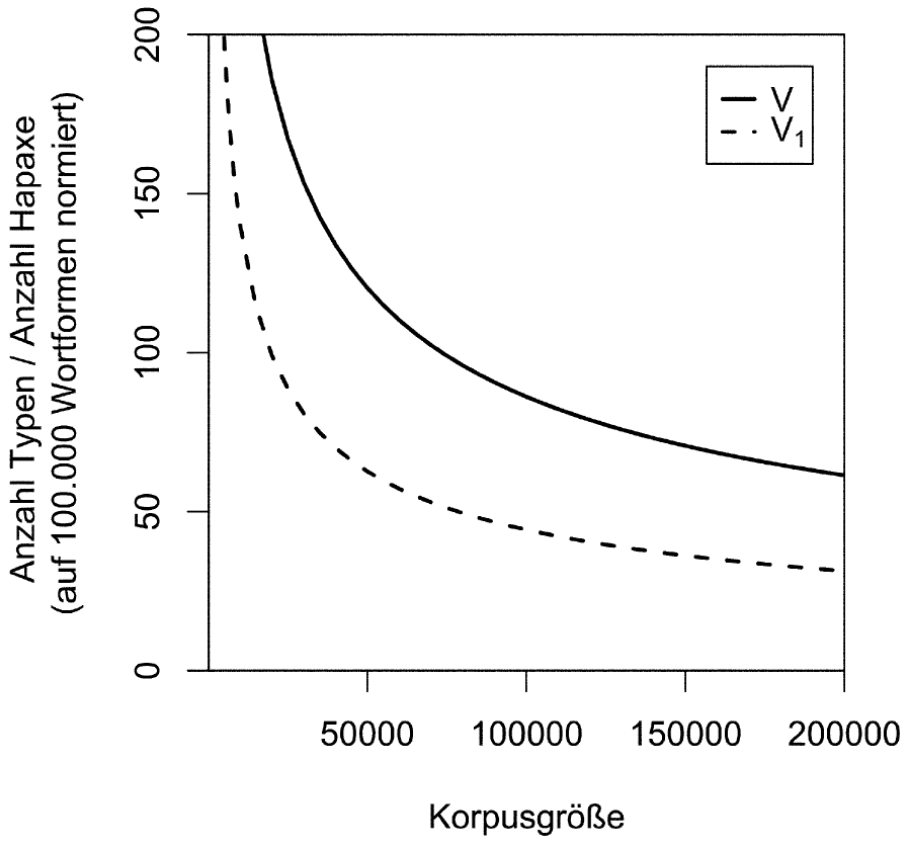
$$\frac{V_n}{K_n} = \frac{V_a}{K_a}$$

$$V_n = \frac{V_a \times K_n}{K_a}$$

V_n	normalised type count
V_a	actual type count
K_n	normalised corpus size
K_a	actual corpus size

The problem with normalisation

Problem: The normalised value is still dependent on the original corpus size.



$$V_n = \frac{V_a \times K_n}{K_a}$$

$$\lim_{K_a \to \infty} \frac{V_a \times K_n}{K_a} = 0$$

(Figure from Lüdeling 2009: 337, Fig. 1)

Recap

Measure	Unsatisfying because...
Type counts	increase with increasing sample size
Normalised counts	decrease with increasing original corpus size

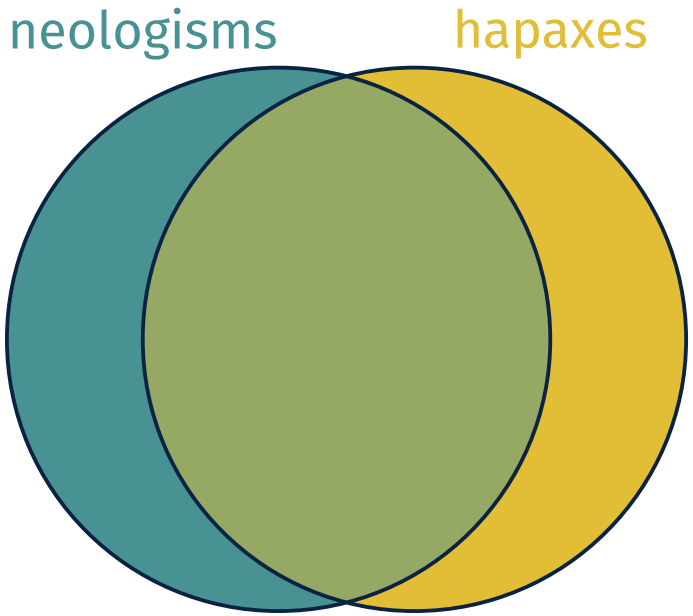
Hapax legomena and potential productivity

Background

Background: Hapax legomena

Hapax legomena (words appearing once) ≈ newly coined words

(Baayen 2001, 2009)



We care about newly coined words because these are the ones that we know have been productively formed.

Background: Potential productivity $\mathscr{P}$

$$\mathscr{P} = \frac{\text{number of hapax legomena}}{\text{number of tokens}}$$

The idea: $\mathscr{P}$ quantifies the potential of a morphological process to coin a new word (intuitive as a probability).

Interpretation: Values range between 0 (totally unproductive, no hapaxes) and 1 (totally productive, only hapaxes).

(Baroni 2009, Baayen 2001, 2009)

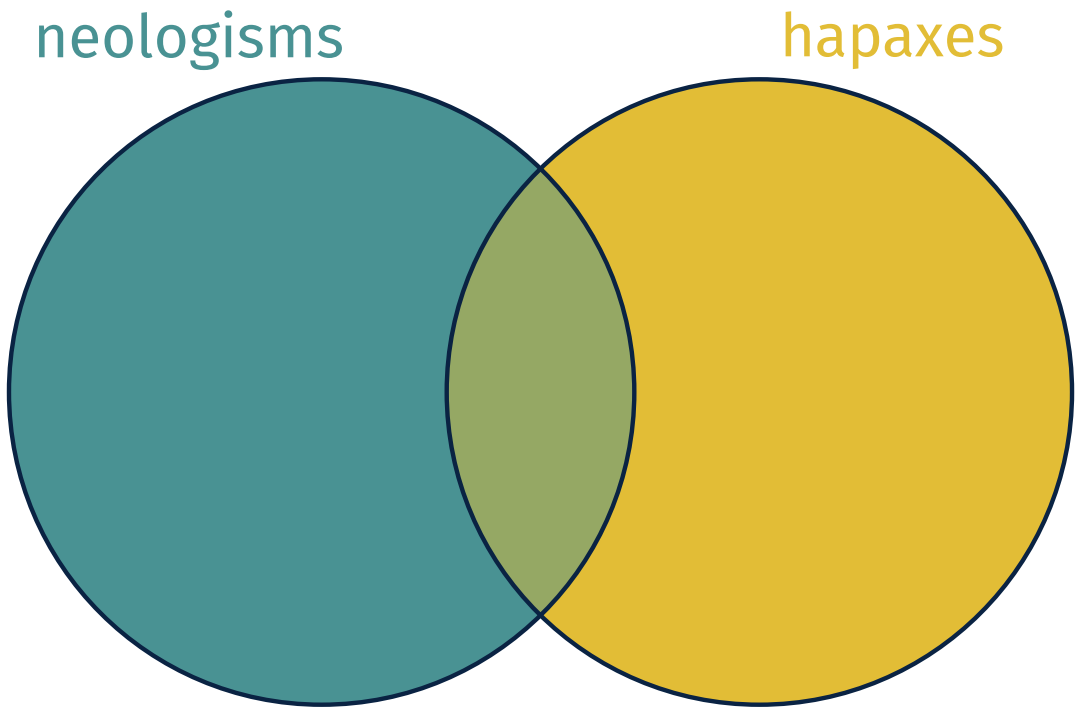
Hapax legomena and potential productivity

Validity

Validity: Hapax legomena

In small corpora, the conceptual link between neologisms and hapax legomena is more tenuous.

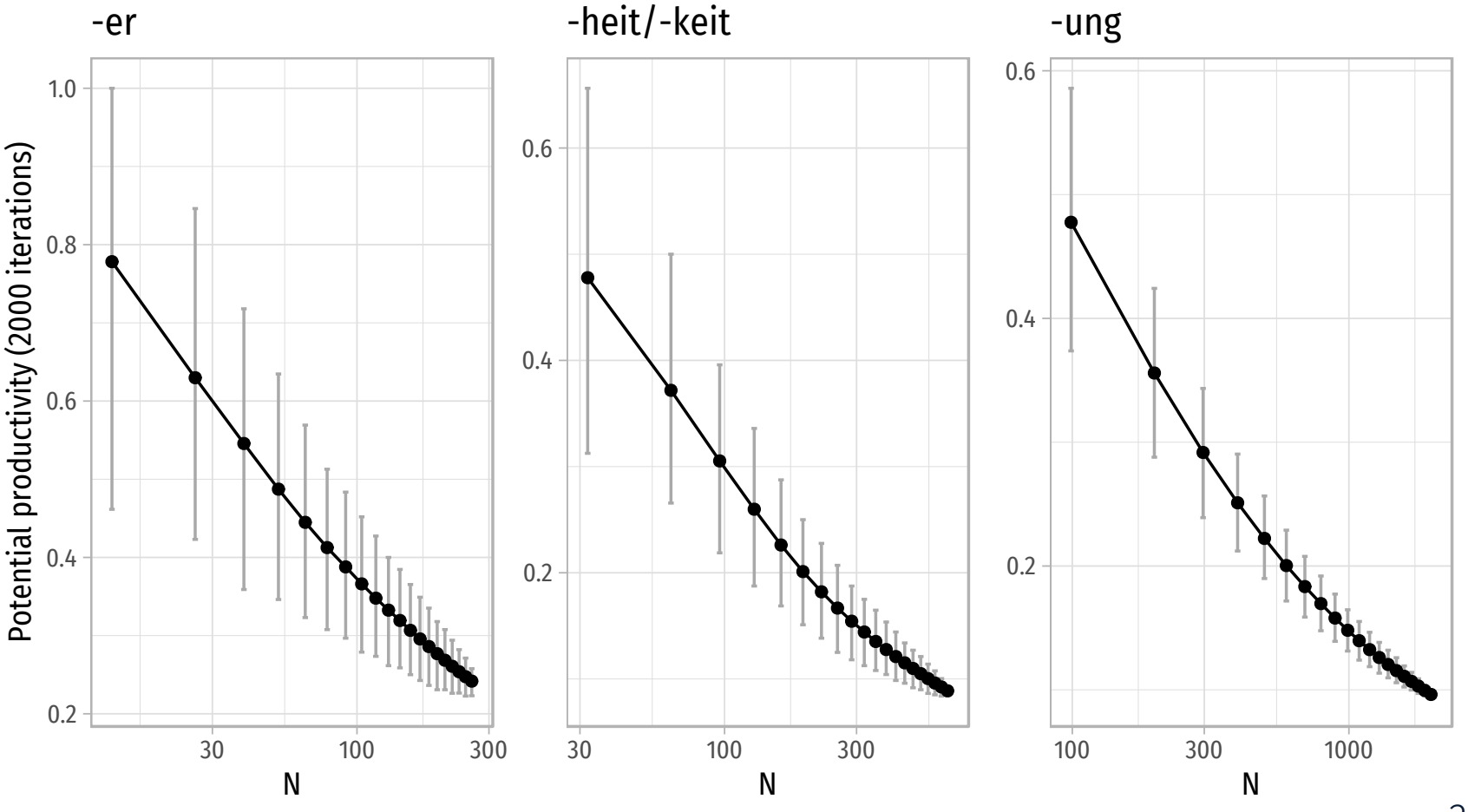
(Cowie 1999, Cowie & Dalton-Puffer 2002)



Validity: Potential productivity à la Tweedie & Baayen (1998)

Problem for comparability: $\mathscr{P}$ is dependent on token count *N*.

(Hartmann 2016, Lüdeling 2009)



Recap

Measure	Unsatisfying because...
Type counts	increase with increasing sample size
Normalised counts	decrease with increasing original corpus size
Hapax legomena	**only correspond loosely to neologisms**
Potential productivity	**decreases with increasing sample size**

LNRE models

Background

Background: Vocabulary growth curves (VGCs)

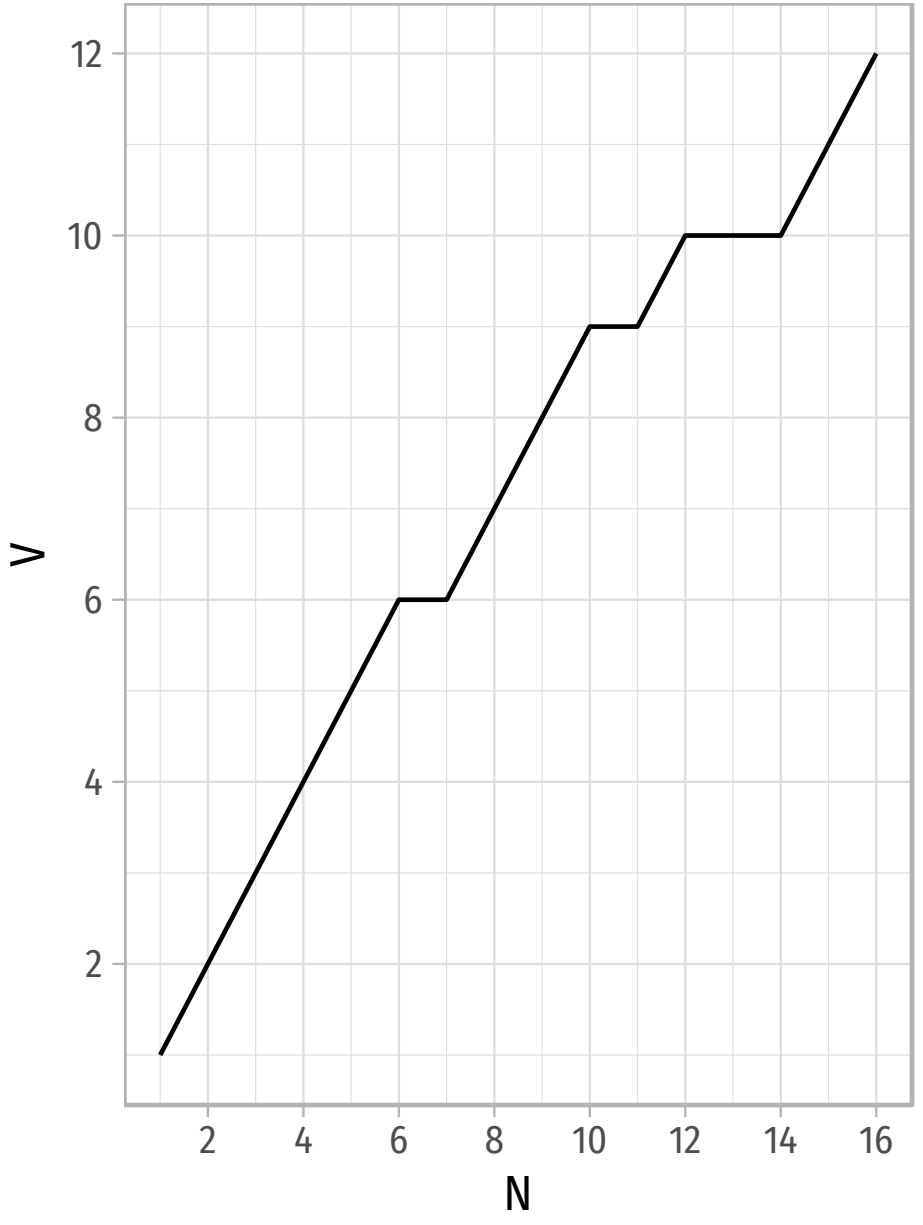
Vocabulary growth curves (VGCs) track how the type-token ratio develops as one moves token by token through a sample.

	The	quick	brown	fox	jumps	over	the	lazy	dog
N	1	2	3	4	5	6	7	8	9
V	1	2	3	4	5	6	6	7	8

	and	the	slow	brown	fox	does	too
N	10	11	12	13	14	15	16
V	9	9	10	10	10	11	12

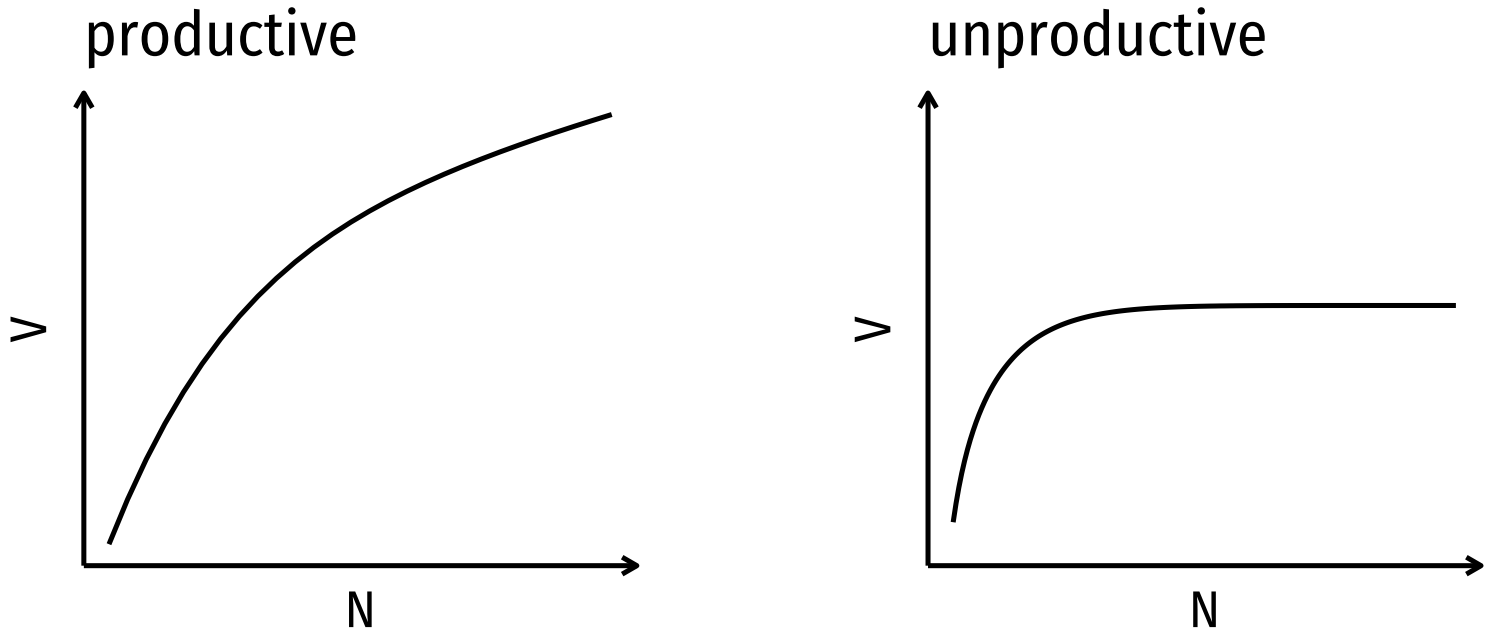
Background: Vocabulary growth curves (VGCs)

N	V
1	1
2	2
3	3
4	4
5	5
6	6
7	6
8	7
9	8
10	9
11	9
12	10
13	10
14	10
15	11
16	12



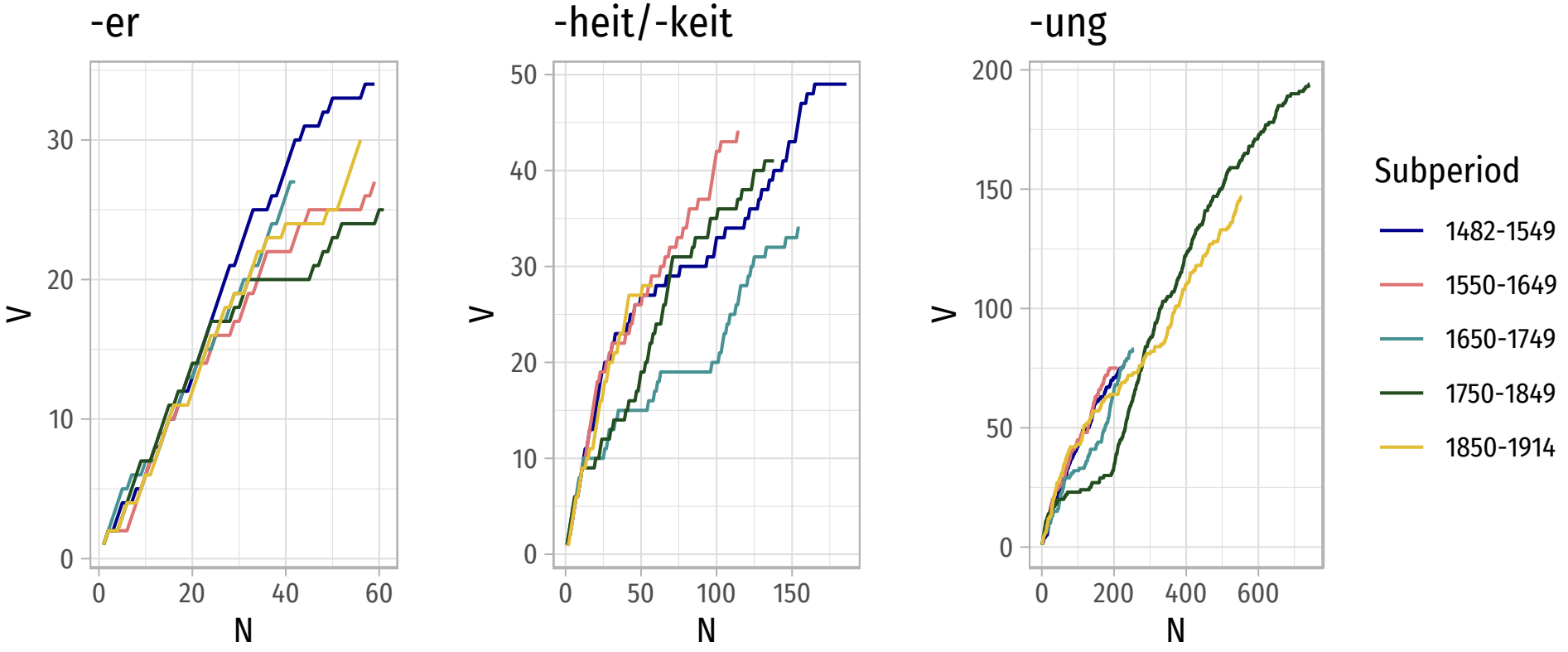
Background: Vocabulary growth curves (VGCs)

For samples of derivational processes:



(Lüdeling, Evert & Heid 2000, Evert & Lüdeling 2001)

Background: Vocabulary growth curves (VGCs)



But, still not comparable – where are we in the larger curve?

Background: LNRE models

Possible solution: Extrapolate data beyond known sample sizes to create a VGC up to any value of N.

Large Number of Rare Events models are designed to do this.

(Baayen 2001)

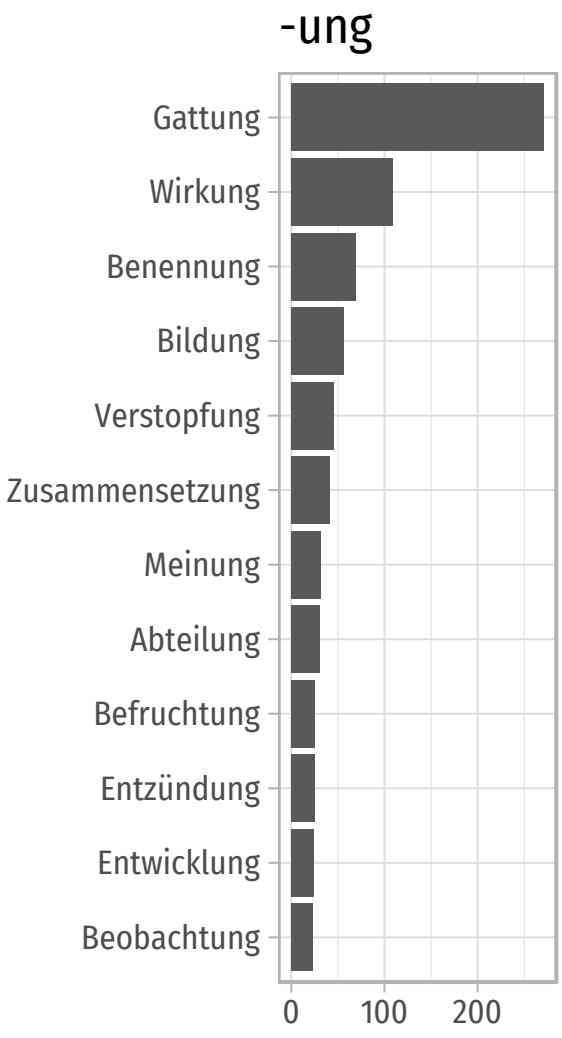
LNRE modelling can be done in R with the `zipfR` package.

(R Core Team 2014, Evert & Baroni 2007, Baroni & Evert 2014)

I will test out the **finite Zipf-Mandelbrot (fZM)** model.

(Evert 2004, Zipf 1949, Mandelbrot 1953)

Aside: Why “Large Number of Rare Events”?



Zipf’s Law:

$$f_z = \frac{C}{z}$$

- z frequency rank
- f_z frequency of word with rank z
- C a constant

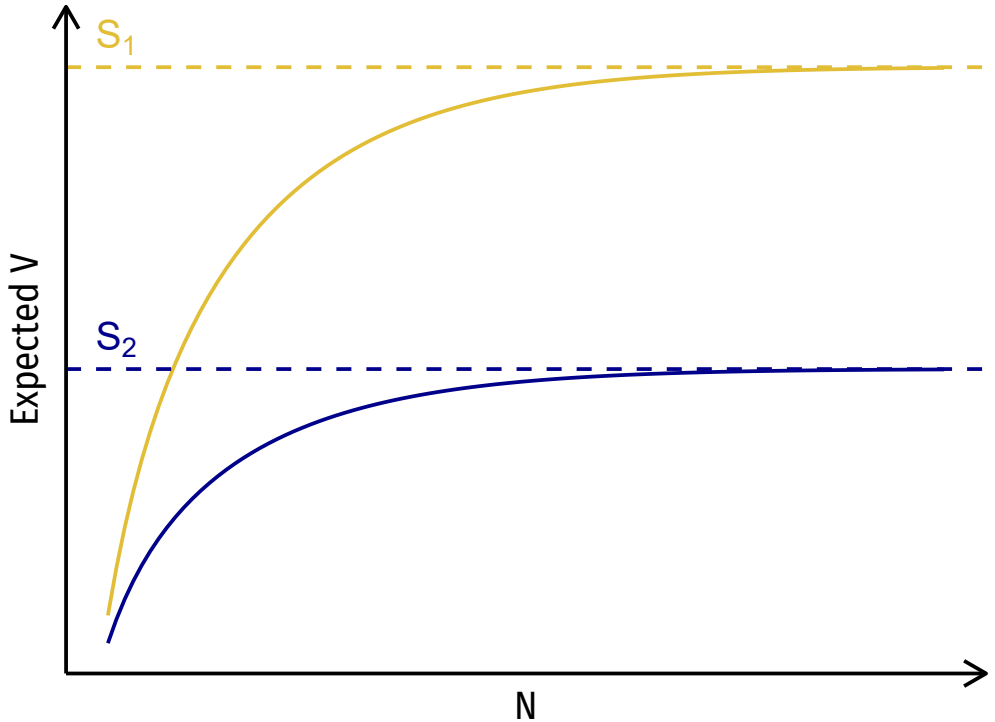
(Zipf 1949)

Background: LNRE models

The idea: The fZM model calculates S, the maximum number of types that the derivational process would form.

Interpretation: A greater S means greater productivity.

(Evert 2004, Zeldes 2012)

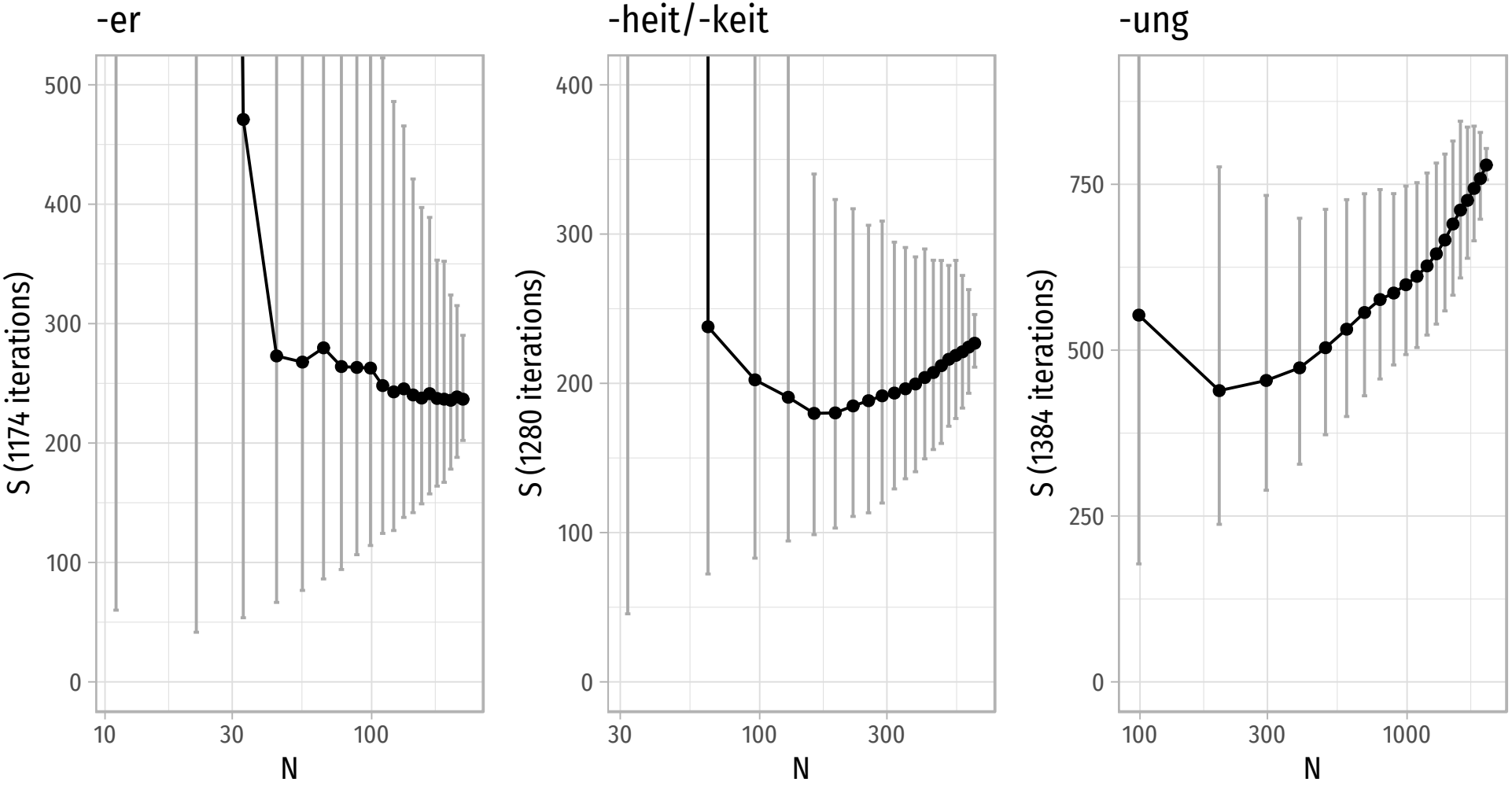


LNRE models

Validity

Validity: LNRE models à la Tweedie & Baayen (1998)

Problem: *S* might not actually be a constant.



Recap

Measure	Unsatisfying because...
Type counts	increase with increasing sample size
Normalised counts	decrease with increasing original corpus size
Hapax legomena	only correspond loosely to neologisms
Potential productivity	decreases with increasing sample size
Observed VGCs	**cannot be compared directly**
S	**increases(?) with increasing sample size**

Recap

Measure	Unsatisfying because...
Type counts	increase with increasing sample size
Normalised counts	decrease with increasing original corpus size
Hapax legomena	only correspond loosely to neologisms
Potential productivity	decreases with increasing sample size
Observed VGCs	**cannot be compared directly**
S	**increases(?) with increasing sample size**

→ **For analyses to work diachronically, token counts must stay constant in samples across time.**

Ways forward

Keeping tokens constant?

Keeping tokens constant: Two ways

	Subperiods?	Data?
Cut subcorpora down to max. common size	firm	some discarded
Make several equally-sized subcorpora	flexible	all maintained

(Dal & Namer 2016, Gaeta & Ricca 2003)

Possible solution: Sliding windows of *x* tokens?

Keeping tokens constant: Sliding windows

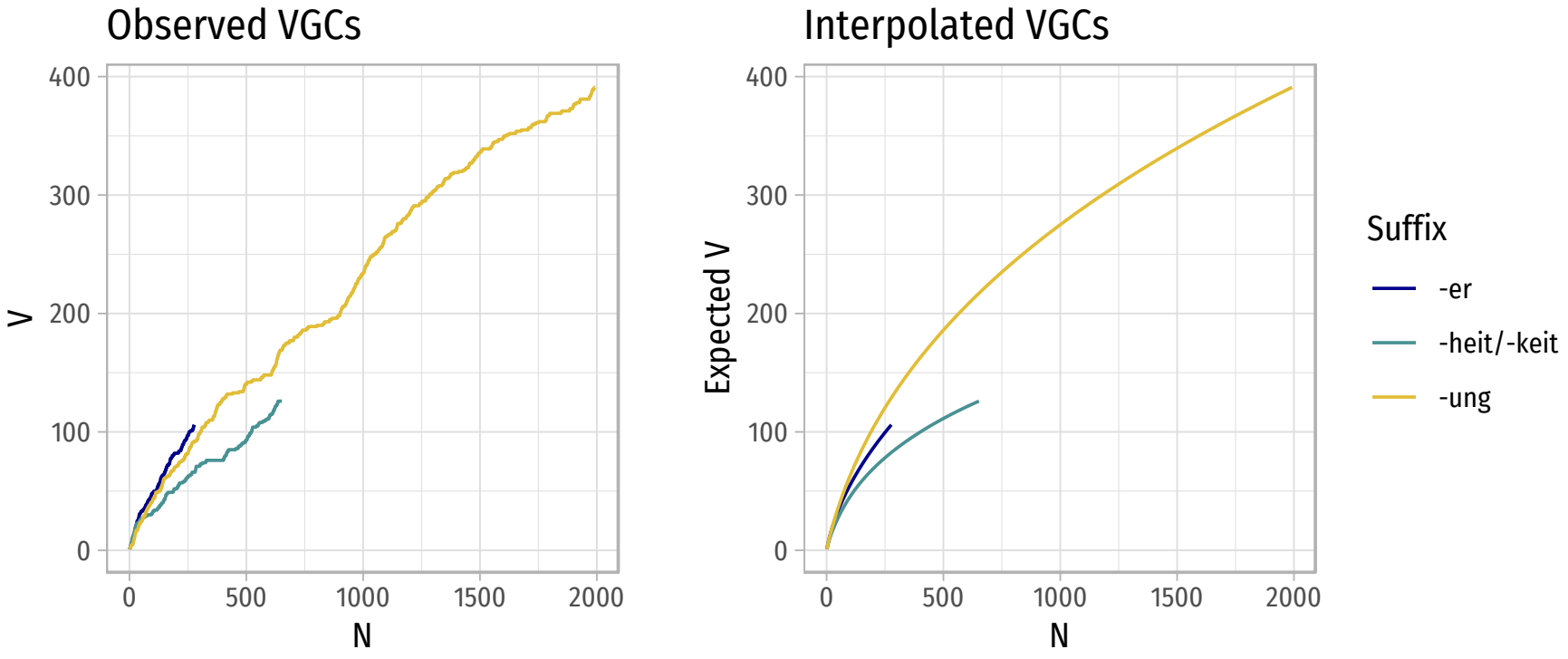
Sliding windows would probably work, **if** we had a reliable measure to apply to those windows.

Type counts' dependency on original sample size is not remedied by sliding windows.

And even beyond their dependency on token counts, $\mathscr{P}$ and S have deeper problems.

The deeper problem with $\mathscr{P}$

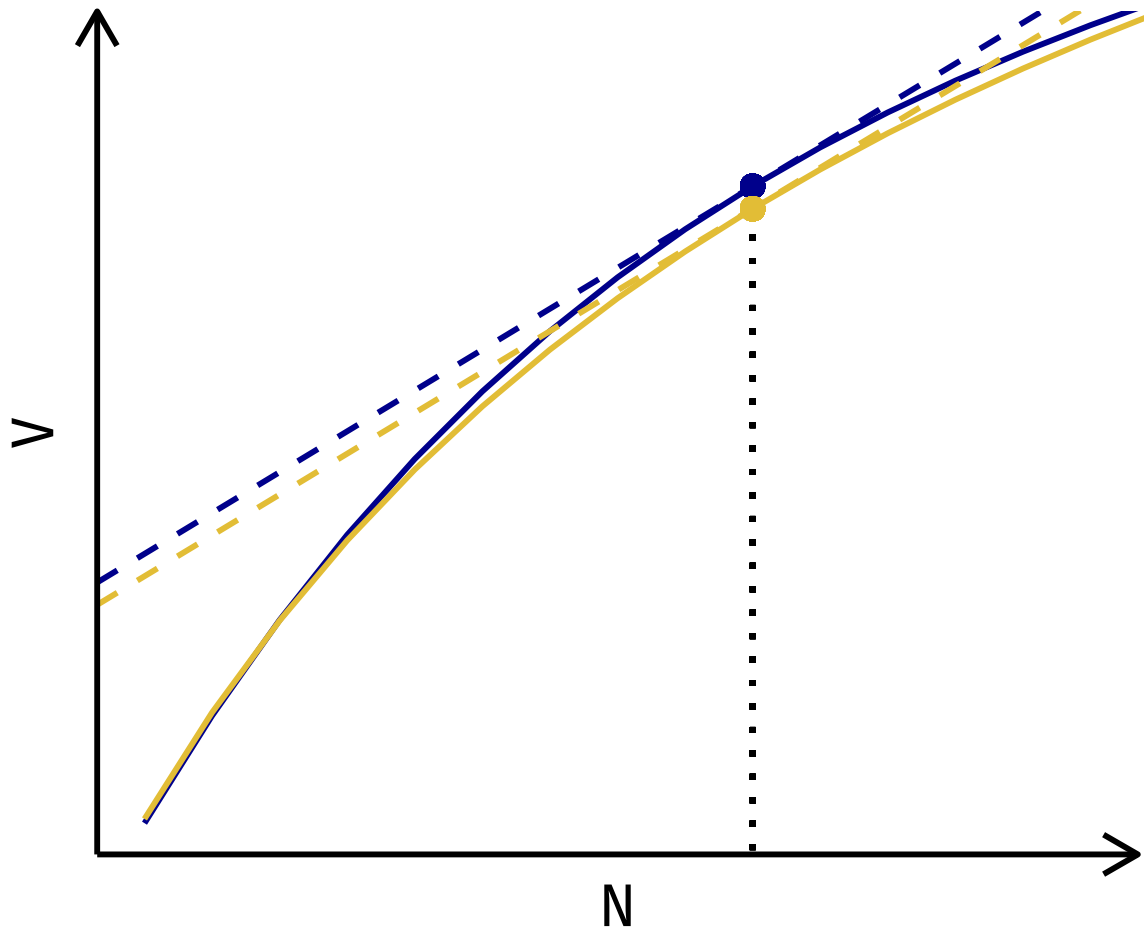
$\mathscr{P}$ is the slope of the interpolated VGC at its endpoint (at the sample's maximum *N*). (Baayen 2001)



The deeper problem with $\mathscr{P}$

Problem: Two different curves can have the same slope at the same value of N, i.e. the same $\mathscr{P}$ for different curves.

So $\mathscr{P}$ actually tells us very little about the curve we're on!



The deeper problem with LNRE models

Problem: For mathematical convenience, LNRE models assume that word choice is random (bag of words/marbles from urn).

But this is not how language works.

→ Word choice for position x is fundamentally informed by words $x - 1$, $x - 2$, ...

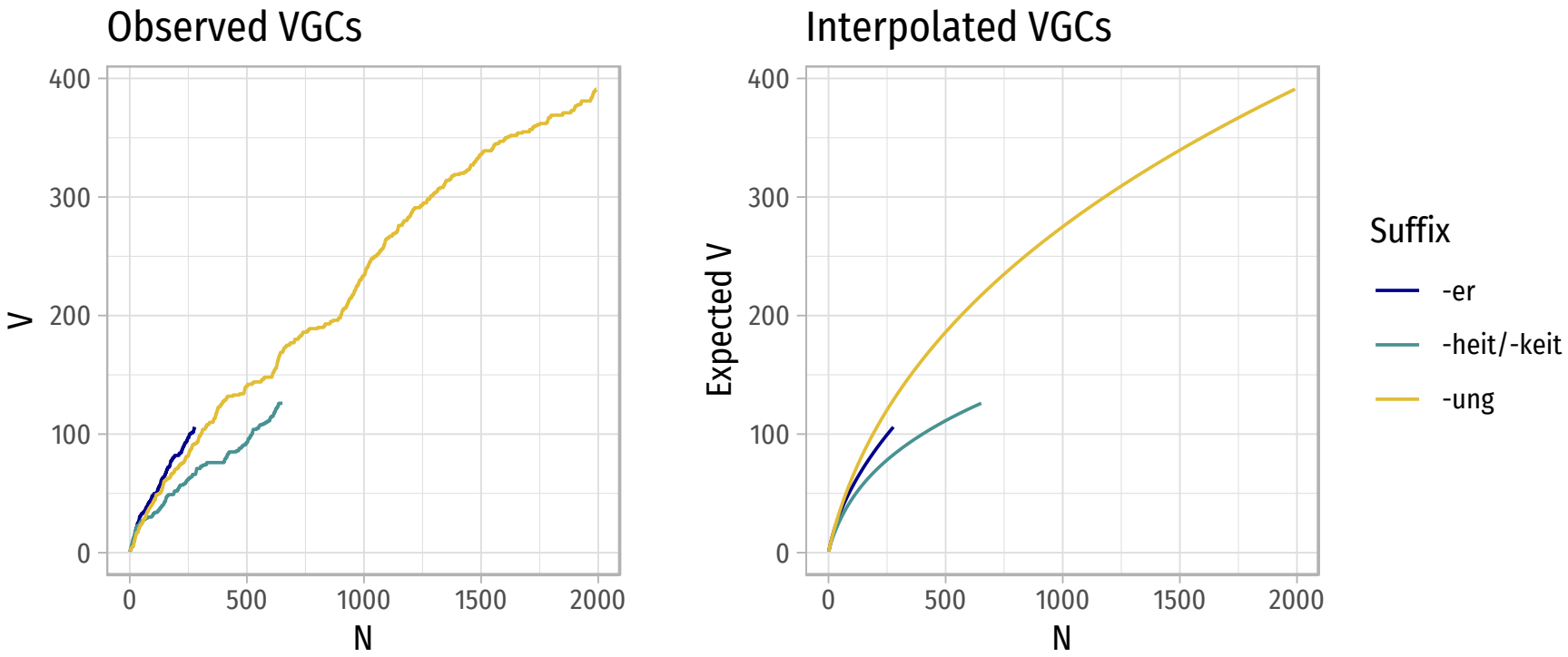
→ Rare words cluster together ("underdispersion")

(Mandelbrot 1961, Zeldes 2012, Baayen 2001, Evert 2004)

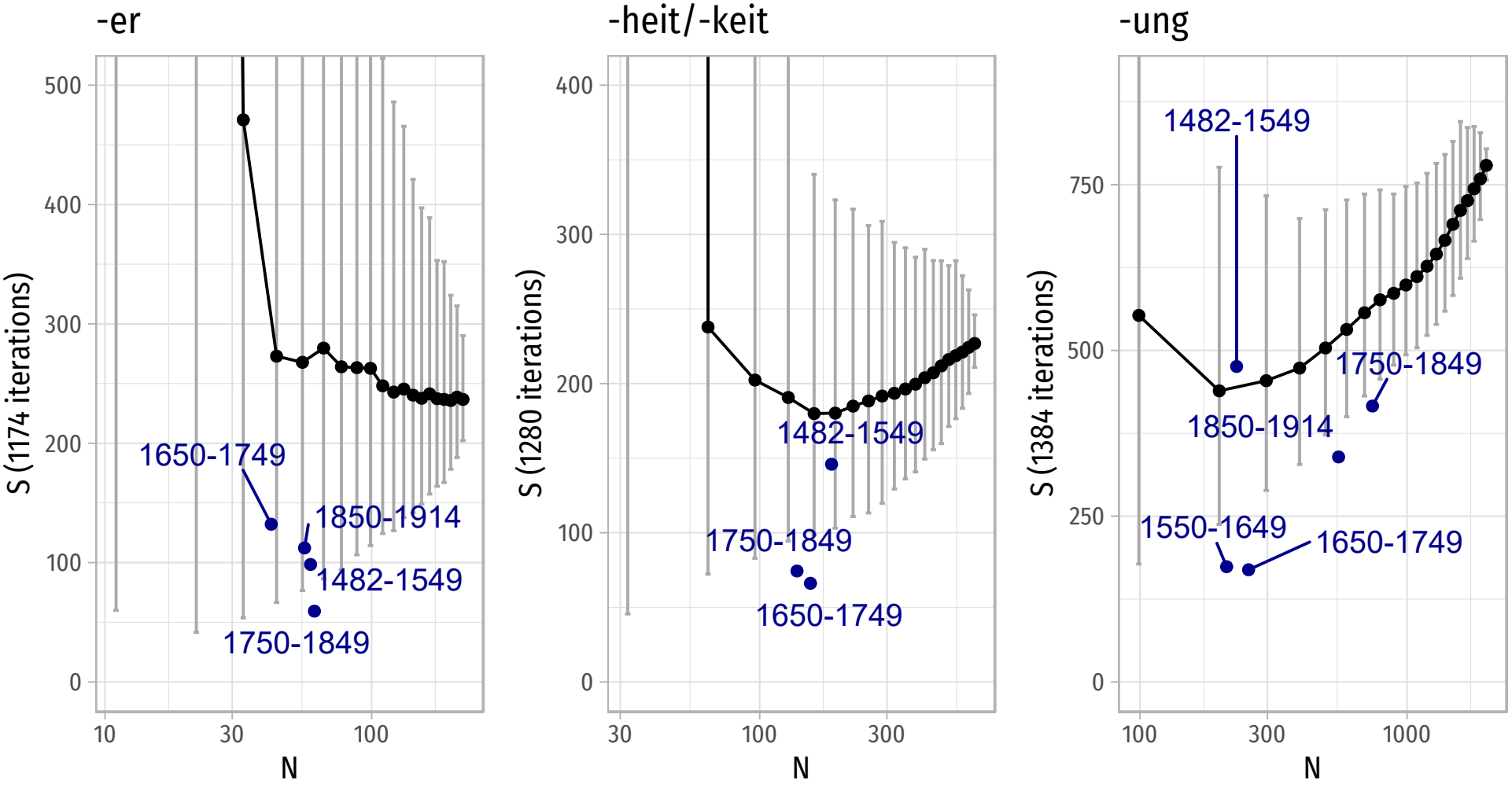
The deeper problem with LNRE models

Randomness assumption in model
Underdispersion in data

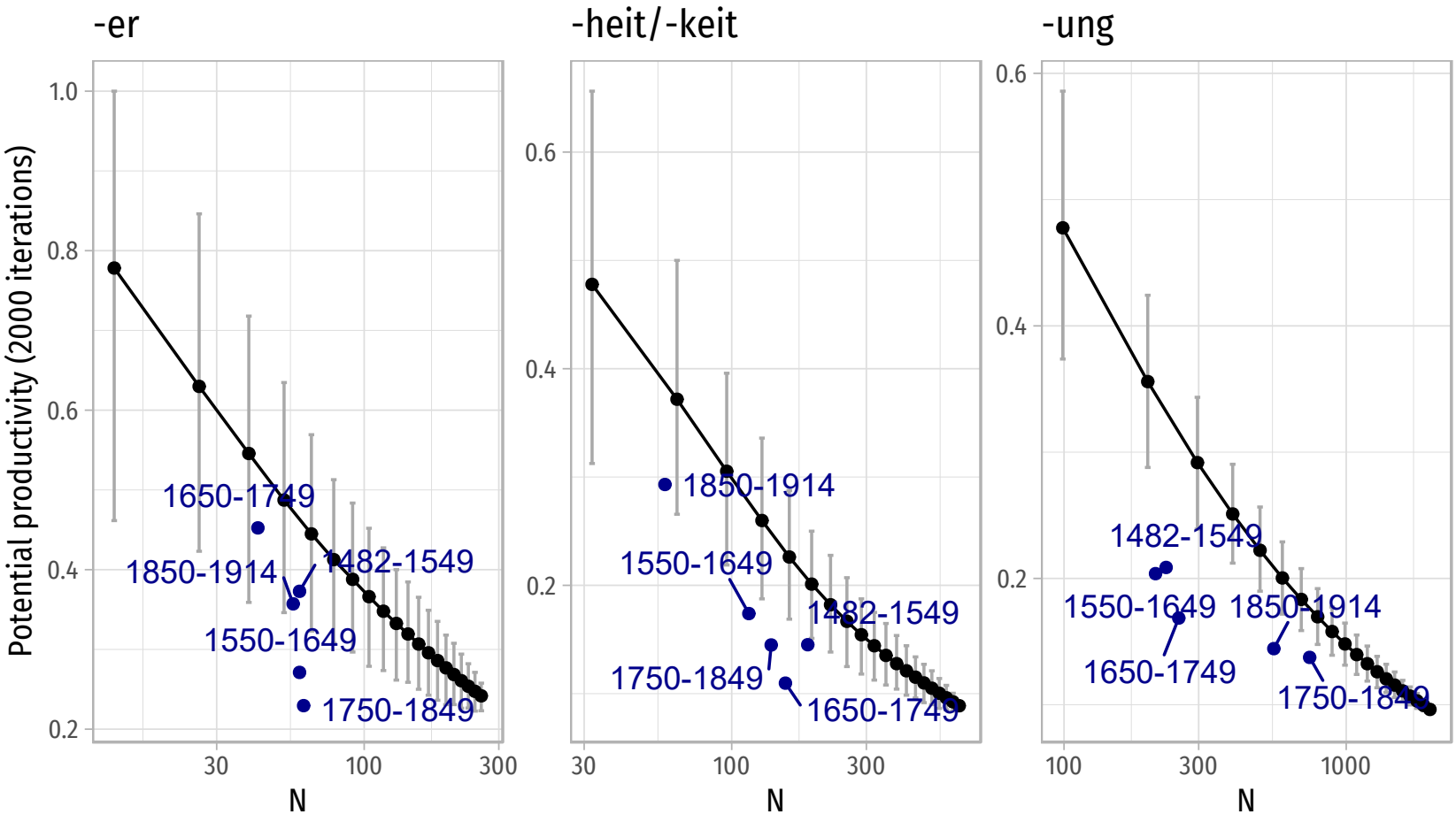
→ Model overestimates *V*



S for randomised vs. natural language data



$\mathscr{P}$ for randomised vs. natural language data



Final recap

Measure	Unsatisfying because...
Type counts	increase with increasing sample size
Normalised counts	decrease with increasing original corpus size
Hapax legomena	only correspond loosely to neologisms
Potential productivity	decreases with increasing sample size, and **can be the same for different curves**
Observed VGCs	cannot be compared directly
S	increases(?) with increasing sample size, and **overestimates *V* because of randomness assumption**

Ways forward

Desiderata

Desiderata

A reliable measure of diachronic productivity must, above all else, be **comparable**.

Ideally, it should:

1. account for original text length,
2. be constant over all sample sizes, and
3. model actual natural language use.

Thank you!

These slides, my data, and my R scripts will be available at:

`github.com/epankratz/diachronic-productivity-thesis`

References i

Aronoff, Mark. 1976. *Word formation in generative grammar*. Cambridge, MA: MIT Press.

Baayen, R. Harald. 2001. *Word frequency distributions*. Dordrecht/Boston/London: Kluwer Academic Publishers.

Baayen, R. Harald. 2009. Corpus linguistics in morphology: morphological productivity. In Anke Lüdeling & Merja Kytö (eds.), *Corpus linguistics. An international handbook*, 900–919. Berlin: De Gruyter.

Baroni, Marco. 2009. Distributions in text. In Anke Lüdeling & Merja Kytö (eds.), *Corpus linguistics. An international handbook*, 803–822. Berlin: De Gruyter.

Baroni, Marco & Stefan Evert. 2014. *The zipfR package for lexical statistics: A tutorial introduction*. Available from `http://zipfr.r-forge.r-project.org/`.

Bauer, Laurie. 2001. *Morphological productivity*. Cambridge/New York/Melbourne: Cambridge University Press.

Cowie, Claire. 1999. *Diachronic word-formation: A corpus-based study of derived nominalizations in the history of English*. University of Cambridge dissertation.

Cowie, Claire & Christiane Dalton-Puffer. 2002. Diachronic word-formation and studying changes in productivity over time: Theoretical and methodological considerations. In Javier E. Díaz Vera (ed.), *A changing world of words: Studies in English historical lexicography, lexicology and semantics*, 410–437. Amsterdam: Rodopi.

Dal, Georgette & Fiammetta Namer. 2016. Productivity. In Andrew Hippisley & Gregory T. Stump (eds.), *The Cambridge Handbook of Morphology* (Cambridge Handbooks in Language and Linguistics), 70–89. Cambridge: Cambridge University Press.

Evert, Stefan. 2004. A simple LNRE model for random character sequences. In *Proceedings of JADT 2004: 7es Journées internationales d'Analyse statistique des Données Textuelles*.

Evert, Stefan & Marco Baroni. 2007. zipfR: word frequency distributions in R. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, 29–32.

Evert, Stefan & Anke Lüdeling. 2001. Measuring morphological productivity: Is automatic preprocessing sufficient? In *Proceedings of Corpus Linguistics 2001*, vol. 168.

References ii

Gaeta, Livio & Davide Ricca. 2003. Italian prefixes and productivity: A quantitative approach. *Acta Linguistica Hungarica*. 89–108.

Hartmann, Stefan. 2016. *Wortbildungswandel: Eine diachronie Studie zu deutschen Nominalisierungsmustern*. (Studia Linguistica Germanica 125). Berlin / Boston: De Gruyter.

Hartweg, Frédéric & Klaus-Peter Wegera. 2005. *Frühneuhochdeutsch. Eine Einführung in die deutsche Sprache des Spätmittelalters und der frühen Neuzeit*. Tübingen: Max Niemeyer Verlag.

Kempf, Luise. 2016. *Adjektivsuffixe in Konkurrenz: Wortbildungswandel vom Frühneuhochdeutschen zum Neuhochdeutschen*. (Studia Linguistica Germanica 126). Berlin / Boston: De Gruyter.

Klein, Wolf Peter. 2010. Die deutsche Sprache in der Gelehrsamkeit der frühen Neuzeit. Von der lingua barbarica zur HaubtSprache. In Herbert Jaumann (ed.), *Diskurse der Gelehrtenkultur in der Frühen Neuzeit. Ein Handbuch*, 465–516. Berlin/New York: De Gruyter.

Lüdeling, Anke. 2009. Carmen Scherer, Wortbildungswandel und Produktivität. *Beiträge zur Geschichte der deutschen Sprache und Literatur (PBB)* 131(2). 333–339.

Lüdeling, Anke, Stefan Evert & Ulrich Heid. 2000. On measuring morphological productivity. In *KONVENS 2000/Sprachkommunikation, Vorträge der gemeinsamen Veranstaltung 5. Konferenz zur Verarbeitung natürlicher Sprache (KONVENS), 6. ITG-Fachtagung Sprachkommunikation*, 57–61. VDE-Verlag GmbH.

Lüdeling, Anke, Carolin Odebrecht, Laura Perlitz & Amir Zeldes. N.d. *RIDGES-Herbology (Version 8.0)*. Humboldt-Universität zu Berlin. `http://korpling.org/ridges/.http://hdl.handle.net/11022/0000-0007-C6A3-1`.

Mandelbrot, Benoit. 1953. An informational theory of the statistical structure of language. *Communication theory* 84. 486–502.

Mandelbrot, Benoit. 1961. On the theory of word frequencies and on related Markovian models of discourse. In Roman Jakobson (ed.), *Structure of language and its mathematical aspects*, 190–219. Providence, Rhode Island: American Mathematical Society.

References iii

Odebrecht, Carolin, Malte Belz, Amir Zeldes, Anke Lüdeling & Thomas Krause. 2017. RIDGES Herbology: designing a diachronic multi-layer corpus. *Language Resources and Evaluation* 51(3). 695–725.

Plag, Ingo. 1999. *Morphological productivity. Structural constraints in English derivation*. Berlin/New York: Mouton de Gruyter.

R Core Team. 2014. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. `http://www.R-project.org/`.

Scherer, Carmen. 2005. *Wortbildungswandel und Produktivität: Eine empirische Studie zur nominalen -er-Derivation im Deutschen*. Tübingen: Niemeyer.

Schneider-Wiejowski, Karina. 2011. *Produktivität in der deutschen Derivationsmorphologie*. Bielefeld: Universität Bielefeld dissertation.

Tweedie, Fiona J. & R. Harald Baayen. 1998. How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities* 32(5). 323–352.

Zeldes, Amir. 2012. *Productivity in argument selection. From morphology to syntax*. Berlin/Boston: De Gruyter Mouton.

Zipf, George K. 1949. *Human behaviour and the principle of the least effort. An introduction to human ecology*. New York: Hafner.

Procedure for data collection in RIDGES

1. Gather initial sample to detect graphematic variation
2. Gather sample for analysis (including inflections and graphematic variants)
3. Manually tidy and lemmatise samples

(Evert & Lüdeling 2001, Lüdeling, Evert & Heid 2000, Hartmann 2016)

Multilevel annotation in RIDGES

dipl	ein	reines	Gl̊aſzlin	erflieſſen	.	Rectifici-	rung	vnd	lewᵴ	terũg	der	waſᵴ
clean	ein	reines	Gläszlin	erfliessen	.	Rectificirung		vnd	lewterumg|lewterung		der	was-
norm	ein	reines	Gläslein	erfließen	.	Rektifizierung		und	Läuterung		der	Wasser

(Odebrecht et al. 2017)

(Link to this example in ANNIS: `http://tiny.cc/rektifizierung`)

Graphematic variants in RIDGES

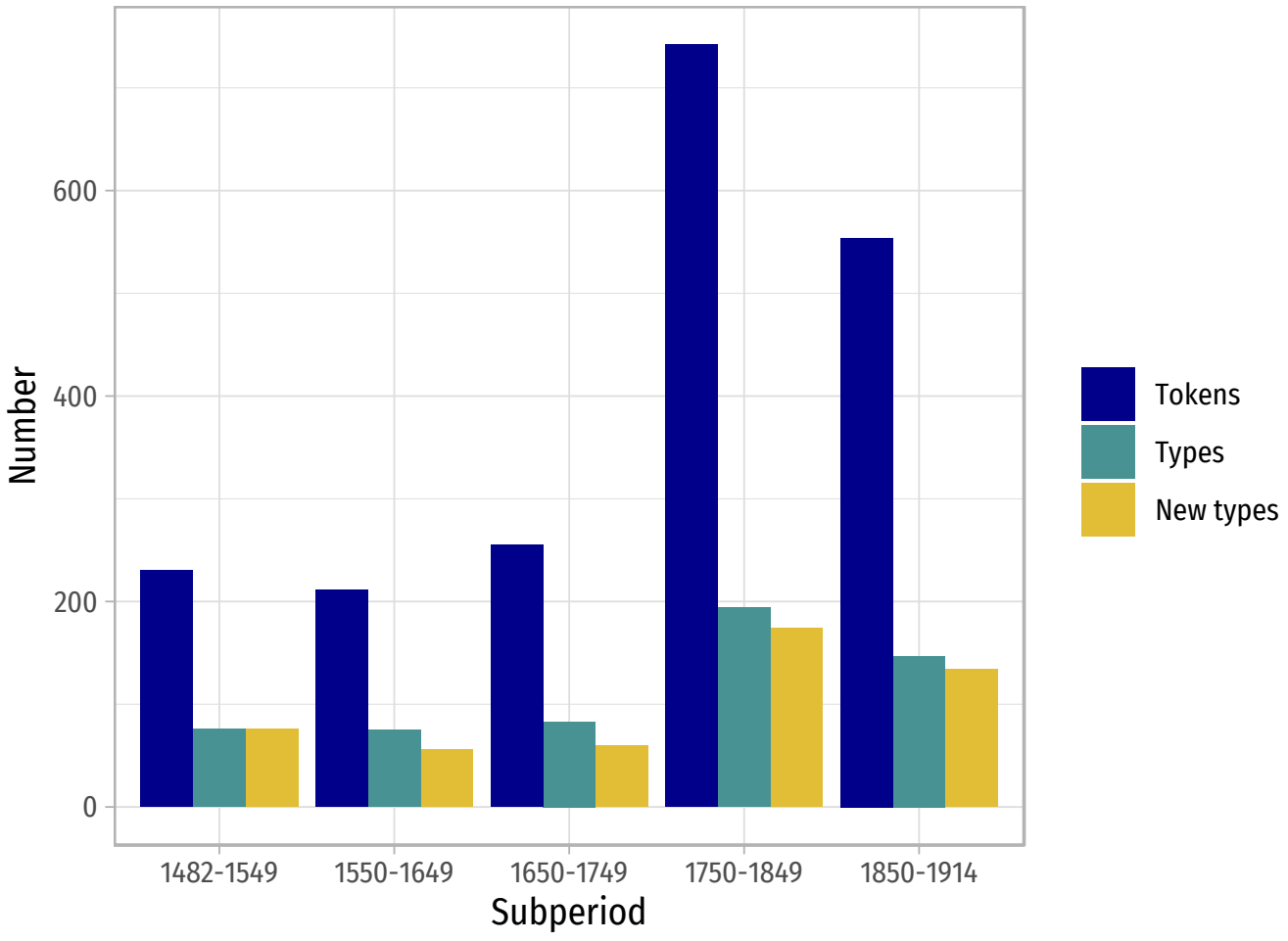
Suffix	Graphematic variants
-er	*-r*
-heit/-keit	*-heyt/-keyt*, *-hait/-kait*
-ung	*-ug*, *-nng*, *-umg*, *-unn*

Lemmatisation

Word form	Lemma
Gattungen	*Gattung*
Ungesundheit	*Gesundheit*
Pflanzengattung	*Gattung*
Herstellung	*Herstellung* (not *Stellung*)
Zusammensetzung	*Zusammensetzung* (not *Setzung*)
Zusammenvermischung	*Vermischung*
Lohnkutscher	*Kutscher*
Däuung, Dauung	*Däuung*

Tokens, types, and new types (shown for *-ung*)

More **tokens** → more **types** (→ more **new types**)



Zipf(-Mandelbrot) Laws

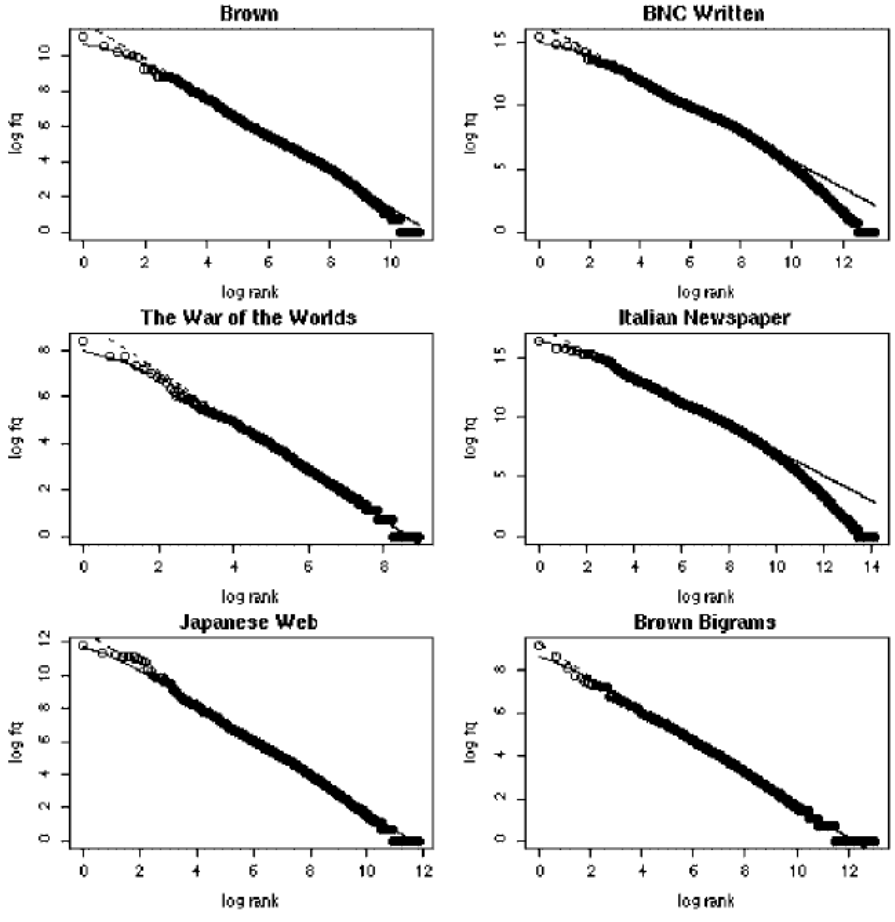


Fig. 37.7: Log rank/log frequency plots with Zipf and Zipf-Mandelbrot fits for the Brown (top left), written BNC (top right), *The War of the Worlds* (middle left), *la Repubblica* (middle right), the Japanese web-page corpus (bottom left), the Brown bigrams (bottom right)

Zipf's Law:

$$f_z = \frac{C}{z}$$

Zipf-Mandelbrot Law:

$$f_z = \frac{C}{(z+b)^a}$$

(Zipf 1949, Mandelbrot 1953; figure from Baroni 2009: 814)