FACULTY OF PSYCHOLOGY
AND EDUCATIONAL SCIENCES

LANGUAGE PRODUCTIVITY @ WORK

PRODUCTIVITY OF THE SPANISH INCHOATIVE CONSTRUCTION:

DOES SEMANTIC DISTANCE INFLUENCE EYE- TRACKING READING TIMES?

Mariia Baltais & Prof. Dr. Robert Hartsuiker
CogLing Days, Tilburg University, 9 December 2022

GHENT
UNIVERSITY

1. BACKGROUND: PRODUCTIVITY & SPANISH INCHOATIVE

2. DESIGN EYE-TRACKING

3. PRELIMINARY RESULTS

4. DISCUSSION

GHENT UNIVERSITY

PRODUCTIVITY IN CORPUS LINGUISTICS

- Syntactic productivity = a construction’s ability to attract new or existing lexical items (Barðdal, 2008)
  - Usage-based approach: productivity = continuum

- Corpus measures of productivity as the **range of attested lexical items** (Baayen, 2009)
  - Token frequency of (co-)occurrence
  - Type frequency
  - Hapax frequency, etc.

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 / echarse / …*** + 'a' + INF

Pedro	*se echa*	*a*	*reír*	*lit. 'Pedro throws himself to laugh'*
[Subj]	[V]	[Prep]	[INF]	**High co-occurrence frequency**

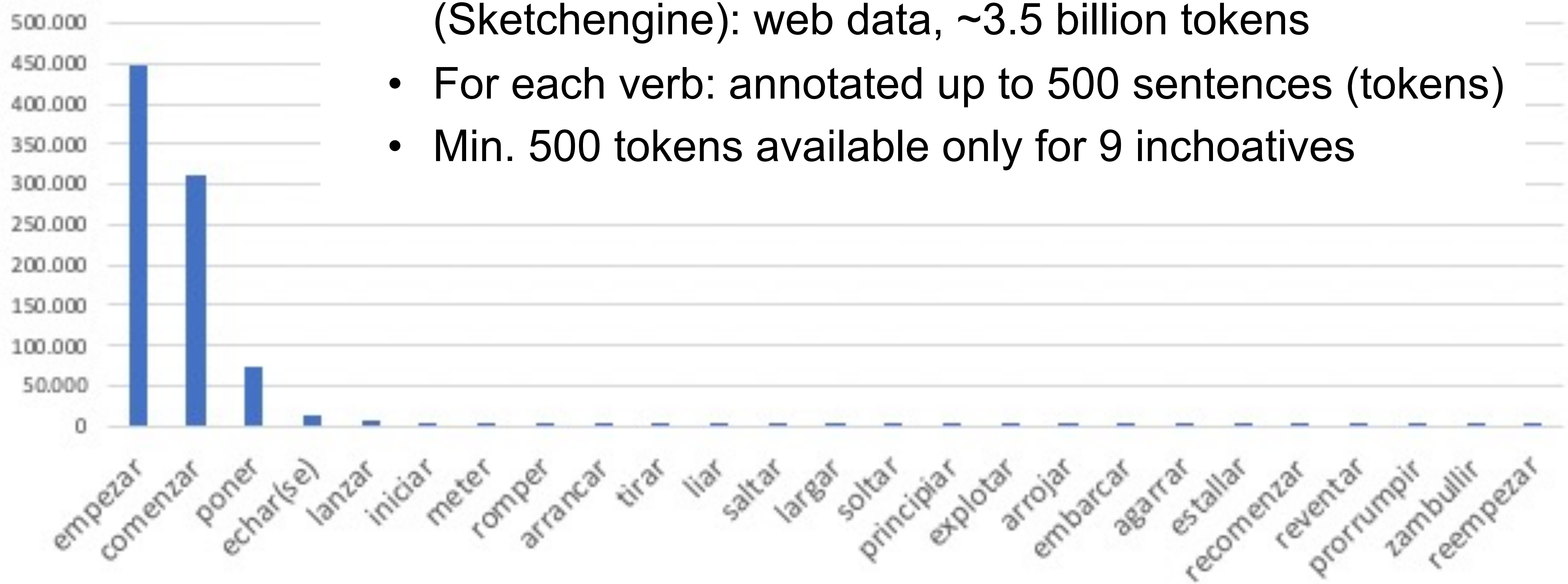
2) e.g., ***echarse*** + 'a' + **different INFs**

?	*Pedro*	*se echa*	*a*	*leer*	*lit. 'Pedro throws himself to read'*
	[Subj]	[V]	[Prep]	[INF]	**Low co-occurrence frequency**

GHENT UNIVERSITY

INCHOATIVE DATASET

- 25 inchoative verbs
- European Spanish subcorpus of esTenTen18 (Sketchengine): web data, ~3.5 billion tokens
- For each verb: annotated up to 500 sentences (tokens)
- Min. 500 tokens available only for 9 inchoatives



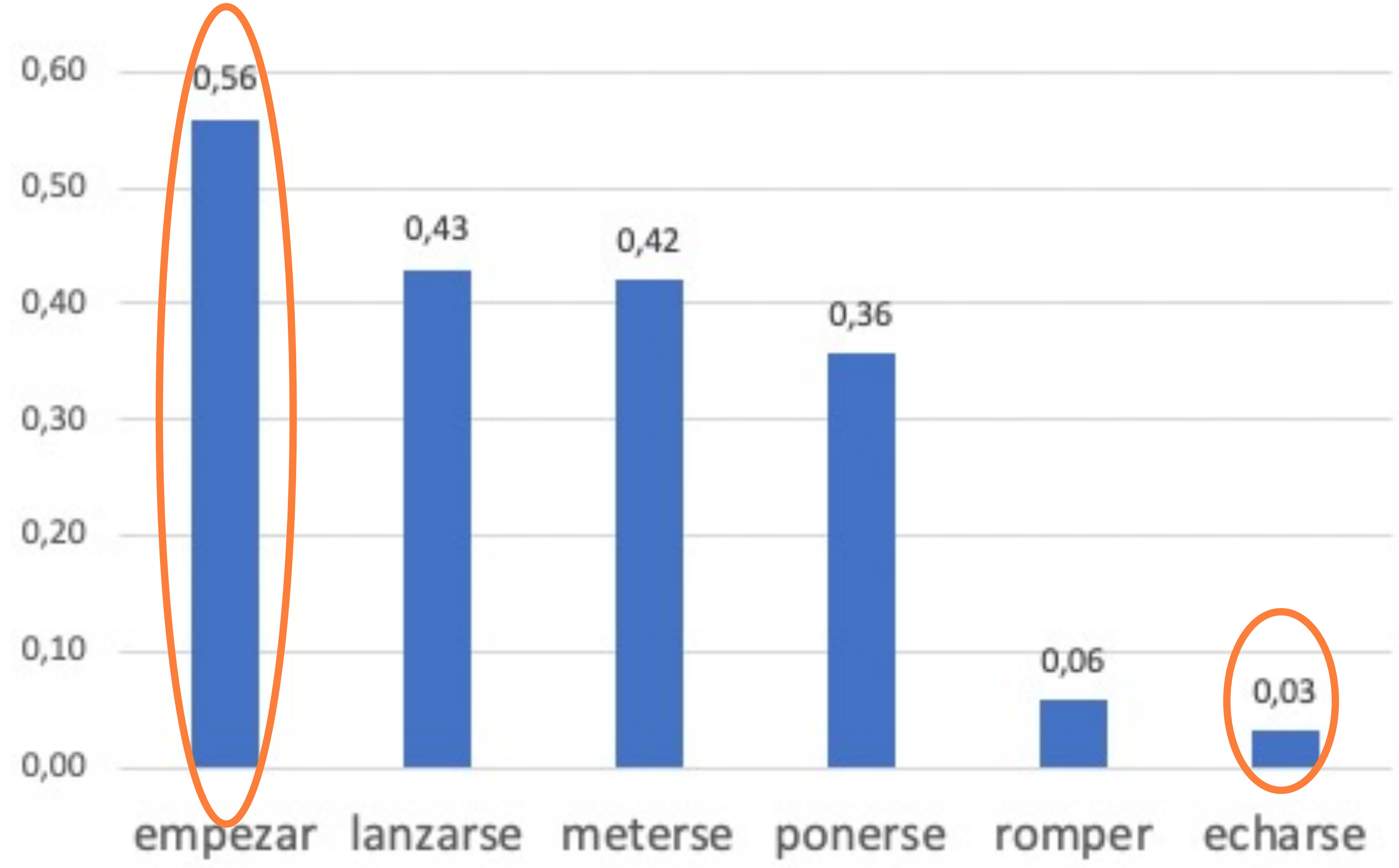
Estimated token frequency in the corpus

GHENT UNIVERSITY

EXAMPLES TYPE FREQUENCY

280 infinitives in the sample of 500 tokens

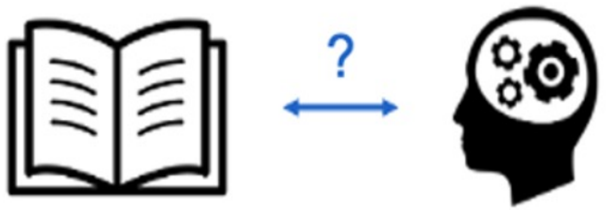
ser	17
trabajar	15
hacer	14
tener	10
dar	9
ver	9
notar	8
poner	8
buscar	7
hablar	7
sonar	7
decir	6
recibir	6
tomar	6
bajar	5
conocer	5
disfrutar	5
jugar	5
llamar	5
caminar	4
construir	4
funcionar	4
etc.	



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

CORPUS AND EXPERIMENTS



- Data sparseness problem (Keller, 2003)
- Constructions are ***extensible*** beyond closed-ended corpora (Barðdal, 2008)
- Speakers' individual characteristics
- Corpus measures of productivity ↔ experimental data?
  - Acceptability ratings: grammaticality-frequency discrepancy (Divjak, 2017)
  - Other experimental research techniques?

GHENT UNIVERSITY

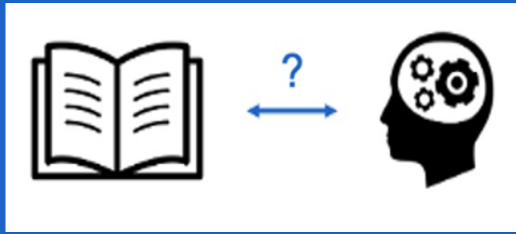
"LANGUAGE PRODUCTIVITY AT WORK"

"Language Productivity @ Work" project

FACULTY OF ARTS AND PHILOSOPHY + FACULTY OF PSYCHOLOGY AND EDUCATIONAL SCIENCES

Combine corpus-based and experimental data to answer the question…

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



…both in comprehension (this study) and in production (poster Anna Jessen)

PREVIOUS STUDY: ACCEPTABILITY RATINGS

- 96 native speakers of European Spanish

- Corpus measures of productivity were predictive of acceptability ratings

- Infrequent infinitives were more acceptable if belonged to a frequently attested semantic class
  - http://adesse.uvigo.es/: creation, perception, displacement, physiology…

GHENT UNIVERSITY

PRODUCTIVITY AND SEMANTICS

- Occurrence of a novel item ↔ its **semantic similarity to previous usage**
    - Argument structure Cxs with novel verbs in Icelandic (Barðdal, 2008)
    - Acceptability ratings
        - Infrequent uses of Spanish V-Adj copular Cxs with verbs of becoming (Bybee & Eddington, 2006)

GHENT UNIVERSITY

DISTRIBUTIONAL SEMANTICS

- Semantic distance between items in a Cx measured through their co-occurrence frequency with other words in the corpus (Erk, 2012; Perek, 2018)

"You shall know a word by the company it keeps"
(Firth, 1957: 11)

- Makes the analysis data-driven and automatic
- A more objective way of grouping lexical items
- Drawback: ignores polysemy

- E.g., Suttle and Goldberg (2011)
  - Acceptability ratings
  - But – artificial language, metalinguistic off-line task

EYE-TRACKING DURING READING

- On-line method: measures participants' unconscious and automatic responses to language stimuli as they unfold
- “Early” measures ↔ lexical access: first fixation duration, gaze duration, probability of skipping, etc.
- “Late” measures ↔ syntactic processing and semantic integration: regression path duration, probability of re-reading, total reading time, etc.

RESEARCH QUESTIONS

- Does co-occurrence frequency influence processing cost when semantic distance is kept constant?

- **Does semantic distance influence processing cost when co-occurrence frequency is kept constant?**

- Participants' individual characteristics?

STUDY DESIGN

- Semantic distance = distance between the semantic vector of the infinitive and the centroid vector of the inchoative (its 10 most frequently attested INFs)

- Three conditions:
  1) BASELINE: highly frequent, semantically “close” INF – baseline condition
  2) CLOSE: low-frequent, semantically “close” INF
  3) DISTANT: low-frequent, semantically “distant” INF

- 15 inchoatives x 3 minimal triplets = 45 triplets

EXAMPLE STIMULI

		n-1	n	n+1	
BASELINE					
	Manuela	se arrancó a	**tocar**	una pieza	de violín
	Manuela	*started to*	***play***	*a violin piece*	
CLOSE					
	Manuela	se arrancó a	**imitar**	una pieza	de arte
	Manuela	*started to*	***imitate***	*a piece of art*	
DISTANT					
	Manuela	se arrancó a	**mover**	una pieza	de ajedrez
	Manuela	*started to*	***move***	*a chess piece*	

	Co-occurrence frequency	Semantic distance
BASELINE	12	0,52
CLOSE	2	0,56
DISTANT	2	0,72

- Critical region of interest: INF
- INFs matched on length in letters, lemma fq in the corpus

GHENT UNIVERSITY

PROCEDURE

- Three presentation lists: 45 critical + 185 fillers = 230 sentences each
- Practice block in the beginning, 36 ‘yes/no’ comprehension questions
- Sociobiographic questionnaire, BFI-2 personality test
- 1-hour session

PARTICIPANTS (SO FAR)

- 36 native speakers of European Spanish (end goal: 60 participants)
- 3 excluded (parents from Latin America, dyslexia)
- No one excluded based on comprehension questions (accuracy > 85%)

→ 33 participants

  - Mean age: 22.4 y, SD: 2.93
  - 11 m, 22 f

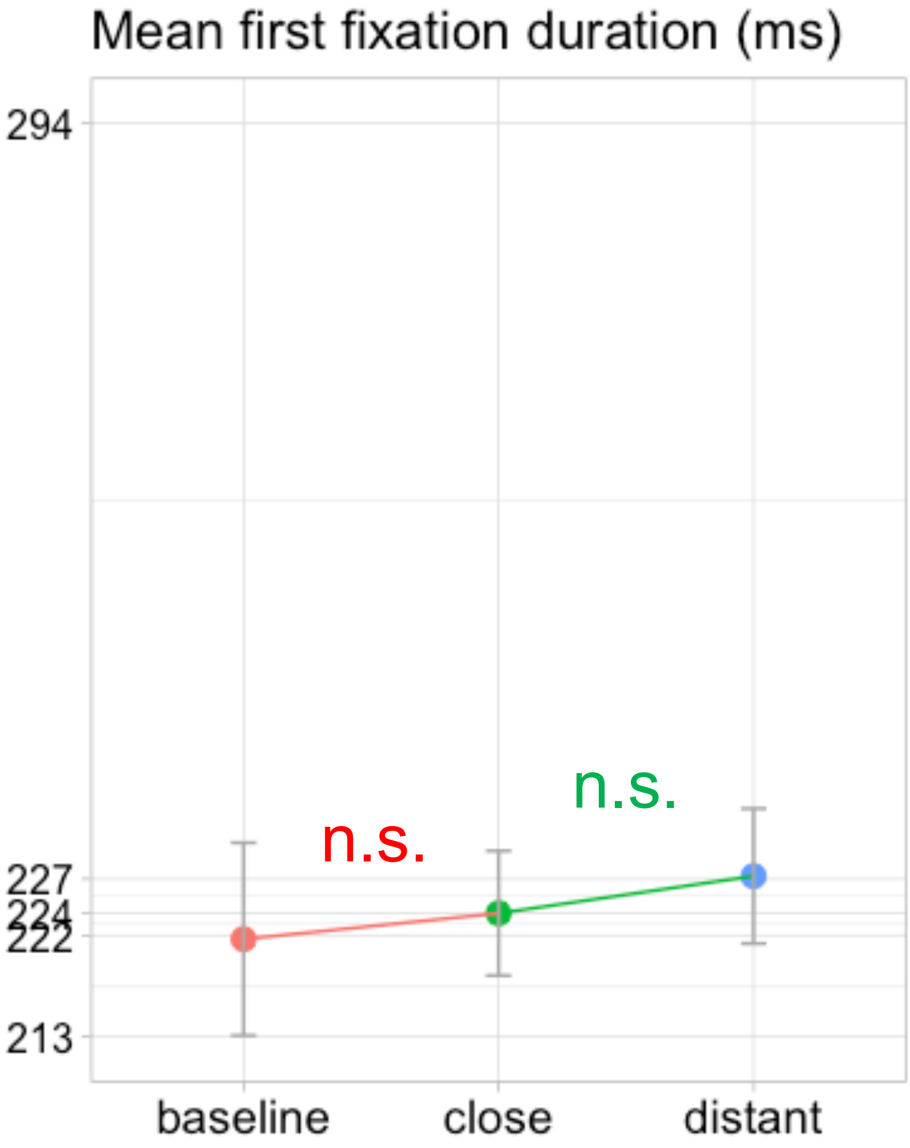
GHENT UNIVERSITY

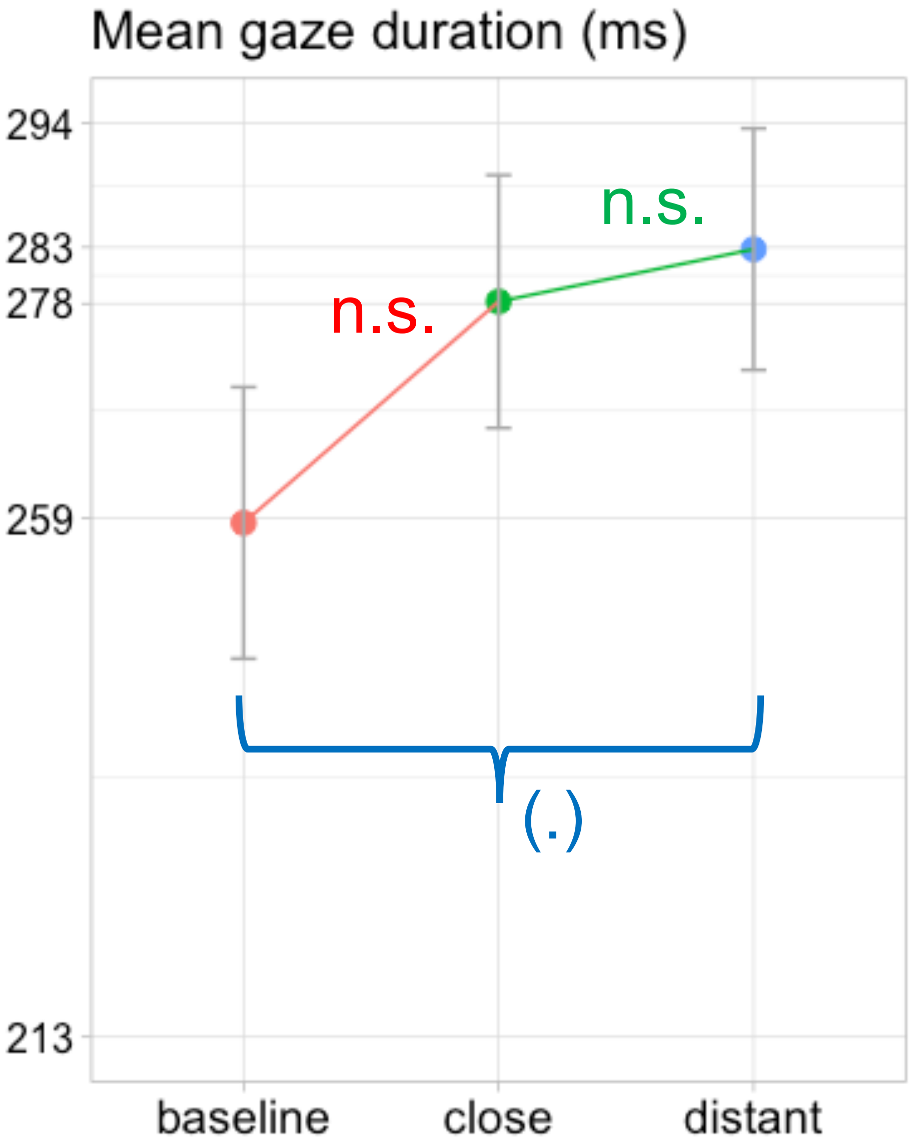
PRELIMINARY ANALYSIS

- 6 triplets were excluded → 39 triplets analyzed

- Generalized linear mixed models
- Tukey method for multiple comparisons

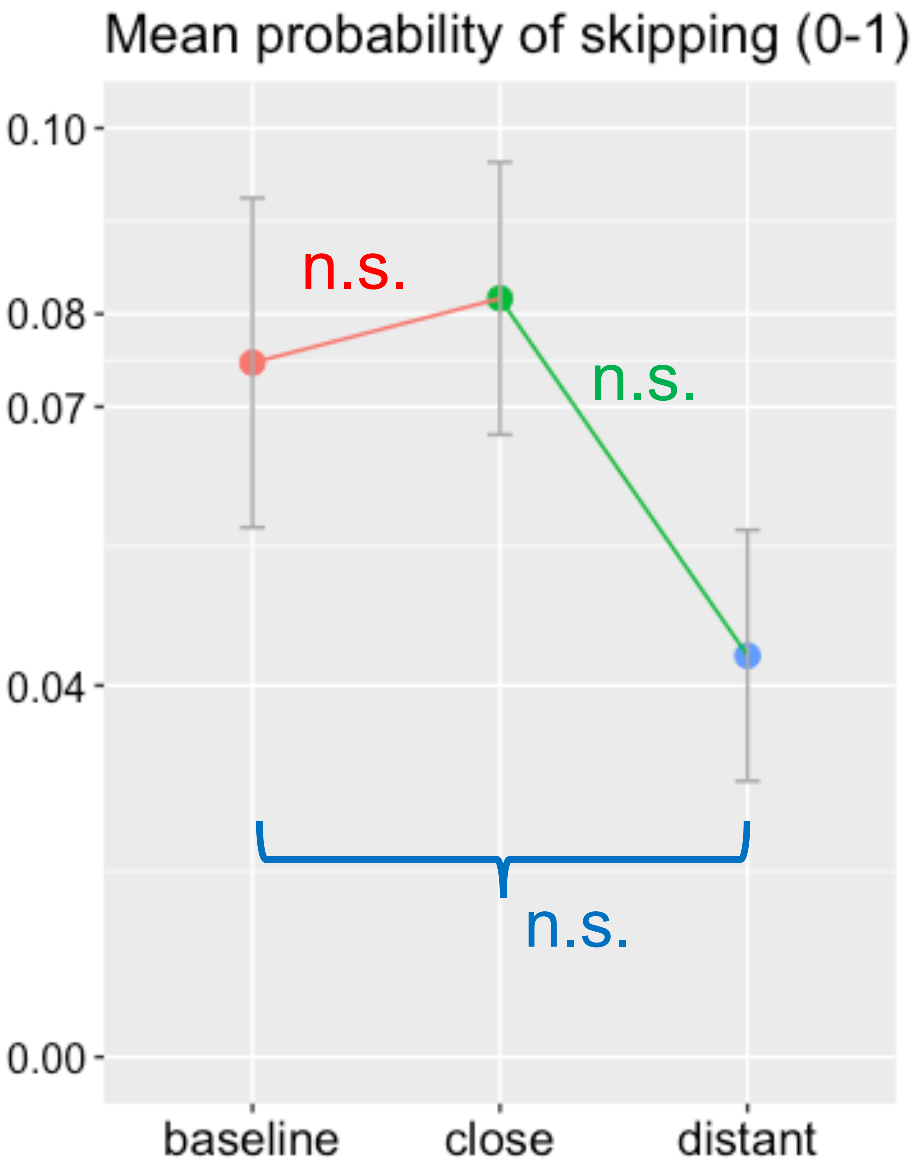
EARLY MEASURES





(Error bars indicate the standard error of the mean)

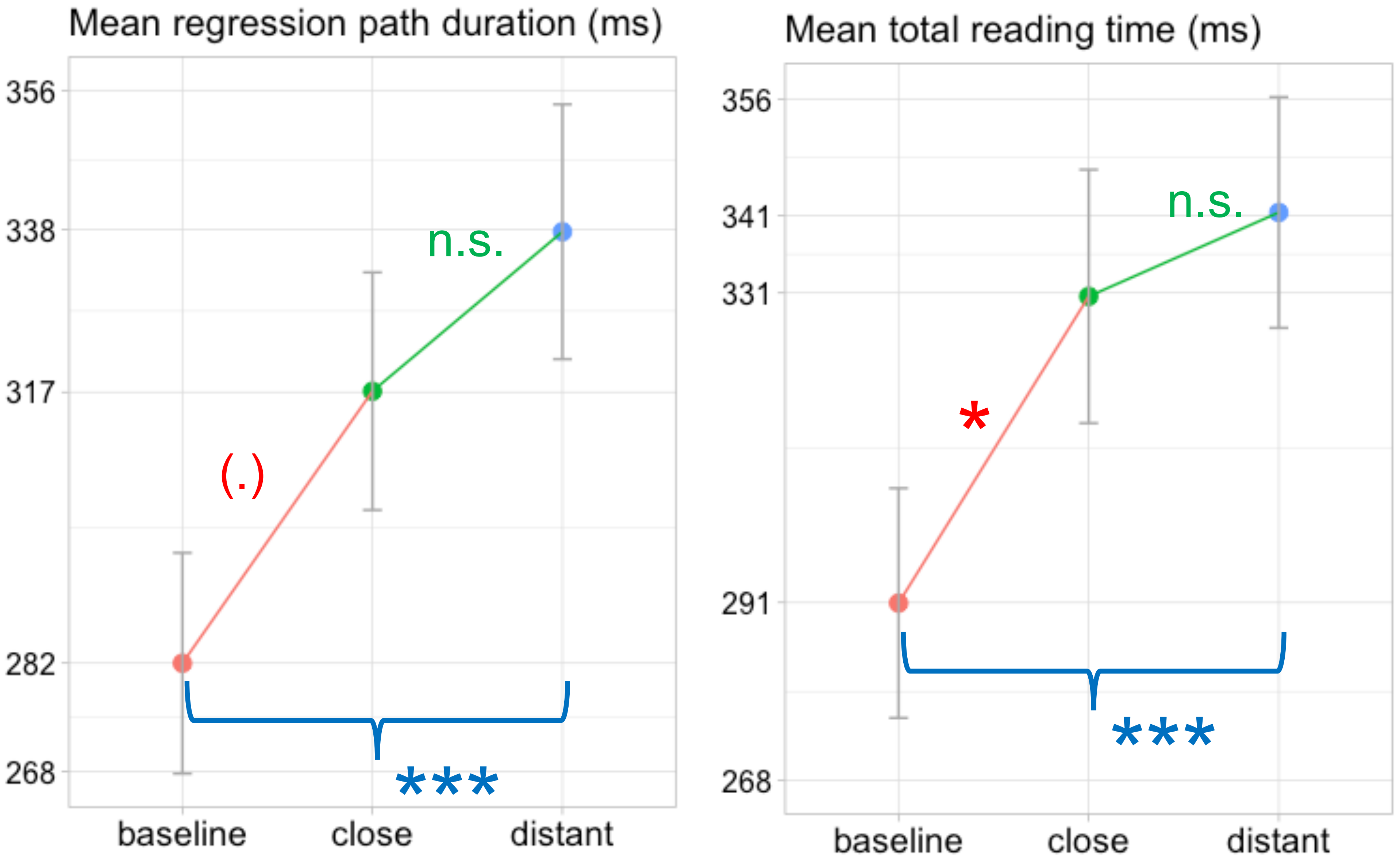
EARLY MEASURES



(Error bars indicate the standard error of the mean)

GHENT UNIVERSITY

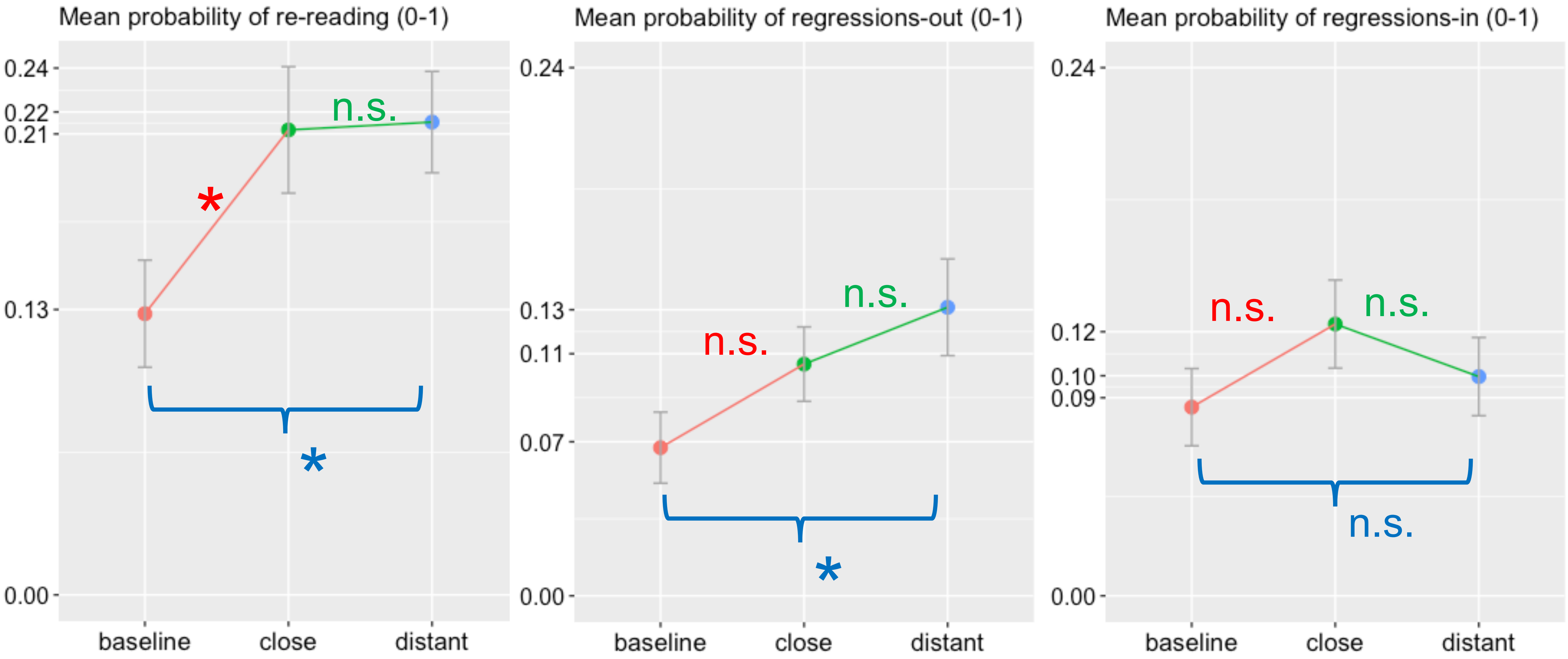
LATE MEASURES



(Error bars indicate the standard error of the mean)

GHENT UNIVERSITY

LATE MEASURES



GHENT UNIVERSITY

(Error bars indicate the standard error of the mean)

SUMMARY

- At short semantic distance, co-occurrence frequency influenced processing cost in late reading measures

- At low co-occurrence frequency, semantic distance didn't have an independent effect (cf. numerical trend in skipping)

- Semantic distance seems to "boost" the effect of co-occurrence frequency (e.g., regression path duration, regressions-out)

- No spillover effect so far

DISCUSSION

- Productivity & semantics: on-line method, natural language

- It seems plausible that semantic similarity would be important
  - We haven’t found evidence (so far)
  - Is the measure of semantic distance inadequate? E.g., polysemy

DISCUSSION

- Data collection still on-going; more differences might be revealed
- We applied a strict correction for multiple comparisons

- Not a 2x2 design (interaction co-occurrence frequency and semantics?)
  - Data suggest that semantics can "boost" the effect of frequency
  - Interaction could be explored with an artificial language

- Future analysis: interactions with individual variables (language background, personality traits…)

MANY THANKS TO…

LANGUAGE PRODUCTIVITY @ WORK





UNIVERSIDAD NEBRIJA


GHENT UNIVERSITY

Mariia Baltais, Prof. Dr. Robert Hartsuiker
Department of Experimental Psychology

Corresponding author:
mariia.baltais@ugent.be

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

Thank you!

CHOICE OF INCHOATIVES

	Inchoative	Nu. triplets	Sample size
1	comenzar	3	500
2	empezar	3	500
3	lanzarse	3	500
4	meterse	3	500
5	iniciar	3	350
6	ponerse	3	500
7	liarse	3	500
8	saltar	3	234
9	principiar	3	140
10	largarse	3	176
11	arrancarse	3	283
12	tirarse	3	81
13	romper	3	500
14	soltarse	3	95
15	echar	1	500
	echarse	2	500

(sample size > 30)

CHOICE OF HIGH-FREQUENT INF

- Criterion: type/token ratio of the inchoative at the maximal common sample size 81
  - TTR > 0.3 → choose from 15% most frequent
  - TTR < 0.3 → choose from 5% most frequent
  - Exception: *largarse* (TTR = 0.5, extended the threshold to 10% most frequent to have 3 INF)

GHENT UNIVERSITY

VSS

- Locate the verbs in a large corpus and count, for all the tokens found, the frequency of co-occurrence with other words within a set context window (e.g., five words to the left and five words to the right) → co-occurrence matrix, with the set of words under consideration as rows, the collocates as columns, and the co-occurrence frequency in each cell
- Transformations to the co-occurrence matrix: weighting and dimensionality reduction
- Each row of the final matrix is a vector representing the distributional profile of a given word. Under the assumption that semantic distance between words is a function of distributional differences, similarity between rows approximates semantic similarity (the cosine measure)

GENERAL DESIGN EYE-TRACKING

- Practice block (6 sentences)
- 3 experimental blocks (230 sentences in total) with 2 breaks in between:
  - 32 (16 x condition) Eyetr_Prod,
  - 45 (15 x condition) Eyetr_FqSem,
  - 40 (20 x condition) Eyetr_Perception
  - 113 fillers

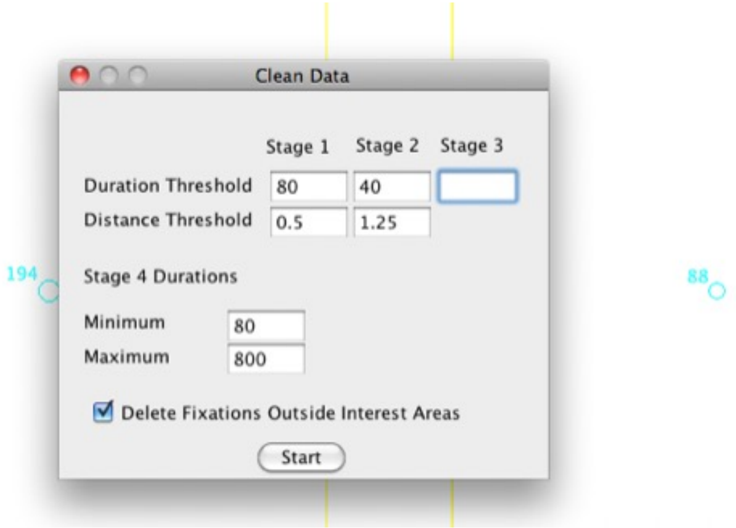
- 36 comprehension questions, 18 “no”, 18 “yes”
- 6 lists + 6 reversed lists

- Calibration in the beginning & after each break

DATA CLEANING

- Exclude trials with horizontal drift / didn't finish reading / started from right to left
- Exclude trials with blinks betw. 2 fix. on the critical region / if skipped because of blinking
- Correct vertical drift (move fix. up/down)
- ”Yellow sticker” issue (first 8 part didn't have it)

- Automatically remove very short (<80 ms) and very long (>800 ms) fixations, fixations outside interest areas
  - Modified default cleaning procedure in Data Viewer

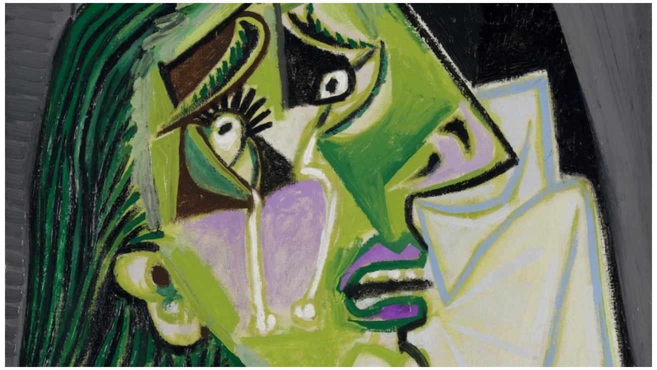


→ **1226 datapoints** (4.7% removed during cleaning)

Baseline	407
Close	416
Distant	403

GHENT UNIVERSITY

ISSUE: CLOZE PROBABILITY



La viuda se echó a…

The widow threw herself to…

"In research on the role of lexical predictability in language comprehension, predictability is generally defined as the probability that a word is provided as a sentence continuation in the cloze task (Taylor, 1953), in which subjects are asked to guess the next word of a sentence." (Staub et al., 2015)

POST-HOC CLOZE TASK

- Prolific: 35 participants (43 recruited)
  - 17 m, 17 f, 1 other
  - Mean age: 25.3 y, SD: 3.26

- 3 options:
  - Remove 9 triplets (t-test)
  - Remove 3 triplets
  - Remove nothing and include cloze probability as covariate

- Decision: remove 3 + include cloze probability

	TripletID	diffbaseatyp	Subject	baselineinf
1	echar1	85.71	La viuda	llorar
2	saltar34	31.43	El abogado	defender
3	meter15	11.43	El biólogo	estudiar
4	soltar19	8.57	El actor	hablar
5	echar3	5.71	El conquistador	correr
6	romper5	5.71	El bebé	reír
7	soltar20	5.71	Mi hija	andar
8	arrancar27	2.86	La ingeniera	hacer
9	lanzar16	2.86	El rey	conquistar
10	largar44	2.86	Carmen	ver
11	largar45	2.86	Roberto	hacer
12	poner12	2.86	Joaquín	buscar
13	romper6	2.86	La concursante	sudar
14	tirar22	2.86	Tatiana	nadar
15	arrancar25	0.00	Manuela	tocar
16	arrancar26	0.00	El hombre	aplaudir
17	comenzar40	0.00	El pintor	desarrollar

GHENT UNIVERSITY

ISSUE: PRINCIPIAR

Eleonora **principió** a ser cada día más reconocida por su talento.
Eleonora started to be increasingly recognized for her talent.

– Several participants (both eye-tracking and cloze task) reported that they were not familiar with this verb

Inchoative	Sample size	TTR	TTR at s.s.81	HTR	ADESSE ratio (out of 57)
principiar	140	0,76	0,83	0,61	0,60

– Decision: remove these 3 triplets

→ **39 triplets instead of 45**

GHENT UNIVERSITY

ANALYSIS

- 6 triplets were excluded → 39 triplets analyzed
- Generalized linear mixed models: condition and cloze probability as fixed effects, item and participant as random effects

Condition	First fixation duration (SD)	Gaze duration (SD)	Skipping(SD)	Re-reading (SD)
Baseline	222 (49)	259 (69)	7.5% (10.2%)	12.8% (14%)
Close	224 (32)	278 (64)	8.2% (8.4%)	21.2% (16.5%)
Distant	227 (34)	283 (62)	4.3% (7.8%)	21.5% (13.3%)

Condition	Regression path duration (SD)	Regressions out (SD)	Regressions in (SD)	Total reading time (SD)
Baseline	282 (82)	6.73% (9.28%)	8.57 % (10.08 %)	291 (85)
Close	317 (88)	10.53% (9.71%)	12.35 % (11.5 %)	331 (94)
Distant	338 (95)	13.11% (12.61%)	9.96 % (10.21 %)	341 (86)

FIRST FIXATION DURATION: N

Fixed effects	Estimate	SE	z-value	Pr(>|z|)
Close vs baseline	3.592	5.793	0.620	0.809
Close vs distant	-1.748	6.195	-0.282	0.957
Distant vs baseline	5.340	5.808	0.919	0.628
Formula: FFD ~ condition + c.(clozeprob) + (1 | item) + (1 | participant)				

GHENT UNIVERSITY

GAZE DURATION: N

Fixed effects	Estimate	SE	z-value	Pr(>|z|)
Close vs baseline	16.465	9.345	1.762	0.1822
Close vs distant	-3.561	10.069	-0.354	0.9333
Distant vs baseline	20.026	9.190	2.179	0.0746 .
Formula: GD ~ condition + c.(clozeprob) + (1 | item) + (1 | participant)				

GHENT UNIVERSITY

SKIPPING: N

Fixed effects	Estimate	SE	z-value	Pr(>|z|)
Close vs baseline	0.0276	0.3562	0.077	0.997
Close vs distant	0.7971	0.3946	2.020	0.107
Distant vs baseline	-0.7694	0.4096	-1.879	0.144
Formula: rate ~ condition + c.(clozeprob) + (1 | item) + (1 | participant)				

GHENT UNIVERSITY

REGRESSION PATH DURATION: N

Fixed effects	Estimate	SE	z-value	Pr(>|z|)
Close vs baseline	24.67	10.56	2.336	0.0501 .
Close vs distant	-17.43	12.53	-1.391	0.3421
Distant vs baseline	42.10	10.12	-4.161	<0.001 ***
Formula: RPD ~ condition + c.(clozeprob) + (1 | item) + (1 | participant)				

GHENT UNIVERSITY

TOTAL READING TIME: N

Fixed effects	Estimate	SE	z-value	Pr(>|z|)
Close vs baseline	29.86	11.64	2.564	0.0275 *
Close vs distant	-10.43	13.44	-0.776	0.7162
Distant vs baseline	40.28	11.23	3.587	<0.001 ***
Formula: TRT ~ condition + c.(clozeprob) + (1 | item) + (1 | participant)				

GHENT UNIVERSITY

RE-READING: N

Fixed effects	Estimate	SE	z-value	Pr(>|z|)
Close vs baseline	0.593	0.232	2.559	0.0284 *
Close vs distant	0.0006	0.213	0.003	1
Distant vs baseline	0.592	0.238	2.485	0.0344 *
Formula: rate ~ condition + c.(clozeprob) + (1 | item) + (1 | participant)				

GHENT UNIVERSITY

REGRESSIONS OUT: N

Fixed effects	Estimate	SE	z-value	Pr(>|z|)
Close vs baseline	0.5680	0.3385	1.678	0.2131
Close vs distant	-0.2821	0.3029	-0.931	0.6197
Distant vs baseline	0.8501	0.3450	2.464	0.0364 *
Formula: rate ~ condition + c.(clozeprob) + (1 | item) + (1 | participant)				

GHENT UNIVERSITY

REGRESSIONS IN: N

Fixed effects	Estimate	SE	z-value	Pr(>|z|)
Close vs baseline	0.4301	0.2797	1.538	0.273
Close vs distant	0.2781	0.2680	1.038	0.553
Distant vs baseline	0.1520	0.2949	0.515	0.864
Formula: rate ~ condition + c.(clozeprob) + (1 | item) + (1 | participant)				

GHENT UNIVERSITY