

The Cube Method Beyond Economics

Abdelkader BOUSABAA

A Balanced Random Sampling Approach to the
Measurement of Arabic Etymology
in the French Lexicon

August 18, 2025

Abstract

The Cube method, introduced by Deville and Tillé (2004) and initially implemented for institutional survey purposes at INSEE (French national statistical institute) in 1999 by Bousabaa, Lieber and Sirolli, has been applied almost exclusively to economic and demographic surveys since its inception. This article proposes a novel application of this balanced random sampling algorithm to the field of computational historical linguistics, approximately thirty years after its original implementation.

The research question is the following: “*what is the true proportion of French common nouns and adjectives that originate, through the complete etymological chain, from Arabic?*” Standard French dictionaries systematically report only the proximate etymological source — often an Italian or Spanish intermediary — thereby concealing the Arabic ultimate origin. We term this phenomenon *etymological masking*.

Using a stratified balanced sample of 500 lemmas drawn from a population of French nouns and adjectives, balanced simultaneously on five auxiliary variables (frequency rank, part of speech, semantic domain, word length and Arabic prefix indicator), the Horvitz-Thompson estimator yields an Arabic-origin rate of 15.4%.

This result substantially exceeds the 1–3% figure commonly reported in standard etymological references, and suggests that Arabic is in fact the third most important source language of French vocabulary, after Latin and Proto-Indo-European. The article discusses both the methodological contribution — a rare application of the Cube method outside its original economic domain — and the linguistic implications of the findings.

Keywords: Cube method, balanced sampling, Horvitz-Thompson estimator, French etymology, Arabic loanwords, etymological masking, historical linguistics

Classification: C83, N7, 91F20.

7. Conclusion

Approximately thirty years after contributing to the first practical implementation of the Cube method for INSEE, we have applied it to an entirely different domain: the measurement of Arabic influence on the French lexicon. The exercise was motivated by a simple observation. Standard etymological dictionaries apply an asymmetric methodology that systematically undercounts words of Arabic origin by stopping the etymological chain at a Romance intermediary.

Using a balanced random sample of $n = 500$ French nouns and adjectives, drawn by the Cube method with five auxiliary balancing variables and analyzed using the Horvitz-Thompson estimator, we find that approximately 15.4% of French content words have an Arabic ultimate origin. This is the third largest source language of French vocabulary, after Latin and Proto-Indo-European.

The result substantially exceeds standard estimates. We attribute the discrepancy primarily to etymological masking. Approximately half of Arabic-origin words in French are described in standard dictionaries as being of Italian, Spanish or Latin origin, without acknowledgment of the Arabic source of those intermediaries.

We hope this article illustrates two lessons. First, the Cube method, three decades after its formalization, remains a flexible and powerful tool that can travel well beyond its original economic context. Second, the French language carries a richer Arabic heritage than is commonly acknowledged — *a heritage that deserves both scholarly attention and, perhaps, a certain linguistic pride.*