DO CORPUS-DERIVED PRODUCTIVITY MEASURES PREDICT LANGUAGE PROCESSING?
THE CASE OF THE SPANISH INCHOATIVE

Mariia Baltais, Anna Jessen, Robert Hartsuiker
Grammar and Corpora 2022, Ghent University

GHENT UNIVERSITY

GHENT
UNIVERSITY

PRODUCTIVITY IN CORPUS LINGUISTICS

Usage-based approach: syntactic productivity = continuum

Realized productivity:

- Token frequency

- Type/token ratio

Potential productivity:

- Hapax/token ratio

(Baayen 2009)

But:

- Data sparseness (Keller 2003)
- Cxs' ***extensibility*** beyond closed-ended corpora (Barðdal 2008)

GHENT
UNIVERSITY

CORPUS VS. EXPERIMENTS

Acceptability ratings based on experience with language in use → direct correlation with corpus?

❗ **Grammaticality-frequency discrepancy**: "corpus frequencies are poor predictors for acceptability ratings, in particular at the lower end of the frequency spectrum, in morphology and syntax" (Divjak 2017)

- High frequency → high rating
- ❗ High rating → <u>moderate</u> to high frequency
- Low rating → low frequency
- ❗ Low frequency → low to <u>moderate</u> ratings
- → **Ratings tend to be more lenient than corpus data**

(Kempen & Harbusch 2005, 2008)

- Extensibility not so well-captured by corpora → leniency of ratings might come in handy
- Some low-frequent expressions have the potential of making their way into production → elicited production experiments might provide additional data on extensibility

PRODUCTIVITY IN PSYCHOLINGUISTICS

Combine corpus-based and experimental data
- to increase the reliability of results
- to add new perspectives to our understanding of productivity

We conducted:
- Acceptability judgment task (comprehension)
- Sentence completion task (production)

How is productivity attested in corpora related to productivity "at work" in the mind of language users?
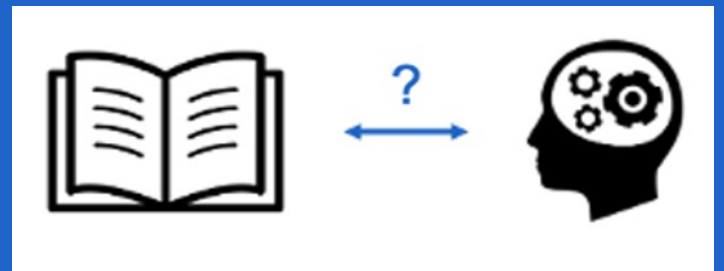
SPANISH INCHOATIVE CONSTRUCTION

- [NP + **V(refl)** + Prep + **INF**]: "agent / cause starts the event of the INF"

Pedro	***empieza***	*a*	***reír***	*'Pedro begins to laugh'*
[Subj]	**[V]**	[Prep]	**[INF]**	

- Two slots of interest: inchoative verb, infinitive

1) *empezar / romper / …* + 'a' + INF

Pedro	*rompe*	*a*	*reír*	*lit. 'Pedro breaks to laugh'*
[Subj]	[V]	[Prep]	[INF]	

2) e.g., *romper* + 'a' + **different INFs**

|					
?	*Pedro*	*rompe*	*a*	*entender*	*lit. 'Pedro breaks to understand'*
	[Subj]	[V]	[Prep]	[INF]	

GHENT
UNIVERSITY

CORPUS DATA

- Sven Van Hulle: data on 25 inchoative verbs
- European Spanish subcorpus of esTenTen18 (Sketchengine): web data, ~3.5 billion tokens
- For each verb: annotated up to 500 sentences
- At least 500 available only for 9 verbs

Estimated token fq INCH in the subcorpus
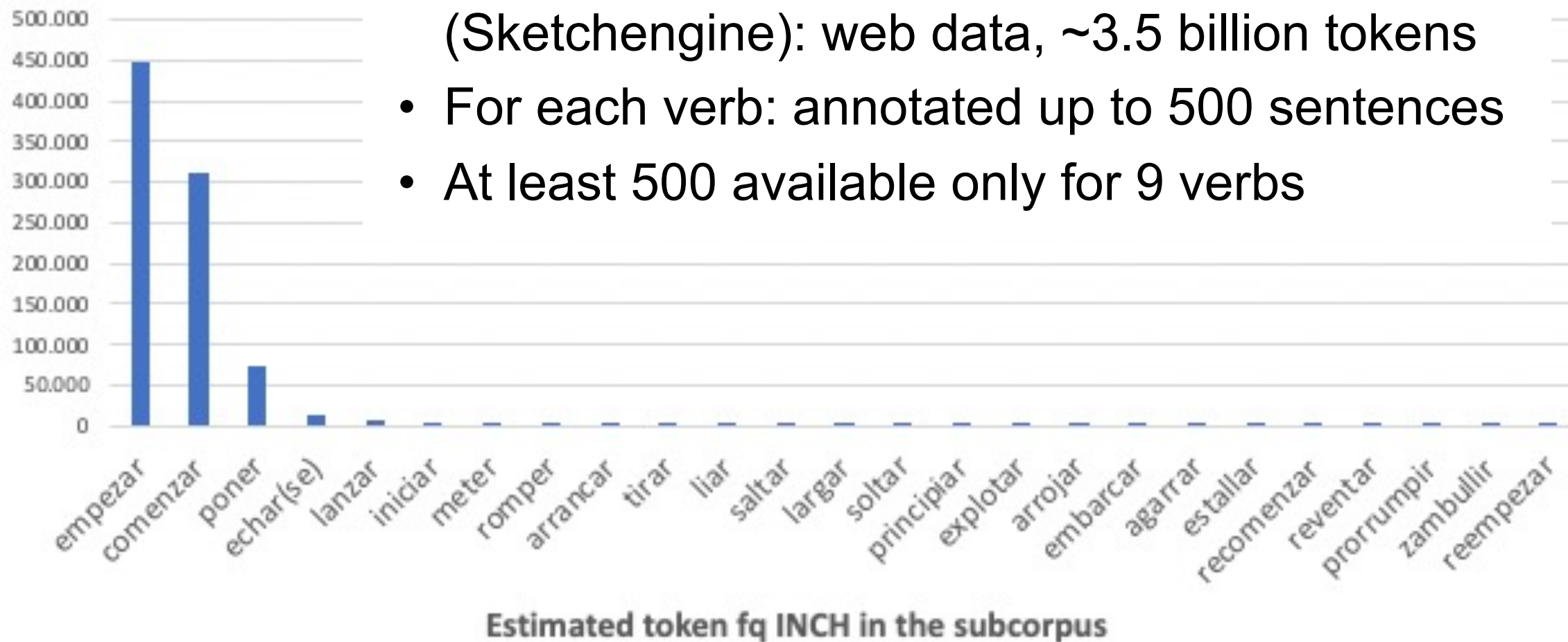
TYPE/TOKEN RATIO: EXAMPLES

280 infinitives

ser	17
trabajar	15
hacer	14
tener	10
dar	9
ver	9
notar	8
poner	8
buscar	7
hablar	7
sonar	7
decir	6
recibir	6
tomar	6
bajar	5
conocer	5
disfrutar	5
jugar	5
llamar	5
caminar	4
construir	4
funcionar	4

etc.
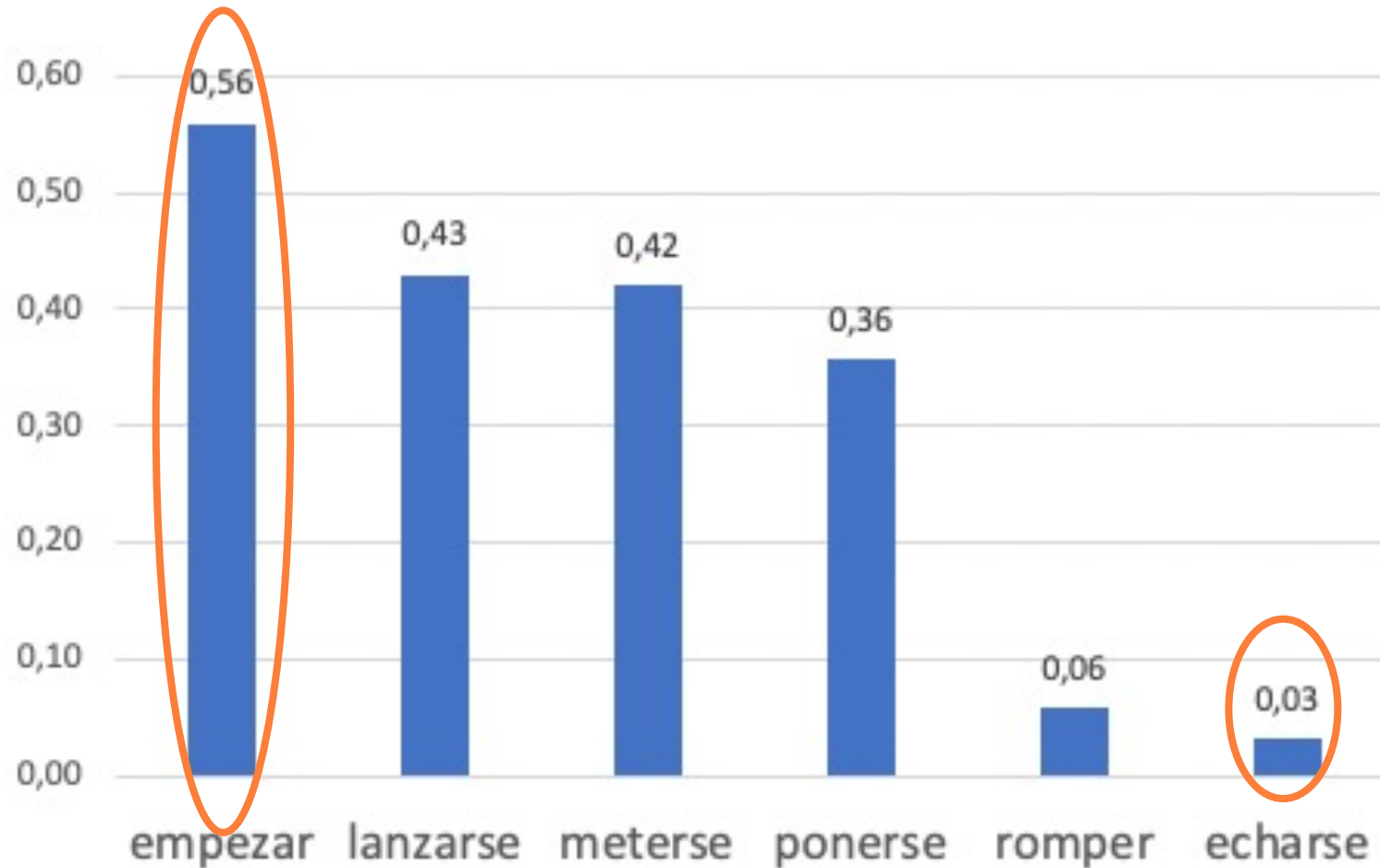
17 infinitives

|	
llorar	189
reír	119
temblar	80
dormir	59
correr	15
andar	14
faltar	8
caminar	3
morir	3
volar	3
arder	1
bailar	1
descansar	1
gemir	1
leer	1
navegar	1
recorrer	1

ACCEPTABILITY STUDY

CHOICE OF INCHOATIVES

- 6 inchoatives with different degrees of productivity
- Equal samples of 500 sentences each

empezar	'to begin, to start'	**very productive**
lanzarse	'to throw, to launch oneself'	**medium productivity**
ponerse	'to put oneself'	
meterse	'to put oneself'	
romper	'to break'	**less productive**
echarse	'to throw oneself'	

CHOICE OF INFINITIVES

empezar (without hapaxes)
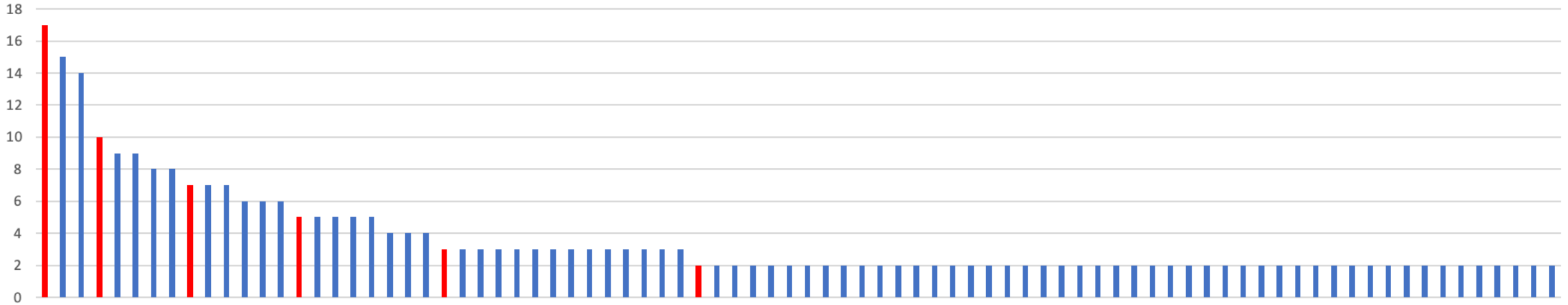
10 infinitives per inchoative

- 6 from high to intermediate to low frequency
- **2 hapaxes: frequent and infrequent semantic class***
- **2 non-attested: frequent and infrequent semantic class***

*http://adesse.uvigo.es/: creation, perception, displacement, physiology…

PARTICIPANTS

- 110 monolingual speakers of European Spanish via https://prolific.co/
  - 10 were excluded due to low accuracy (<80%) for 'yes/no' comprehension questions
  - 4 were excluded because grew up in Latin America

- Ratings from 96 participants were analyzed (37 women, 59 men; mean age: 29 y.o., SD: 10.4 y.)

MATERIALS & PROCEDURE

- 6 inchoatives x 10 infinitives = 60 critical sentences

  + 140 distractor sentences = 200 sentences in total

- Authentic (simplified) corpus sentences

- 21 'yes/no' comprehension questions
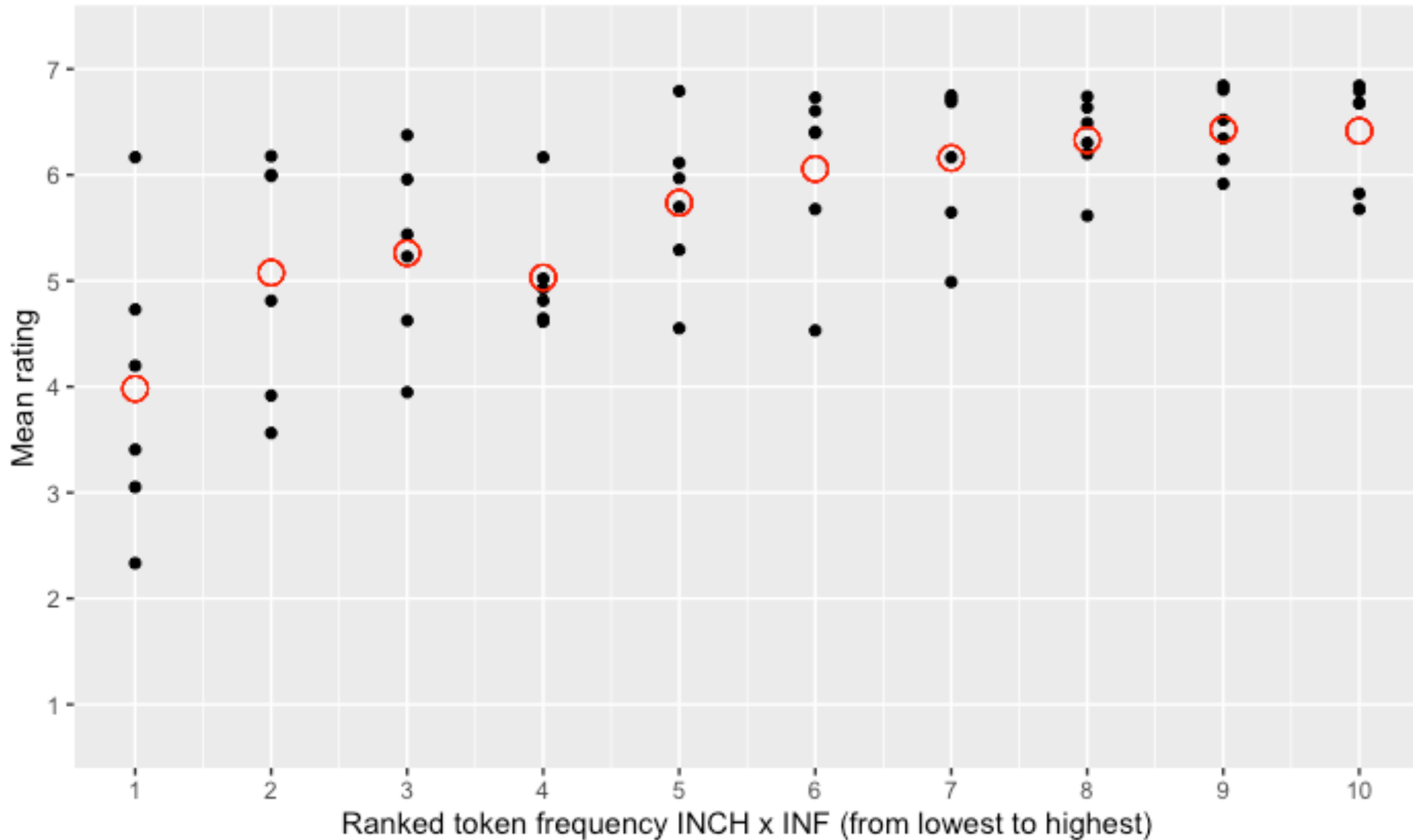
- 7-point Likert scale   unacceptable ○ ○ ○ ○ ○ Ⓧ ○ acceptable
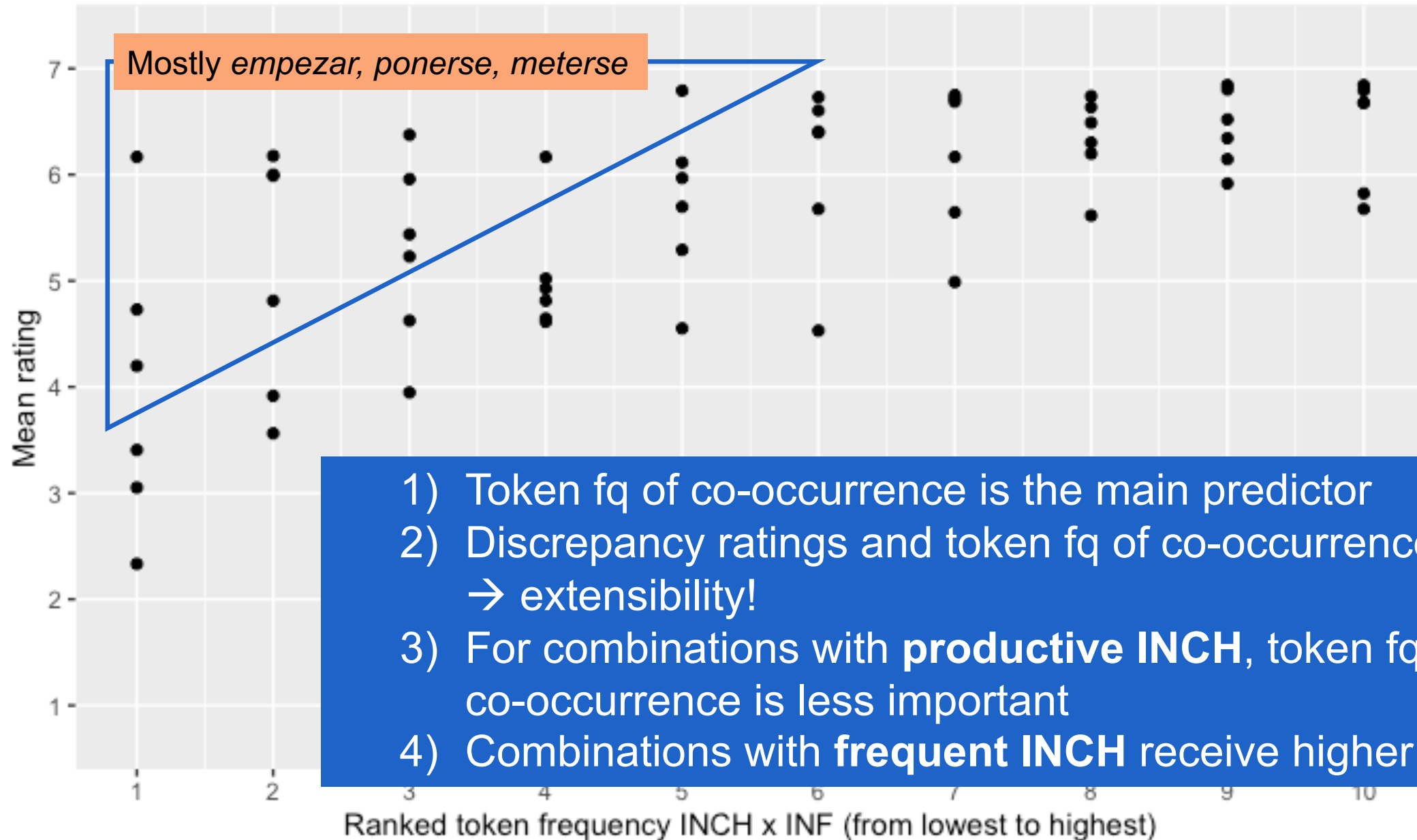
- Linear mixed models

RESEARCH QUESTIONS

- ## Are corpus measures predictive of acceptability ratings?

  - Does productivity of the INCH verb influence the ratings?

  - Discrepancy between ratings and token fq of co-occurrence (INCH x INF)?

- ## When do speakers extend constructions to new items?

  - Does productivity of the INCH verb reflect its extensibility?

  - Are low-frequent INF more acceptable if they are semantically "compatible"
    with the INCH (belong to a frequent semantic class)?

RATINGS ~ TOKEN FQ CO-OCCURRENCE

RATINGS ~ TOKEN FQ CO-OCCURRENCE

Mostly *empezar, ponerse, meterse*

Mean rating

Ranked token frequency INCH x INF (from lowest to highest)

1) Token fq of co-occurrence is the main predictor
2) Discrepancy ratings and token fq of co-occurrence → extensibility!
3) For combinations with **productive INCH**, token fq of co-occurrence is less important
4) Combinations with **frequent INCH** receive higher ratings
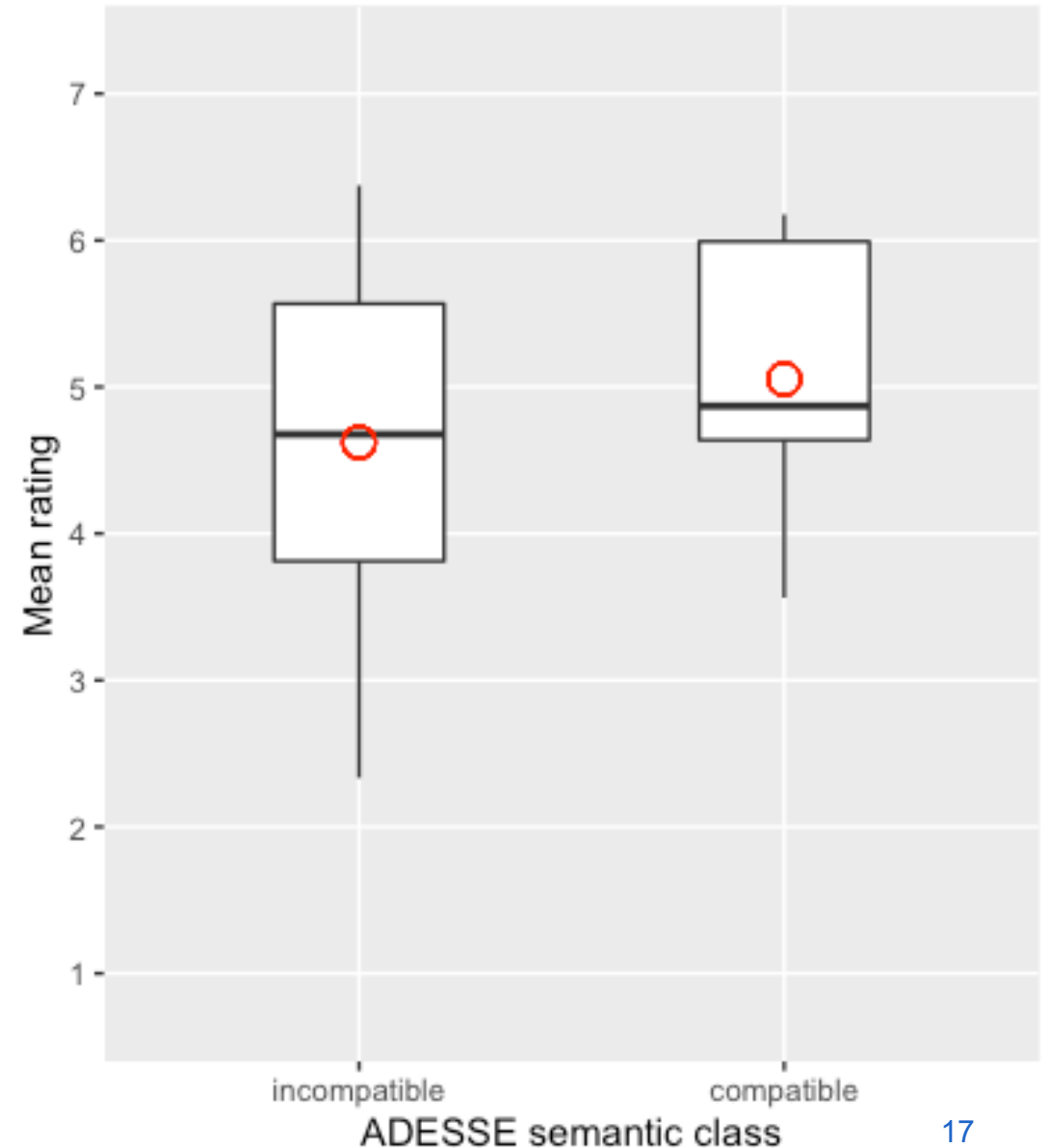
EXTENSIBILITY TO NEW ITEMS

- Subset hapaxes and non-attested INFs
- **Productive INCH** are more extensible (get higher ratings)
- **Semantically "compatible"** combinations are preferred

Compatible: physiology

echarse a respirar *'throw oneself to breath'*

Incompatible: phase

echarse a desarrollar *'throw oneself to develop smth'*

INTERIM CONCLUSION

- **Token frequency of co-occurrence** is the main predictor of acceptability ratings
- But – discrepancy: low frequency ≠ only low ratings (see e.g., Kempen & Harbusch 2008)

<u>Productivity influences ratings:</u>
- The **more productive** the inchoative (high type/token, hapax/token ratio), the less influence token frequency of co-occurrence has on the ratings
- Overall **token frequency of the inchoative** in the corpus also has an effect

<u>Lower end of the fq spectrum:</u>
- **Productive** inchoatives are more **extensible** to new items
- Low-frequent INFs are more acceptable if they are **semantically "compatible"** with the inchoative (are similar to frequent INFs)

PRODUCTION STUDY

GHENT
UNIVERSITY

RESEARCH QUESTIONS

- Are corpus measures predictive of elicited production?
  - Do the INCH verbs show similar 'behaviour' regarding productivity between experiment and corpus?
    - E.g., do more productive inchoatives elicit a higher number of types?

- Do speakers extend constructions to new items?
  - If so, under what circumstances?

GHENT UNIVERSITY

PARTICIPANTS

- 110 monolingual native speakers of European Spanish via
  https://prolific.co/
  - 10 excluded (developmental disorder (n=1), and not living in Spain for more than 6 months at the time of testing (n=9))

- 100 participants remained (37 f, 62 m, 1 other), mean age 29 y, range 18-62 y

GHENT
UNIVERSITY

MATERIALS & PROCEDURE

- 27 sentence beginnings with animate subject, followed by inchoative and preposition 'a':

Termine la frase

Inez empieza a

- 45 distractor sentences with varying structures of comparable length

RESULTS

- 2,700 responses in total

- 1,625 responses with a "true" inchoative meaning (= 60.2%)
  - Other responses mainly with literal meaning (e.g., 'throw themselves into the pool' instead of 'throw themselves to laugh' – *Federico se echó a la piscina – Federico se echó a reír*)

- One INCH did not elicit a single inchoative sentence completion (*agarrar* 'to grab/catch')
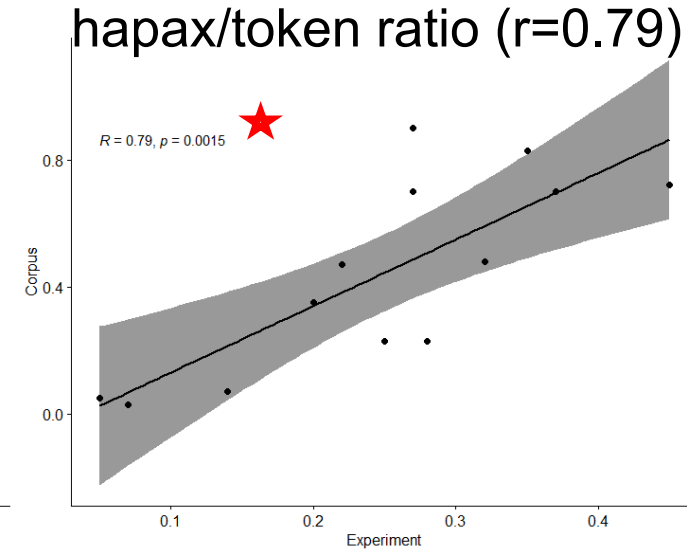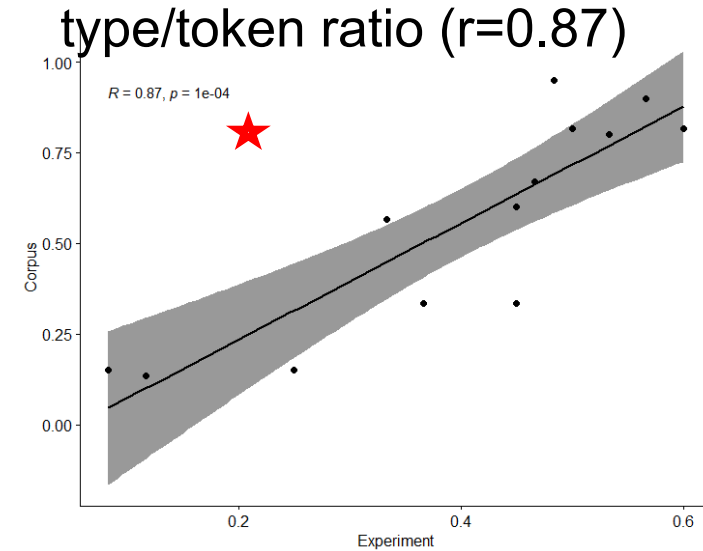
RESULTS (26 INCHOATIVES):

Frequency of 100

token

Frequency 8-16

GHENT
UNIVERSITY

24
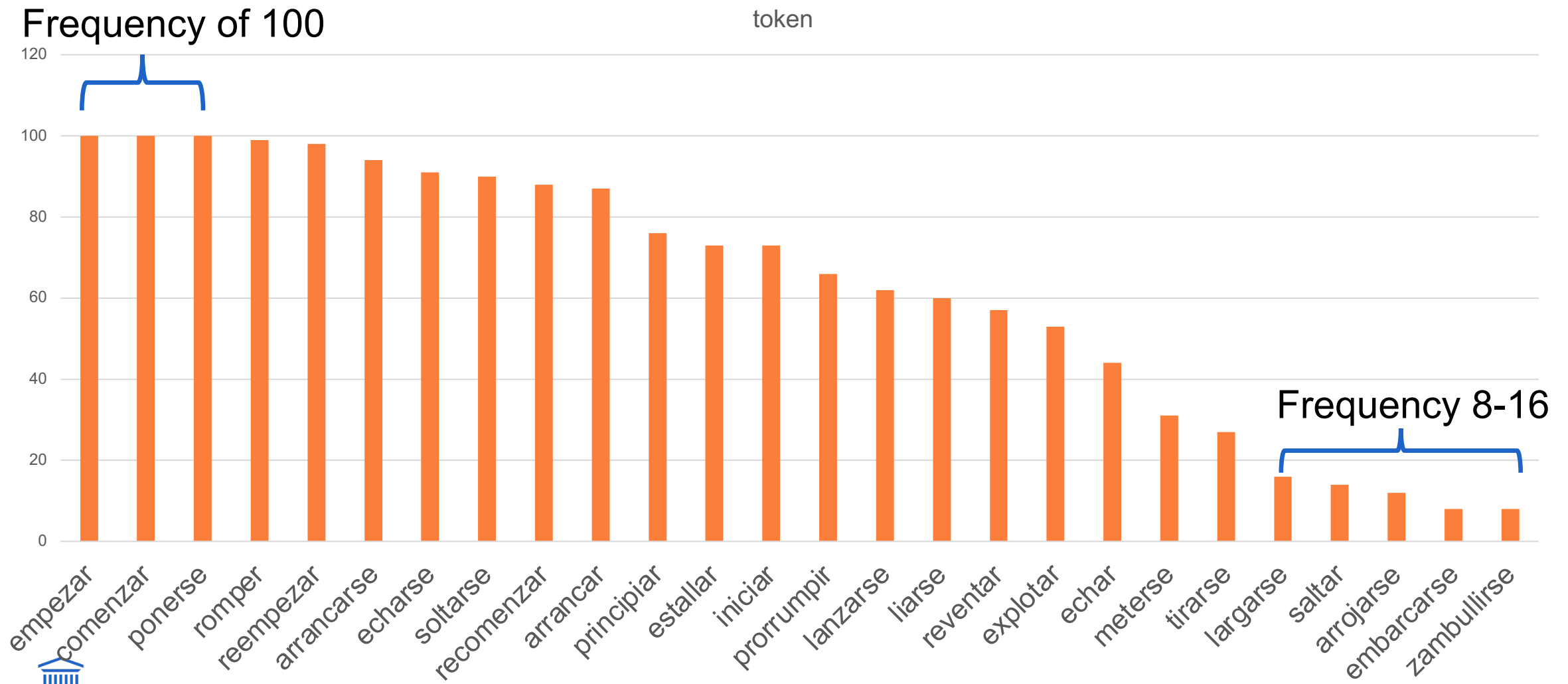
ANALYSIS

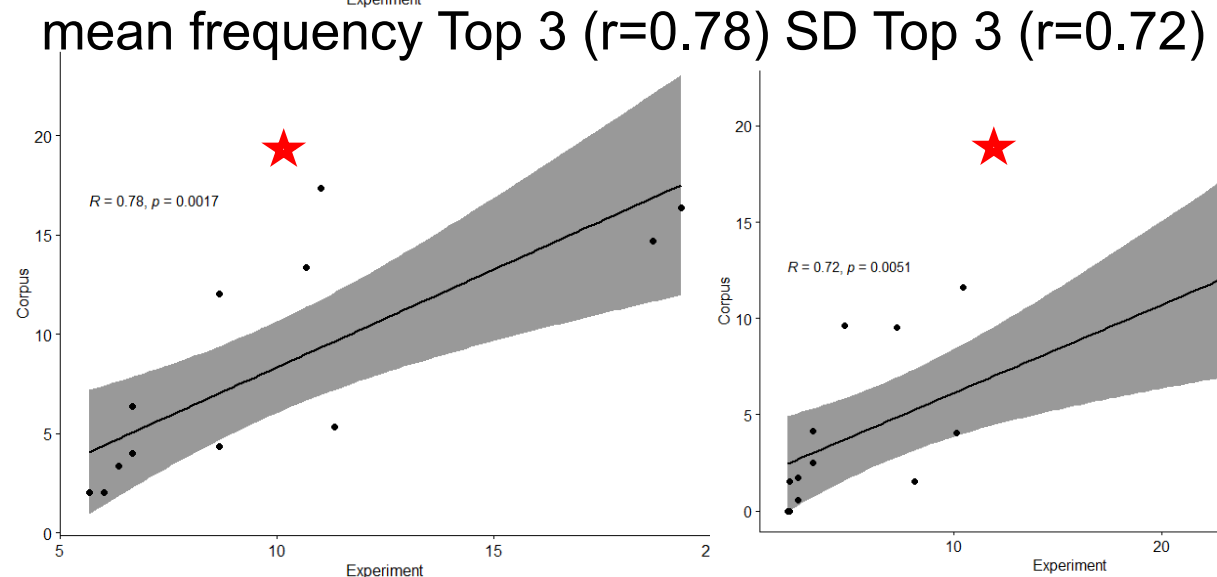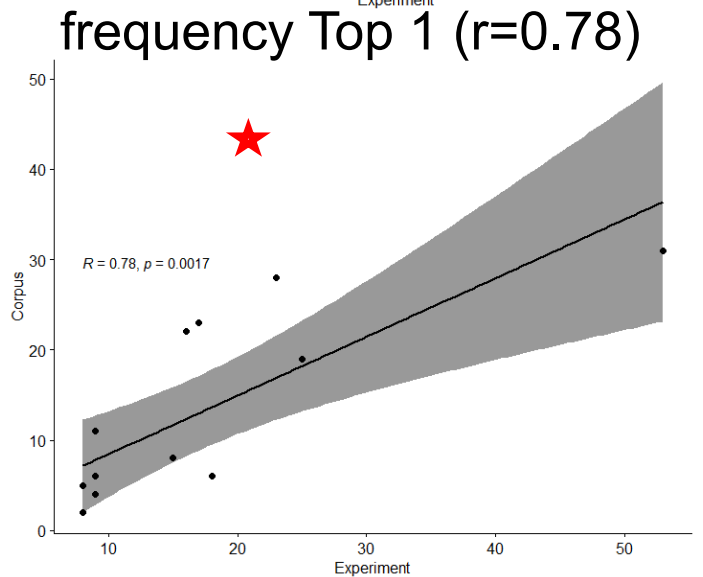| Productivity	Anti-Productivity
Type / Token Ratio	Frequency Top 1
Hapax / Token Ratio	Mean Frequency Top 3
	Standard Deviation Top 3

- Very large differences in sample size from corpus (range from 3 to 500)
- Large difference in sample size from experiment (range from 8 to 100)
- The productivity (and anti-productivity) measures are dependent of sample size

→ Subset of 13 INCH that reached a minimum of 60 tokens in production and corpus
→ Random combinations of 60 from both experiment and corpus data
→ Correlations of (anti)productivity measures

GHENT
UNIVERSITY

CORRELATIONS: CORPUS AND PRODUCTION

type/token ratio (r=0.87)

hapax/token ratio (r=0.79)

Productivity

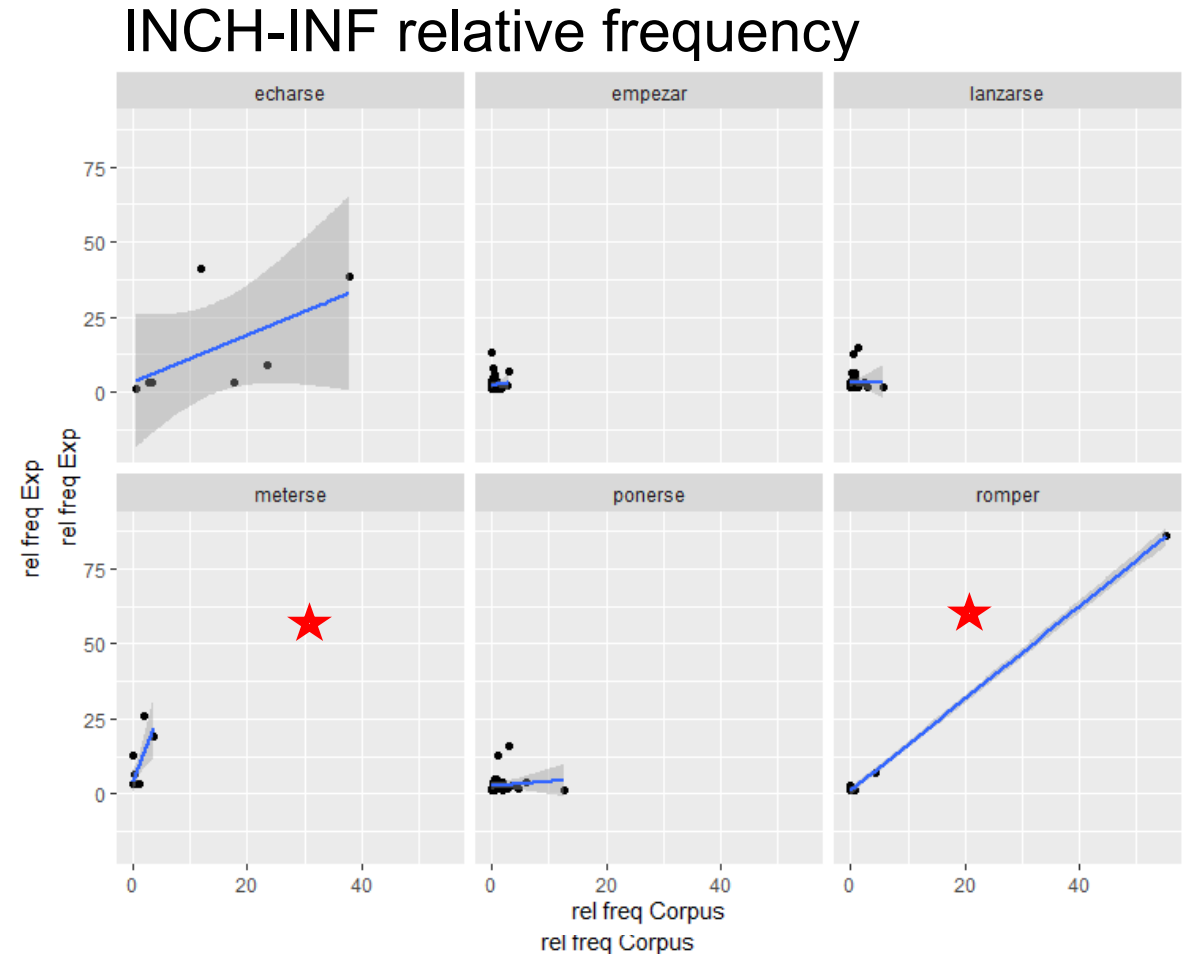frequency Top 1 (r=0.78)

mean frequency Top 3 (r=0.78) SD Top 3 (r=0.72)

Anti-productivity

ZOOMING IN: CO-OCCURENCES

- Relative frequency of combinations (INCH x INF) correlates between corpus and experiment only for some INCH

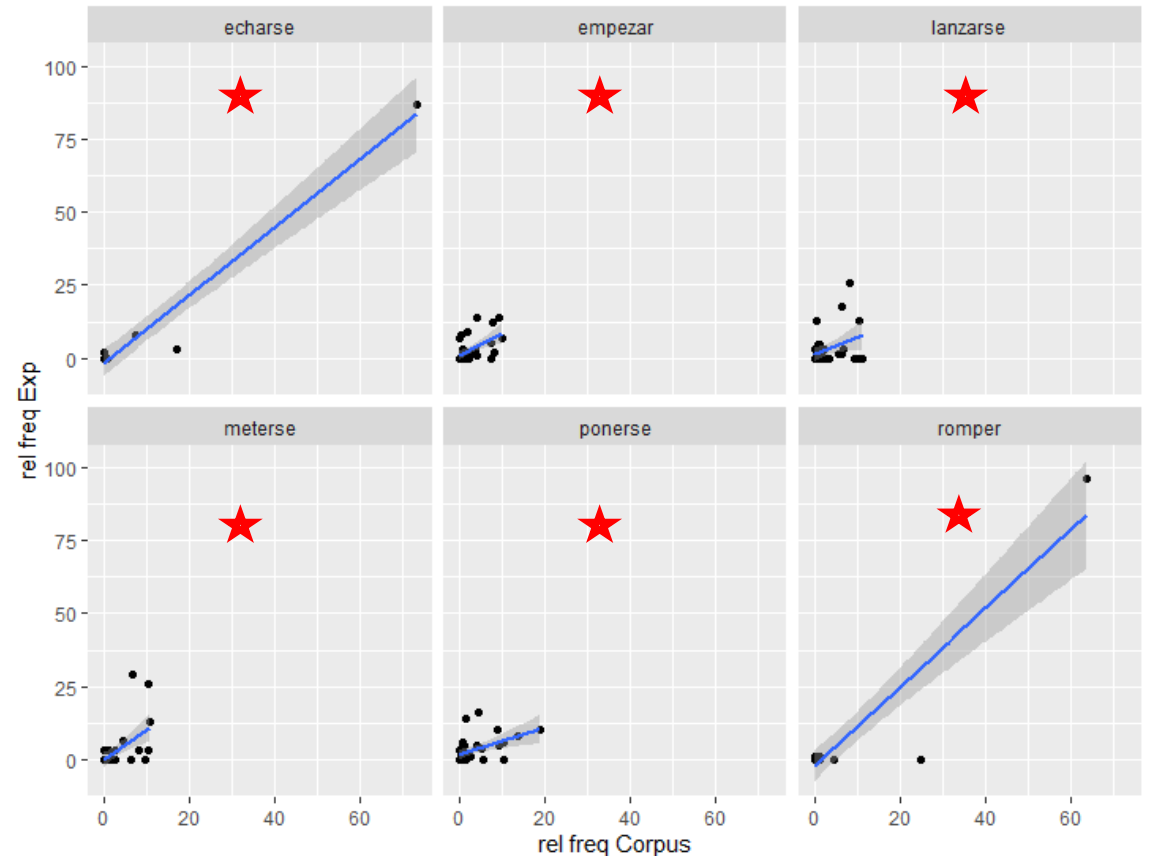INCH-INF relative frequency

Representative sample of 6 inchoatives

ZOOMING IN: SEMANTICS

- Relative frequency of INCH x ADESSE class does correlate for almost all INCH (only 2 exceptions)

→ Combinations with individual INFs might be coincidental (especially with very productive INCH), but their semantic class is not

INCH-ADESSE semantic classes

WHAT ABOUT EXTENSIBILITY?

Several low-frequent inchoative verbs elicited a high number of inchoative sentence continuations:

	corpus	experiment
— *Reempezar* (restart):	4 tokens	98 tokens
— *Recomenzar* (restart):	28 tokens	88 tokens
— *Estallar* (explode/break out):	16 tokens	73 tokens
— *Prorrumpir* (explode/burst):	11 tokens	66 tokens

→ When motivated by the experimental setting, speakers are willing/able to creatively extend the constructions to new items (and semantic classes)

CLOSING REMARKS

How is productivity attested in corpora related to productivity "at work" in the mind of language users?

- Corpus measures are predictive of acceptability ratings and production data
- Acceptability is influenced not only by token fq of co-occurrence but also by productivity
- Elicited production correlates with corpus in terms of productivity of inchoatives
- Experimental data provide insights on **extensibility** of inchoative verbs

- Experimental data can be useful to study productivity in languages for which sufficiently large corpora are not available yet

CHOICE OF INCHOATIVES

- 6 inchoatives with different productivity characteristics
- Equal samples of 500 sentences each

Inchoative	Estimated token frequency in the subcorpus	Type/token ratio	Hapax/token ratio	Ratio of semantic classes*
empezar	429583	0.56	0.39	0.84
ponerse	60728	0.36	0.24	0.66
lanzarse	7476	0.43	0.27	0.75
meterse	1648	0.42	0.28	0.75
romper	1976	0.06	0.03	0.29
echarse	4242	0.03	0.01	0.16

* http://adesse.uvigo.es/ : creation, perception, displacement, modification, physiology…

GHENT UNIVERSITY

ACCEPTABILITY INSTRUCTION

Hereafter, we will present you with a number of sentences. Please read each sentence carefully and rate it based on how "acceptable" it sounds <u>to you</u>.
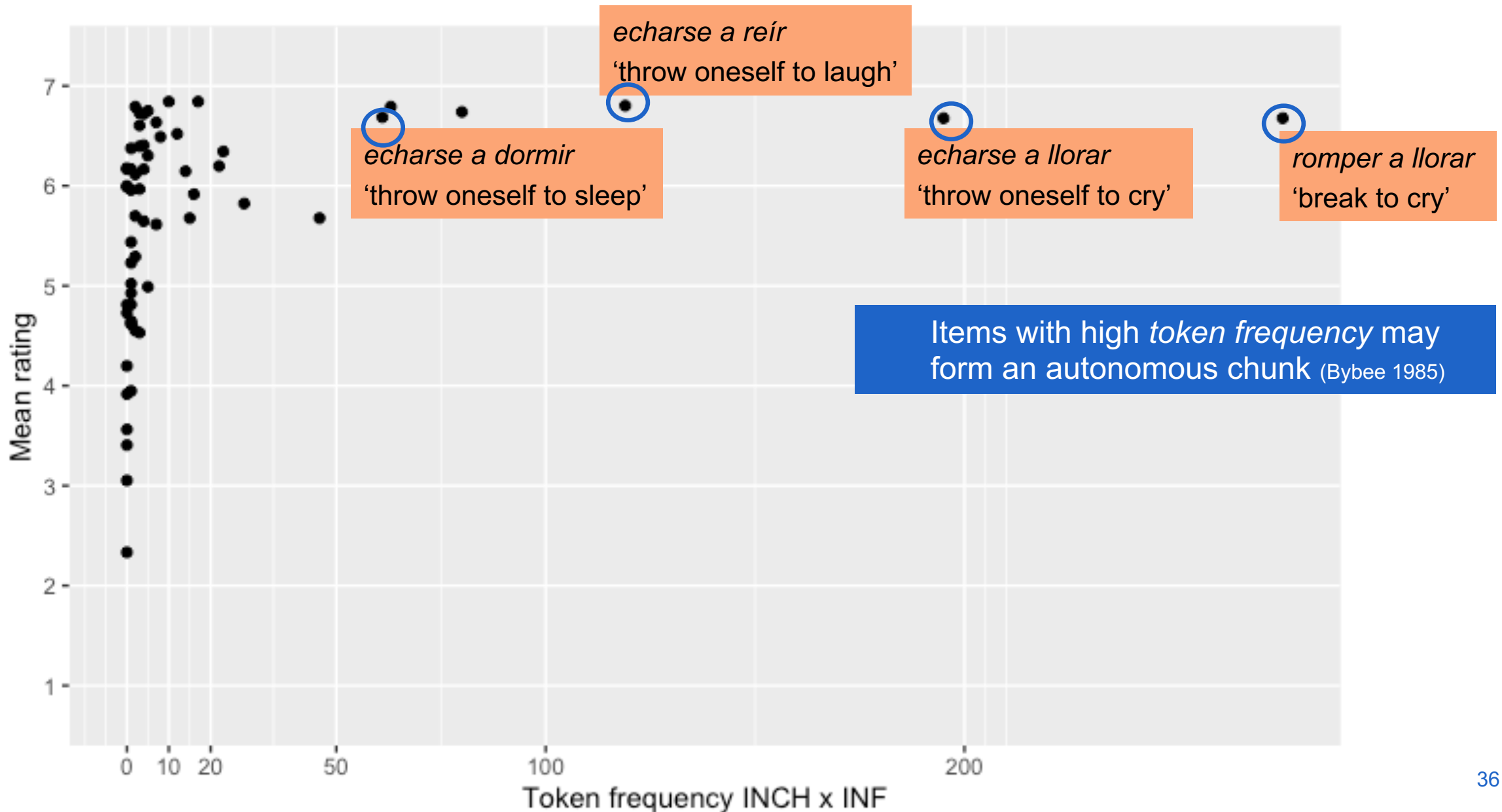
In this study, an *acceptable* sentence is one that would seem **natural** for a native speaker of Spanish to say or to write.

In contrast, an *unacceptable* sentence is one that would seem **unnatural** for a native speaker of Spanish to say or to write.

DATA INSPECTION

Inchoative	Mean rating	SD	Hapax/token ratio
empezar	6.33	1.34	0.39
ponerse	6.02	1.57	**0.24**
meterse	5.66	1.63	0.28
lanzarse	5.57	1.73	0.27
echarse	5.50	1.95	**0.01**
romper	5.01	2.07	0.03

RATINGS ~ TOKEN FQ CO-OCCURRENCE

echarse a reír
'throw oneself to laugh'

echarse a dormir
'throw oneself to sleep'

echarse a llorar
'throw oneself to cry'

romper a llorar
'break to cry'

Items with high *token frequency* may form an autonomous chunk (Bybee 1985)

Mean rating

Token frequency INCH x INF

LMEM: FULL DATA SET

– Absolute token fq INCH x INF

| Fixed effects	Estimate	SE	*t-value*
(Intercept)	5.6477	0.1381	40.895
Token fq INCH x INF	**0.5089**	0.1270	**4.006***
Hapax/token ratio INCH	**0.4866**	0.1273	**3.821***
AR ~ Token fq INCH x INF + Hapax/token ratio + (1 | item) + (1 + Token fq INCH x INF + Hapax/token ratio | participant)			

– Main effect *token frequency INCH x INF*
– Main effect *hapax/token ratio INCH*

INTERACTION PRODUCTIVITY
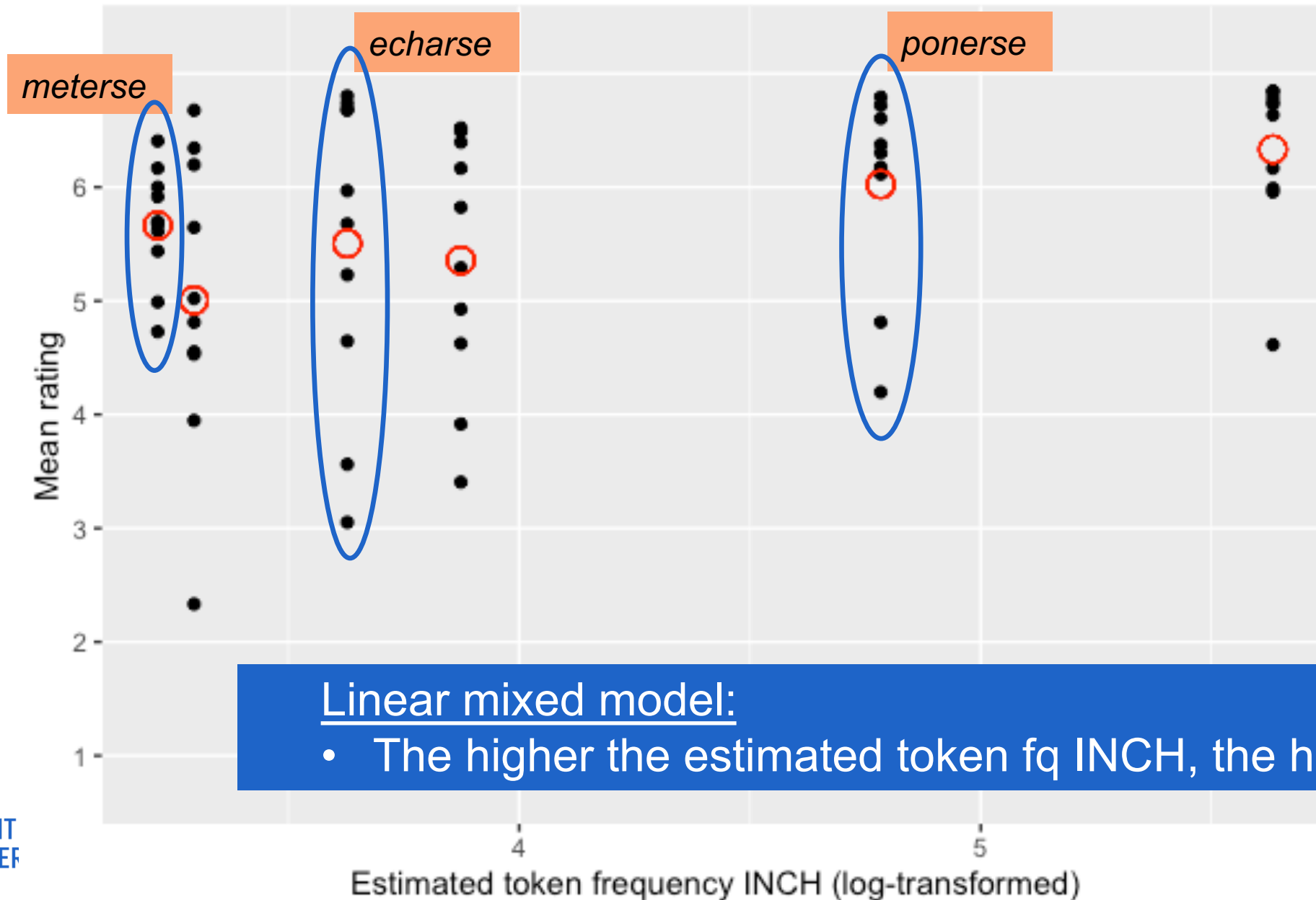
Less productive ⟶ **More productive**
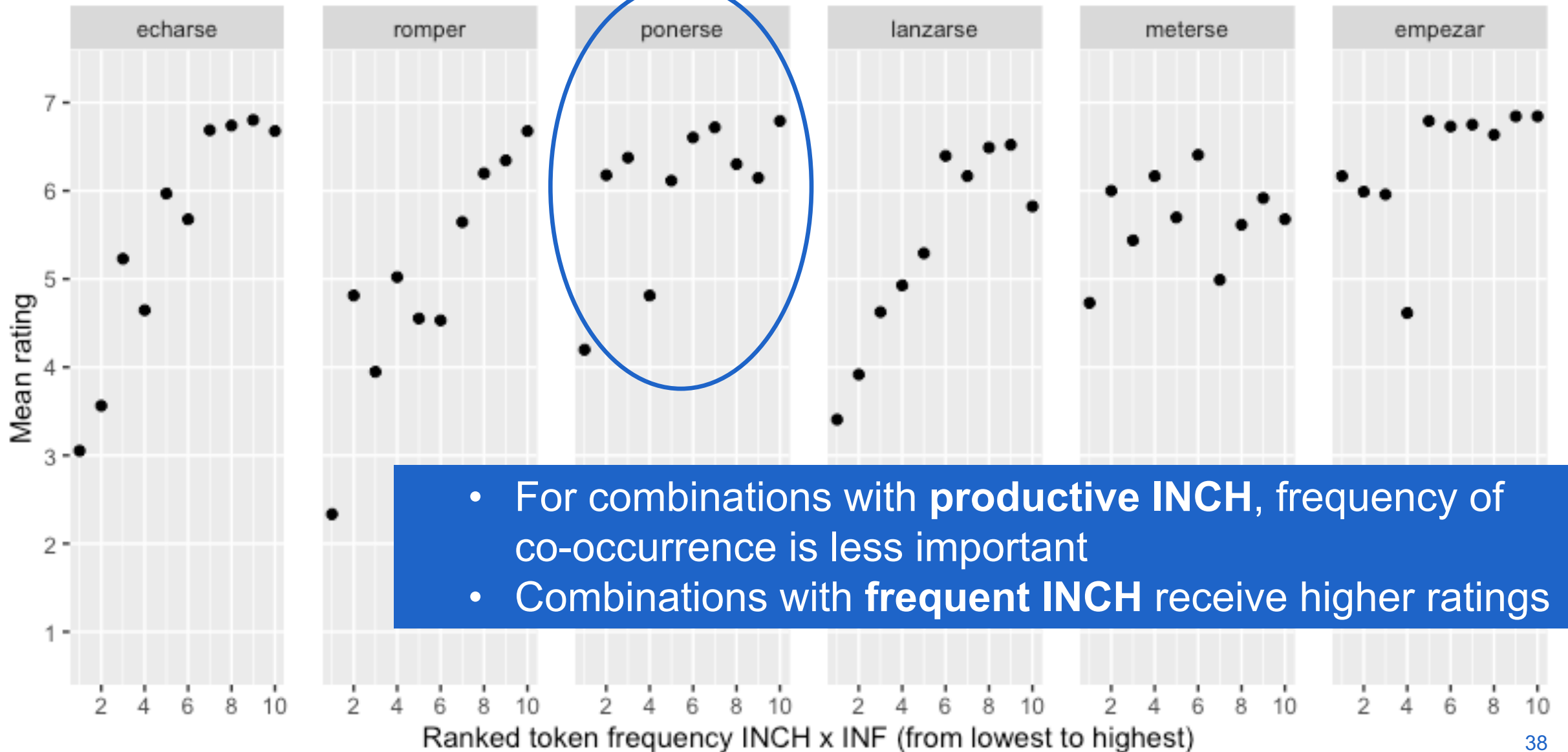
- For combinations with **productive INCH**, frequency of co-occurrence is less important
- Combinations with **frequent INCH** receive higher ratings

RATINGS ~ ESTIMATED TOKEN FQ INCH

Linear mixed model:
- The higher the estimated token fq INCH, the higher the rating

LMEM: FULL DATA SET

– Ranked token fq INCH x INF

Fixed effects	Estimate	SE	*t-value*
(Intercept)	5.64774	0.10989	51.395
Ranked token fq INCH x INF	**0.70366**	0.08626	**8.158***
Hapax/token ratio INCH	0.13266	0.11114	1.194
Estimated token fq INCH	**0.28833**	0.11300	**2.551***
Ranked token fq INCH x INF: Hapax/token ratio INCH	**-0.31554**	0.08483	**-3.720***
*AR ~ Ranked token fq INCH x INF * Hapax/token ratio + Estimated token fq INCH + (1 | item) + (1 + Ranked token fq INCH x INF * Hapax/token ratio + Estimated token fq INCH | participant)*			

– Main effect *ranked token frequency INCH x INF*
– No main effect *hapax/token ratio INCH*
– Main effect *estimated token fq INCH*
– Significant interaction *ranked token frequency INCH x INF* and *hapax/token ratio INCH*
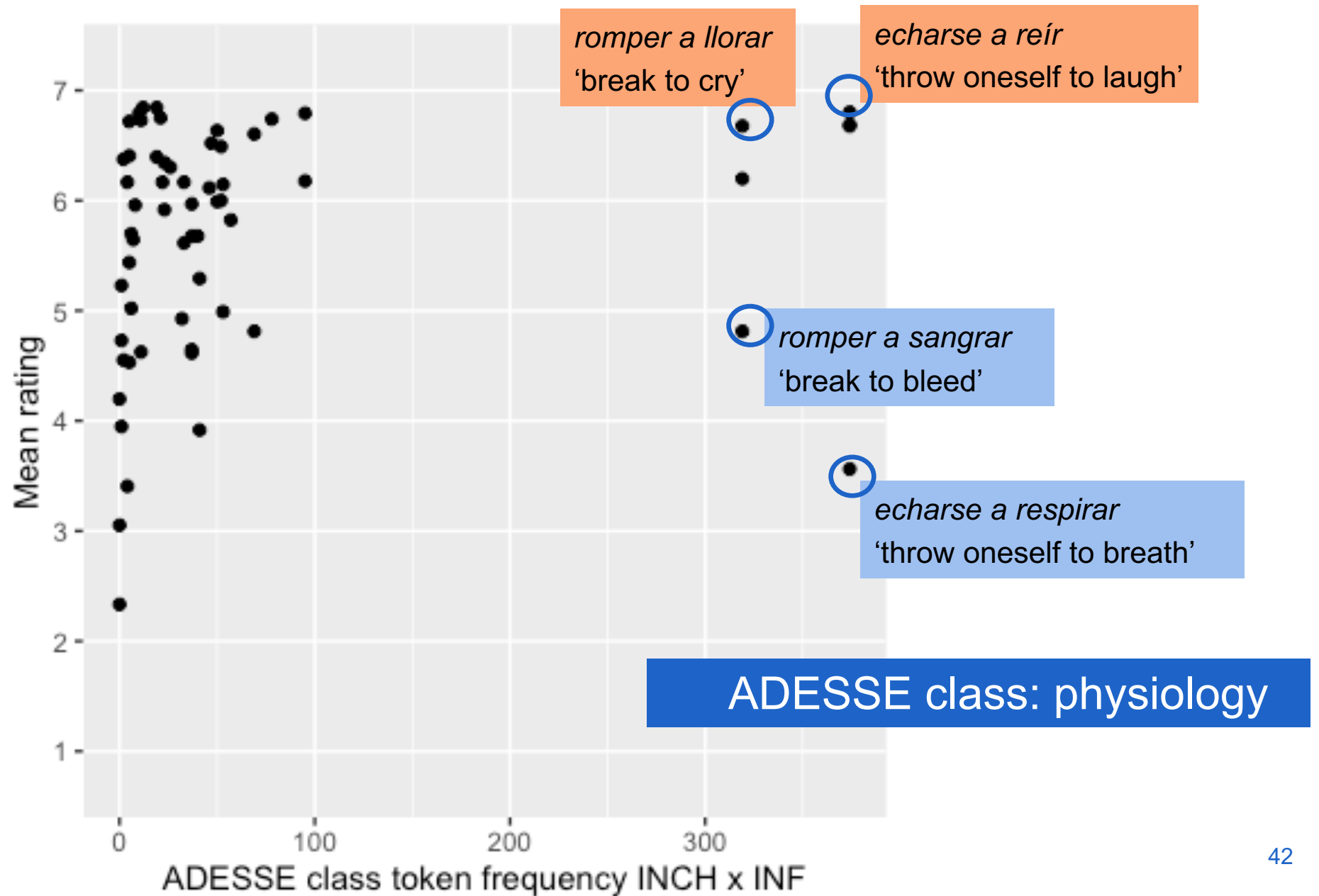
EXTENSIBILITY FACTORS

Frequency

- The higher the *type frequency* of a Cx, the more likely it is to occur with a novel item (Bybee 1985; Baayen 1993)
- Anti-productivity: Items with *high token frequency* may form an autonomous chunk (Bybee 1985)
- *Hapax frequency* reflects a construction's ability to attract new or existing lexical items (Barðdal 2008)

Semantics

- New items are often only acceptable if they are *semantically similar* to already attested ones (Barðdal 2008; Suttle & Goldberg 2011)

SEMANTICS?

romper a llorar 'break to cry'

echarse a reír 'throw oneself to laugh'

romper a sangrar 'break to bleed'

echarse a respirar 'throw oneself to breath'

ADESSE class: physiology

LMEM: RANKS 1-4

– Binary token fq INCH x INF (0 or 1)
– Binary ADESSE frequent/ infrequent

| Fixed effects	Estimate	SE	*t-value*
(Intercept)	5.2037	0.2476	21.018
Hapax/token ratio	**0.5909**	0.1950	**3.031***
Infrequent ADESSE	**-0.7320**	0.3433	**-2.132***
Lemma fq INF	**-0.4823**	0.1683	**-2.866***
Nu. words in sentence	-0.3783	0.1962	-1.928
AR ~ Hapax/token ratio + Infrequent ADESSE + Lemma fq INF + Nu. words in sentence + (1 | item) + (1 + Hapax/token ratio + Infrequent ADESSE + Lemma fq INF + Nu. words in sentence | participant)			

GHENT
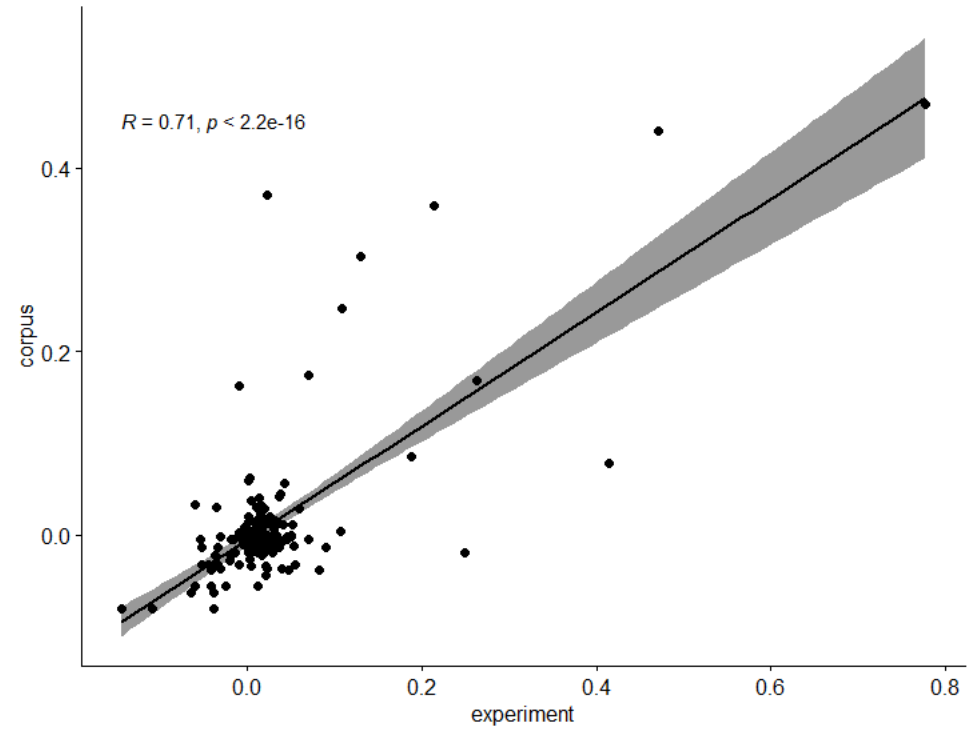UNIVERSITY

LMEM: RANKS 1-5

– Absolute token fq INCH x INF

– Binary ADESSE frequent/ infrequent

Fixed effects	Estimate	SE	*t-value*
(Intercept)	5.0174	0.1691	29.677
Token fq INCH x INF	**0.5830**	0.1495	**3.900***
Hapax/token ratio	**0.4980**	0.1469	**3.389***
Lemma fq INF	**-0.3346**	0.1534	**-2.181***
AR ~ Token fq INF x AUX + Hapax/token ratio + Lemma fq INF + (1 | item) + *(1 + Token fq INF x AUX + Hapax/token ratio + Lemma fq INF | participant)*			

GHENT
UNIVERSITY
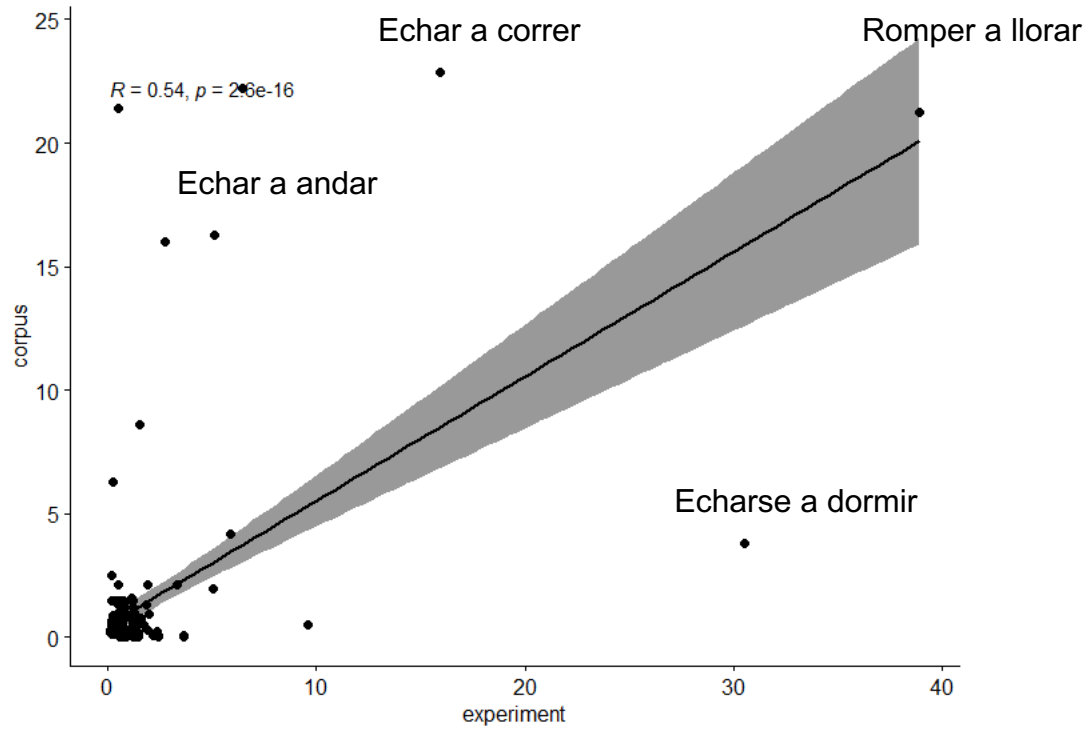
Collocation strength

delta p (constr to word)