FACULTY OF PSYCHOLOGY AND EDUCATIONAL SCIENCES

FACULTY OF ARTS AND PHILOSOPHY

LANGUAGE PRODUCTIVITY @ WORK

PRODUCTIVITY OF THE SPANISH INCHOATIVE CONSTRUCTION:

FROM CORPUS TO ACCEPTABILITY JUDGMENTS

Mariia Baltais, Sven Van Hulle and Robert J. Hartsuiker

LinGhentian Doctorials / 14.12.2021

GHENT UNIVERSITY

1. LANGUAGE PRODUCTIVITY & LP@W PROJECT
2. SPANISH INCHOATIVE CONSTRUCTION
3. ACCEPTABILITY STUDY
    - DESIGN & RESEARCH QUESTIONS
    - DATA INSPECTION
    - FIRST CONCLUSIONS & FUTURE PLANS

GHENT UNIVERSITY

1. LANGUAGE PRODUCTIVITY & LP@W PROJECT

2. SPANISH INCHOATIVE CONSTRUCTION

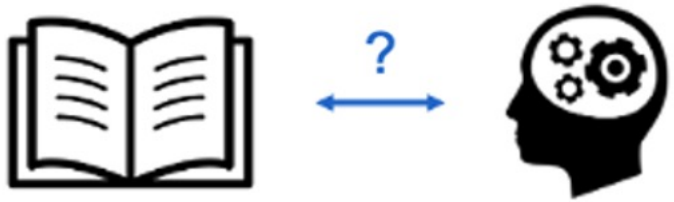
3. ACCEPTABILITY STUDY

- DESIGN & RESEARCH QUESTIONS
- DATA INSPECTION
- FIRST CONCLUSIONS & FUTURE PLANS

SYNTACTIC PRODUCTIVITY IN LINGUISTICS

- **Realized productivity: as attested in language corpora**
  - →TYPE FREQUENCY: “the range of lexical items that may fill the slots of constructions” (lexical scope) (Perek 2016)
  - →TOKEN FREQUENCY: “the total occurrences of either one or all the types of a construction in a text or corpus” (Barðdal 2008)
  - →HAPAX FREQUENCY: “a construction’s ability to attract new or existing lexical items” (extensibility) (Barðdal 2008)

PRODUCTIVITY ”AT WORK” IN THE MIND



- **Potential productivity: beyond corpora**
  - → “Off-line” and “on-line” language processing
  - → Comprehension and production

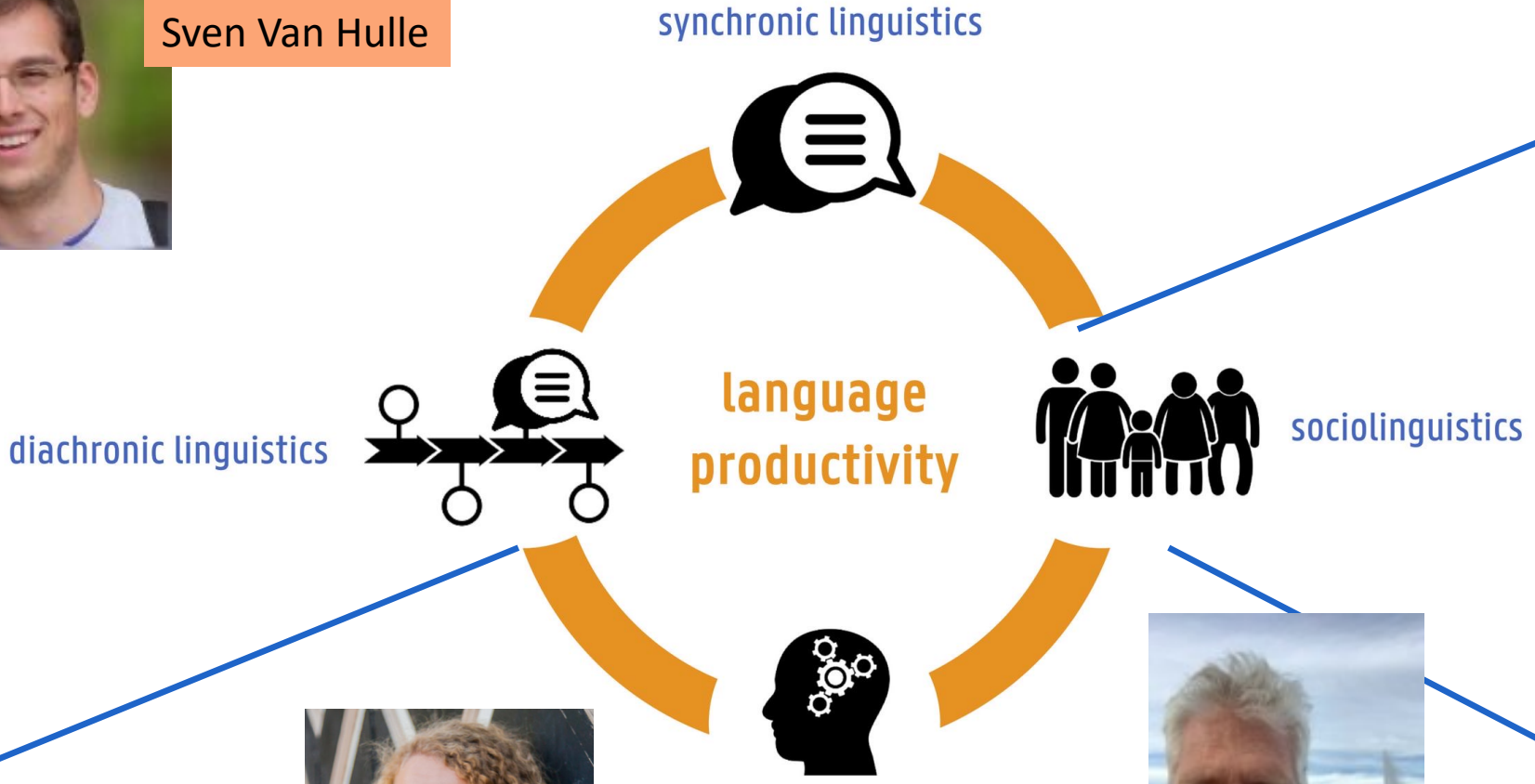
How is realized productivity of a construction related to potential productivity “at work” in the mind of speakers?

GHENT UNIVERSITY

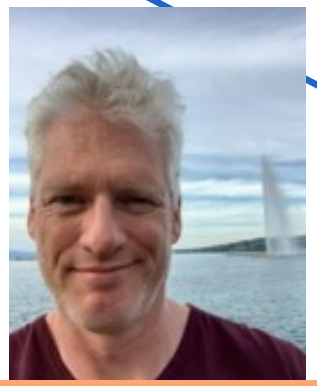
“LANGUAGE PRODUCTIVITY @ WORK” PROJECT

https://www.languageproductivity.ugent.be/

Sven Van Hulle



Mariia Baltais



Prof. Dr. Robert Hartsuiker

GHENT UNIVERSITY

1. LANGUAGE PRODUCTIVITY & LP@W PROJECT

2. SPANISH INCHOATIVE CONSTRUCTION

3. ACCEPTABILITY STUDY

- DESIGN & RESEARCH QUESTIONS
- DATA INSPECTION
- FIRST CONCLUSIONS & FUTURE PLANS

GHENT UNIVERSITY

SPANISH INCHOATIVE CONSTRUCTION

- **Macro**-level: **AUXILIARY** + ’a’ + INFINITIVE

Pedro	*empieza*	*a*	*reír*	*(‘Pedro starts to laugh’)*
[Subj]	[AUX]	[Prep]	[INF]	

- **Micro**-level: *empezar* + ’a’ + **INF**, *romper* + ’a’ + **INF** etc.
  = 25 different auxiliaries

Pedro rompe a reír *(‘Pedro breaks into laughing’)*

GHENT UNIVERSITY

‘25 WAYS TO EXPRESS THE START OF AN EVENT IN SPANISH’

poner

meter

echar

arrojar

tirar

comenzar

agarrar

largar

embarcar

empezar

reempezar

soltar

explotar

iniciar

liar

recomenzar

arrancar

reventar

romper

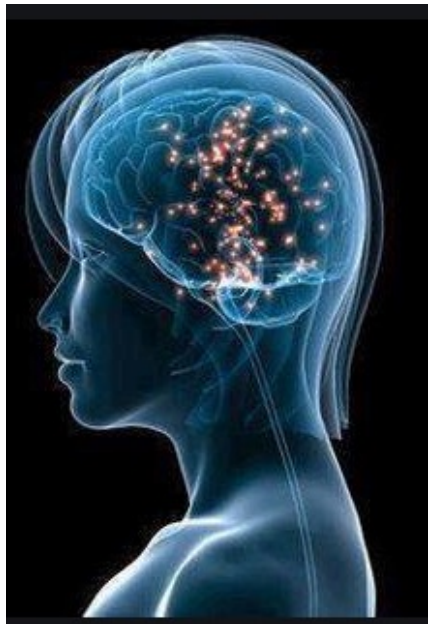
principiar

zambullir

estallar

saltar

prorrumpir



5 SEMANTIC DOMAINS

superlexicals

empezar	“to start”
comenzar	“to start”
reempezar	“to restart”
recomenzar	“to restart”
iniciar	“to initiate”
principiar	“to start”

positioning

poner	“to put”
meter	“to put”

throwing

arrojar	“to throw”
echar	“to throw”
lanzar	“to launch
tirar	“to fire”

movement

agarrar	“to grab”
embarcar	“to embark”
largar	“to let go”
liar	“to bind”
saltar	“to jump”
soltar	“to loosen”
zambullir	“to dive”

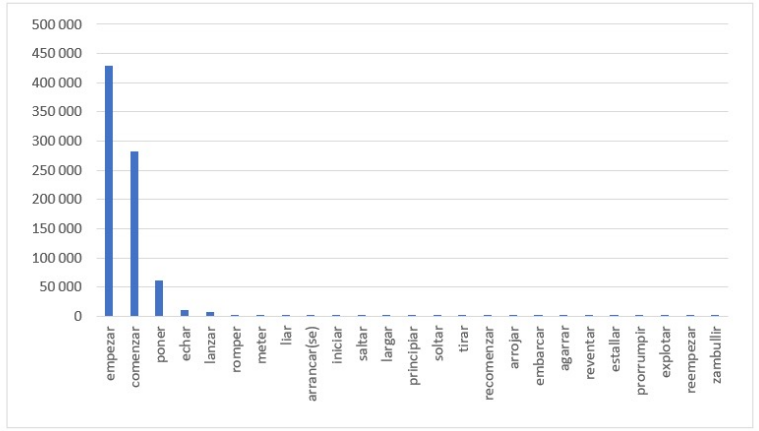
destruction

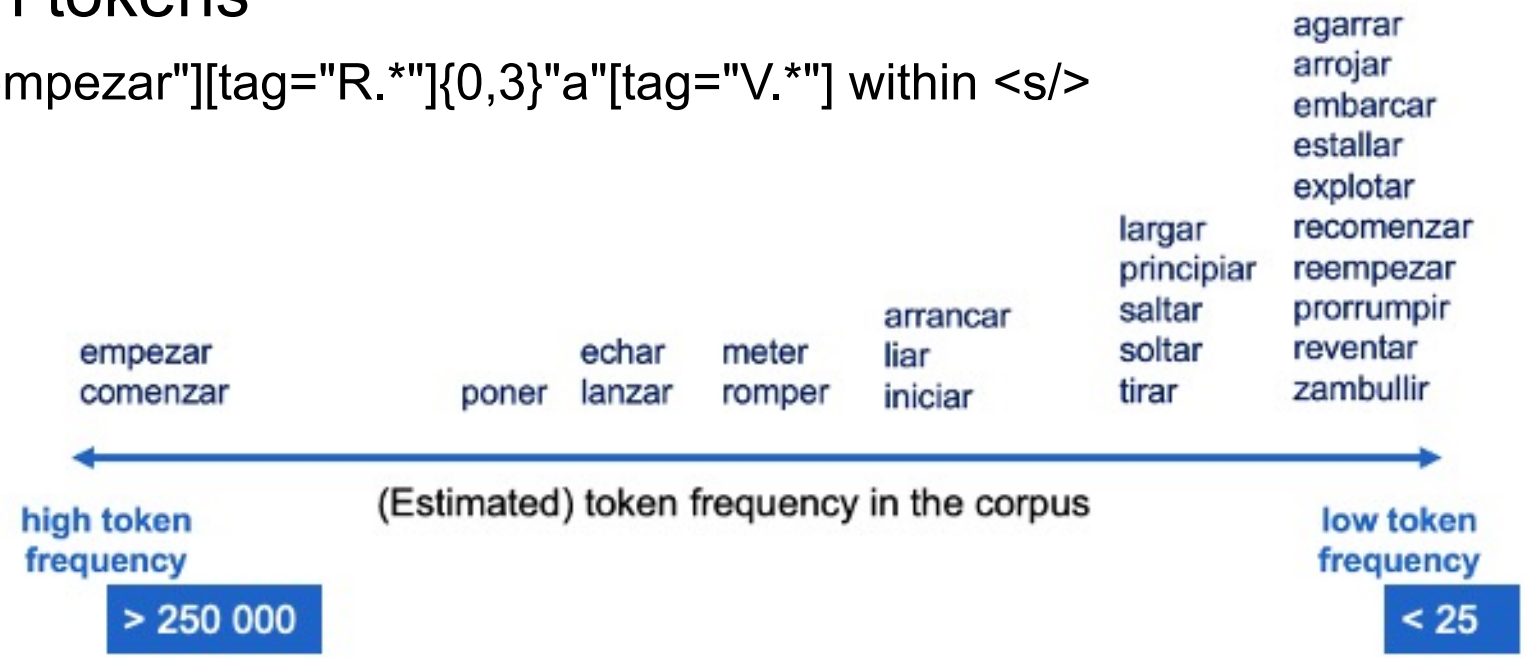
arrancar	“to tear off”
estallar	“to explode”
explotar	“to explode”
prorrumpir	“to break loose”
reventar	“to shatter”
romper	“to break”

GHENT UNIVERSITY

CORPUS DATA

- Empirical study of contemporary peninsular Spanish
- European Spanish subcorpus of esTenTen18 (Sketchengine)
  → Web data, ~3.5 billion tokens
- Search syntax: [lemma="empezar"][tag="R.*"]{0,3}"a"[tag="V.*"] within <s/>





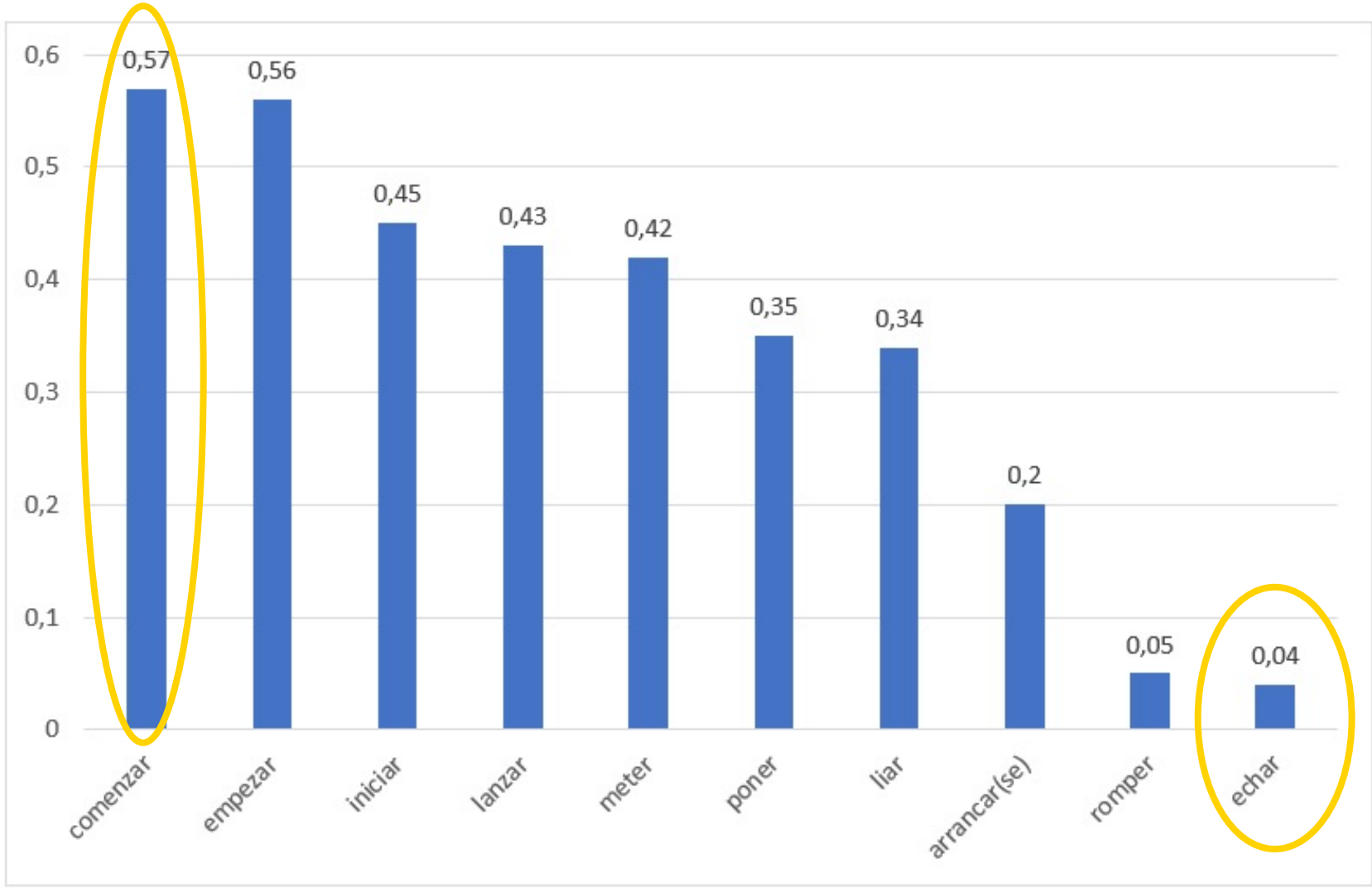
- For each auxiliary: randomized sample of 10 000 sentences
  → Sven cleaned and annotated up to 500 "true positives"
  → Equal samples of 500 are available for 9 auxiliaries

GHENT UNIVERSITY

TYPE / TOKEN RATIO: EXAMPLES

287 infinitives

trabajar	22
ser	15
hacer	12
dar	10
introducir	6
mostrar	6
tomar	6
aparecer	5
desarrollar	5
leer	5
operar	5
recibir	5
utilizar	5
actuar	4
ahorrar	4
caer	4
construir	4
construirse	4
crecer	4
formar	4
funcionar	4
jugar	4
llegar	4



23 infinitives

andar	156
correr	87
llorar	66
temblar	57
reír	33
rodar	32
dormir	20
volar	15
faltar	14
caminar	5
morir	3
arder	2
conocer	1
fundir	1
gemir	1
girar	1
lavar	1
leer	1
navegar	1
pasear	1
pelear	1
recorrer	1

1. LANGUAGE PRODUCTIVITY & LP@W PROJECT

2. SPANISH INCHOATIVE CONSTRUCTION

3. ACCEPTABILITY STUDY

- DESIGN & RESEARCH QUESTIONS
- DATA INSPECTION
- FIRST CONCLUSIONS & FUTURE PLANS

WHAT INFLUENCES PRODUCTIVITY?

- **Frequency**
  - The higher the *type frequency* of a Cx, the more likely it is to occur with a novel item (Bybee 1985; Baayen 1993)
  - Items with high *token frequency* may form an autonomous chunk (Bybee 1985)
  - Number of *hapaxes*
- **Semantics**
  - New fillers are often only acceptable if they are *semantically similar* to already attested ones (Barðdal 2008; Suttle & Goldberg 2011)

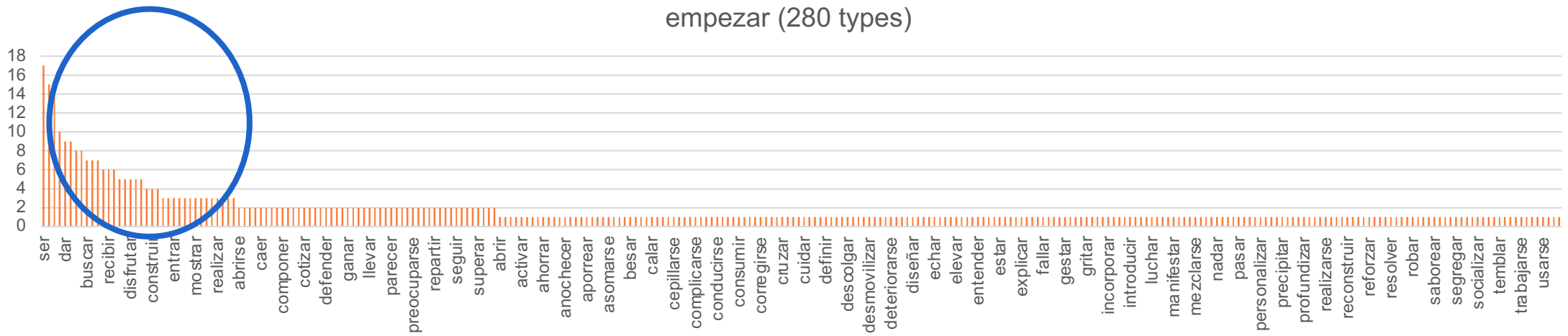
CHOICE OF AUXILIARIES

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

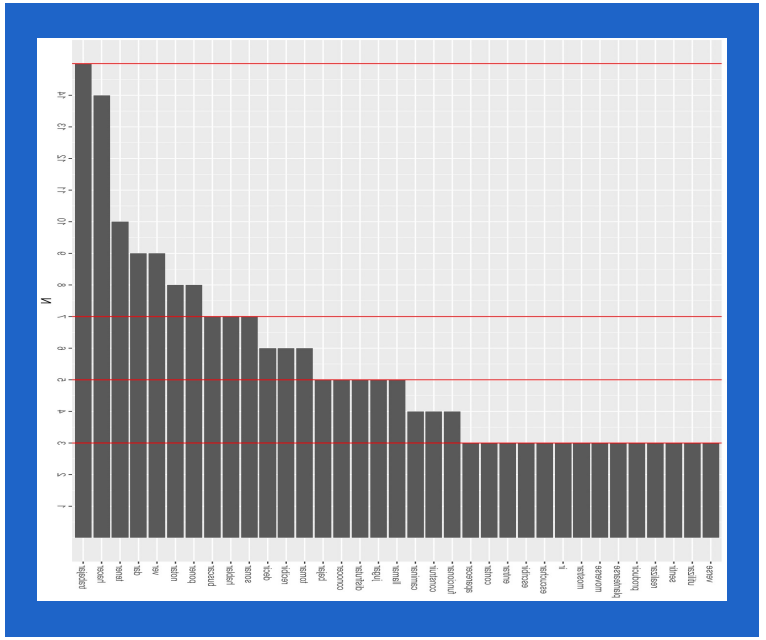
Auxiliary	Estimated token frequency	Nu. INF types	Nu. hapaxes	Nu. semantic classes*	Type/token ratio	Hapax/token ratio	Ratio of semantic classes	Hapax/type ratio
empezar	429.583	280	196	47	0,56	0,39	0,84	0,70
ponerse	60.728	179	119	37	0,36	0,24	0,66	0,66
lanzarse	7.476	215	136	42	0,43	0,27	0,75	0,63
meterse	1.648	210	140	42	0,42	0,28	0,75	0,67
romper	1.976	29	17	16	0,06	0,03	0,29	0,59
echarse	4242	16	7	8	0,03	0,01	0,16	0,44

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

CHOICE OF INFINITIVES



- 10 infinitives per auxiliary
  - 1 most frequent
  - 4 of different frequency (quartiles)
  - 1 least frequent (not hapax)
  - 2 hapaxes: semantically ”typical” and ”atypical”
  - 2 non-attested: semantically ”typical” and ”atypical”



GHENT UNIVERSITY

MATERIALS & PROCEDURE

- Authentic (simplified corpus sentences), 6-12 words, V Present / Preterit
- 6 auxiliaries x 10 infinitives = 60 inchoative sentences
- 140 filler sentences
  - 77 random sentences
  - 63 ”weird” sentences to increase rating variability

= 200 sentences in total

  - (Implicit) training block of 10 filler sentences
  - 10 blocks in random order (19 sentences each)
- 21 yes/no comprehension questions

- 7-point Likert scale unacceptable ○○○○○⊗○ acceptable
- 110 monolingual native speakers of European Spanish via https://prolific.co/

- Sociobiographic questionnaire
- Personality test (Big Five Inventory)

GHENT UNIVERSITY

RESEARCH QUESTIONS

- Are usage data predictive of acceptability?
  - INF frequency & AUX “productiveness” ←→ ratings
- To what extent can speakers extend constructions and under which circumstances are they willing to do so? (How do frequency and semantics interact?)
  - Are low-frequent INF more acceptable if they are semantically “typical”?
- Influence of participants’ individual characteristics?

DATA INSPECTION

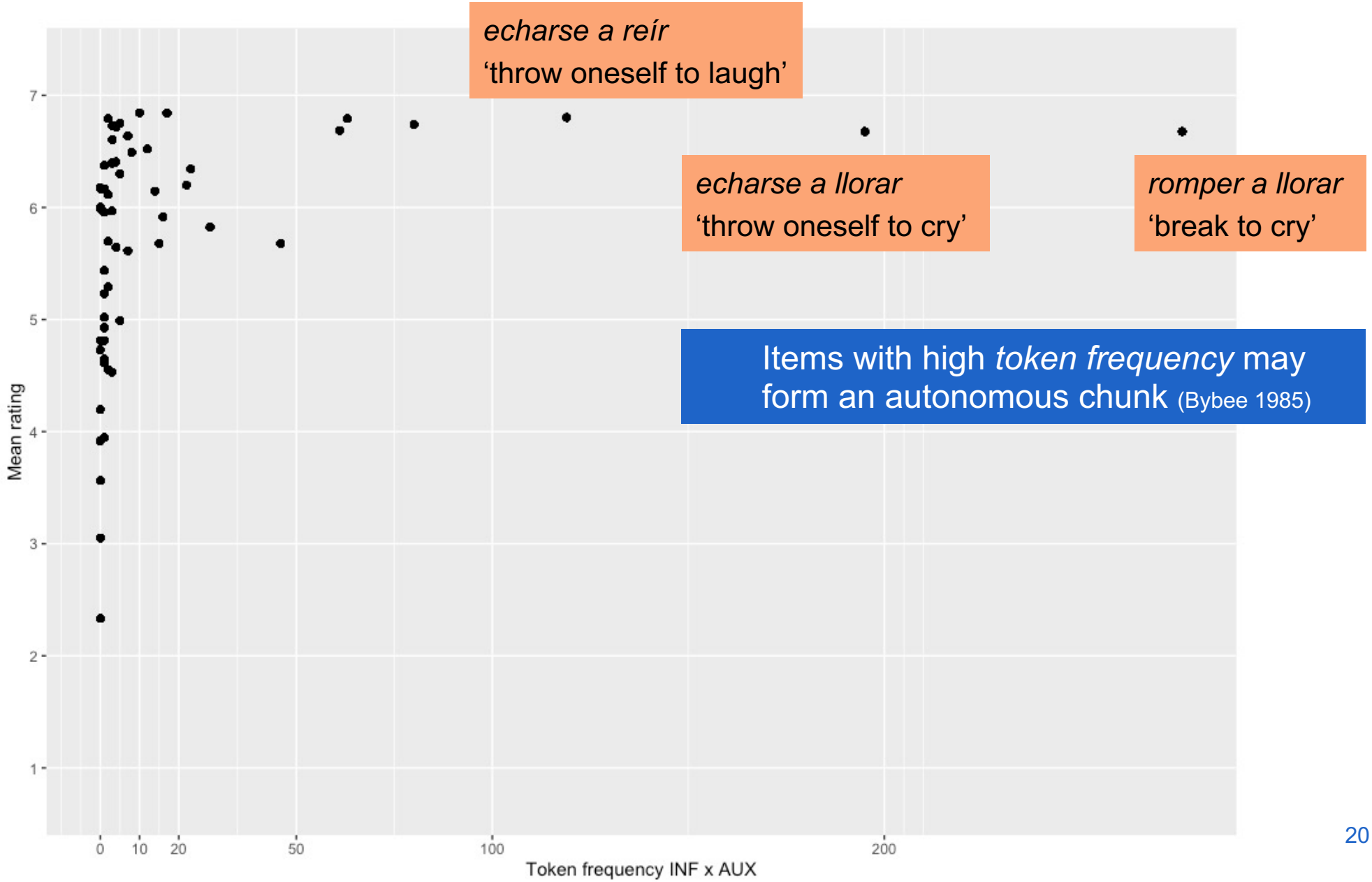
- Ratings of 96 participants (37 women, 59 men; mean age 29, SD 10.4)
  - 10 out of 110 were excluded due to accuracy <80% for yes/no questions
  - 4 were excluded because grew up in Latin America

Sentence type	Mean rating	SD
Random fillers	6.42	1.19
Inchoative sentences	5.65	1.81
“Weird” sentences	4.77	2.02

Auxiliary	Mean rating	SD
empezar	6.33	1.34
ponerse	6.02	1.57
meterse	5.66	1.63
lanzarse	5.57	1.73
echarse	5.50	1.95
romper	5.01	2.07

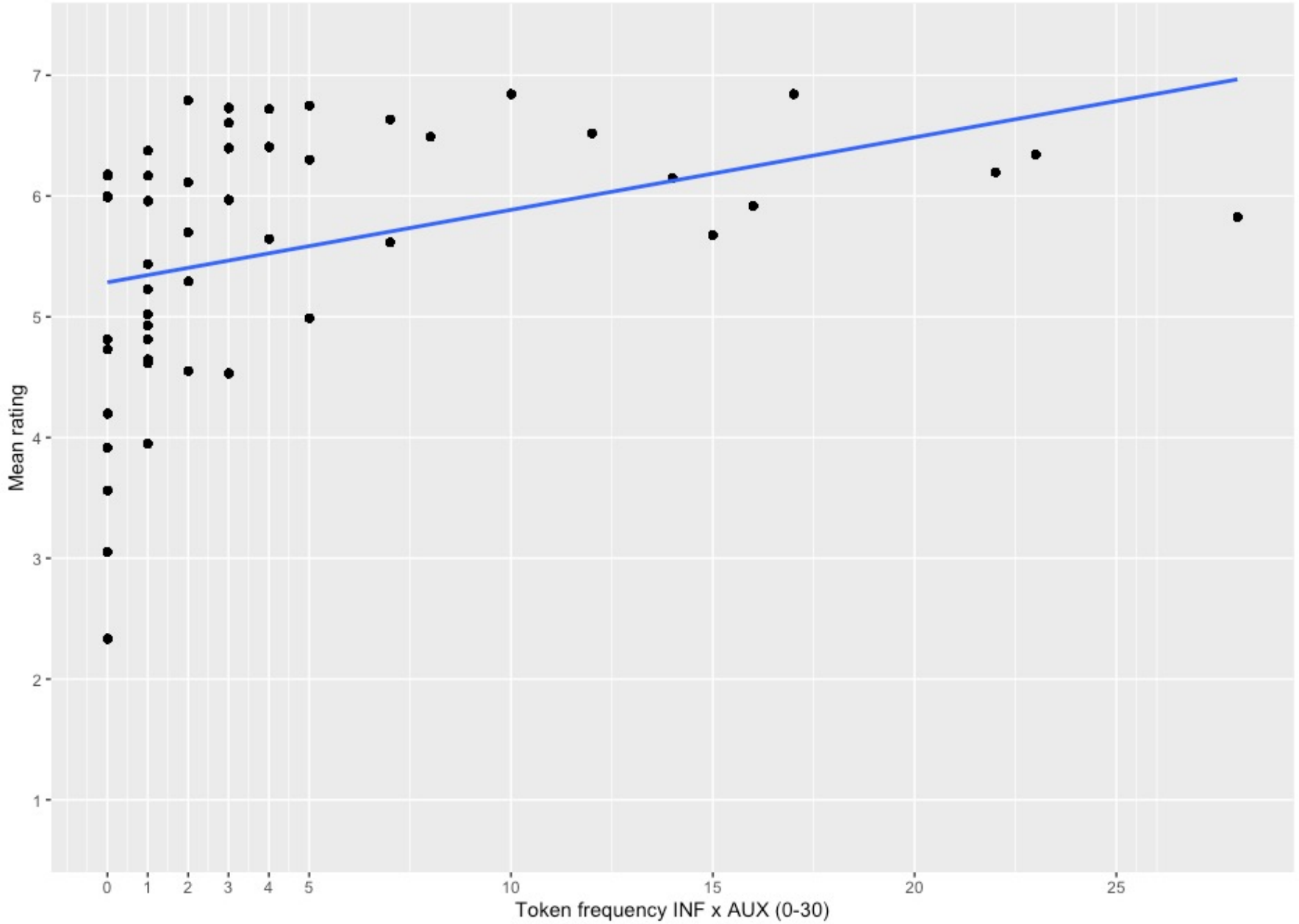
GHENT UNIVERSITY

TOKEN FREQUENCY INF & AUX



GHENT UNIVERSITY

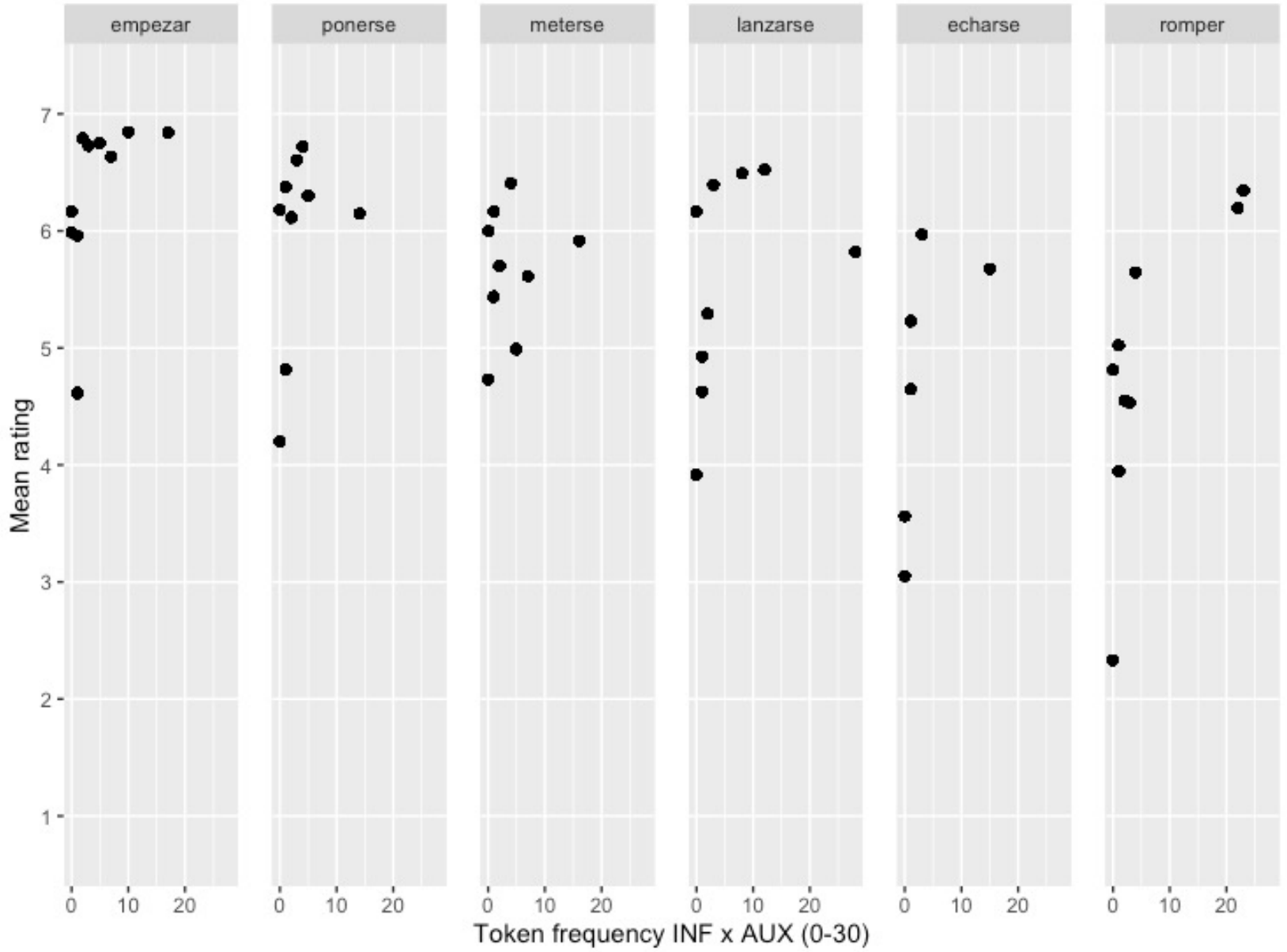
TOKEN FREQUENCY INF & AUX (0-30)



GHENT UNIVERSITY

TOKEN FREQUENCY INF & AUX (0-30)

Mean rating ↘



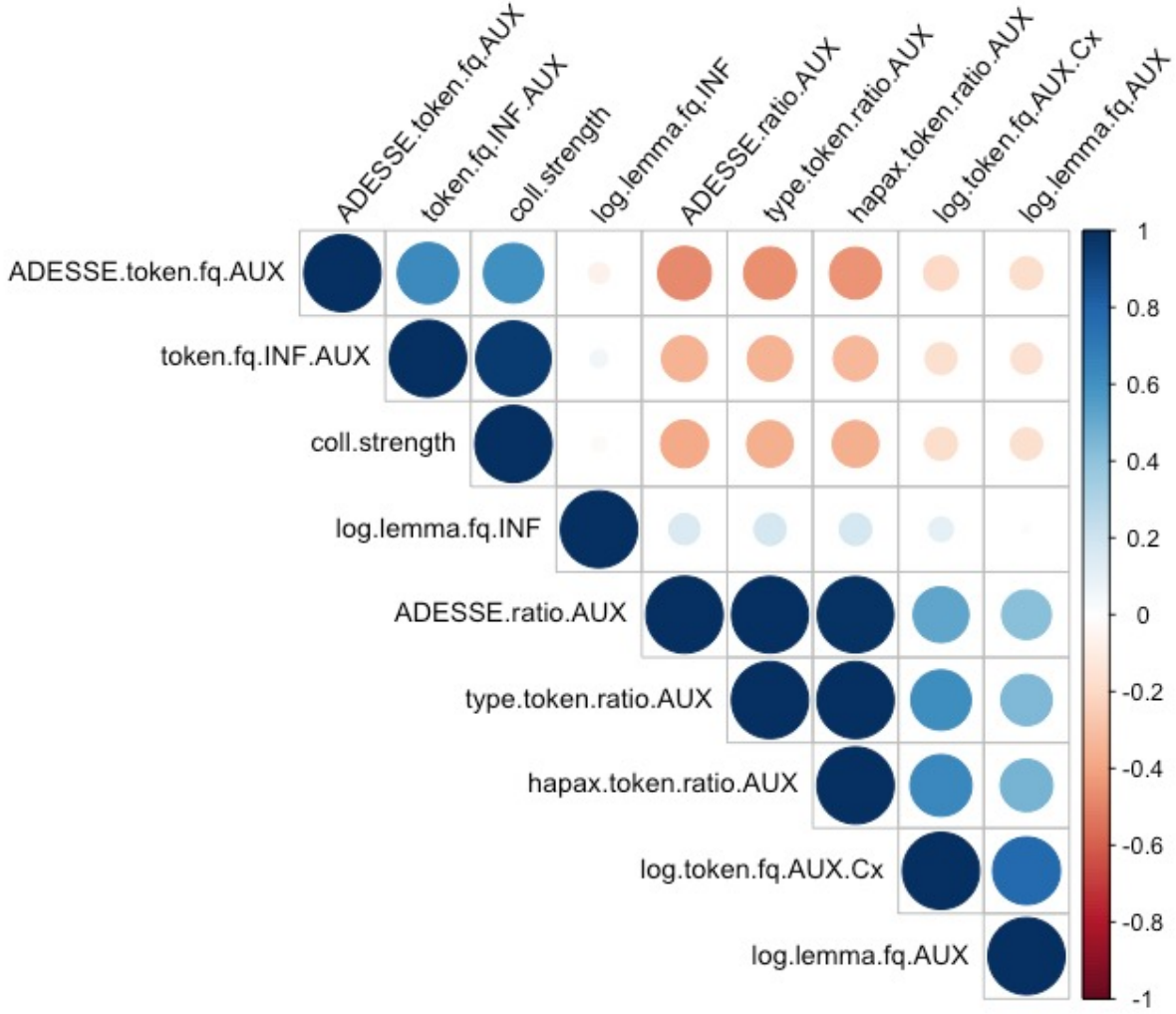
GHENT UNIVERSITY

POTENTIAL VARIABLES

Micro-level	
Token frequency INF x AUX	
Collocation strength INF x AUX	
ADESSE class token frequency	
Type/token ratio AUX	productivity measures AUX
Hapax/token ratio AUX	
ADESSE ratio AUX (semantic classes)	
Macro-level	
Estimated token frequency AUX	
Other	
Lemma frequency AUX in corpus	
Lemma frequency INF in corpus	

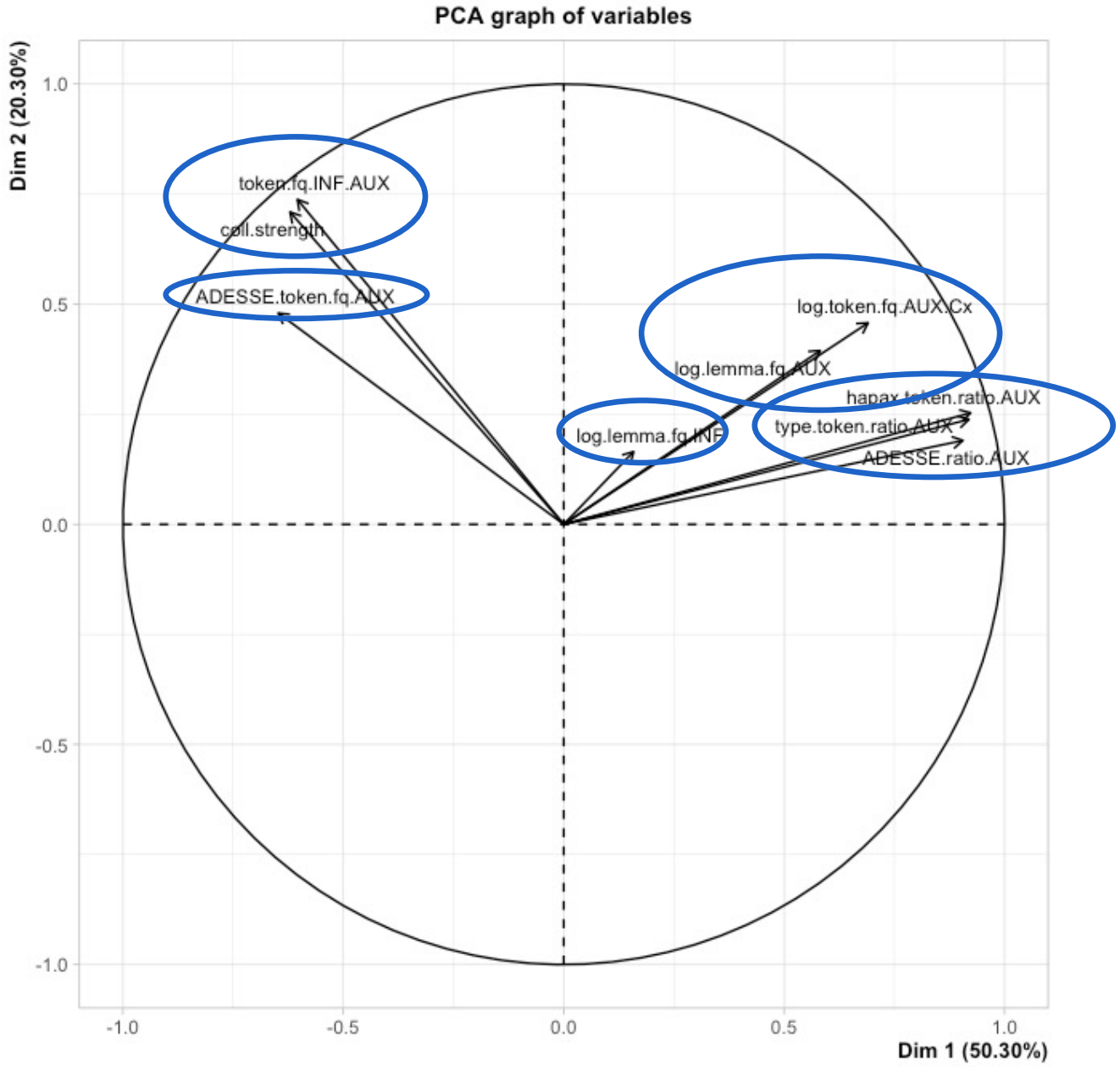
GHENT UNIVERSITY

CORRELATION PLOT (PCA)



GHENT UNIVERSITY

3 DIMENSIONS (PCA)



GHENT UNIVERSITY

POTENTIAL VARIABLES

Micro-level
Token frequency INF x AUX
Collocation strength INF x AUX
Type/token ratio AUX
Hapax/token ratio AUX
ADESSE ratio AUX (semantic classes)
ADESSE class token frequency
Macro-level
Estimated token frequency AUX
Other
Lemma frequency AUX in corpus
Lemma frequency INF in corpus

GHENT UNIVERSITY

LINEAR REGRESSION: FULL DATA SET

Fixed effects	Estimate	SE	*t-value*
(Intercept)	5.6857	0.1326	42.870
token.fq.INF.AUX	0.5041	0.1208	4.173*
type.token.ratio.AUX	0.4029	0.1325	3.041*
lemma.fq.AUX	0.2149	0.1258	1.708
AR ~ token.fq.INF.AUX + type.token.ratio.AUX + log(lemma.fq.AUX) + (1 | item) + (1 + token.fq.INF.AUX + type.token.ratio.AUX + log(lemma.fq.AUX) | participant)			

- Main effects of token frequency INF x AUX and of type/token ratio of the AUX
- No significant effects if interactions are included (?)

GHENT UNIVERSITY

LINEAR REGRESSION: LOW-FREQUENT INF

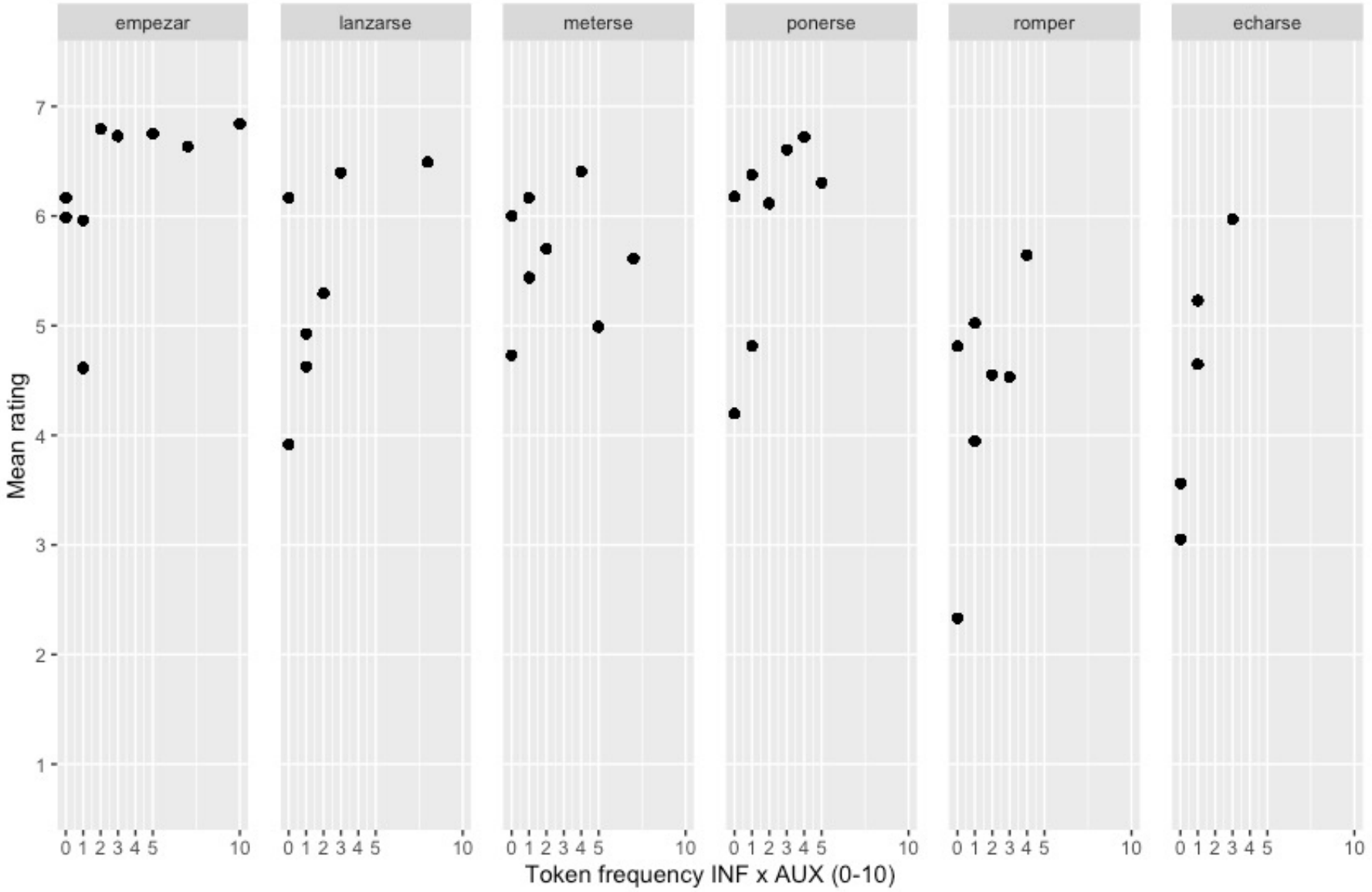
– Subset token frequency INF x AUX 0-10

Fixed effects	Estimate	SE	*t-value*
(Intercept)	5.5405	0.1438	38.530
token.fq.INF.AUX	0.6979	0.1681	4.152*
type.token.ratio.AUX	0.3559	0.1353	2.630*
lemma.fq.AUX	0.2220	0.1247	1.781
lemma.fq.INF	-0.1475	0.1503	-0.981
ADESSE.token.fq.AUX	0.0674	0.1250	0.539
token.fq.INF.AUX : type.token.ratio.AUX	-0.3329	0.1643	-2.026*
*AR ~ **token.fq.INF.AUX** * **type.token.ratio.AUX** + log(lemma.fq.INF) + log(lemma.fq.AUX) + ADESSE.token.fq.AUX + (1 | item) + (1 + token.fq.INF.AUX * type.token.ratio.AUX + log(lemma.fq.INF) + log(lemma.fq.AUX) + ADESSE.token.fq.AUX | participant)*			

GHENT UNIVERSITY

TOKEN FREQUENCY INF & AUX (0-10)

Type/token ratio ↘



GHENT UNIVERSITY

CONCLUSIONS & PLANS

✓ INF frequency & AUX “productiveness” ←→ ratings

? Interaction frequency x semantics

Future plans:

- Add more macro-level variables and semantic variables (VSS)
- Explore methods of analyzing ordinal data
- Include sociobiographic variables and personality traits
- Zoom in on the low end of frequency spectrum (eye-tracking)

GHENT UNIVERSITY

FACULTY OF PSYCHOLOGY AND EDUCATIONAL SCIENCES

FACULTY OF ARTS AND PHILOSOPHY

LANGUAGE PRODUCTIVITY @ WORK

Mariia Baltais
PhD student
Department of Experimental Psychology
mariia.baltais@ugent.be

“Language Productivity @ Work” project:
https://www.languageproductivity.ugent.be/

Thank you!

GHENT UNIVERSITY