# Correlated grammaticalization
# The broad dataset

David Goldstein

17 September 2021

## 1 Introduction

This file reports the results from the broad dataset.

## 2 Article inventories in Indo-European

Relaxing the criteria for articlehood increases the number of articles recognized in the dataset (Figure 1) as well as the number of ancestral classes (Tables 1 and 2). In Figure 1, the relaxed criteria lead to an increase of languages with both articles.

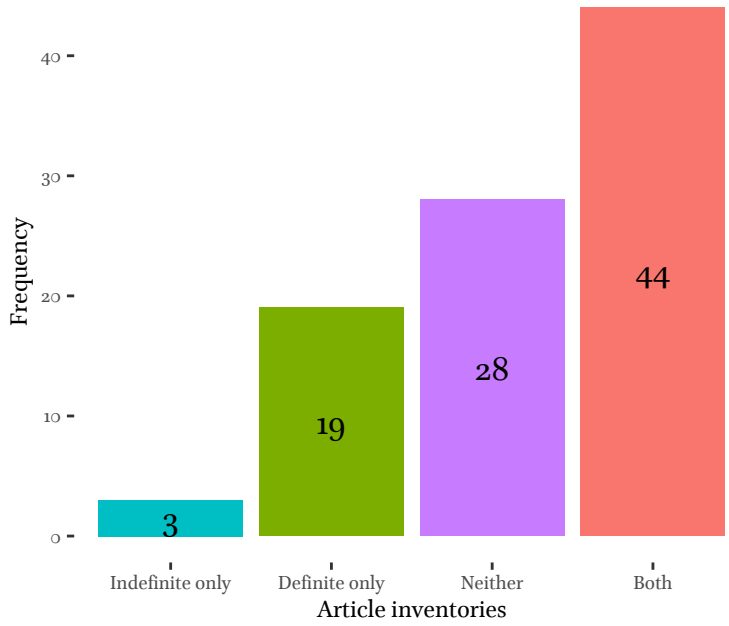


Figure 1: Frequency distribution of definite and indefinite articles in Indo-European (n = 94)

Table 1: Minimal ancestral classes for the definite article

| Ancestral Class | Clade | Languages |
|---|---|---|
| 1 | Celtic | Old Irish, Irish, Scots Gaelic, Welsh, Breton, Cornish |
| 2 | Italic | Latin |
| 2 | Romance | Nuorese, Cagliari |
| 3 | Romance | Romanian, Arumanian, Catalan, Portuguese, Spanish, French, Provencal, Walloon, Ladin, Romansh, Friulian, Italian |
| 4 | Germanic | Gothic |
| 5 | Germanic | Old West Norse, Icelandic, Faroese, Norwegian, Swedish, Danish |
| 6 | Germanic | Old English, English |
| 7 | Germanic | Frisian |
| 8 | Germanic | Old High German, German, Luxembourgish, Swiss German |
| 9 | Germanic | Dutch, Flemish, Afrikaans |
| 11 | Albanian | Tosk, Arvanitika |
| 12 | Greek | Ancient Greek, Modern Greek |
| 13 | Armenian | Classical Armenian, Eastern Armenian, Adapazar |
| 14 | Baltic | Old Prussian |
| 15 | Baltic | Latvian |
| 16 | Slavic | Lower Sorbian, Upper Sorbian |
| 17 | Slavic | Macedonian, Bulgarian |
| 18 | Iranian | Sogdian, Wakhi, Baluchi |
| 19 | Iranian | Kurdish |
| 20 | Iranian | Shughni, Sariqoli |
| 21 | Indic | Assamese, Oriya, Bengali |
| 22 | Indic | Kashmiri |
| 23 | Indic | Singhalese |
| 24 | Indic | Romani |

Table 2: Minimal ancestral classes for the indefinite article

| Ancestral Class | Clade | Languages |
|---|---|---|
| 1 | Celtic | Breton |
| 2 | Romance | Nuorese, Cagliari, Romanian, Arumanian, Catalan, Portuguese, Spanish, French, Provencal, Walloon, Ladin, Romansh, Friulian, Italian |
| 3 | Germanic | Faroese, Norwegian, Swedish, Danish |
| 4 | Germanic | English, Frisian |
| 5 | Germanic | German, Luxembourgish, Swiss German |
| 6 | Germanic | Dutch, Flemish, Afrikaans |
| 7 | Albanian | Tosk, Arvanitika |
| 8 | Greek | Modern Greek |
| 9 | Armenian | Eastern Armenian, Adapazar |
| 10 | Slavic | Lower Sorbian, Upper Sorbian |
| 11 | Iranian | Tajik, Persian |
| 12 | Iranian | Baluchi, Kurdish, Zazaki |
| 13 | Iranian | Wakhi, Shughni, Sariqoli |
| 14 | Indic | Assamese, Oriya, Bengali |
| 15 | Indic | Singhalese |
| 16 | Indic | Romani |

## 2.1 Establishing precedence

Table 3: Languages with both a definite and an indefinite article

| Clade | Languages |
|---|---|
| Celtic | Breton |
| Romance | Nuorese, Cagliari |
| Romance | Romanian, Arumanian, Catalan, Portuguese, Spanish, French, Provencal, Walloon, Ladin, Romansh, Friulian, Italian |
| Germanic | Faroese, Norwegian, Swedish, Danish |
| Germanic | English |
| Germanic | Frisian |
| Germanic | German, Luxembourgish, Swiss German |
| Germanic | Dutch, Flemish, Afrikaans |
| Albanian | Tosk, Arvanitika |
| Greek | Modern Greek |
| Armenian | Eastern Armenian, Adapazar |
| Slavic | Lower Sorbian, Upper Sorbian |
| Iranian | Wakhi, Baluchi |
| Iranian | Kurdish |
| Iranian | Shughni, Sariqoli |
| Indic | Assamese, Oriya, Bengali |
| Indic | Singhalese |
| Indic | Romani |

Table 4: Languages in which an indefinite article developed after a definite article

| Evidence | Clade | Languages |
|---|---|---|
| Parsimony | Celtic | Breton |
| Textual | Romance | Nuorese, Cagliari |
| Textual | Romance | Romanian, Arumanian, Catalan, Portuguese, Spanish, French, Provencal, Walloon, Ladin, Romansh, Friulian, Italian |
| Textual | Germanic | Old West Norse, Icelandic, Faroese, Norwegian |
| Textual | Germanic | English |
| Textual | Germanic | German |
| Textual | Germanic | Dutch |
| Textual | Greek | Modern Greek |
| Textual | Armenian | Eastern Armenian, Adapazar |

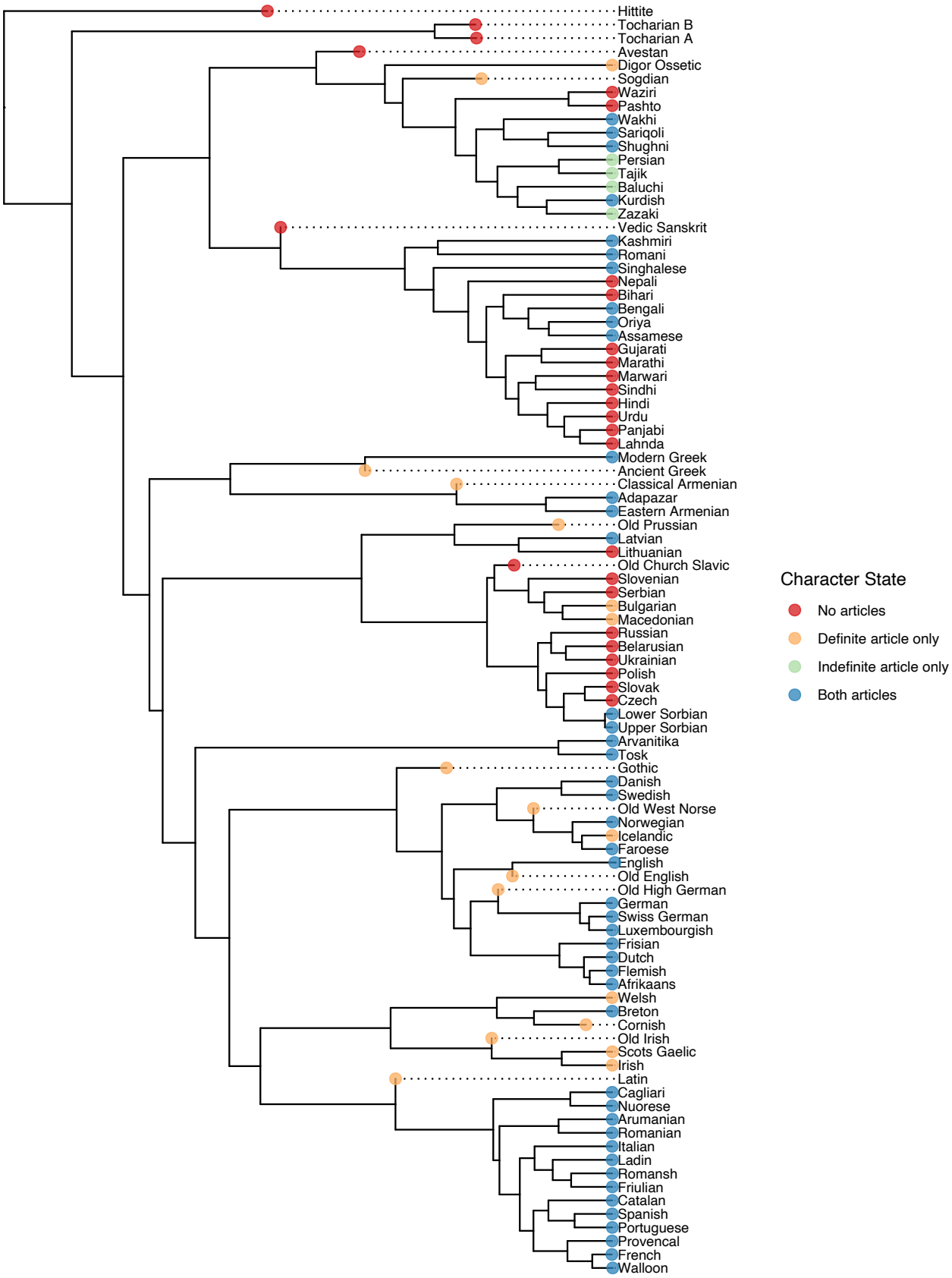


Figure 2: Maximum a posteriori tree of the A3 dataset and model of Chang et al. 2015 with observed character states in the broad dataset

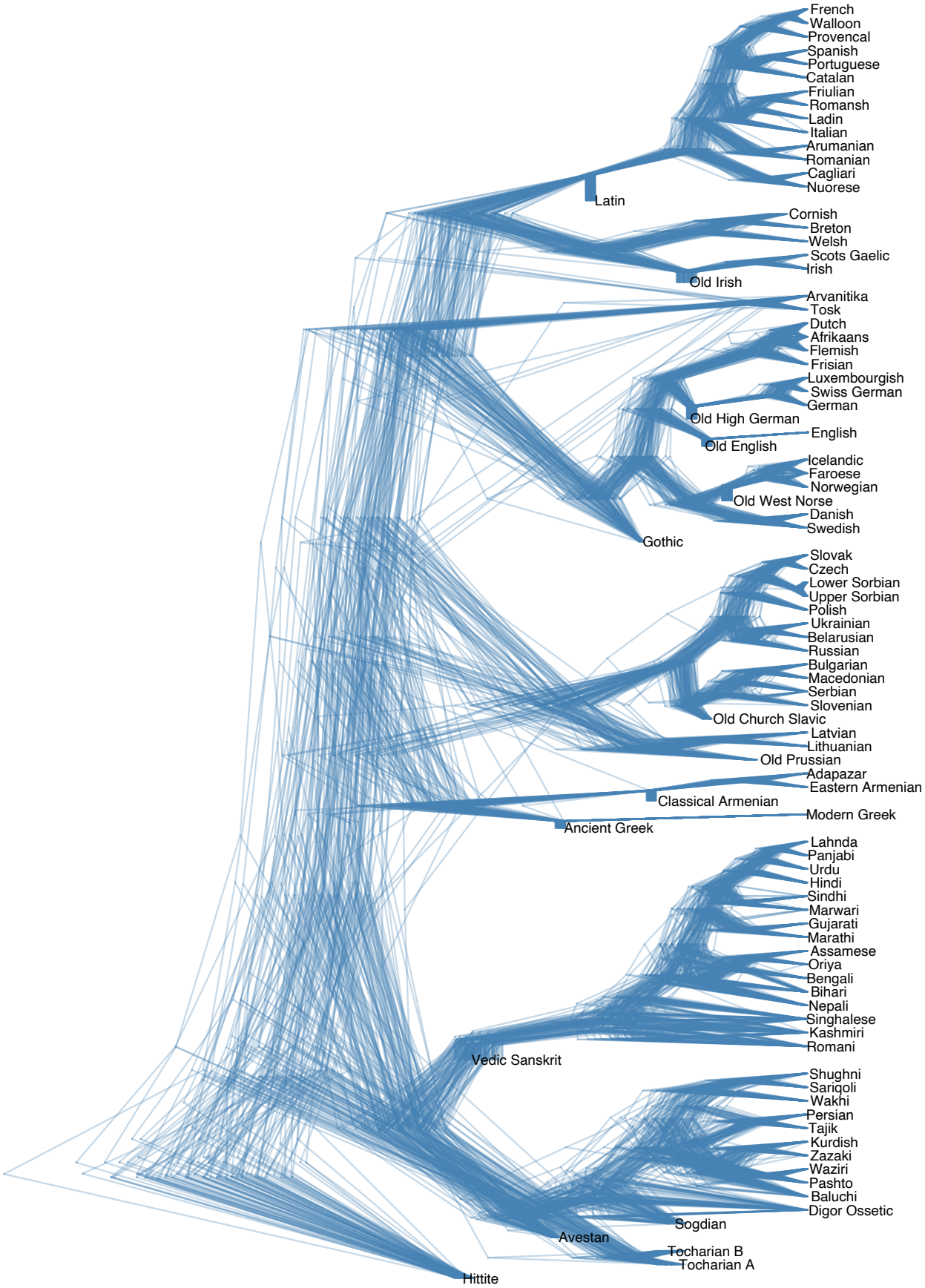


Figure 3: One hundred phylogenetic trees from the A3 dataset and model of Chang et al. 2015

# 3 Results

## 3.1 Posterior distributions of the rate parameters

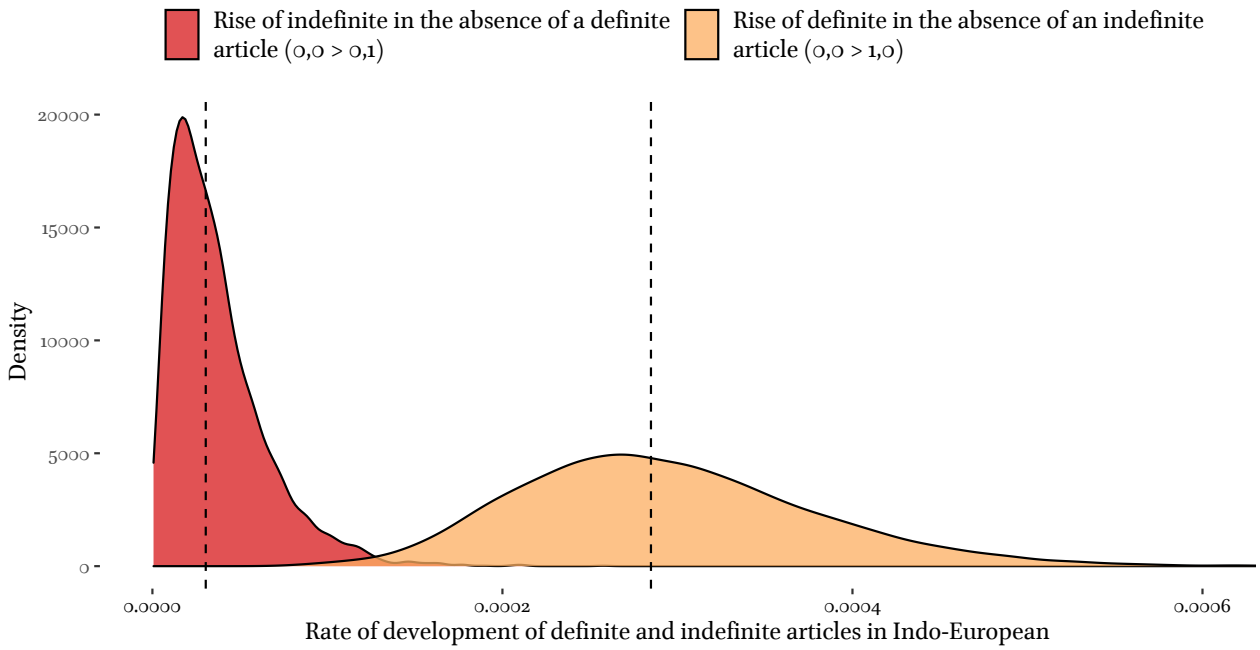


Figure 4: Posterior distributions of the rate parameters for definite and indefinite articles

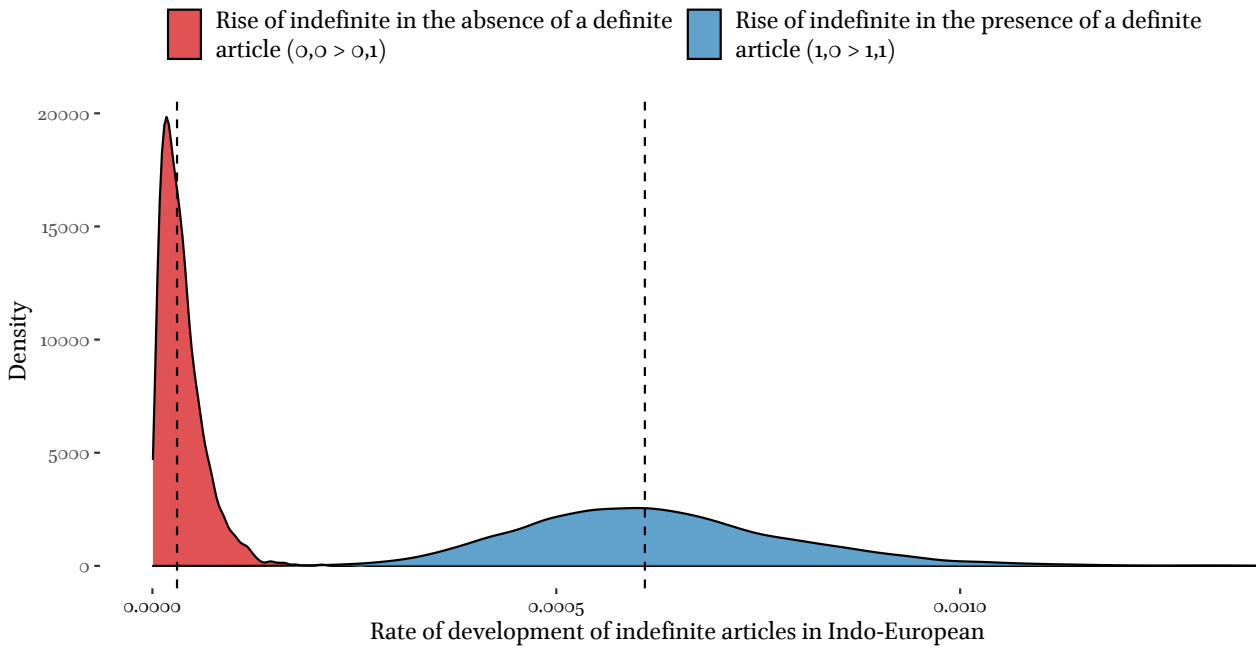


Figure 5: Posterior distributions of the rate parameters for indefinite articles

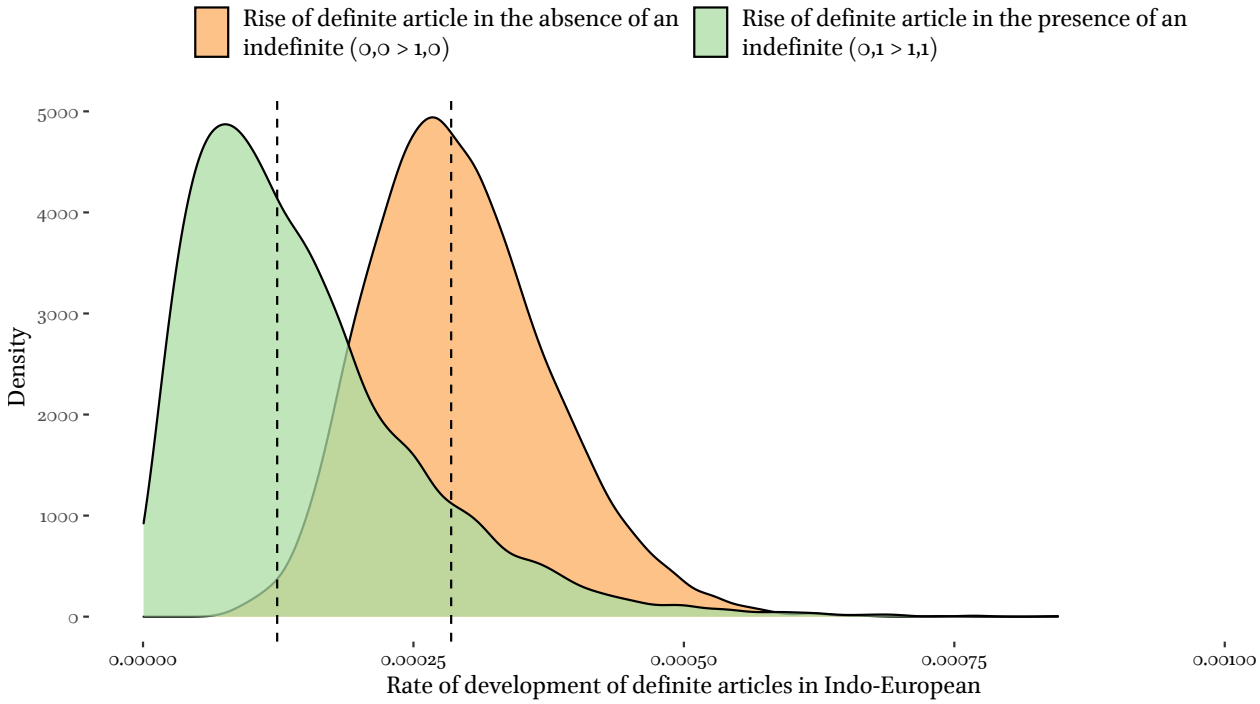


Figure 6: Posterior distributions of the rate parameters for definite articles

# 4 Model comparison

Support for the dependent model (with four parameters) over the three-parameter model is decisive:

Table 5: Model comparison

| Model 1 | Model 0 | Log-$BF_{10}$ | Support |
|---|---|---|---|
| Two indefinite rate parameters | Single indefinite rate parameter | 5.95 | Decisive |

# 5 Language contact

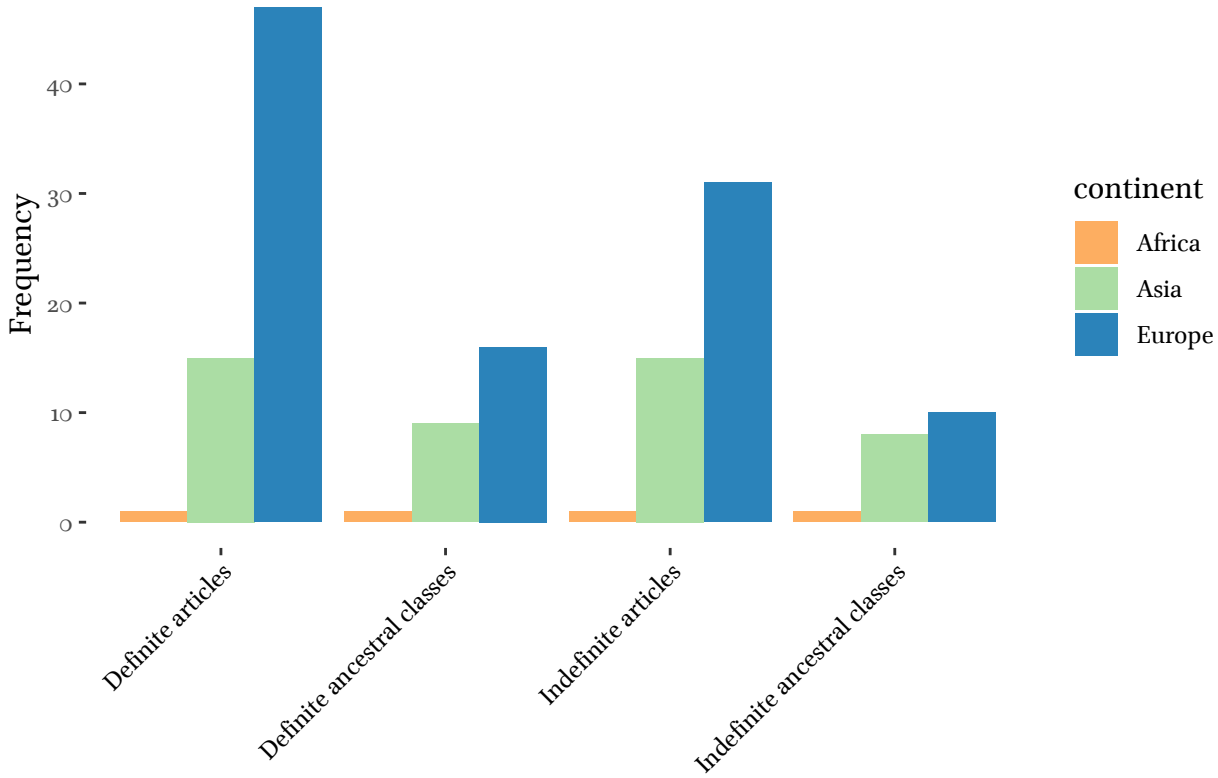


Figure 7: The frequency of articles and ancestral classes according to continent