

Predictor	Estimate	SE	<i>z</i> value	<i>p</i> value
(Intercept)	-0.26	0.64	-0.4	0.67
constr=person-marking	-0.42	0.07	-5.9	<0.001*
group=West&South	-2.48	0.74	-3.3	<0.001*
constr=person-marking x group=West&South	0.77	0.10	7.7	<0.001*

Table 1: Summary of the logistic-regression model: presence of a pronominal subject as predicted by construction and language group with by-VERB and by-LANGUAGE random effects; within clauses in simple sentences and first conjuncts only. Asterisks denote significance at the 0.05 level.

In the article, we report a regression model that is fit to data from all clauses (Table 3, Figure 4). It is, however, reasonable to expect that the means of encoding subject will strongly vary across clause types (main, simple, subordinate, coordinate). However, adding `CLAUSE TYPE` as a predictor leads to severe convergence problems, rendering the models unusable. Instead, we opt for a simpler way of controlling for a potential effect of `CLAUSE TYPE`. We fit a separate model with the same specification, but use only clauses from simple sentences (i.e. excluding all clauses from complex sentences: subordinate, superordinate and coordinate). The total number of observations is 24,507; the number of unique verb lemmas is 4,962.

The model yields similar results, see Table 1. The absolute values of the coefficients are different, but the general picture is the same.

$C = 0.91$ , Somers'  $D_{xy} = 0.82$ .