Figure 3: PCA of the language vectors resulting from the concatenation of similarity judgments for all pairs.

Littell (2017) proposed to construct such representations by training language-identifying vectors end-to-end as part of neural machine translation models.

The vector for similarity judgments and the vector of linguistic features for a given language have different dimensionality. Hence, we first construct a distance matrix for each vector space, such that both columns and rows are language indices, and each cell value is the cosine distance between the vectors of the corresponding language pair. Given a set of L languages, each resulting matrix S has dimensionality of $\mathbb{R}^{|L|\times|L|}$ and is symmetrical. To estimate the correlation between the matrix for similarity judgments and each of the matrices for linguistic features, we run a Mantel test (Mantel 1967), a non-parametric statistical test based on matrix permutations that takes into account inter-dependencies among pairwise distances.

The results of the Mantel test reported in Table 3 show that there exist statistically significant correlations between similarity judgments and geography, family, and syntax, given that $p < 0.05$ and $z > 1.96$. The correlation coefficient is particularly strong for geography ($r = 0.647$) and syntax ($r = 0.649$). The former result is intuitive, because languages in contact easily borrow and loan lexical units, and cultural interactions may result in similar cognitive categorizations. The result for syntax, instead, cannot be explained so easily, as formal properties of language do not affect lexical semantics. Instead, we conjecture that, while no causal relation is present, both syntactic features and similarity judgments might be linked to a common explanatory variable (such as geography). In fact, several syntactic properties are not uniformly spread across the globe. For instance, verbs with Verb–Object–Subject word order are mostly concentrated in