

BIG SYNTAX DOESN'T WANT YOU TO KNOW

MAL is universal? Hardly.

The Menzerath–Altmann law on all 180 UD treebanks

Pegah Faghiri^{1,2} Kim Gerdes² Sylvain Kahane^{1,3}
¹Modyco, CNRS & Paris Nanterre University ²LISN, CNRS & Paris-Saclay University ³Institut Universitaire de France

YOUR LANGUAGE SECRETS ARE HERE

→

typometrics.elizia.net/
menzerath

Interactive maps & curves

HOW WE COMPUTE MAL

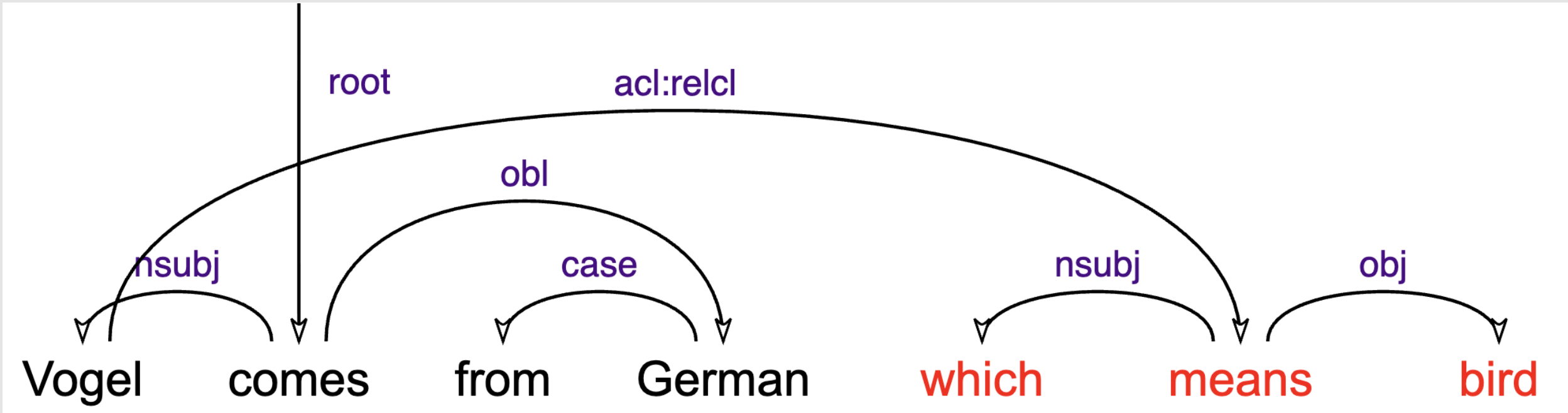
Real constituents only. Fake dependents eliminated!

The **Menzerath–Altmann law** (MAL) predicts that “the greater the number of components of a given unit, the shorter their respective length” (Menzerath, 1928). In the verbal domain: as the number of dependents n grows, their average size decreases.

Altmann (1980) formalised this as a **power law**: $MAL_n \approx n^{-\beta}$, i.e. $\log(MAL_n) \approx -\beta \cdot \log(n)$. The slope β measures the **MAL effect**. Previous studies: Macutek et al. (2017) for Czech; Tanaka-Ishii (2021) for UD 2.3 without filtering.

NEW! Our data extraction is filtered: we exclude `punct`, `discourse`, `parataxis`, `conj`, `cc`, `vocative`, `aux`, `compound:prt`, `mark`, `case` from the verbal construction. Only **real syntactic constituents** are retained.

What about “bastards”? A constituent’s subtree may spill across the verb. We count the spilled part as a separate dependent of the verb:



The relative clause “which means bird” (**acl:reld**) syntactically depends on ‘Vogel’ (left of verb) but appears right of verb – a **bastard**. It is counted as a right-side constituent of **comes**. Source: `en_gum`.

NEW! LEFT VS. RIGHT

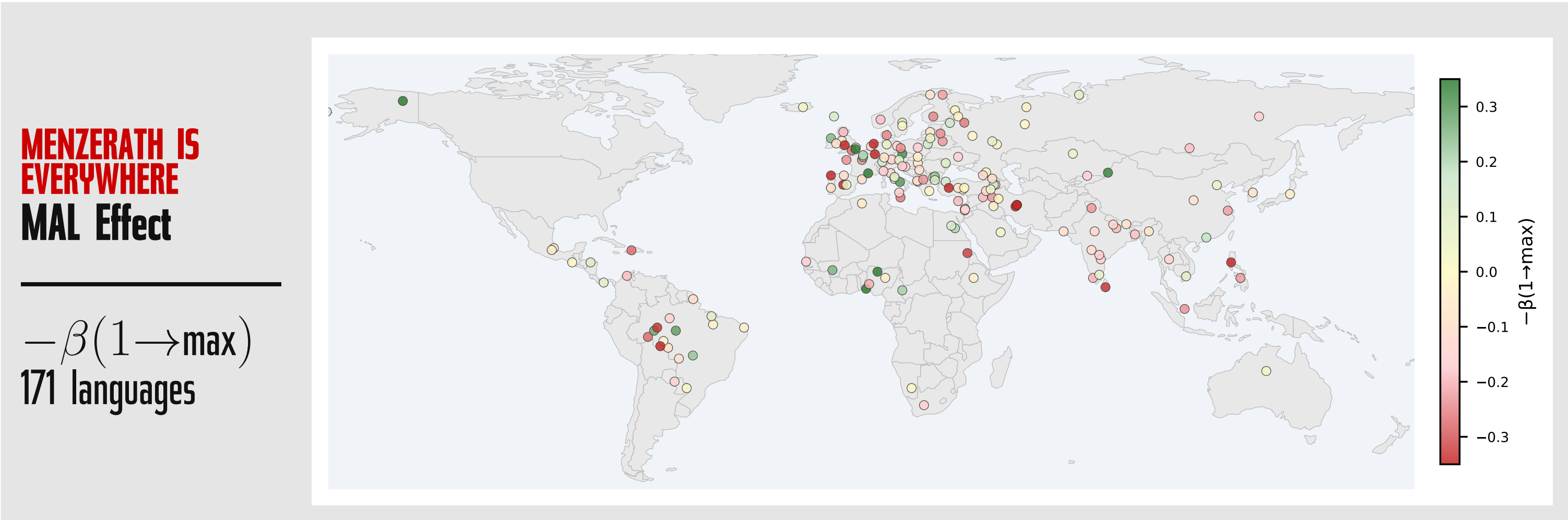
Two sides of the verb, two different worlds

MAL
We compute three variants of the MAL effect. $MAL_n(L)$ = mean size of constituents in all verbal constructions of L with n constituents (left + right combined). Standard MAL as in the literature.

LMAL & RMAL NEW!
 $LMAL_n(L)$ = mean size of constituents in constructions with n **preverbal** (left) constituents, regardless of the right side.
 $RMAL_n(L)$ = mean size of constituents in constructions with n **postverbal** (right) constituents, regardless of the left side.

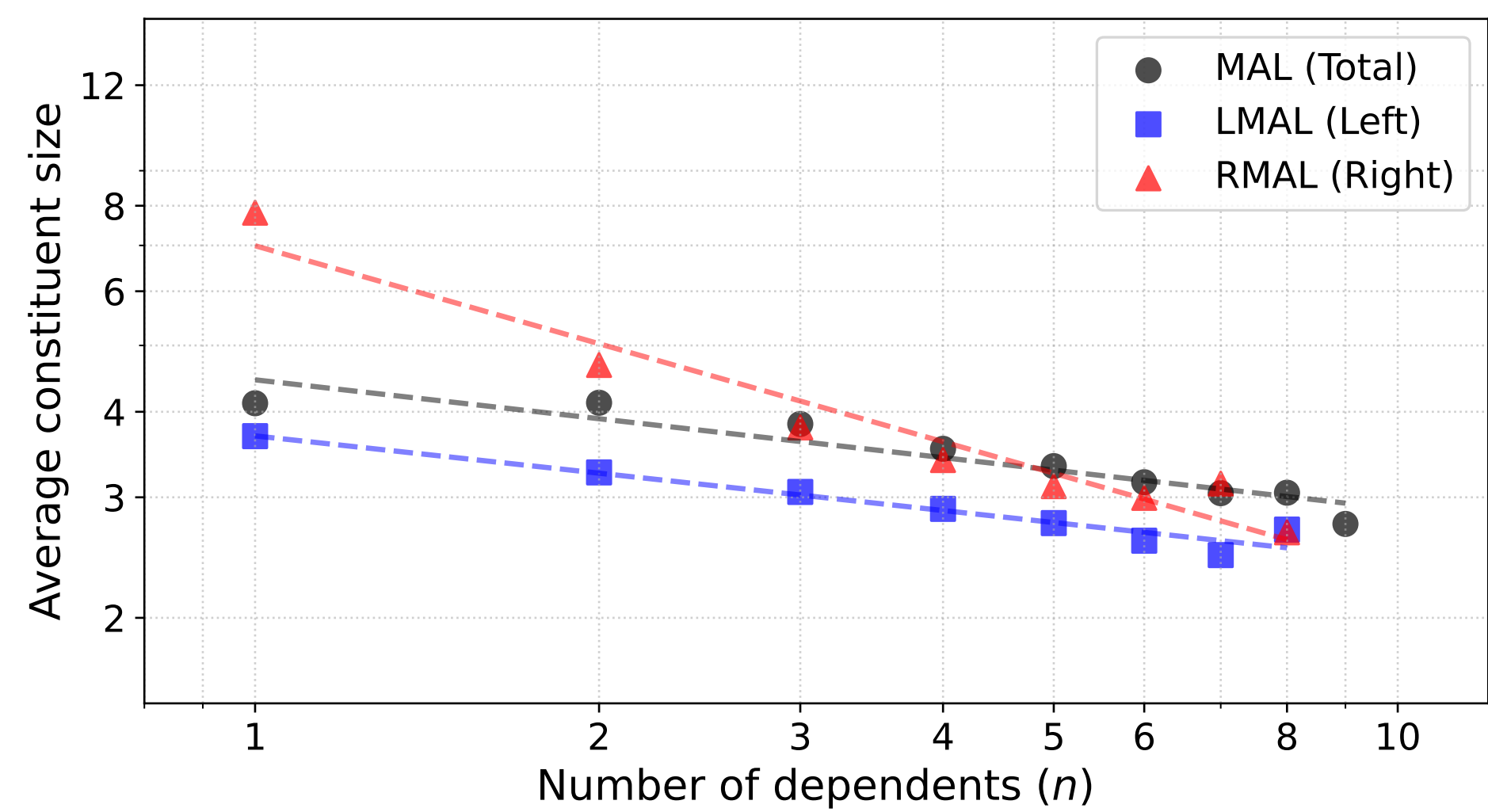
MAL effect = β , the slope of the regression line in log-log space.
 $\beta > 0.1$: **MAL language**. $\beta < -0.1$: **Anti-MAL**. In between: **grey zone**.

Why does this matter? DLM (Dependency Length Minimization) predicts short-before-long postverbally and long-before-short preverbally. Unlike DLM, MAL is direction-free. So: **does MAL hold equally on both sides?**
Spoiler: **no**.

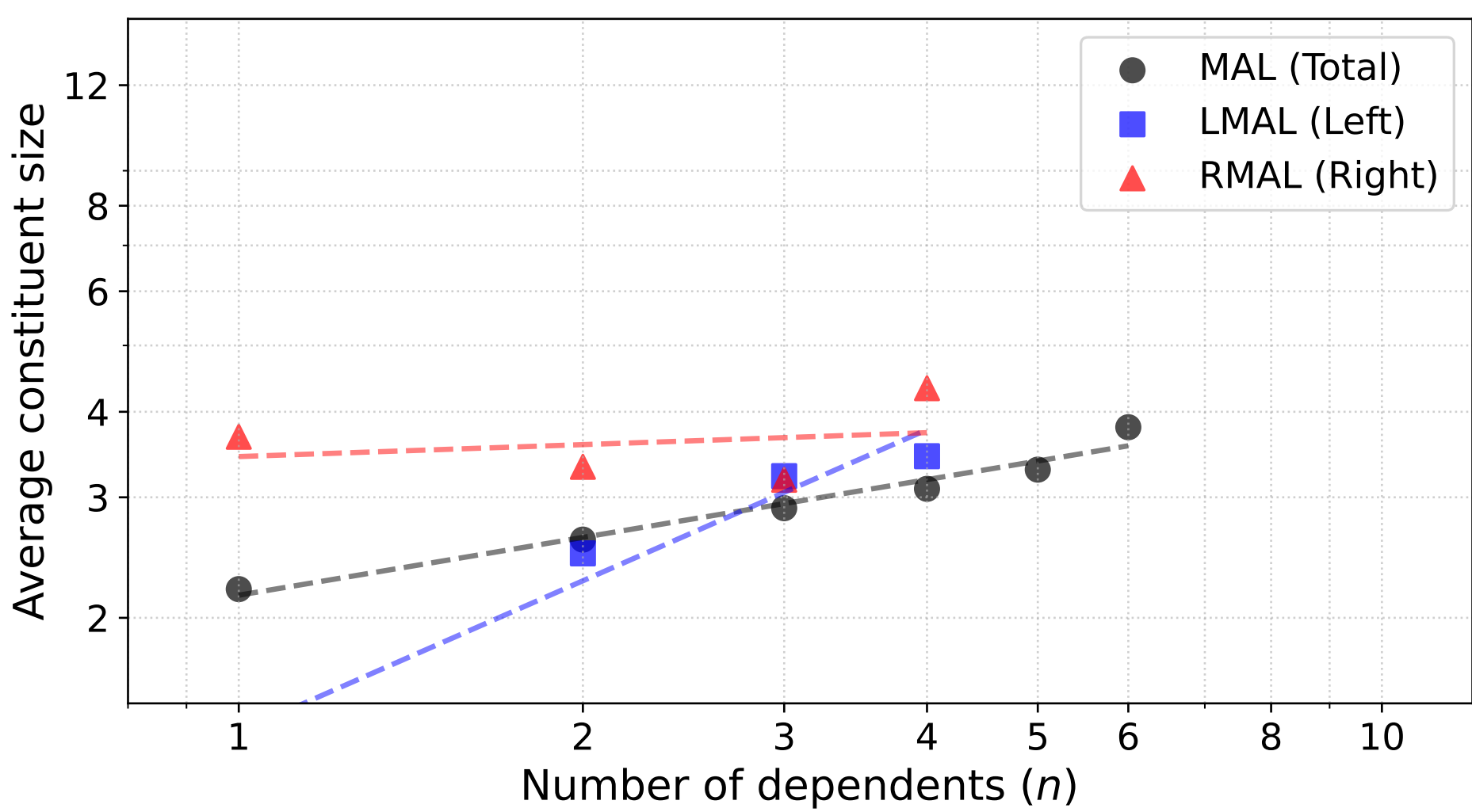


AS USUAL

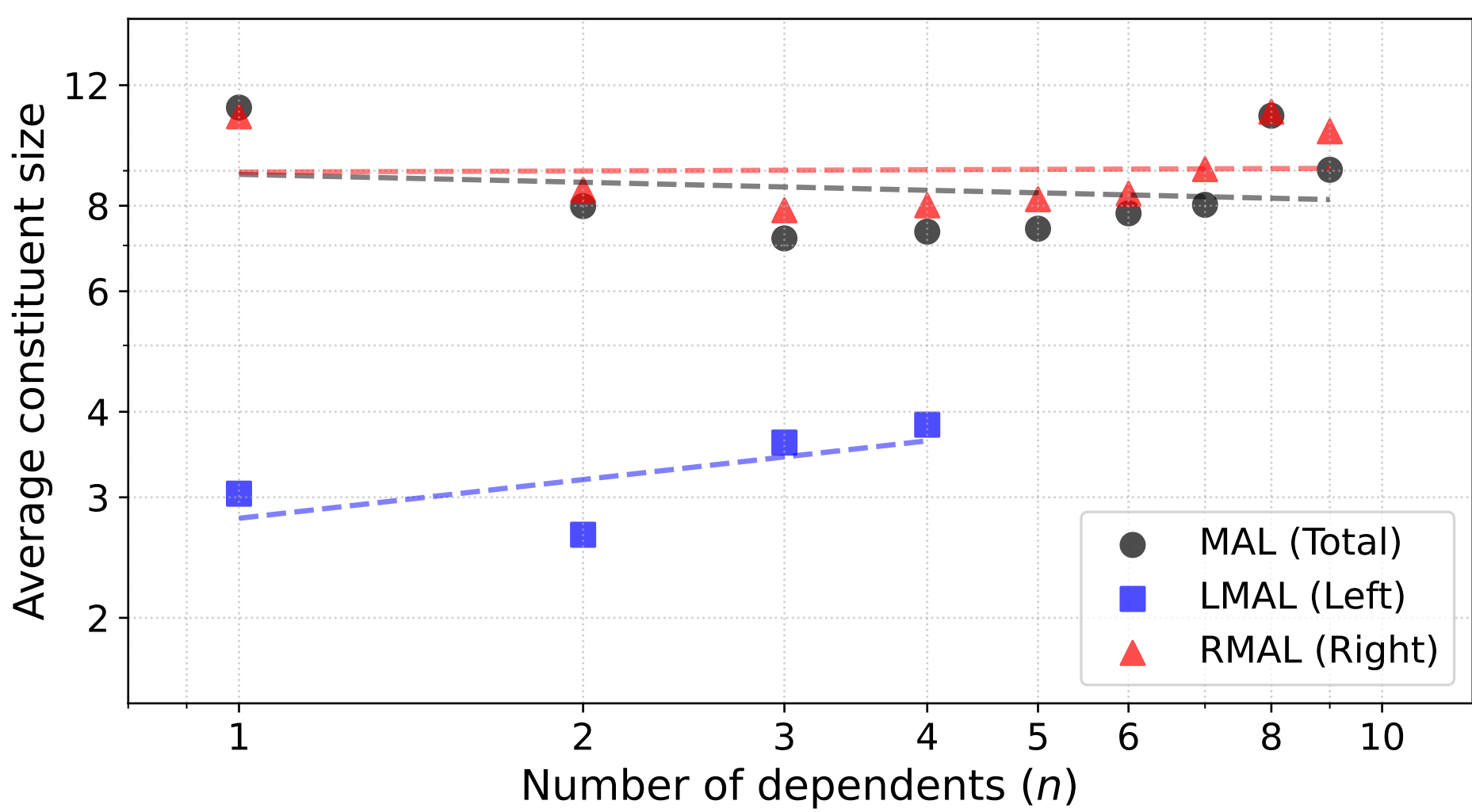
German nails it. Naija wobbles. Arabic surprises.



German – textbook MAL
slope = **-0.19**, **R² ~ 0.89** on the log-log fit. A near-perfect power law: every extra dependent shrinks the average constituent. German is MAL, RMAL and LMAL as expected. Is anyone surprised?



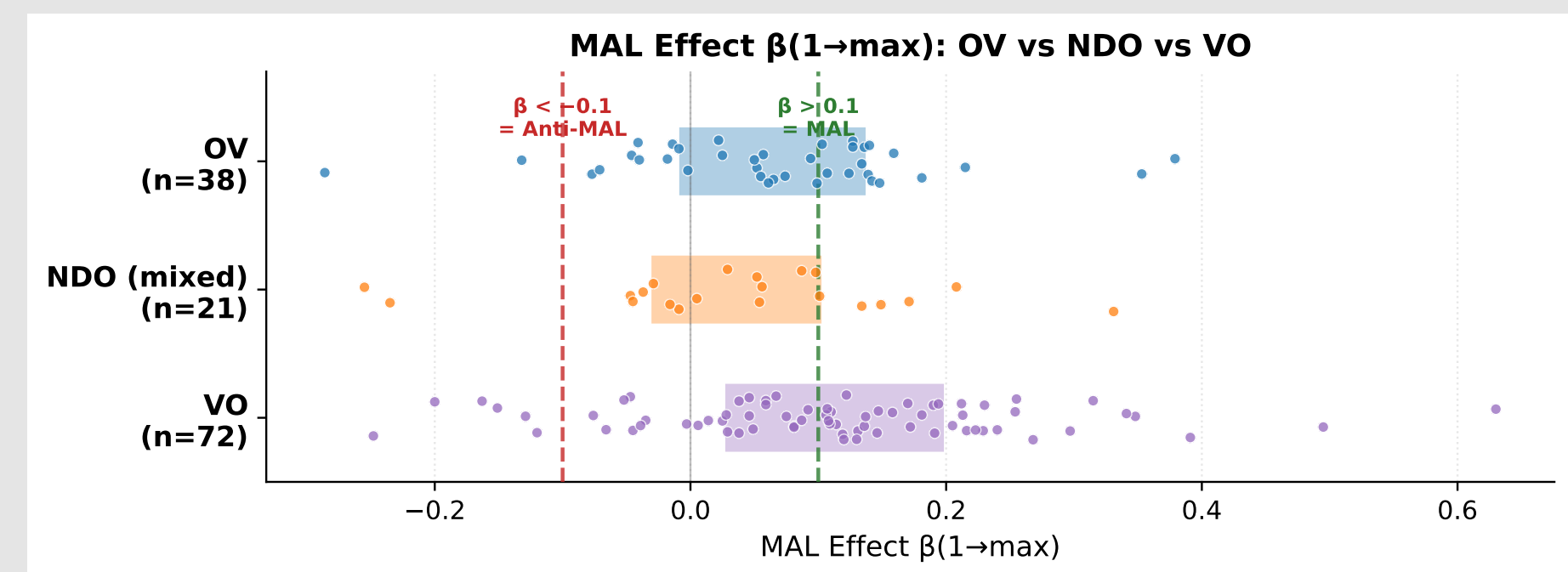
Naija – mixed, asymmetric
Left and right dependents diverge wildly. A young pidgincreole refusing to fit one curve – in fact the **strongest Anti-LMAL in the whole sample** ($\beta = -0.73$ preverbally).



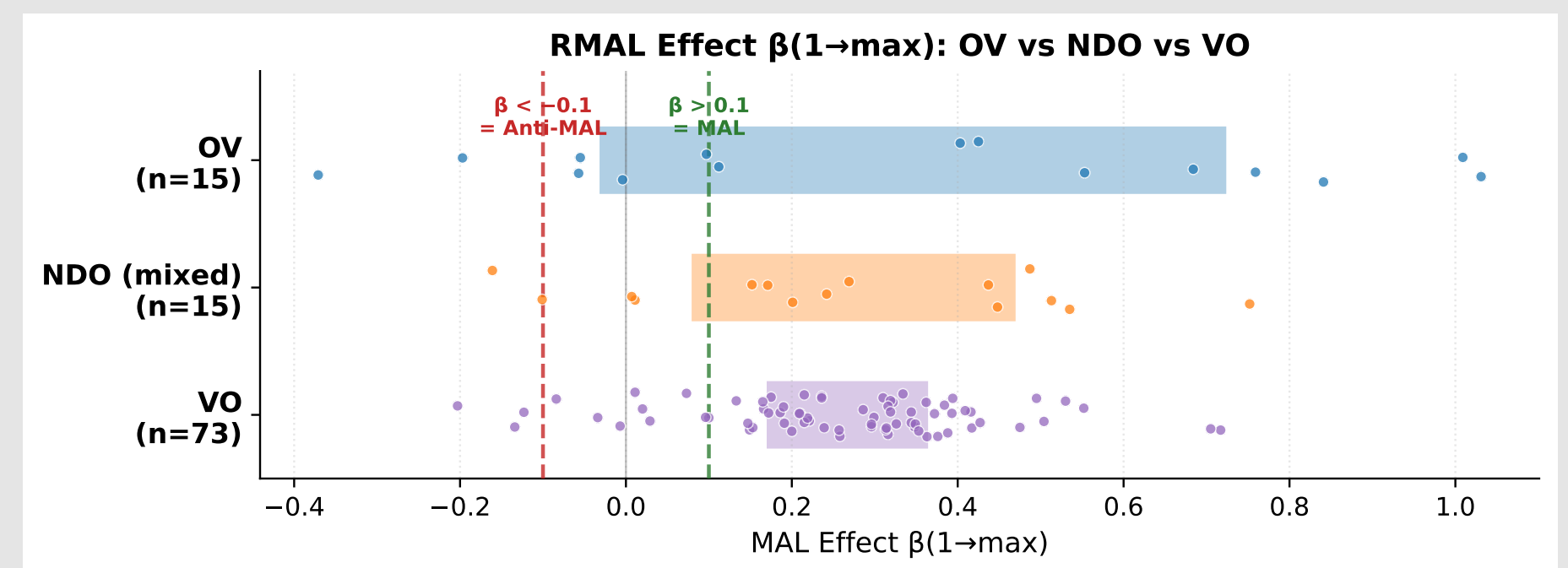
Arabic – complex behavior
Arabic has a particularly low **R²**. It follows MAL from $n = 1$ to 3, but becomes Anti-MAL from $n = 3$ onwards.
And you? Check your language at typometrics.elizia.net/menzerath!

THE NUMBERS NOBODY TALKS ABOUT

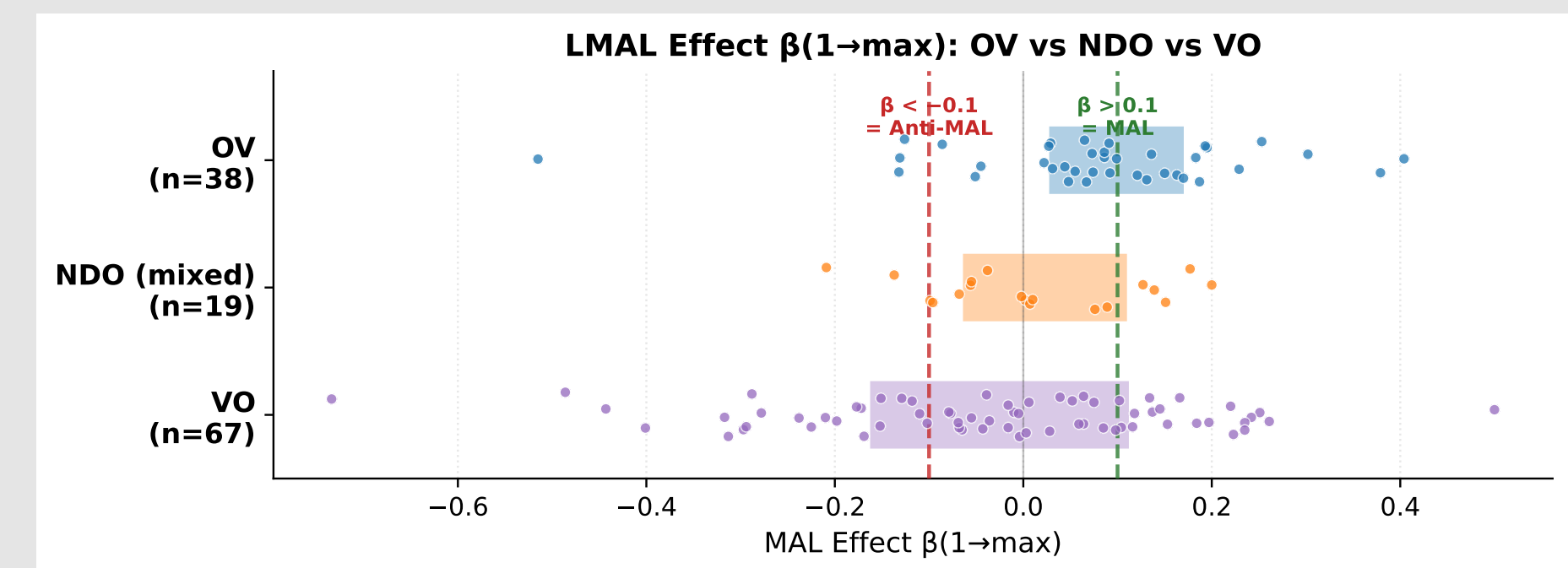
MAL is not ambidextrous. The verb leans right.



The Word Order Illusion
They want you to think word order is irrelevant. Look at the global averages: MAL and Anti-MAL types are distributed with eerie equality across VO and OV languages. It’s the perfect cover-up. But once you split the verb in half, their entire symmetry falls apart...



RMAL: postverbal domain
79% of UD languages are RMAL (81/103) and only **7% are Anti-RMAL** (7/103). The right side of the verb strongly compresses constituent size as n grows. VO languages lead: 84% are RMAL vs. 63% for OV.



LMAL: preverbal domain
Only **31% are LMAL** (39/124) and **23% are Anti-LMAL** (29/124). The left side of the verb resists compression – or even reverses it. OV languages are better: 39% are LMAL vs. 29% for VO. VO languages are **3x** more likely to be Anti-LMAL.