

Correlated Grammaticalization

The rise of articles in Indo-European

David Goldstein
University of California, Los Angeles

October 15, 2021

Abstract

Grammaticalization is characterized by robust directional asymmetries (e.g., Kuteva et al. 2019). For instance, body-part nominals develop into spatial adpositions; minimizers develop into negation markers; and subject pronouns become agreement markers. Changes in the opposite direction are either rare or unattested (Garrett 2012: 52). Such robust cross-linguistic asymmetries have led some scholars to reify grammaticalization trajectories as universal mechanistic forces (Heath 1998: 729). One consequence of such a view is that the ambient morphosyntax of a language has little or even no relevance for grammaticalization. The main objective of this paper is to demonstrate the critical role that pre-existing morphosyntax can play in grammaticalization. The empirical basis for this claim is the grammaticalization of definite and indefinite articles in the history of Indo-European: indefinite articles developed at a faster rate among languages in which a definite article had already emerged compared to those lacking a definite article. The two changes are thus correlated. The results of this case study suggest that there is much more to be learned about when and why grammaticalization occurs by investigating its relationship to the pre-existing linguistic system (cf. Reinöhl & Himmelmann 2017: 381).

Keywords: Grammaticalization, Bayesian phylogenetics, definite article, indefinite article, historical morphosyntax, Indo-European

1 Introduction

One of the hallmark features of grammaticalization is its robust directional asymmetries (e.g., Kuteva et al. 2019). For instance, body-part nominals often develop into spatial adpositions; minimizers develop into negation markers; and subject pronouns become agreement markers. Changes in the opposite direction are either rare or unattested (Garrett 2012: 52). The cross-linguistic robustness of these asymmetries has led some scholars to codify them as universal pathways of change (see, e.g., Heath 1998: 728–729). Under such a view, pre-existing linguistic systems have little or even no role to play in grammaticalization. Although scholars in the early twentieth century, such as Kuryłowicz, Meillet, and Benveniste recognized the crucial role that the morphosyntax of a language played in its future trajectory (Heath 1998: 731–733), subsequent research has tended to neglect this aspect. As a result, little is known about the role that pre-existing morphosyntax

plays in grammaticalization. Recognizing this fact, Reinöhl & Himmelmann (2017: 381) recently declared that “what is needed ... is a substantial revision of grammaticalization theory so as to explicitly take into account the fact that grammatical change does not happen in a vacuum.” The objective of this paper is to demonstrate the crucial role that the pre-existing morphosyntax of a language can play in grammaticalization. I substantiate this claim on the basis of the grammaticalization of articles in Indo-European.

Within Indo-European, there are four basic inventories of articles. Languages such as English have both a definite and indefinite article:

- (1) *English*
 - a. **the** farmer
 - b. **a** farmer

At the other end of the spectrum lie languages such as Russian, which lacks articles altogether:

- (2) *Russian* (Bailyn 2012: 45)
 - krasivyj* *novyj* *derevjannyj* *dom*
 - beautiful.NOM.SG new.NOM.SG wooden.NOM.SG house.NOM.SG

‘**a/the** beautiful new wood **house**’

Out of context, a noun phrase in Russian can have either a definite or indefinite reading. Between these two extremes lie languages such as Old Irish and Persian, which have just a single article:

- | | |
|---|--|
| <ul style="list-style-type: none"> (3) <i>Old Irish</i> <ul style="list-style-type: none"> a. <i>in</i> <i>macc</i> DEF.ART boy ‘the boy’ b. <i>macc</i> boy ‘a boy’ | <ul style="list-style-type: none"> (4) <i>Persian</i> <ul style="list-style-type: none"> a. <i>ketab</i> book ‘The book’ b. <i>ketab-i</i> book-INDEF.ART ‘A book’ |
|---|--|

In Old Irish, indefinite nouns are typically bare (example 3b), but in Persian it is definite noun phrases that usually occur without a determiner (4a).¹

Typologists have observed a tendency among languages that possess an indefinite article to also have a definite article (Moravcsik 1969: 87, Heine 1997: 69), which has been attributed to the emergence of definite articles before indefinite articles (Crisma 2011: 175, de Mulder & Carlier 2011: 524). This implicational relationship raises the question of whether the grammaticalization of an indefinite article is more likely in languages that already have a definite article. This question is the central focus of my study.

Although this question has not been investigated for Indo-European as a whole, it has been addressed in particular clades. For instance, Irslinger (2013: 46) in her survey of articles in the

¹Definiteness marking in Persian is of course more complex than this statement betrays. See further section 6.3 below.

languages of Europe (in particular Romance and Celtic) concludes that that definite and indefinite articles “seem to develop independently, i.e. the emergence of a definite article does not entail the simultaneous emergence of an indefinite one.” Her claim rests on languages such as Irish, which has a definite article, but lacks an indefinite article. In an examination of articles in Romance and Germanic, Vincent (2018) also claims that definite and indefinite articles develop independently, but offers no definition of independence. For as important as these studies are, they illustrate a critical problem, which is that there is no recognized method for assessing correlations among linguistic changes.

This study uses a phylogenetic method originally developed by Pagel (1994) and Pagel & Meade (2006) to investigate correlated evolution in biology. Although this method is not yet common in linguistics (for examples, see Haynie & Bower 2016, Dunn et al. 2011, Dunn et al. 2017, Cathcart et al. 2020), it brings with it crucial advantages. First and foremost, it offers an explicit definition of correlation among linguistic changes: if the occurrence of one change impacts the rate at which a second change occurs, that is evidence for correlation. Second, it enables researchers to *test* hypotheses of dependence. In section 4.4 below, I compare models of dependent and independent grammaticalization to determine which one accounts for the data better.

1.1 Claims

This study advances the following two claims:

(5) *Central claims*

a. *Precedence*

Definite articles predominantly emerged before indefinite articles in the history of Indo-European.

b. *Correlation*

Indefinite articles became more likely to emerge in the wake of definite articles.

These claims result in the following diachronic trajectory for articles in Indo-European:

(6) *Article trajectory*

No articles > Definite Article > Definite Article, Indefinite Article

Since definite referents are more prominent than indefinite referents (section 6.1 below), this trajectory can be formulated even more simply: articles in Indo-European emerged among successively weaker referent types.

The empirical foundation of this study is a sample of ninety-four Indo-European languages, of which seventy-eight are contemporary and sixteen are ancient.² These are the same ninety-four languages used in the broad dataset of Chang et al. (2015a). This study is the first to investigate the diachrony of article inventories across Indo-European and builds on previous research, which has focused on the history of articles in specific languages or specific clades (e.g., Aebischer 1948, Kovari 1984, Manolassou 2001, Manolassou & Horrocks 2007, Carlier & de Mulder 2010, Guardiano 2012, Kraiss 2014, Börjars et al. 2016, Carlier & Lamiroy 2018).

²The term *language* is used in the sense of ‘taxonomic unit’ in this study. Whether the taxa under investigation here are more accurately described as languages or dialects is irrelevant for my purposes.

The remainder of the paper is structured as follows. Section 2 lays out three criteria that distinguish articles from other determiners. On the basis of these criteria, section 3 presents an overview of articles in Indo-European and establishes that their history has been one exclusively of gain. Section 4 then introduces the methods and representation of the data. Section 5 presents the results, which are then discussed in section 6. Section 7 brings the paper to a close with a summary of the main claims and concluding thoughts.

2 What are articles?

Although the existence of definite and indefinite articles in a language such as English is beyond dispute, establishing the inventory of articles in other languages can be anything but trivial. The main challenge is that articles can be homophonous with or orthographically identical to other determiners. For instance, segmental identity between demonstratives and definite articles is well known (e.g., Masica 1986: 134, Estrada Fernández 1996: 19, Dahl 2003, Bashir 2009: 841, Dryer 2013a), as we see in the following pair of examples from German:

- (7) a. *Demonstrative*³ (Wemke et al. 1995: §373)
- Mit 'dem Kerl will ich nichts mehr zu tun*
 with DEM.DAT.SG guy.DAT.SG want.1SG.PRES.ACT 1SG.PRO nothing more to do.INF
haben
 have.INF
- 'I don't want anything more to do with **that** guy.'
- b. *Definite article*
- Ich habe dem Bub das Spielzeug*
 1SG.PRO have.1SG.PRES.ACT DEF.ART.DAT.SG boy.DAT.SG DEF.ART.ACC.SG toy.ACC.SG
gegeben.
 give.PERF.PART
- 'I gave the toy to **the** boy.'

In example (7a), the determiner *dem* is used as a demonstrative meaning 'that', whereas in example (7b) it serves as a definite article, which is its predominant use.

Indefinite articles can be homophonous with the cardinal 'one' (e.g., Masica 1986: 134, Dryer 2013b), as illustrated by Modern Greek:

- (8) *Modern Greek* (Holton et al. 2012: 53)
- a. *Numeral*
- īða éna skíli sto ḑrómo.*
 see one.ACC.SG dog.ACC.SG in.DEF street
- 'I saw **one** dog in the street.'
- b. *Indefinite article*

³Although the demonstrative *dem* in this example is orthographically indistinguishable from the definite article *dem* in example (7b), prosodically it differs from the definite article in that it must be stressed (cf. Lockwood 1968: 87).

īḍa éna *skilí* *sto* *ḍrómo.*
 see INDEF.ART.ACC.SG dog.ACC.SG. in.DEF street

‘I saw a **dog** in the street.’

In example (8a) *éna* is parsed as a numeral, but in (8b) it is an indefinite article. The homophony in examples (7) and (8) reflects the precursors from which each type of article emerged: definite articles commonly emerge from demonstratives (e.g., Kuteva et al. 2019: 137–139) and indefinite articles from the numeral ‘one’ (e.g., Kuteva et al. 2019: 299–301).

Articles often undergo phonetic reduction and end up segmentally distinct from their diachronic sources. When such reduction does not occur, it can be difficult to determine whether or not a determiner has become an article. In German and Modern Greek, it is possible to distinguish demonstrative from definite article and numeral from indefinite article on the basis of distributional patterns and semantics. In both languages, articles are by and large obligatory for definite and indefinite noun phrases. When articles are not used systematically, however, it is far more difficult to distinguish them from their sources. Upper Sorbian (Slavic) and Kashmiri (Indic) illustrate this challenge:

- (9) a. *Upper Sorbian* (Berger 2007: 1)
 Na, něk je mi wona pokazawa **tón puć won**, hale blows přede wsu.
 ‘Na, now she showed me **the way out**, but only out of the village.’
- b. *Kashmiri* (Wali et al. 1997: 101)
akh ləḍki
 ‘{**some, one, a**} boy’

According to Berger (2007: 1), the use of Upper Sorbian *tón* in example (9a) resembles that of definite articles in English and German. In a similar vein, Wali et al. (1997: 100) maintain that *akh* ‘one’ can serve as an indefinite article.

The data in example (9) prompt the questions of whether Sorbian has a definite article and whether Kashmiri has an indefinite article. In fact, the question of whether or not a particular language has an article has been extensively debated for a number of languages in my dataset, including Latin (e.g., Abel 1971, Kurzová-Jedličková 1963, Hertzenberg 2015: 10), Old English (e.g., Wood 2003, Sommerer 2012), Old High German (e.g., Leiss 2007: 91, Kraiss 2014, Flick 2020), and Gothic (e.g., Miller 2018: 65). These debates have often been fruitless, because scholars have approached them with different—and sometimes implicit—criteria for articlehood (Hawkins 1978: 154–155, Selig 1992: 25–26, Hertzenberg 2015: 22, Sommerer 2012: 176–177, 184–185). One therefore needs an explicit and justifiable way of determining what is and is not an article.

2.1 Three necessary and sufficient conditions for articlehood

Becker (2018: 23–41) advances the following three necessary and sufficient criteria for articlehood:

(10) *Criteria for articlehood*

a. *Semantics*

Articles mark referential types.

- b. *Syntactic distribution*
Articles have to occur together with nouns.
- c. *Systematic distribution*
Articles occur systematically (allowing for exceptions) in the contexts matching their referential functions.

Criterion three in example (10c) is the most significant for my purposes. As we will see in section 2.3 below, the requirement of systematic distribution has a crucial effect on which determiners are deemed articles (cf. Greenberg 1978: 252, Guardiano 2013: 77). The judgments presented below as to whether or not a particular determiner meets the criteria for articlehood are based on grammatical descriptions and (in the case of the corpus languages) textual investigation. The grammars consulted are typically based on a standardized form of the language and take no account of patterns found in colloquial registers (such as described in, e.g., van Gelderen 2011: 224).

2.2 Referential types

Criterion (10c) stipulates that articles occur systematically in the contexts matching their referential functions. The question then arises of what the referential functions of definite and indefinite articles are. This section answers that question.

Definite referential types fall into two classes, pragmatic and semantic definites (a distinction originally formulated by Löbner 1985: 298–299 and later adopted by, e.g., Himmelmann 1997: 38, Napoli 2009, Müth 2011: 13–15, Szczepaniak 2011: 71–73, Flick 2019: 159, Flick 2020: 87–92).⁴ A pragmatically definite referent is identifiable only on the basis of contextual information, as in the following four examples:

- (11) *Pragmatic definites*
 - a. *Deictic*
Hand me **the hammer**.
 - b. *Anaphoric* (Schwarz 2013: 535)
John bought a book and a magazine. **The book** was expensive.
 - c. *Recognitional*
Did you buy **the car** (that you were telling me about)?
 - d. *Establishing*
Did you hear **the news** that they are going to cut the budget again?

A deictic referent is identifiable based on shared physical context. An anaphoric referent is identifiable on the basis of its identity with a referent previously mentioned in the discourse. A recognitional referent is identifiable on the basis of shared knowledge (Himmelmann 1997: 61–82, Becker 2018: 24). As a result, such the use of the definite article is felicitous with such referents even on their first mention. An establishing referent is identifiable on the basis of the information supplied in an adjunct, such as a restrictive relative clause.

By contrast, a semantically definite referent is identifiable independent of the immediate situation or context of utterance, as in the following examples:

⁴The classification of referential types below follows those of Löbner (1985) and Szczepaniak (2011: 73).

- (12) *Semantic definites*
- a. *Contextually unique*
What's the best way to **the center**?
 - b. *Absolutely unique*
The Earth is round.
 - c. *Bridging*
Eva bought a book. **The author** is French.

A contextually unique referent is identifiable on account of its being the only salient referent of its kind in the discourse. An absolutely unique referent is more general: it is the only salient referent of its kind in any discourse. I include in this category the use of a definite article with superlatives and ordinals, e.g., *the hardest problem* or *the the second stop* (Müth 2011: 14). A bridging referent is identifiable on account of its relationship with a previously identified referent.

Following Himmelmann (1997: 41), Wendtland (2011: 24) contends that, since demonstratives can only be used in pragmatically definite contexts, it is the use of definite articles in semantically definite contexts that distinguishes them from demonstratives (cf. Becker 2018: 165). This view is consistent with that of Becker (2018: 70), who asserts that a definite article must be able to encode the anaphoric, establishing, bridging, and contextually unique referential types. In the languages sampled for this study, the descriptive resources do not always enable one to establish whether a definite article is used systematically in these contexts. I have accordingly adopted a slightly different approach: I counted as a definite article any determiner that was used systematically in at least one pragmatically definite context and at least one semantically definite context.

There are two indefinite referential types, specific and non-specific:

- (13) *Indefinites*
- a. *Specific*
We saw **an amazing house** yesterday.
 - b. *Non-specific* (Becker 2018: 77)
Do you have **a pen**? Any pen will do.

A specific indefinite referent cannot be identified but does refer to a particular referent of its kind. A non-specific indefinite is also unidentifiable but refers to any referent of its kind and not a particular one. I classified as indefinite articles determiners that met the conditions in example (10) above and were used to encode at least one of the two indefinite referent types in example (13). The vast majority of the indefinite articles in my sample occur with both indefinite specific and non-specific referents.⁵

⁵The indefinite article in Balochi (Jahani & Korn 2009: 667) and Wakhi (Bashir 2009: 841), for instance, can only be used with indefinite specific referents. I am aware of no language in my sample that uses an indefinite article only with indefinite non-specific referents. It may be the case that indefinite articles emerge first among indefinite specific referents and only later are extended to non-specific referents (see Heine 1997: 71–76 for a diachronic trajectory of indefinites).

2.3 Disputed determiners

As mentioned above in section 2.1, the systematic-distribution criterion is the one that has the greatest impact on my classification of determiners. This criterion resulted in the exclusion of determiners whose status as articles is debated (e.g., Latin, Old High German), since the source of the dispute is typically the fact that a determiner is not used systematically. For instance, in a number of Iranian languages, a demonstrative is described as being able to be used as a definite article, but does not systematically co-occur with definite referents. Table 1 presents the languages whose determiners were not considered articles on account of the systematic-distribution criterion.⁶

Table 1: Languages with determiners whose status as articles is disputed

Article	Clade	Language	Reference
Definite	Italic	Latin	Abel (1971), Fruyt (2003), Hertenberg (2015), Adams & Vincent (2016: 482–527)
Definite	Germanic	Gothic	Sauvageot (1929), Leiss (2007: 91), Pimenova (2017), Miller (2018: 63 n. 2)
Definite	Germanic	Old High German	Kraiss (2014), Flick (2020)
Definite	Germanic	Old English	McCull Millar (2000: 279), Crisma (2011), Wood (2003: 10), Sommerer (2015)
Definite	Baltic	Old Prussian	Maþiulis (n.d.: §158), Holvoet (2018: 1668)
Definite	Baltic	Latvian	Bielenstein (1863: 257, 416), Mathiassen (1997), Prauliņš (2012)
Definite	Slavic	Upper Sorbian	Schaarschmidt (1984: 76), Berger (2007)
Definite	Slavic	Lower Sorbian	Schaarschmidt (1984: 76), Berger (2007)
Definite	Iranian	Wakhi	Bashir (2009: 841, 858)
Definite	Iranian	Shughni	Edelman & Dodykhudoeva (2009b: 794)
Definite	Iranian	Sariqoli	Edelman & Dodykhudoeva (2009a: 781)
Definite	Iranian	Balochi	Gilbertson (1923: 12), Axenov (2006: 65–66), Jahani & Korn (2009: 667)
Definite	Iranian	Digor Ossetic	Erschler (2019: 880)
Definite	Indic	Kashmiri	Wali et al. (1997: 100)
Indefinite	Baltic	Latvian	Bielenstein (1863: 255, 258, 415), Mathiassen (1997), Prauliņš (2012)
Indefinite	Slavic	Upper Sorbian	Schaarschmidt (1984: 76), Breu (2012)
Indefinite	Slavic	Lower Sorbian	Schaarschmidt (1984: 76)
Indefinite	Iranian	Sariqoli	Edelman & Dodykhudoeva (2009a: 781)
Indefinite	Iranian	Shughni	Edelman & Dodykhudoeva (2009b: 794, 822–823)
Indefinite	Indic	Romani	Matras (2002: 98)
Indefinite	Indic	Kashmiri	Wali et al. (1997: 100)

⁶In the supplementary files, one can find an expanded dataset in which the languages in Table 1 are all represented as having a definite and indefinite article. The results of the analyses based on this dataset also support the claim of correlated grammaticalization.

2.3.1 Sinhalese

In the languages in Table 1 above, a determiner is typically not classified as an article on account of usage. Sinhalese stands out from every other language in my dataset in that the debate over whether it has a definite article centers on morphosyntax:

- (14) *Sinhalese* (Garland 2006: 7)
- a. *laməy-a maawə dækka*
child-SG.DEF 1SG.ACC see.PST
'The child saw me.'
 - b. *laməy-ek maawə dækka*
child-SG.INDEF 1SG.ACC see.PST
'A child saw me.'
 - c. *lamai maawə dækka*
child.PL 1SG.ACC see.PST
'(The) children saw me.'

This triplet illustrates two properties of Sinhalese that are widely agreed upon. The first is that definiteness is only realized on singular nouns. The second is that Sinhalese has an indefinite determiner (e.g., Gair 2003: 788, Chandralal 2010: 79), which is an article according to the criteria in section 2.1. The indefinite article has two allomorphs, *ak* and *ek*. The former is restricted to inanimate nouns, while the latter occurs with both animate and inanimate nouns (Chandralal 2010: 79).

Analyses of singular definite forms such as *laməy-a* in example (14a) differ. The gloss above is from Garland (2006: 7), who analyzes the suffix *-a* as an exponent of both number and definiteness (Paranavitana 1956: §348 seems to have a similar analysis in mind, even though his description is less explicit). Masica (1991: 248) and Chandralal (2010: 79), however, argue that Sinhalese does not have a devoted suffix for marking definiteness. An analysis of the inflectional morphology of Sinhalese lies beyond the scope of this investigation, but I have followed Paranavitana (1956: §348) and Garland (2006: 8) in recognizing a definite suffix. Referential marking in Sinhalese does not resemble the data from languages that lack a definite article, such as Persian in example (4) above, since singular definite nouns bear a morphological exponent that singular indefinite nouns do not.

2.4 Cross-linguistic comparison

Nichols (2003: 291) observes that cross-linguistically only gross structural features and abstract categories can be meaningfully compared. As a result, cross-linguistic comparison will inevitably simplify linguistic reality. This study is no exception. The criteria in example (10) above provide an explicit way of evaluating whether or not a determiner in a particular language is an article or not, but it must be borne in mind that the determiners deemed to be articles do not exhibit the same distributional patterns. For instance, in some of the languages in my sample, an article is prenominal and in others it is postnominal (see further Carling 2019: 75–81). Both linearization patterns are found in Romanian:

- (15) a. *Postnominal definite article*

frate-le
brother-DEF.ART

‘the brother’

b. *Prenominal indefinite article*

un frate
INDEF.ART brother

‘a brother’

The definite article is postnominal, but the indefinite article is prenominal. This pair also illustrates the variable prosodic realization of articles. The definite article in example (15a) is bound, but the indefinite article in example (15b) is free.

Some languages have multiple definite and indefinite articles. Certain dialects of Fering (Ebert 1971a, Ebert 1971b) and Armenian (Wendtland 2011: 24–40), for instance, have multiple articles for definite referents. Persian has two exponents of indefiniteness (Jasbi 2016), as does Tajik (Perry 2005). It is also possible for speakers to mark a referent with more than one marker of definiteness, such as we find in Scandinavian languages (e.g., Dahl 2003).

The distribution of articles can also exhibit fine-grained lexical distinctions from one language to the next:

- (16) *Er ist mit dem Bus gekommen.*
3SG.MASC be.PRES.ACT.3SG with DEF.DAT.SG bus.DAT.SG come.PERF.PART
‘He came **by bus**.’

German speakers standardly use a definite article to refer to the bus in this type of sentence, but English speakers do not (Wendtland 2011: 20).

3 Article inventories in Indo-European

Against this background, it is now possible to tally the number of languages in my dataset with definite and indefinite articles. Figure 1 presents the distribution of article inventories of the ninety-four languages in my sample. The past several millenia of Indo-European linguistic history have resulted in a skewed distribution, with most languages either having no articles or having both (Lakoff 1972: 174).

The distribution in Figure 1 does not, however, offer us any insight into the question of how this skewed distribution arose. What sort of article inventory did Proto-Indo-European have? If articles have been lost over the past six millenia, is there a pattern to their loss? If they have been gained, was one type of article gained first? In the next section, I demonstrate that the history of articles in my dataset is exclusively one of gain. In section 3.2, I argue that definite articles predominantly emerged before indefinite articles.

3.1 A history of gain

Two crucial generalizations reveal that the history of articles in Indo-European has been one of gain. The first is that the earliest attested languages of the family lack articles. These include Hittite (Anatolian), Mycenaean Greek (Greek), Vedic Sanskrit (Indic), and Avestan (Iranian). This absence suggests that articles were not present in Proto-Indo-European (Schwyzer 1936: 156, Lakoff

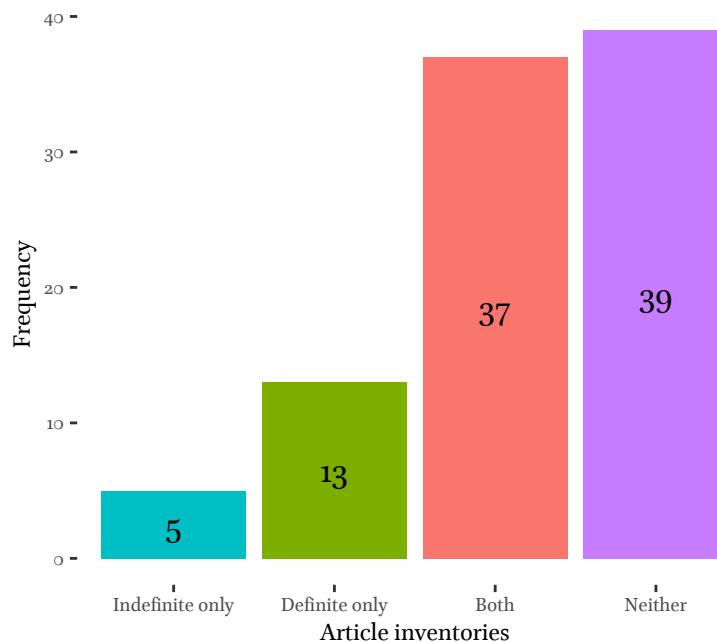


Figure 1: Frequency distribution of definite and indefinite articles in Indo-European (n = 94)

1972: 174, Mallory & Adams 2006: 107, van Gelderen 2007: 290, Napoli 2009: 570, de Mulder & Carlier 2011: 523, Cathcart et al. 2018: 21, Napoli 2019: 17). Indeed, among the languages in my dataset, the first language to develop an article is probably Greek, which by 500 BCE systematically used a definite article. (It is possible that the definite article arose in Proto-Insular-Celtic around the same time.)

The second generalization, which to the best of my knowledge has thus far remained unacknowledged, is that cognate articles (i.e., articles that descend from a common ancestral definite article) are found exclusively within the so-called major clades (Anatolian, Tocharian, Indo-Iranian, Greek, Armenian, Baltic, Slavic, Albanian, Germanic, Celtic, Italic). The definite articles in Celtic, which are presented in Table 2, exemplify this distribution. They all descend from an inferred ancestor **sindos*, which is itself assumed to have been a definite article. I refer to sets of articles that descend from a common ancestral article as *ancestral classes*.⁷ The crucial property of Table 2 is that the descendants of **sindos* are limited to Celtic. In other words, the ancestral class in Table 2 contains no non-Celtic languages and no Celtic languages are found in any other ancestral class.

This is in fact true of all the ancestral classes for both definite and indefinite articles, which are presented in Tables 3 and 4. The languages in each row uniformly belong to a single major clade.

⁷It is worth explicating how an ancestral class differs from a cognate class. In short, the former is more restrictive than the latter. Cognates are standardly defined as linguistic units that share a common ancestor. So the Bulgarian definite article *-t* and the Greek (neuter nominative-accusative singular) definite article *tó* are cognate, since they share a common ancestor, which is reconstructed as **tó*. Crucially, however, this **tó* is not reconstructed as a definite article, but as a demonstrative. So, despite being cognate, the Bulgarian and Greek articles do not belong to the same ancestral class. The reason for using ancestral classes as opposed to cognate classes is that it approximates the number of times articles arose. Definite articles developed independently in Bulgarian and Greek and assigning them to different ancestral classes registers this independence.

Table 2: Ancestral class for the definite article in Celtic

Language	Article
Proto-Insular-Celtic	* <i>sindos</i>
Old Irish	<i>in</i>
Irish	<i>an</i>
Scots Gaelic	<i>an</i>
Welsh	<i>y</i>
Cornish	<i>an</i>
Breton	<i>an</i>

Within the major clades, there is in several cases more than one ancestral class. Rows two and three of Table 3, for instance, reveal that there are (at least) two ancestral classes for the definite article in Romance. The first contains the Sardinian languages Nuorese and Cagliari, whose definite articles descend from Latin *ipse* ‘self’ (e.g., Vincent 1997). The other class contains Romance languages whose definite articles continue the Latin distal demonstrative *ille* ‘that’. Other clades that witnessed multiple instances of the emergence of a definite article include Germanic, Indic, and Iranian. The ancestral classes in Tables 3 and 4 are the minimum number of such classes that one needs to posit. It could well be the case, for instance, that one has to recognize more than two ancestral classes for the Romance languages.

The distribution of languages among the ancestral classes in Tables 3 and 4 buttresses the view that articles within the history of Indo-European are innovations, since it suggests that articles arose only after the major clades had already formed. Had articles emerged before Indo-European started to diversify, we would expect ancestral classes to contain languages from more than one major clade. In addition, Tables 3 and 4 provide crucial insight into the rate at which definite and indefinite articles developed. The 17 rows in Table 3, for instance, indicate that definite articles emerged at least 17 times across the tree. The 15 rows in Table 4 indicate that indefinite articles emerged at least 15 times. I return to the issue of transition rates below in section 4.3 below.

It is of course possible for a language to lose an article (e.g., Greenberg 1978). Articles have in fact been lost in the history of Indo-European (e.g., Matras 2002: 96 on loss of articles in certain dialects of Romani). Among the languages in my dataset, however, no evidence suggests that there has been any loss events.⁸

3.2 Establishing precedence

According to Table 1 above, 37 languages in my dataset contain both a definite and indefinite article. These languages are presented according to clade and ancestral class in Table 5. The crucial question for my investigation is which of the articles emerged first. Two types of evidence offer insight into this question: textual records and parsimony analysis.

In some languages, one article appears before the other in the textual record. In Greek, for

⁸If one follows Kent (1944) in the view that Old Persian had a definite article, this would be an example of the loss of a definite article in the history of Western Iranian. The Old Persian determiner does not, however, meet the criteria in (10) above for articlehood. Ringe (2017: 195) hypothesizes that the *n*-stem suffix of Germanic weak adjectives was originally a definite article. If correct, this would be a case of article loss among the languages in my dataset. Since the origin of the *n*-stem suffix is still debated, I have not adopted this view.

Table 3: Minimal ancestral classes for the definite article

Ancestral Class	Clade	Languages
1	Celtic	Old Irish, Irish, Scots Gaelic, Welsh, Breton, Cornish
2	Romance	Nuorese, Cagliari
3	Romance	Romanian, Arumanian, Catalan, Portuguese, Spanish, French, Provençal, Walloon, Ladin, Romansh, Friulian, Italian
4	Germanic	Old West Norse, Icelandic, Faroese, Norwegian, Swedish, Danish
5	Germanic	English
6	Germanic	Frisian
7	Germanic	German, Luxembourgish, Swiss German
8	Germanic	Dutch, Flemish, Afrikaans
9	Albanian	Tosk, Arvanitika
10	Greek	Ancient Greek, Modern Greek
12	Armenian	Classical Armenian, Eastern Armenian, Adapazar
12	Slavic	Macedonian, Bulgarian
13	Iranian	Sogdian
14	Iranian	Kurdish
15	Indic	Assamese, Oriya, Bengali
16	Indic	Romani
17	Indic	Singhalese

instance, the emergence of the definite article preceded that of the indefinite article, since in the classical language there is only a definite article (e.g., Horrocks 2010: 292). Indefinite articles appear only at later stages of the language:

- (17) a. *Definite article (Classical Greek)* (Lys. 12.10)

eiselthōn eis tō dōmátion tēn
 enter.AOR.PART.SG.ACT into DEF.ART.ACC.SG bedroom.ACC.SG DEF.ART.ACC.SG
kibōtōn anoígnymi.
 chest.ACC.SG open.1SG.PRES.ACT.IND

‘I went into the bedroom and opened **the chest**.’

- b. *Bare indefinite (Classical Greek)* (Lys. 12.9)

eípon hóti tálanton argyríou hétoimos
 say.1SG.AOR.ACT.IND COMP talent.ACC.SG silver.GEN.SG prepared.NOM.SG
eíēn doûnai.
 be.1SG.PRES.ACT.OPT give.INF.AOR.ACT

‘I said that I was prepared to give (him) **a talent** of silver.’

- (18) a. *Definite article (Modern Greek)*

Table 4: Minimal ancestral classes for the indefinite article

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

ðen éxō *ðiavási* *to* *vivlío*
 NEG have.1SG.PRES.ACT read.PERF.PART DEF.ART.ACC.SG book.ACC.SG

‘I have not read **the book**.’

b. *Indefinite article (Modern Greek)*

Agórasa *éna* *vivlío*.
 buy.1SG.AOR.ACT.IND INDEF.ART.ACC.SG book.ACC.SG

‘I bought **a book**.’

Similar evidence reveals that the definite article also emerged first in Armenian (e.g., Klein 1996, Clackson 2008: 140–141, Dum-Tragut 2009: 102, 105–111), Dutch (van der Horst 2008: 388–392, Van de Velde 2010: 271, Van de Velde & van der Horst 2013: 70), English (Lockwood 1968: 90, Van de Velde 2010: 271), German (Lockwood 1968: 87, 90, Keller 1978: 206, Flick 2020: 16), Frisian (e.g., Bremmer 2009: 54), and Old Norse and its descendants (e.g., Faarlund 2004: 56–59).

In some languages with both a definite and indefinite article, we can infer the precedence of the former on the basis of parsimony. For instance, all of the Celtic languages in my sample have a definite article that descend from a common ancestor. The ancestor of the definite articles is likely to have existed before the Celtic languages in my sample began to diversify. By contrast, only one of the Celtic languages in my sample has an indefinite article, Breton. According to the most parsimonious history, this lone indefinite article arises after the definite article comes into existence. Such an account posits just two changes: an early one for the definite article, and a later one that takes place only in Breton. Any model in which the indefinite article of Breton

Table 5: Languages with both a definite and an indefinite article

Clade	Languages
Celtic	Breton
Romance	Nuorese, Cagliari
Romance	Romanian, Arumanian, Catalan, Portuguese, Spanish, French, Provençal, 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
Iranian	Kurdish
Indic	Assamese, Oriya, Bengali
Indic	Singhalese

precedes the rise of the definite article will entail more than two changes and will not be maximally parsimonious.

Parsimony does not always yield sound inferences, however. Among the Romance languages, for instance, there are two ancestral classes for the definite article (one based on Latin *ille* ‘that’, the other on Latin *ipse* ‘self’) but only one for the indefinite article (based on Latin *unus* ‘one’). According to the most parsimonious history, the indefinite article develops first, followed by the definite article. Both of these changes would predate the speciation of Romance. Such an account is at odds with the evidence, however, since there is no evidence for *unus* ‘one’ as an indefinite article at any stage of Latin (Ledgeway 2016: 766; *pace* de la Villa Polo 2011: 226–233). Indeed, the *communis opinio* is that the indefinite article emerges only in medieval Romance (Ledgeway 2012: 85, with further references). By contrast, *ille* and *ipse* have both been argued to exhibit some of the behavior of a definite article even in Latin (e.g., Trager 1932, Nocentini 1990).

Table 6 lists the languages for which we have evidence that a definite article emerged before an indefinite article.⁹ In most languages from my sample with two articles, the emergence of the definite article preceded that of the indefinite. We cannot assume that this was the only pathway of development, however, since there are languages in which the evidence does not permit one

⁹Tosk and Arvanitika in all likelihood also belong on this list. The definite article in these languages originates in the **so-/to-* demonstrative. In the accusative singular, definite masculine nouns end in *-n(ë)*, which ultimately descends from a string **-om#tom* (Matzinger 2006: 74, 94–95, Schumacher 2009: 56–58). At some point *-n(ë)* must have had a geminate nasal, i.e., *-nn(ë)*, which blocked rhotacism of the *n* in Tosk. The definite article thus had to develop very early in the history of Albanian (exactly how early is difficult to say, however). Given the early emergence of the definite article, it is hard to imagine that the indefinite article could have preceded it, but, strictly speaking, there is no way to know for certain which article developed first. Both articles are attested in the earliest Albanian corpora.

to determine which article developed first. It is in principle possible for an indefinite article to develop before a definite article, since 5 languages in my sample possess just an indefinite article: Baluchi, Wakhi, Zazaki, Persian, and Tajik (which are discussed in section 6.3 below).

Table 6: 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, Provençal, 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

4 Data and methods

4.1 Coding the data

Two binary states were used to model the history of definite and indefinite articles. If a determiner met the criteria described above in example (10) for articlehood, the language was assigned a value of one. Otherwise it was assigned a value of zero. Table 7 presents the four possible combinations of character states, which are all attested in my sample.¹⁰ This representation of the data focuses exclusively on the presence or absence of definite and indefinite articles. So long as the necessary and sufficient criteria for articlehood presented above in section 2.1 were met, a language was counted as having a definite or indefinite article. The pairs of character states in Table 7 were then transformed into a single character with four possible states (corresponding to 00, 10, 01, and 11). The character state for each language is presented in Figure 2.

Table 7: Representation of the data as two binary characters

Definite Article	Indefinite Article	Inventory	Example
0	0	No articles	Russian
1	0	Definite article only	Old Irish
0	1	Indefinite article only	Persian
1	1	Definite and indefinite article	English

¹⁰One can compare the more extensive coding scheme of Carling (2019: 75–81), which takes into consideration the linear position of articles and definiteness marking on adjectives.

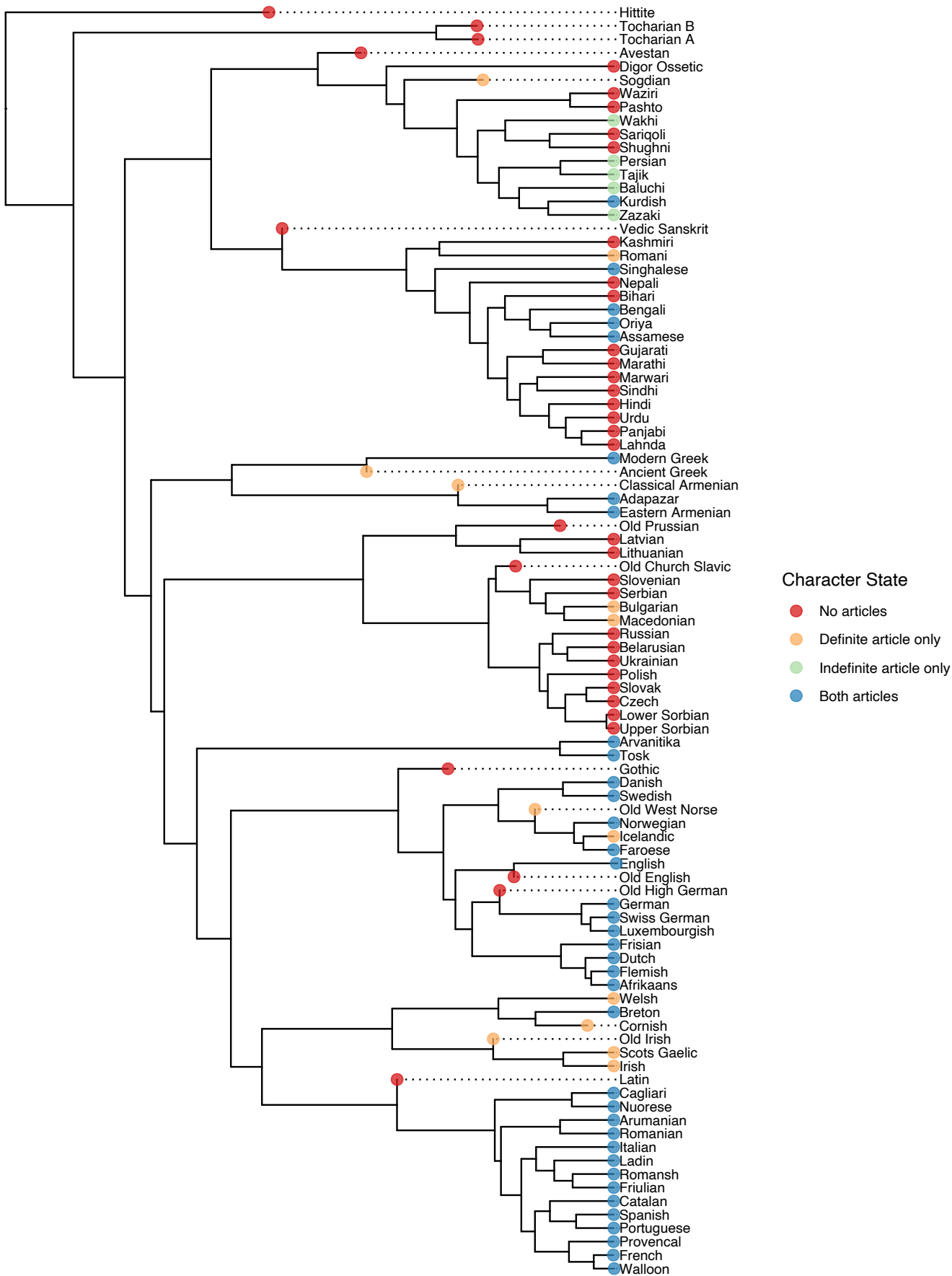


Figure 2: Maximum a posteriori tree of the A₃ dataset and model of Chang et al. 2015 with observed character states

4.2 Phylogenetic trees

Despite over a century of investigation, the true Indo-European tree remains unknown (Widmer 2018: 374). To account for phylogenetic uncertainty, the posterior rate estimates are based on a sample of one hundred trees from the A_3 dataset and model of Chang et al. (2015a). This model and dataset are characterized by the following properties:

- (19) *A₃ dataset and model*
 - a. Ancestry constraints
 - b. Broad dataset, no time constraints on splits
 - c. Amended IELEX dataset
 - d. 197 meaning classes
 - e. Loan inclusion
 - f. Continuous-time Markov Chain with gamma-distributed among site rate variation (4 classes)
 - g. Generalized skyline coalescent

The maximum a posteriori (MAP) tree from the A_3 dataset and model is presented in Figure 2. One hundred trees were randomly sampled from the output of this model and dataset, which are presented in Figure 3. The median age of the sampled trees is 5920 years (cf. Chang et al. 2015b: 6).

4.3 Transition rates

As noted above in section 1, there is no standard method for testing hypotheses of dependence between linguistic changes. Pagel (1994) and Pagel & Meade (2006) present a method for evaluating claims of correlated evolution among two discrete binary traits. This method is not common in historical linguistics, although there is now a handful of studies that make use of it (Dediu & Levinson 2012, Haynie & Bowerman 2016, Dunn et al. 2017, Widmer et al. 2017, Blasi et al. 2019). The following questions lie at the heart of this framework:

- (20) *Establishing correlation*
 - a. Does the rate at which an indefinite article develops differ if a definite article is present in the language?
 - b. Does the rate at which a definite article develops differ if an indefinite article is present in the language?

To estimate the rates at which definite and indefinite articles emerged, transitions between the four character states in Table 7 are modeled as continuous-time Markov chains or CTMCs (Cathcart 2018, Jäger 2019a: 166–167). CTMCs model language change as a stochastic process, that is, the transitions among states are probabilistic. In addition, they bring with them the following assumptions about language change. First, they are memoryless: the probability of a transition depends only on the current state. So if a language is in state (1,0), the probability of transitioning to state (1,1) is determined solely by that state. States of the language prior to (1,0) are irrelevant. Second, transitions between character states are assumed to be independent. Transitions

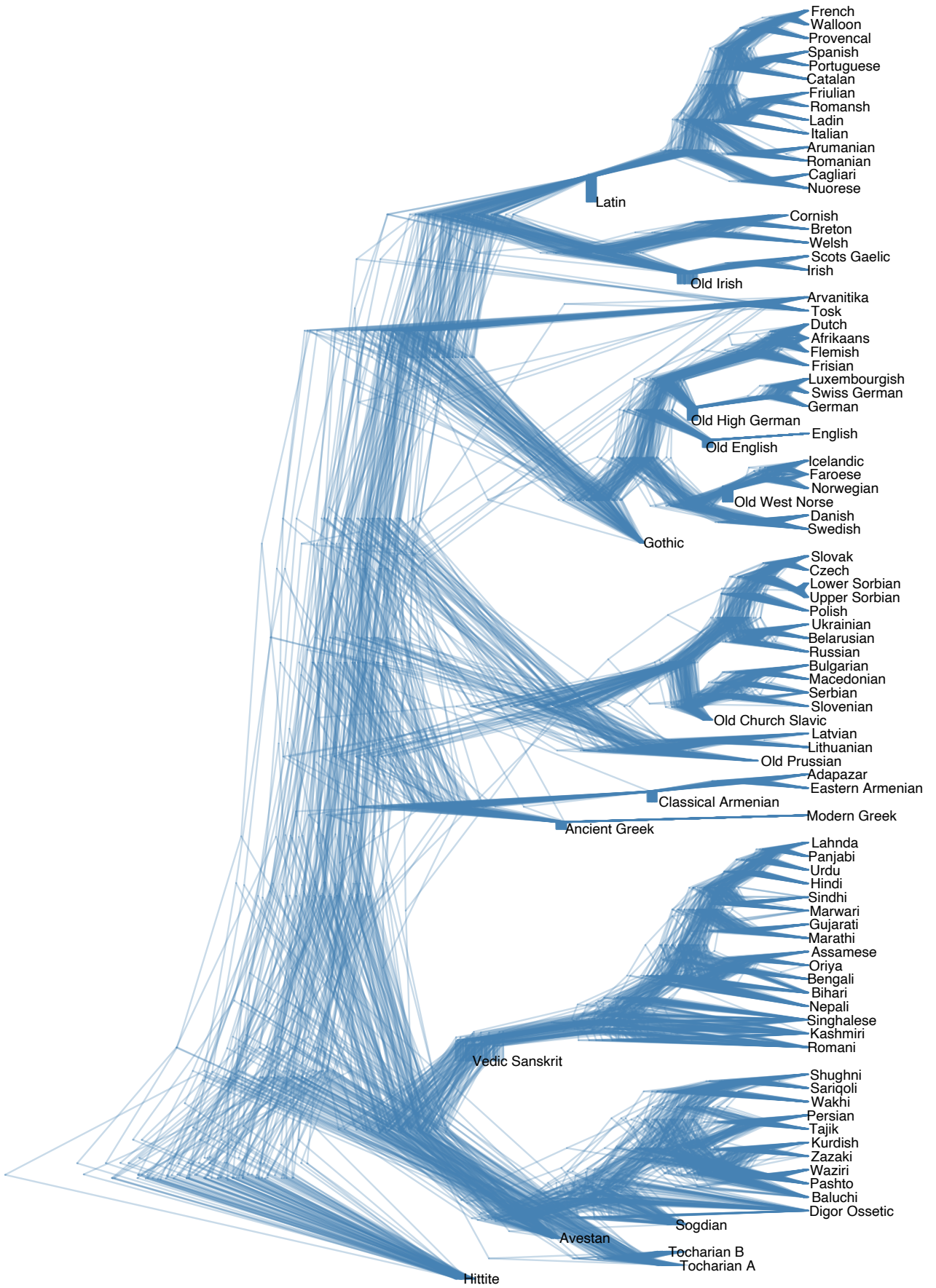


Figure 3: One hundred phylogenetic trees from the A3 dataset and model of Chang et al. 2015

in one region of the tree are thus assumed to have no impact on transitions in another part of the tree. CTMCs assume a constant rate of change for each rate parameter across the entire tree and thus do not countenance the possibility that the rate of a particular transition is faster along some branches than others. Finally, CTMCs simplify linguistic reality in that the emergence of an article is a discrete event (e.g., a change from state 0,0 to state 1,0) and not a gradual phenomenon, which is typically how grammaticalization is believed to proceed (e.g., Greenberg 1978: 252, de Mulder & Carlier 2011: 530).

Transition rates are estimated with the following rate matrix:

$$\begin{array}{c}
 (0, 0) \quad (1, 0) \quad (0, 1) \quad (1, 1) \\
 \begin{array}{c}
 (0, 0) \\
 (1, 0) \\
 (0, 1) \\
 (1, 1)
 \end{array}
 \begin{bmatrix}
 - & q_{12} & q_{13} & 0.0 \\
 0.0 & - & 0.0 & q_{24} \\
 0.0 & 0.0 & - & q_{34} \\
 0.0 & 0.0 & 0.0 & -
 \end{bmatrix}
 \end{array}$$

The rows represent starting states and the columns end states. The rate q_{12} in the first row and second column is the rate at which a language transitions from character state (0,0) to (1,0). In other words, it is the rate of transitioning from having no articles to having a definite article. The values in the cells with dashes are determined by the negated sum of the values in each row. Transitions with a rate of zero cannot occur. Two types of transitions are set to zero in the above matrix, dual transitions and loss events. A dual transition involves a simultaneous change of both characters, e.g., (0,0) to (1,1). Although a transition from (0,0) to (1,1) is of course possible over a longer duration, it is highly unlikely that such a change ever took place simultaneously. The rate parameters for loss events are set to zero because above in section 3.1 I established that the history of articles in my sample of languages is one of gain. So once a character transitions to 1, it cannot transition back to 0 (i.e., the model is irreversible).

Table 8: Descriptions of the rate parameters

Rate Parameter	Description	Change	Background State
q_{13}	0,0 > 0,1	Gain of indefinite article	Definite article absent
q_{24}	1,0 > 1,1	Gain of indefinite article	Definite article present
q_{12}	0,0 > 1,0	Gain of definite article	Indefinite article absent
q_{34}	0,1 > 1,1	Gain of definite article	Indefinite article present

Table 8 presents descriptions of each of the rate parameters. As noted above in example (20), the crucial question is the extent to which the background states affect the rates of gain. In essence, this is a question of how similar the rates for the gain of each type of article are (i.e., how similar q_{13} and q_{24} are and how similar q_{12} and q_{34} are).

The similarity of the rate parameters are assessed in two ways. First, their posterior distributions can be visually compared, as is done in section 5.1 below. Second, the model above can be compared to a model that has only a single rate parameter for the development of indefinite articles:

$$\begin{array}{c}
(0, 0) \quad (1, 0) \quad (0, 1) \quad (1, 1) \\
\begin{array}{c}
(0, 0) \\
(1, 0) \\
(0, 1) \\
(1, 1)
\end{array}
\begin{bmatrix}
- & q_{12} & q_{\alpha} & 0.0 \\
0.0 & - & 0.0 & q_{\alpha} \\
0.0 & 0.0 & - & q_{34} \\
0.0 & 0.0 & 0.0 & -
\end{bmatrix}
\end{array}$$

The rate parameters q_{13} and q_{24} have been merged into a single parameter q_{α} . This model is less complex than the one presented above since it has one less rate parameter. If the rate at which an indefinite article is gained is not sensitive to the presence of a definite article, then this simpler model will be a better fit for the data. If the rate at which an indefinite article is gained is sensitive to presence of a definite article, then the additional rate parameter will be justified.

4.4 Model comparison

To compare the two models, I estimate their marginal likelihoods using path sampling and stepping stone analysis. The marginal likelihood is the probability of the data given a particular model. In Bayesian phylogenetics, models are standardly (although not uncontroversially) compared by means of Bayes factors (BF). The Bayes factor is the ratio of the marginal likelihoods of competing models:

$$BF_{10} = \frac{p(y|M_1)}{p(y|M_0)}$$

BF_{10} denotes the extent to which the data support M_1 over M_0 . For model comparison, I use the discrete categories of log Bayes factors presented in Table 9. It is worth noting that Bayes factors measure the *relative* fit of a model to data. They do not measure model adequacy (see further Jäger 2019b).

Table 9: Interpreting log Bayes Factors (Höhna et al. 2017: 27)

Strength of Evidence	Log-BF(M_1, M_0)
Negative (supports M_0)	< 0
Barely worth mentioning	0 to 1.16
Substantial	1.16 to 2.3
Strong	2.3 to 4.6
Decisive	> 4.6

To estimate the marginal likelihoods, path-sampling and stepping-stone analyses were run for ten thousand generations with a warm-up phase of two thousand generations. Analyses were run multiple times to ensure the stability of the results. The values were nearly identical so in the interest of space I report only the stepping stone values below. Both analyses were based on a power posterior analysis with 127 categories.

In most real-world applications of Bayes' Theorem, it is not possible to calculate the posterior probability analytically. The standard practice is instead to use Markov Chain Monte Carlo

to sample from the posterior distribution. The results presented in the next section are based on analyses in which six independent chains were run for one million generations, with samples being taken every one thousandth generation. The six chains were then merged into one and the first twenty-five percent of these samples were discarded as burn-in for the calculation of the posterior probabilities. Convergence was confirmed through visual inspection. The analyses were carried out with RevBayes version 1.1.1 (Höhna et al. 2016).

4.5 Informative priors

Bayesian inference is distinguished from frequentist inference by the inclusion of priors, which provide probability distributions for parameter values before any data has been observed. Vague priors distribute prior probability over a wide range of parameter values, whereas informative priors focus on specific values. I use informative priors for the probability of the character state at the root and the hyperpriors over the rate parameters. On the basis of the evidence in section 3.1 above, the root of the tree is assumed to be in the state (0,0), i.e., to have neither definite nor indefinite articles. Second, the values for the rate parameters are drawn from exponential distributions, whose hyperparameter is the sum of the branch lengths of the tree divided by the approximate number of changes n that took place on the tree:

$$q_{ij} \sim \text{Exp}(\lambda)$$

$$\lambda = \text{Length}(\text{Tree})/n$$

In sections 3.1 and 3.2 above, I estimated the number of times definite and indefinite articles developed. The estimates provide evidence for two rate classes, a faster one and a slower one. The faster transitions are (0,0) > (1,0) and (1,0) > (1,1). On the basis of the ancestral classes in Table 3 above, the transition (0,0) > (1,0) occurred at least 17 times. Tables 5 and 6 offer an estimate for number of times the transition (1,0) > (1,1) occurred. There are 14 ancestral classes in Table 5, in at least 9 of which a definite article emerged first (as laid out in Table 6). The slower transitions are (0,0) > (0,1) and (0,1) > (1,1). Since a definite article emerged first in at least 9 of 14 ancestral classes in Table 5, the transition (0,1) > (1,1) occurred around 5 times. The transition (0,0) > (0,1) happened at least 3 times (ancestral classes 11, 12, and 13 of Table 4) and may have happened as many as 8.

On the basis of these estimates, I used the following discrete uniform hyperprior distributions for the number of transitions in each rate class:

$$n_{\text{fast}} \sim \mathcal{U}(9, 17)$$

$$n_{\text{slow}} \sim \mathcal{U}(3, 8)$$

These are strong prior distributions, but they are justified on the basis of the empirical evidence presented in sections 3.1 and 3.2 above. Using more diffuse priors would be tantamount to disavowing the insights that can be gleaned from traditional methods.

5 Results

5.1 Posterior distributions of the rate parameters

The posterior distributions of the rate parameters are presented in Figures 4 through 6 below (dashed vertical lines signal the median). The posterior distributions in Figure 4 reflect the rates at which definite and indefinite articles emerge among languages with no articles. The red indefinite distribution (which reflects the change $0,0 > 0,1$) has a median value of 0.000036 . Its 95% highest posterior density (HPD) interval is $[0.000002, 0.000090]$. The orange definite distribution (which reflects the change $0,0 > 1,0$) has a median value of 0.00023 and 95% highest posterior density (HPD) interval of $[0.00012, 0.00036]$. On average, definite articles are emerging faster than indefinite articles from languages that lack articles.

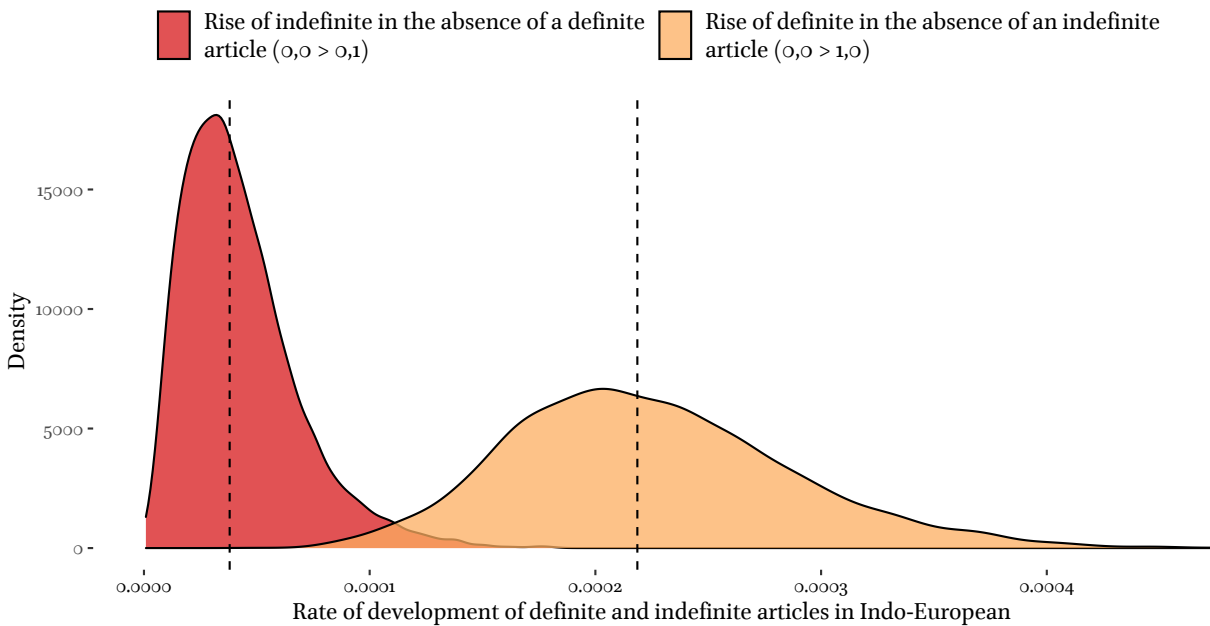


Figure 4: Posterior distributions of the rate parameters for definite and indefinite articles

The posterior distributions in Figure 5 provide strong support for the claim of correlated grammaticalization. The two distributions reflect the rise of indefinite articles in languages with a definite article (blue distribution) and without a definite article (red distribution). The median and 95% HPD of the latter were presented above. The median of the blue distribution is 0.00064 and its 95% HPD interval is $[0.0003, 0.0010]$. Crucially, these distributions do not overlap and we infer from them that indefinite articles emerged faster among languages with a definite article. The blue distribution is considerably wider than the red distribution, which reflects the greater uncertainty in the estimation of this parameter. Although it is clear that indefinite articles emerged at a faster rate among languages with a definite article, it is harder to say exactly how much faster.

Figure 6 presents the posterior distributions for the rate of the emergence of definite articles in languages with no article (orange) and in languages with an indefinite article (green). The median and 95% HPD of the former was presented above. The median of the green distribution (which represents the change $1,0 > 1,1$) is 0.00007 and its 95% HPD interval is $[0.0000009, 0.0002452]$. In contrast to the distributions above, the 95% HPD intervals of the distributions in Figure 6 overlap.

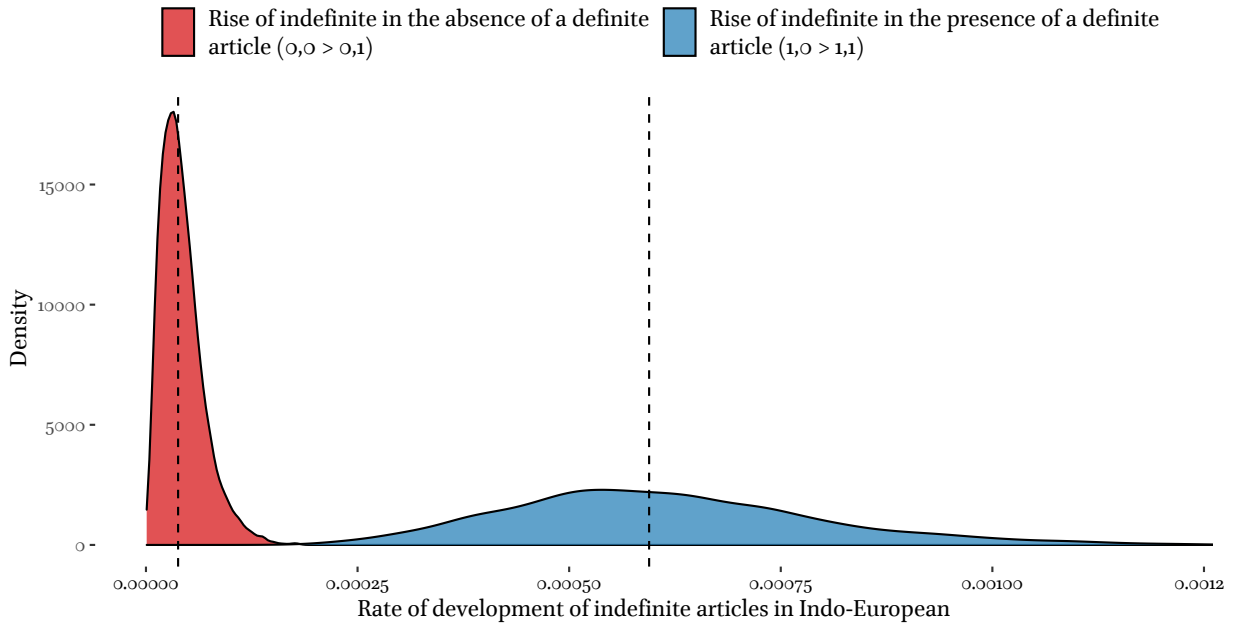


Figure 5: Posterior distributions of the rate parameters for indefinite articles

This suggests that rate at which definite articles emerge may well not be sensitive to the presence of an indefinite article.

5.2 Model comparison

The marginal log-likelihood of the proposed model and a simpler model with a single rate parameter for the gain of an indefinite article are presented in Table 10. Linguistically the difference between these two models boils down to the question of whether the existence of a definite article impacts the rate at which indefinite articles emerged. The proposed model has decisive support over the simpler model, which buttresses the claim that the grammaticalization of indefinite articles is sensitive to the existence of a definite article.

Table 10: Model comparison

Model 1	Model 0	Log-BF ₁₀	Support
Two indefinite rate parameters	Single indefinite rate parameter	5.31	Decisive

6 Discussion

The posterior rate distributions support the following diachronic trajectory for Indo-European:

(21) *Diachronic trajectory*

No articles > Definite Article > Definite Article, Indefinite Article

This diachronic trajectory results in the following synchronic generalization:

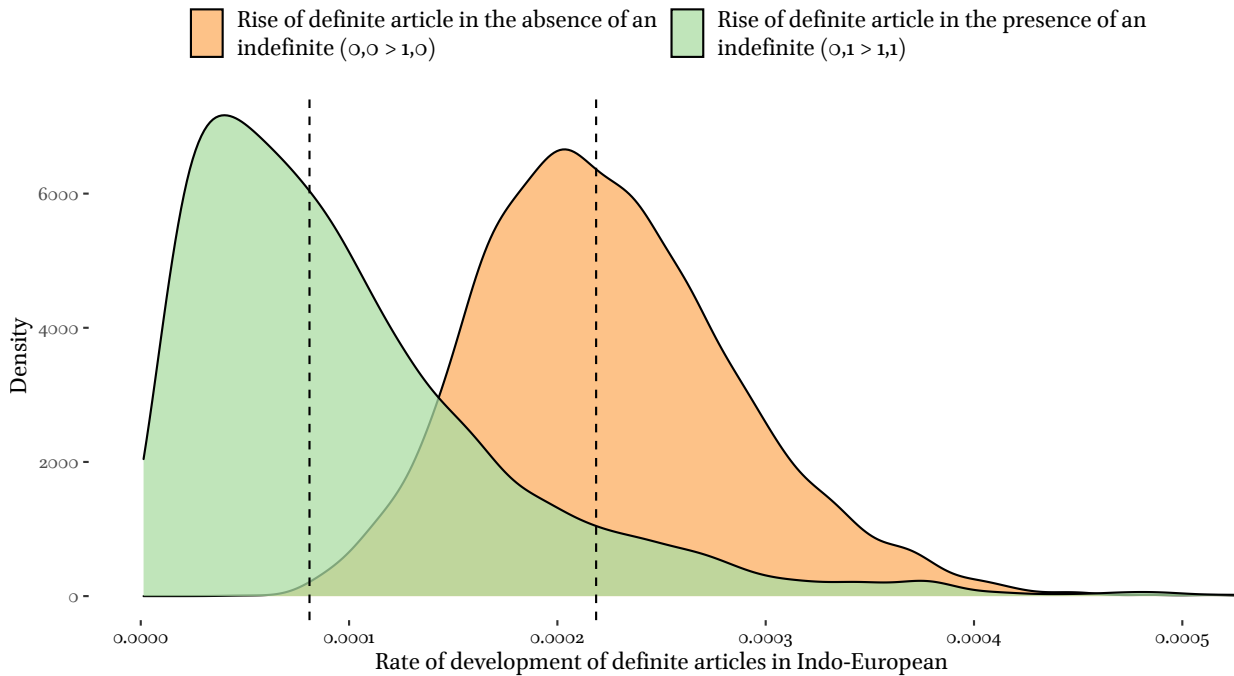


Figure 6: Posterior distributions of the rate parameters for definite articles

(22) *Synchronic generalization*

A language that marks indefinite referents with an article marks definite referents with an article (but not vice versa).

This is thus a prime example of how diachronic change creates synchronic typological patterns (cf. Greenberg 1966, Bybee 1988, Bybee 2006, Kiparsky 2008).

These results prompt the following questions, which I take up in this section:

(23) a. *The directionality question*

Why do definite articles emerge before indefinite articles?

b. *The correlation question*

Why does the presence of a definite article increase the likelihood of the emergence of an indefinite article?

c. *The exceptions*

Why do some languages counter the predominant diachronic trend?

d. *The role of language contact*

To what extent has the emergence of articles in Indo-European been due to language contact?

In section 6.1 below, I argue that the emergence of articles in Indo-European typically follows a scale of referential prominence, according to which more prominent referents are marked before less prominent referents. As for correlation, I argue in section 6.2 that indefinite articles emerge by analogy to definite articles. Once a language begins to use articles to mark definite referents, speakers extend that pattern to indefinite articles. Section 6.3 then discusses the exceptional patterns of definiteness marking in Iranian and section 6.4 discusses the role of language contact in

the grammaticalization of articles. I conclude by highlighting the implications of this study for our understanding of grammaticalization.

6.1 The directionality question

Investigations of definiteness reach back well over a century. During this time, two prominent approaches have emerged (Birner & Ward 1994: 93, Becker 2018: 49). One maintains that the central feature of definiteness is uniqueness (e.g., Russell 1905, Hawkins 1991, Birner & Ward 1994, Abbott 1999, Kadmon 2001, Abbott 2008). According to this view, a definite description has just one referent that satisfies the description in a definite noun phrase (Lyons 1999: 8, Kadmon 2001: 79–80). The second approach takes familiarity to be the crucial property of definiteness (e.g., Christophersen 1939, Heim 1991, Heim 2011). The basic intuition of this approach is that definite articles reflect shared knowledge of a referent.

Becker (2018: 53) argues that neither uniqueness or familiarity alone is sufficient to account for the distribution of definite expressions (a point that has also been made by, e.g., Birner & Ward 1994: 93 and Epstein 2002: 40). Familiarity is too restrictive to capture the range of contexts in which definite articles occur since its focus lies squarely on anaphoric definites. It is consequently unclear how to bring non-anaphoric uses of definite expressions into the remit of this approach. The uniqueness approach, according to Becker, is less restrictive, but brings with it a different challenge: uniqueness has to be relativized to make it compatible with different discourse scenarios. Consider the following pair of examples:

- (24) a. **The sun** is about 93 million miles from **the earth**.
b. Follow this road to get to **the city center**.

The sun and the earth are absolutely unique, but city centers are unique only with respect to particular cities. Following insights from Christophersen (1939: 72–73) and Löbner (1998: 12), Becker (2018: 53) adopts a modified definition of uniqueness rooted in mutual and unambiguous identifiability. Referents that are mutually and unambiguously identifiable are guaranteed to be the only salient referents of their kind.

(25) *Definite referent*

A definite referent is the maximal number of referents of a given kind that are mutually and unambiguously identifiable by the speaker and the hearer (Becker 2018: 57).

Becker (2018: 57–65) then explicates how the sub-types of definite referents that she recognizes fall out from mutual and unambiguous identifiability.

Turning to indefinite referents, Becker (2018: 55) contends that indefiniteness is distinguished from definiteness by not being a primitive referential type. It is instead the union of two referent types, specific and non-specific referents. Specific referents are single and particular referents from their kind set, as in the following example:

(26) *Indefinite specific*

Sarah wanted to talk to **a colleague of mine**, but I forgot which one.

Neither the speaker nor the hearer can identify the referent of *a colleague of mine*, but crucially this phrase is associated with a particular referent. In other words, the referent of the phrase cannot simply be any member of the kind set ‘colleague of mine’.

Indefinite non-specific expressions do not refer to any particular member of the kind set denoted by the referential expression, as in the following example from Becker (2018: 68):

(27) *Indefinite Non-specific*

Lea wants to buy a bike, but she doesn't know if she will find one.

The expression *a bike* can potentially refer to any member of the kind set 'bike', since the hearer is given no information that could be used to identify the specific member (Becker 2018: 68).

The three referential types just introduced can be ranked according to their referential prominence. Definite referents are most prominent because they are mutually identifiable by the speaker and the hearer. Indefinite specific and non-specific referents do not meet this threshold. Indefinite specific referents are in turn more prominent than indefinite non-specifics because they are linked to a particular referent of a kind set, whereas indefinite non-specifics are not.

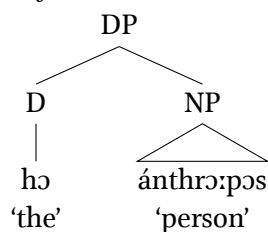
The grammaticalization trajectory proposed in section 5 can now be recast along a cline of referential prominence. Articles emerge first among referents that are mutually and unambiguously identifiable and only later among those that are not. Although the results in section 5 above pertain only to the precedence of definite before indefinite articles, there is limited evidence that suggests that articles emerge first among indefinite specific referents and only later among indefinite non-specifics.

6.2 The correlation question

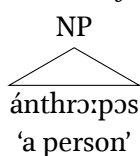
The results in section 5 support the claim that the grammaticalization of definite and indefinite articles in Indo-European is correlated: once a definite article emerges, an indefinite article becomes more likely to develop. As a result, more languages end up with both articles than simply a definite article (cf. Lakoff 1972: 174). In this section, I argue that the development of an indefinite article comes about by analogy, which may ultimately be motivated by consistency in referential marking.

Among languages with two articles, the grammaticalization of the first article has more of an impact on the language. It introduces the category of definiteness into a language (Haspelmath 2018: 113). When the second article is recruited, by contrast, definiteness already exists as a category in the language. The grammaticalization of the first article also introduces a morphosyntactic asymmetry into the language. Referential marking in Classical Greek, for instance, is asymmetric in that an article is used for definite referents, but not for indefinite referents:

(28) *Definite article*

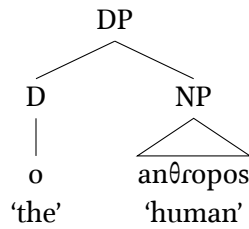


(29) *Bare indefinite*

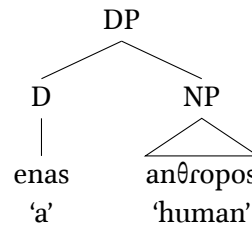


The recruitment of an indefinite article removes this asymmetry (at least for singular nouns), as illustrated by Modern Greek:

(30) *Definite article*



(31) *Indefinite article*



With the grammaticalization of an indefinite article, the erstwhile asymmetry is gone and referential marking is now uniform for both definite and indefinite referents. Pre-existing morphosyntax thus paves the way for the grammaticalization of the indefinite article.¹¹

It is not yet clear if the structural consistency brought about by the analogical change proposed in this section is a cause or an effect. That is, the grammaticalization of an indefinite article could have come about specifically for the purpose of making the referential marking in examples (28) and (29) consistent. Alternatively, the structural consistency that results from the analogical change could be an effect of a broader set of morphosyntactic changes. To take the example of Latin and the Romance languages, besides articles the latter also witnessed an increase in the use of auxiliary verbs and the loss of case endings (e.g., Baldinger 1968, Wandruszka 1980, Vincent 1997, 1999). It is possible that the causes of these changes are also at work in the grammaticalization of articles.

6.3 The exceptions

Five languages in my sample are represented as having exclusively an indefinite article: Baluchi, Wakhi, Zazaki, Persian, and Tajik. These are all Iranian languages. In fact, with the exception of Wakhi, they are all West Iranian languages. These languages counter the diachronic trajectory in example (21) above, since they acquired an indefinite article in the absence of a definite article. Although these languages lack definite articles, they do mark definite referents. Baluchi (Korn & Nourzaei 2018: 665), Wakhi (SanGregory 2018: 72), Persian (Jasbi 2020), and Tajik (Perry 2005: 66–71, 287–288) have definite-/specificity-based differential object marking. The following examples of differential object marking from Persian are illustrative:¹²

(32) *Differential object marking in Persian*

(Jasbi 2020: 127)

¹¹The proposal above requires two caveats. First, I have assumed that the grammaticalization of a definite article introduces DP structures into a language. There a number of scholars who subscribe to the Universal DP-Hypothesis, according to which DPs would exist even before articles developed. Giusti (2001), Giusti & Iovino (2016), for instance, contend that Latin already has DPs and Pereltsvaig (2007) makes the same argument for Russian. If DPs are universal, my proposal above would have to be reformulated as follows. The asymmetry that a definite article introduces would not be syntactic (i.e., DP versus NP), but rather an asymmetry in overt versus null exponence. The grammaticalization of an indefinite article would result in consistent overt exponence in referential marking. Second, not all syntactic accounts of definite and indefinite articles locate them in the same syntactic projection. In examples (30) and (31), both the definite and indefinite article project a DP. Wood (2003), for instance, argues that the indefinite article in English heads a projection NumP. Under such an analysis, my proposal would again have to be reformulated in terms of consistency of exponence (as opposed to syntactic consistency) to allow for articles to head different projections.

¹²In example (32b), I have replaced the colloquial form of the object marker *-o* in the original example with its formal variant *-rā*.

- a. *Ali ketāb xarid*
 Ali book buy.3SG.PST
 ‘Ali bought **one or more books.**’
- b. *Ali ketāb-rā xarid*
 Ali book-OM buy.3SG.PST
 ‘Ali bought **the book.**’

The bare noun *ketāb* ‘book’ in example (32a) has the indefinite interpretation ‘one or more books’. It is the addition of the object marker *-rā* in example (32b) that is responsible for the definite reading ‘the book’. Becker (2019) notes that the combination of differential object marking plus indefinite marker is common in Turkic and Indo-Iranian.

Referential marking in Zazaki is less clear. Although it has been described as lacking differential object marking (e.g., Haig 2016: 465), the oblique case marker is nevertheless sensitive to definiteness (Paul 1998: 21–26):

- (33) *Southern Zazaki oblique case*
- a. *Indefinite non-specific* (Todd 2008: 39)
- tu bergir ramenê t’stor ramenê*
 2SG.PRO.DIR stallion ride mare ride
 ‘Do you ride **a stallion or a mare?**’¹³
- b. *Definite* (Todd 2008: 40)
- ma bergir-an ramenê t’stor-an ramenê*
 1PL.PRO.DIR stallion-OBL.PL ride mare-OBL.PL ride
 ‘Shall we ride **the stallions or the mares?**’

The bare nouns *bergir* ‘stallion’ and *t’stor* ‘mare’ in example (33a) appear to be indefinite and non-specific. The addition of the oblique suffix *-an* in example (33b) yields a definite reading. Todd (2008: 39) describes the distribution thus: “Nouns do not take the Oblique case markers when the reference is indefinite and nonspecific.” It is also worth noting that the indefinite article *-ê* cannot co-occur with an oblique case suffix:

- (34) *Zazaki indefinite* (Paul 2009: 549)
- šew-ê*
 night-INDEF
 ‘for a night’

The indefinite article has blocked the appearance of an oblique case marker.

Although more research is required to understand how exactly definiteness is realized in Zazaki, the five exceptional languages in my dataset appear to be only superficial exceptions to the proposed diachronic trajectory, since definiteness is a category in these languages. The question that remains to be answered is whether the marking of definite referents emerged before the indefinite articles. This question cannot be pursued here, but it raises the prospect that the diachronic trajectory in example (21) should perhaps be formulated in terms of definite and indefinite marking more broadly and not specifically in terms of definite and indefinite articles.

¹³This example may be more accurately rendered ‘Do you ride stallions or mares?’

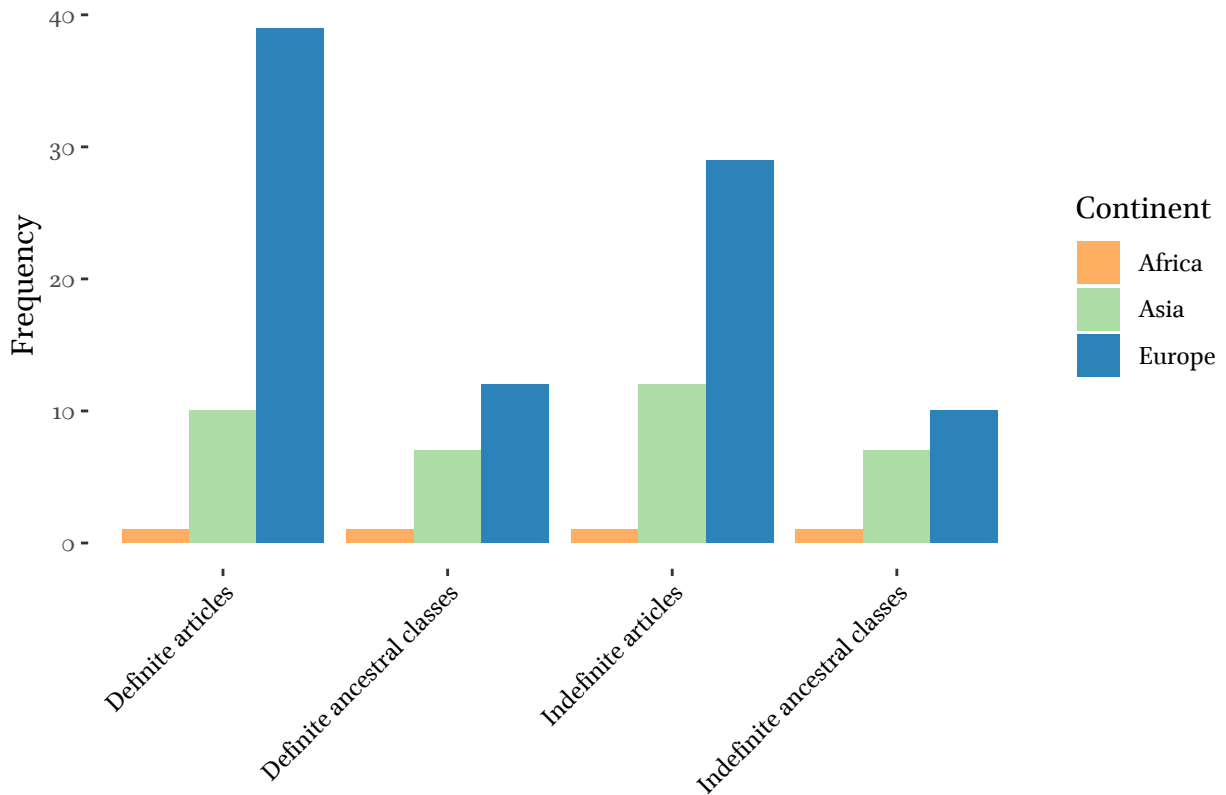


Figure 7: The frequency of articles and ancestral classes according to continent

6.4 The role of language contact

The areal distribution of articles in Indo-European is known to be skewed (e.g., Schwyzer 1936: 146–147, de Mulder & Carlier 2011: 522, Dryer 2014: e244), as articles preponderate in Western and Central Europe (i.e., the among Romance and Germanic languages) and in the Balkans (Greek, Albanian, Bulgarian, Macedonian, and Romanian). Figure 7 presents the frequency distributions of definite articles, definite ancestral classes, indefinite articles, and indefinite ancestral classes according to continent in my dataset. (There are 36 languages in my sample from Asia, 57 from Europe, and 1 from Africa.) The proportion of languages with definite and indefinite articles is higher in Europe than in Asia, although the differences between these two regions are not as pronounced when one considers the estimated number of ancestral classes.

It unclear how much of the areal disparity is due to language contact or to parallel independent development (i.e., homoplasy). To be sure, language contact is widely believed to play a role in the grammaticalization of articles (e.g., Schaarschmidt 1984: 76, Matras 2002: 96, Heine & Kuteva 2011: 293–297, van Gelderen 2011: 224–225) and in my dataset there are languages whose articles do indeed seem to be due to language contact, such as the definite articles in the Balkans (Paliga 2019, Friedman & Joseph 2022) and in Assamese, Oriya, and Bengali (Masica 1991: 250). Elsewhere, however, it is less clear whether language contact is really responsible for the development of articles. Dahl (2004: 127), for instance, contends that the definite article spread to the languages of Western Europe from the Mediterranean during the medieval period. It is true that among the Romance and Germanic languages, definite articles emerge during the medieval pe-

riod. These developments could be due to language contact, but this is just one possibility among other competing hypotheses. For example, the emergence of articles in both Romance and Germanic could have been caused by internal factors, which would make the grammaticalization of articles in these two clades homoplastic.

The contact hypothesis faces at least three challenges. First, if language contact was so pervasive, why were definite articles so rarely borrowed? The only languages in my dataset in which definite articles appear to have been borrowed are Assamese, Oriya, and Bengali (Masica 1991: 250). Elsewhere contact results in pattern borrowing (that is, speakers do not borrow an article, but create an article within their own language on the basis of articles in another language). Second, every Romance language has both a definite and indefinite article. The ubiquity of articles is surprising under the language-contact hypothesis. It is unclear how contact led to the spread of articles throughout the entire clade. Finally, why did the spread of the definite article happen only in the medieval period? To the best of my knowledge, proponents of the language-contact hypothesis have not offered an answer to this question. Given the existence of the definite article in Classical Greek and the prestige of the language, it seems that a definite article could in principle have made its way to Western Europe centuries earlier. In sum, although the possibility of contact-induced grammaticalization cannot be excluded, it is anything but clear that it is the best hypothesis of the emergence of articles in Romance and Germanic.

As described in section 4.3 above, one of the assumptions of the transition models used in this study is that the character-state transitions are independent. Languages in which an article arose by contact would thus violate the assumption of independence described above in section 4.3. The rate parameters in section 5 above could well be inflated. Despite this, the main claim of this paper—that indefinite articles emerge at a faster rate among languages that already have a definite article—remains intact, since it concerns what happens after a language acquires a definite article (cf. Levinson et al. 2011: 526).

7 Conclusion

The central objective of this study was to demonstrate the critical role that pre-existing morphosyntax can play in grammaticalization. I substantiated this claim by showing that in the history of Indo-European indefinite articles emerged at a faster rate among languages that already possessed a definite article compared to those that did not. I proposed that this correlation is the result of an analogical change that extends the use of referential marking with definite articles to indefinite referents. What emerges from my analysis is a diachronic trajectory of referential marking according to which more prominent referents are typically marked first and less prominent referents later.

These results have significant consequences in a number of domains. First and foremost, they show that grammaticalization does not occur in a vacuum. All grammaticalization takes place in a context (Garrett 2012: 71) and we can only understand why such changes take place—and why they take place when they do—by taking that context into account. There is more to grammaticalization beyond its robust cross-linguistic asymmetries. Second, diachronic trajectories are probabilistic and not deterministic. Although definite articles usually emerged before indefinite articles in the history of Indo-European, this was not always the case. Although languages that acquired a definite article typically went on to develop an indefinite article, this did not always happen. Finally, correlated grammaticalization suggests that phylogenetics and diachronic change

have a central role to play in typological reasoning (cf. Cysouw 2011: 429–430, Levinson et al. 2011: 509). The implicational generalization in example (22) above is inferred not from static frequency distributions, but rather from the estimates of transition rates among linguistic properties in the context of a phylogenetic tree.

The results of this study also open up new questions. First, to what extent is the diachrony of articles in Indo-European lineage-specific? If we investigate other linguistic families, will we find similar correlations? The answers to these questions will shed light on the question of whether the causes of the correlations in Indo-European are due to clade-specific factors or whether these are more general historical trends of natural language (cf. Dunn et al. 2011). Second, is it possible to establish a more fine-grained diachronic trajectory for articles? In this study, I have focused on the broad categories of definite and indefinite readings, but it is entirely plausible that definite articles themselves develop along a common diachronic path (as has in fact been proposed by, e.g., Greenberg 1978: 247, 252–264, Hawkins 2004: 84–86). Finally, how do other forms of referential marking, such as differential object marking and definite adjective marking, figure into the diachronic trends investigated here? In section 6.3 above, I suggested that the rise of indefinite articles in the absence of definite articles in Iranian may have been facilitated by definiteness-based differential object marking. If this turns out to be true, the diachronic trajectory proposed in this study could be reformulated in terms of definite and indefinite marking in general and not specifically in terms of articles.

Acknowledgements

I am grateful to Andrew Garrett, Mark Hale, Ian Hollenbaugh, Stephanie Jamison, Shuan Karim, Sorin Paliga, Giuseppina Silvestri, Brent Vine, Anthony Yates, Dmitrii Zelenskii, and two anonymous reviewers for comments on earlier versions of this paper that significantly improved its content. Jessica DeLisi, Will Freyman, Brian Joseph, Jared Klein, Agnes Korn, Martin Kümmel, Kaspar Ozoliņš, Georges-Jean Pinault, Paul Roberge, Benjamin Slade, and Michiel de Vaan kindly shared their expertise on a range of issues with me. Nick Weiss generously discussed a range of presentational issues with me. I alone am responsible for all remaining faults.

Abbreviations

3: third person, ABL: ablative, ACC: accusative, ACT: active, CONJ: conjunction, DAT: dative, DEF: definite article, DEM: demonstrative, GEN: genitive, INDEF: indefinite, MASC: masculine, NOM: nominative, PART: participle, PASS: passive, PERF: perfect, PRES: present, PTCL: particle, REL: relative pronoun, SG: singular, SUBJ: subjunctive.

References

- Abbott, Barbara. 1999. Support for a unique theory of definite descriptions. In Tanya Matthews & Devon Strolovitch (eds.), *Proceedings from Semantics and Linguistic Theory 9*, 1–15. Ithaca: CLC Publications.
- Abbott, Barbara. 2008. Issues in the semantics and pragmatics of definite descriptions in English.

- In Jeanette K. Gundel & Nancy Hedberg (eds.), *Reference: Interdisciplinary perspectives*, 61–72. Oxford University Press.
- Abel, Fritz. 1971. *L'adjectif démonstratif dans la langue de la Bible latine : Étude sur la formation des systèmes déictiques et de l'article défini des langues romanes*. Tübingen: Niemeyer.
- Adams, James N. & Nigel Vincent (eds.). 2016. *Early and late Latin: Continuity or change?* Cambridge: Cambridge University Press. doi:10.1017/CBO9781316450826.
- Aebischer, Paul. 1948. Contribution à la protohistoire des articles *ille* et *ipse* dans les langues romanes. *Cultura neolatina* 8. 181–203.
- Axenov, Serge. 2006. *The Balochi language of Turkmenistan: A corpus-based grammatical description*. Uppsala: Acta Universitatis Upsaliensis.
- Bailyn, John F. 2012. *The syntax of Russian*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511984686.
- Baldinger, Kurt. 1968. Post- und Prädeterminierung im Französischen. In Kurt Baldinger (ed.), *Festschrift walther von wartburg zum 80. Geburtstag*, 87–106. Tübingen: Niemeyer.
- Bashir, Elena. 2009. Wakhi. In Gernot Windfuhr (ed.), *The Iranian languages*, 825–858. London: Routledge.
- Becker, Laura. 2018. Articles in the world's languages. Ph.D. Thesis, Universität Leipzig. <https://laurabecker.gitlab.io/papers/articles-manuscript.pdf>.
- Becker, Laura. 2019. The crosslinguistics distribution and development of articles. Universität Bamberg. <https://laurabecker.gitlab.io/presentations/pres-bamberg.pdf>.
- Berger, Tilman. 2007. Tendencies of the evolution of a definite article in the Sorbian languages. Slavic Linguistic Society. <https://homepages.uni-tuebingen.de/tilman.berger/Handouts/BerlinSLS2007.pdf>.
- Bielenstein, August Johann Gottfried. 1863. *Lettische Grammatik*. Mitau: Fr. Lucas.
- Birner, Betty J. & Gregory Ward. 1994. Uniqueness, familiarity, and the definite article in English. In Susanne Gahl, Andy Dolby & Christopher Johnson (eds.), *Proceedings of the 20th annual meeting of the Berkeley Linguistics Society*, 93–102. Berkeley: Berkeley Linguistics Society. doi:10.3765/bls.v20i1.1479.
- Blasi, Damián E., Steven Moran, Scott R. Moisik, Paul Widmer, Dan Dediu & Balthasar Bickel. 2019. Human sound systems are shaped by post-neolithic changes in bite configuration. *Science* 363. 1–10. doi:10.1126/science.aav3218.
- Börjars, Kersti, Pauline Harries & Nigel Vincent. 2016. Growing syntax: The development of a DP in North Germanic. *Language* 92(1). e1–e37. doi:10.1353/lan.2016.0002.
- Bremmer, Rolf H., Jr. 2009. *An introduction to old Frisian*. Amsterdam: John Benjamins. doi:10.1075/z.147.

- Breu, Walter. 2012. The grammaticalization of an indefinite article in Slavic micro-languages. In Björn Wiemer (ed.), *Grammatical replication and borrowability in language contact*, 275–322. Berlin: de Gruyter Mouton. doi:10.1515/9783110271973.275.
- Bybee, Joan L. 1988. The diachronic dimension in explanation. In John A. Hawkins (ed.), *Explaining language universals*, 350–379. Oxford: Blackwell.
- Bybee, Joan L. 2006. Language change and universals. In Ricardo Mairal & Juana Gil (eds.), *Linguistic universals*, 179–194. Cambridge University Press.
- Carlier, Anne & Béatrice Lamiroy. 2018. The emergence of the grammatical paradigm of nominal determiners in French and in Romance: Comparative and diachronic perspectives. *Canadian Journal of Linguistics/Revue canadienne de linguistique* 63(2). 141–166. doi:10.1017/cnj.2017.43.
- Carlier, Anne & Walter de Mulder. 2010. The emergence of the definite article: *ille* in competition with *ipse* in Late Latin. In Hubert Cuyckens, Lieven van de Lanotte & Kristin Davidse (eds.), *Subjectification, intersubjectification and grammaticalization*, 241–275. The Hague: Mouton. doi:10.1515/9783110226102.3.241.
- Carling, Gerd (ed.). 2019. *The Mouton atlas of languages and cultures: Europe and west, central, and south Asia*, vol. 1. Berlin: de Gruyter. doi:10.1515/9783110367416.
- Cathcart, Chundra Aroor. 2018. Modeling linguistic evolution: A look under the hood. *Linguistics Vanguard* 4(1). 1–11. doi:10.1515/lingvan-2017-0043.
- Cathcart, Chundra Aroor, Gerd Carling, Filip Larsson, Niklas Johansson & Erich R. Round. 2018. Areal pressure in grammatical evolution: An Indo-European case study. *Diachronica* 35(1). 1–34. doi:10.1075/dia.16035.cat.
- Cathcart, Chundra Aroor, Andreas Hölzl, Gerhard Jäger, Paul Widmer & Balthasar Bickel. 2020. Numeral classifiers and number marking in Indo-Iranian. *Language Dynamics and Change* 11(2). 1–53. doi:10.1163/22105832-bja10013.
- Chandralal, Dileep. 2010. *Sinhala*. Amsterdam: John Benjamins.
- Chang, Will, Chundra Aroor Cathcart, David P. Hall & Andrew J. Garrett. 2015a. Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language* 91(1). 194–244. doi:10.1353/lan.2015.0005.
- Chang, Will, Chundra Aroor Cathcart, David P. Hall & Andrew J. Garrett. 2015b. Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis: Supplementary materials. *Language* 91(1). doi:10.1353/lan.2015.0007.
- Christophersen, Paul. 1939. *The articles: A study of their theory and use in English*. Copenhagen: Munksgaard.
- Clackson, James P. T. 2008. Classical Armenian. In Roger D. Woodard (ed.), *The ancient languages of Asia Minor*, 124–144. Cambridge: Cambridge University Press.

- Crisma, Paola. 2011. The emergence of the definite article in English: A contact-induced change? In Petra Sleeman & Harry Perridon (eds.), *The noun phrase in Romance and Germanic: Structure, variation, and change*, 175–192. Amsterdam: John Benjamins. doi:10.1075/la.171.13cri.
- Cysouw, Michael. 2011. Understanding transition probabilities. *Linguistic Typology* 15(2). 415–431. doi:10.1515/lity.2011.028.
- Dahl, Östen. 2003. Definite articles in Scandinavian: Competing grammaticalization processes in standard and non-standard varieties. In Bernd Kortmann (ed.), *Dialectology meets typology*, 147–180. Berlin: de Gruyter. doi:10.1515/9783110197327.147.
- Dahl, Östen. 2004. *The growth and maintenance of linguistic complexity*. Amsterdam: John Benjamins.
- Dediu, Dan & Stephen C. Levinson. 2012. Abstract profiles of structural stability point to universal tendencies, family-specific factors, and ancient connections between languages. *PLoS ONE* 7(09-20). doi:10.1371/journal.pone.0045198.
- Dryer, Matthew S. 2013a. Definite articles. In Matthew S. Dryer & Martin Haspelmath (eds.), *The world atlas of language structures online*, Leipzig: Max Planck Institute for Evolutionary Anthropology. <https://wals.info/chapter/37>.
- Dryer, Matthew S. 2013b. Indefinite articles. In Matthew S. Dryer & Martin Haspelmath (eds.), *The world atlas of language structures online*, Leipzig: Max Planck Institute for Evolutionary Anthropology. <https://wals.info/chapter/38>.
- Dryer, Matthew S. 2014. Competing methods for uncovering linguistic diversity: The case of definite and indefinite articles (Commentary on Davis, Gillon, and Matthewson). *Language* 90(4). 232–249. doi:10.1353/lan.2014.0070.
- Dum-Tragut, Jasmine. 2009. *Armenian*. Amsterdam: John Benjamins.
- Dunn, Michael, Tonya Kim Dewey, Carlee Arnett, Thórhallur Eythórsson & Jóhanna Barðdal. 2017. Dative sickness: A phylogenetic analysis of argument structure evolution in Germanic. *Language* 93(1). e1–e22.
- Dunn, Michael, Simon J. Greenhill, Stephen C. Levinson & Russell D. Gray. 2011. Evolved structure of language shows lineage-specific trends in word-order universals. *Nature* 473. 79–82. doi:10.1038/nature09923.
- Ebert, Karen H. 1971a. *Referenz, Sprachsituation und die bestimmten Artikel in einem nordfriesischen Dialekt*, vol. 4. Bredstedt: Nordfriisk Instituut.
- Ebert, Karen H. 1971b. Zwei Formen des bestimmten Artikels. In Dieter Wunderlich (ed.), *Probleme und Fortschritte der Transformationsgrammatik: Referate des 4. Linguistischen Kolloquiums Berlin 6. bis 10. Oktober 1969*, 159–174. München: Hueber.
- Edelman, Joy I. & Leila R. Dodykhudoeva. 2009a. The Pamir languages. In Gernot Windfuhr (ed.), *The Iranian languages*, 773–786. London: Routledge.

- Edelman, Joy I. & Leila R. Dodykhudoeva. 2009b. Shughni. In Gernot Windfuhr (ed.), *The Iranian languages*, 787–824. London: Routledge.
- Epstein, Richard. 2002. The definite article, accessibility, and the construction of discourse referents. *Cognitive Linguistics* 12(4). 333–378.
- Erschler, David. 2019. A new argument for existence of the DP in languages without articles. *Journal of Linguistics* 55(4). 879–887. doi:10.1017/S0022226719000288.
- Estrada Fernández, Zarina. 1996. *Pima Bajo*. München: LINCOM Europa.
- Faarlund, Jan Terje. 2004. *The syntax of old Norse*. Oxford: Oxford University Press.
- Flick, Johanna. 2019. „Alte“ Daten, neue Methoden. *Jahrbuch für Germanistische Sprachgeschichte* 10(1). 151–175. doi:10.1515/jbgs-2019-0010.
- Flick, Johanna. 2020. *Die Entwicklung des Definitartikels im Althochdeutschen*. Berlin: Language Science Press. doi:10.5281/zenodo.3932780.
- Friedman, Victor A. & Brian D. Joseph. 2022. *The Balkan languages*. Cambridge: Cambridge University Press.
- Fruyt, Michèle. 2003. Anaphore, cataphore et déixis dans l'Itinerarium d'Égérie. In Heikki Solin, Martti Leiwo & Hilla Halla-Aho (eds.), *Latin vulgaire / latin tardif: Actes du vie colloque international sur le latin vulgaire et tardif, Helsinki, 29 août – 2 septembre 2000*, vol. 6, 99–119. Hildesheim: Olms.
- Gair, James W. 2003. Sinhala. In George Cardona & Dhanesh Jain (eds.), *The Indo-Aryan languages*, 766–817. London: Routledge.
- Garland, Jennifer. 2006. Morphological typology and the complexity of nominal morphology in Sinhala. In Robert Englebretson & Carol Genetti (eds.), *Santa Barbara Papers in Linguistics*, vol. 17, 1–19. Santa Barbara: Department of Linguistics at the University of California, Santa Barbara.
- Garrett, Andrew J. 2012. The historical syntax problem: Reanalysis and directionality. In Dianna Jonas, John Whitman & Andrew J. Garrett (eds.), *Grammatical change: Origins, nature, outcomes*, 52–72. Oxford: Oxford University Press.
- van Gelderen, Elly. 2007. The definiteness cycle in Germanic. *Journal of Germanic Linguistics* 19(4). 275–308. doi:10.1017/S147054270700013X.
- van Gelderen, Elly. 2011. *The linguistic cycle: Language change and the language faculty*. Oxford: Oxford University Press. doi:10.1093/acprof:oso/9780199756056.001.0001.
- Gilbertson, George Waters. 1923. *The Balochi language: A grammar and manual*. Hertford: Stephen Austin & Sons.

- Giusti, Giuliana. 2001. The birth of a functional category: From Latin ILLE to the Romance article and personal pronoun. In Guglielmo Cinque & Giampaolo Salvi (eds.), *Current studies in Italian syntax: Essays offered to Lorenzo Renzi*, 157–171. Amsterdam: North Holland.
- Giusti, Giuliana & Rossella Iovino. 2016. Latin as a split-DP language. *Studia Linguistica* 70(3). 221–249. doi:10.1111/stul.12045.
- Greenberg, Joseph H. 1966. *Language universals: With special reference to feature hierarchies*. The Hague: Mouton.
- Greenberg, Joseph H. 1978. How does a language acquire gender markers? In Joseph H. Greenberg, Charles A. Ferguson & Edith A. Moravcsik (eds.), *Universals of human language*, vol. 3, 47–82. Palo Alto: Stanford University Press.
- Guardiano, Cristina. 2012. Parametric changes in the history of the Greek article. In Dianna Jonas, John Whitman & Andrew J. Garrett (eds.), *Grammatical change: Origins, nature, outcomes*, 179–197. Oxford: Oxford University Press. doi:10.1093/acprof:oso/9780199582624.003.0009.
- Guardiano, Cristina. 2013. The Greek definite article across time. *Studies in Greek Linguistics* 33. 76–91.
- Haig, Geoffrey L. J. 2016. Deconstructing Iranian ergativity. In Jessica Coon, Diane Massam & Lisa Travis (eds.), *The Oxford handbook of ergativity*, 465–500. Oxford: Oxford University Press.
- Haspelmath, Martin. 2018. Revisiting the anasynthetic spiral. In Heiko Narrog & Bernd Heine (eds.), *Grammaticalization from a typological perspective*, 97–115. Oxford University Press. doi:10.1093/oso/9780198795841.001.0001.
- Hawkins, John A. 1978. *Definiteness and indefiniteness: A study in reference and grammaticality*. London: Croom Helm.
- Hawkins, John A. 1991. On (in)definite articles: Implicatures and (un)grammaticality prediction. *Journal of Linguistics* 27. 405–442.
- Hawkins, John A. 2004. *Efficiency and complexity in grammars*. Oxford: Oxford University Press.
- Haynie, Hannah J. & Claire Bower. 2016. Phylogenetic approach to the evolution of color term systems. *Proceedings of the National Academy of Sciences of the United States of America* 113(48). 13666–13671. doi:10.1073/pnas.1613666113.
- Heath, Jeffrey. 1998. Hermit crabs: Formal renewal of morphology by phonologically mediated affix substitution. *Language* 74. 728–759. doi:10.2307/417001.
- Heim, Irene. 1991. Artikel und Definitheit. In Arnim von Stechow & Dieter Wunderlich (eds.), *Semantik: Ein internationales Handbuch der zeitgenössischen forschung / An international handbook of contemporary research*, 487–535. Berlin: de Gruyter.
- Heim, Irene. 2011. Definiteness and indefiniteness. In Klaus von Heusinger, Claudia Maienborn & Paul H. Portner (eds.), *Semantics: An international handbook of natural language meaning*, vol. 2, 996–1025. Berlin: de Gruyter.

- Heine, Bernd. 1997. *Cognitive foundations of grammar*. Oxford: Oxford University Press.
- Heine, Bernd & Tania Kuteva. 2011. The areal dimension of grammaticalization. In Bernd Heine & Heiko Narrog (eds.), *The Oxford handbook of grammaticalization*, 291–301. Oxford: Oxford University Press.
- Hertzenberg, Mari Johanne Bordal. 2015. *Third person reference in Late Latin: Demonstratives, definite articles and personal pronouns in the Itinerarium Egeriae*. Berlin: de Gruyter.
- Himmelman, Nikolaus P. 1997. *Deiktikon, Artikel, Nominalphrase: Zur Emergenz syntaktischer Struktur*. Tübingen: Niemeyer. doi:10.1515/9783110929621.
- Höhna, Sebastian, Michael J. Landis & Tracy A. Heath. 2017. Phylogenetic inference using RevBayes. *Current Protocols in Bioinformatics* 57(1). 6.16.1–6.16.34. doi:10.1002/cpbi.22.
- Höhna, Sebastian, Michael J. Landis, Tracy A. Heath, Bastien Boussau, Nicolas Lartillot, Brian R. Moore, John P. Huelsenbeck & Fredrik Ronquist. 2016. Revbayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Systematic Biology* 65(4). 726–736. doi:10.1093/sysbio/syw021.
- Holton, David, Peter A. Mackridge & Irene Philippaki-Warbuton. 2012. *Greek: A comprehensive grammar*. London: Routledge 2nd edn.
- Holvoet, Axel. 2018. The syntax of Baltic. In Jared S. Klein, Brian D. Joseph & Matthias A. Fritz (eds.), *Handbook of comparative and historical Indo-European linguistics: An international handbook of language comparison and the linguistic reconstruction of Indo-European*, vol. 3 (Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science 41), 1668–1681. Berlin: de Gruyter. doi:10.1515/9783110542431-011.
- Horrocks, Geoffrey C. 2010. *Greek: A history of the language and its speakers*. Malden, MA: Blackwell 2nd edn.
- van der Horst, Joop. 2008. *Geschiedenis van de Nederlandse syntaxis*. Leuven: Leuven University Press.
- Irslinger, Britta S. 2013. Standard Average European and the western fringe: A reconsideration. *Historische Sprachforschung* 126(1). 33–88. doi:10.13109/hisp.2013.126.1.33.
- Jäger, Gerhard. 2019a. Computational historical linguistics. *Theoretical Linguistics* (3–4). 151–182. doi:10.1515/tl-2019-0011.
- Jäger, Gerhard. 2019b. Model evaluation in computational historical linguistics. *Theoretical Linguistics* 45(3–4). 299–307. doi:10.1515/tl-2019-0020.
- Jahani, Carina & Agnes Korn. 2009. Balochi. In Gernot Windfuhr (ed.), *The Iranian languages*, 634–692. London: Routledge.

- Jasbi, Masoud. 2016. Three types of indefinites in Persian: Simple, complex, and antidefinite. In Mary Moroney, Carol-Rose Little, Jacob Collard & Dan Burgdorf (eds.), *Proceedings of the 26th Semantics and Linguistic Theory Conference, held at the University of Texas at Austin May 12–15, 2016*, 244–263. Linguistic Society of America. doi:10.3765/salt.v26i0.3807.
- Jasbi, Masoud. 2020. The meaning of the Persian object marker *rā*: What it is not, and what it (probably) is. In Richard K. Larson, Sedigheh Moradi & Vida Samiiian (eds.), *Advances in iranian linguistics*, 119–135. Amsterdam: John Benjamins. doi:10.1075/cilt.351.07jas.
- Kadmon, Nirit. 2001. *Formal pragmatics: Semantics, pragmatics, presupposition, and focus*. Oxford: Blackwell.
- Keller, Rudolf Ernst. 1978. *The German language*. London: Faber & Faber.
- Kent, Roland G. 1944. The Old Persian relative and article. *Language* 20. 1–10. doi:10.2307/410376.
- Kiparsky, Paul. 2008. Universals constrain change; change results in typological generalizations. In Jeff Good (ed.), *Linguistic universals and language change*, 23–53. Oxford: Oxford University Press.
- Klein, Jared S. 1996. *On personal deixis in classical Armenian: A study of the syntax and semantics of the n-, s-, and d- demonstratives in manuscripts E and M of the Old Armenian Gospels*. Dettelbach: Röhl.
- Korn, Agnes & Maryam Nourzaei. 2018. “Those were the hungry years”: A glimpse of Coastal Afro-Balochi. *Journal of the Royal Asiatic Society* 28(4). 661–695. doi:10.1017/S1356186318000238.
- Kovari, Geoffrey. 1984. *Studien zum germanischen Artikel: Entstehung und Verwendung des Artikels im Gotischen*. Wien: Halosar.
- Kraiss, Andrew. 2014. The evolution of the definite article in Old High German. *Journal of Germanic Linguistics* 26(2). 127–155. doi:10.1017/S1470542714000038.
- Kurzová-Jedličková, Helena. 1963. Die Demonstrativa im Vulgärlatein (4.–6. Jahrhundert). *Acta Antiqua Academiae Scientiarum Hungaricae* 11(1–2). 121–143.
- Kuteva, Tania, Bernd Heine, Bo Hong, Haiping Long, Heiko Narrog & Seongha Rhee. 2019. *World lexicon of grammaticalization*. Cambridge: Cambridge University Press 2nd edn. doi:10.1017/9781316479704.
- Lakoff, Robin T. 1972. Another look at drift. In Robert P. Stockwell & Ronald K. S. Macauley (eds.), *Linguistic change and generative theory*, 172–198. Bloomington: Indiana University Press.
- Ledgeway, Adam. 2012. *From Latin to Romance: Morphosyntactic typology and change*. Oxford: Oxford University Press. doi:10.1093/acprof:oso/9780199584376.001.0001.
- Ledgeway, Adam. 2016. Functional categories. In Adam Ledgeway & Martin Maiden (eds.), *The Oxford guide to the Romance languages*, 761–771. Oxford: Oxford University Press. doi:10.1093/acprof:oso/9780199677108.003.0046.

- Leiss, Elisabeth. 2007. Covert patterns of definiteness/indefiniteness and aspectuality in Old Icelandic, Gothic and Old High German. In Elisabeth Stark, Elisabeth Leiss & Werner Abraham (eds.), *Nominal determination: Typology, context constraints and historical emergence*, 73–102. Amsterdam: John Benjamins. doi:10.1075/slcs.89.06lei.
- Levinson, Stephen C., Simon J. Greenhill, Russell D. Gray & Michael Dunn. 2011. Universal typological dependencies should be detectable in the history of language families. *Linguistic Typology* 15(2). 509–534. doi:10.1515/lity.2011.034.
- Löbner, Sebastian. 1985. Definites. *Journal of Semantics* 4(4). 279–326. doi:10.1093/jos/4.4.279.
- Löbner, Sebastian. 1998. Definite associative anaphora. In Simon Botley, Julia Glass, Tony McEnery & Andrew Wilson (eds.), *Approaches to discourse anaphora: Proceedings of DAARC96 – Discourse Anaphora and Resolution Colloquium, Lancaster University July 17th–18th 1996*, Lancaster: University Centre for Computer Corpus Research on Language.
- Lockwood, William B. 1968. *Historical German syntax*. London: Oxford University Press.
- Lyons, Christopher. 1999. *Definiteness*. Cambridge: Cambridge University Press.
- Mallory, James P. & Douglas Q. Adams. 2006. *The Oxford introduction to Proto-Indo-European and the Proto-Indo-European world*. Oxford: Oxford University Press.
- Manolessou, Io. 2001. The evolution of the demonstrative system in Greek. *Journal of Greek Linguistics* 2(1). 119–148. doi:10.1075/jgl.2.05man.
- Manolessou, Io & Geoffrey C. Horrocks. 2007. The development of the definite article in Greek. *Studies in Greek Linguistics* 27. 224–236.
- Masica, Colin P. 1986. Definiteness marking in South Asian languages. In Bhadriraju Krishnamurti (ed.), *South Asian languages: Structure, convergence and diglossia*, 123–146. Delhi: Motilal Barnarsidass.
- Masica, Colin P. 1991. *The Indo-Aryan languages*. Cambridge: Cambridge University Press.
- Mathiassen, Terje. 1997. *A short grammar of Latvian*. Columbus, OH: Slavica.
- Matras, Yaron. 2002. *Romani*. Cambridge: Cambridge University Press.
- Matzinger, Joachim. 2006. *Der altalbanische Text Mbsuame e krështerë (Dottrina cristiana) des Lekë Matrënga von 1592: Eine Einführung in die albanische Sprachwissenschaft*. Dettelbach: J. H. Röhl.
- Mažiulis, Vytautas. n.d. Historical grammar of Old Prussian. <http://donelaitis.vdu.lt/prussian/mazgr.pdf>.
- McColl Millar, Robert. 2000. Some suggestions for explaining the origin and development of the definite article in English: Grammaticalization in English. In Olga Fischer, Anette Rosenbach & Dieter Stein (eds.), *Pathways of change*, 275–310. Amsterdam: John Benjamins. doi:10.1075/slcs.53.14mil.

- Miller, D. Gary. 2018. *The Oxford Gothic grammar*. Oxford: Oxford University Press.
- Moravcsik, Edith A. 1969. Determination. In *Working Papers in Language Universals* 1, 63–98. Stanford: Stanford University.
- de Mulder, Walter & Anne Carlier. 2011. The grammaticalization of definite articles. In Bernd Heine & Heiko Narrog (eds.), *The Oxford handbook of grammaticalization*, 522–534. Oxford: Oxford University Press. doi:10.1093/oxfordhb/9780199586783.013.0042.
- Müth, Angelika. 2011. Categories of definiteness in Classical Armenian. In Erik Welo (ed.), *Indo-European syntax and pragmatics: Contrastive approaches*, 11–25. Oslo: University of Oslo. doi:10.5617/osla.37.
- Napoli, Maria. 2009. Aspects of definiteness in Greek. *Studies in Language* 33(3). 569–611. doi:10.1075/sl.33.3.03nap.
- Napoli, Maria. 2019. Functions of the definite article from classical Greek to New Testament Greek. In Daniel King (ed.), *The article in post-Classical Greek*, 15–55. Dallas: SIL International.
- Nichols, Johanna. 2003. Diversity and stability in language. In Brian D. Joseph & Richard D. Janda (eds.), *The handbook of historical linguistics*, 283–310. Malden, MA: Blackwell. doi:10.1002/9781405166201.ch5.
- Nocentini, Alberto. 1990. L'uso dei dimostrativi nella *Peregrinatio Egeriae* e la genesi dell'articolo romanzo. In Alberto Fatucchi (ed.), *Atti del convegno internazionale sulla Peregrinatio Egeriae*, 137–158. Arezzo: Accademia Petrarca.
- Pagel, Mark. 1994. Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters. *Proceedings of the Royal Society B: Biological Sciences* 255(1342). 37–45. doi:10.1098/rspb.1994.0006.
- Pagel, Mark & Andrew Meade. 2006. Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov Chain Monte Carlo. *The American Naturalist* 167(6). 808–825. doi:10.1086/503444.
- Paliga, Sorin. 2019. The definite article as a reference point in defining the Balkansprachbund. In Valeriu Sîrbu, Alexandra Comşa & Dumitru Hortopan (eds.), *Digging in the past of Old Europe: Studies in honor of Christian Schuster at his 60th anniversary*, Târgu Jiu: Editura Istros a Muzeului Brăilei „Carol I”.
- Paranavitana, Senarath. 1956. *Sigiri graffiti: Being Sinhalese verses of the eighth, ninth and tenth centuries*, vol. 1. London: Oxford University Press.
- Paul, Ludwig. 1998. *Zazaki: Grammatik und Versuch einer Dialektologie*. Reichert.
- Paul, Ludwig. 2009. Zazaki. In Gernot Windfuhr (ed.), *The Iranian languages*, 545–586. London: Routledge.
- Pereltsvaig, Asya. 2007. The universality of DP: A view from Russian. *Studia Linguistica* 61(1). 59–94. doi:10.1111/j.1467-9582.2007.00129.x.

- Perry, John R. 2005. *A Tajik Persian reference grammar*. Leiden: Brill.
- Pimenova, Natalie. 2017. Дистрибуция протоартиклевого sa 'этот, тот' и jains 'тот' на ранней стадии грамматикализации готского артикля. In Nikolai N. Kazansky (ed.), *Индоевропейское языкознание и классическая филология*, vol. 21, 661–671. Санкт-Петербург: Наука.
- Prauliņš, Dace. 2012. *Latvian*. London: Routledge.
- Reinöhl, Uta & Nikolaus P. Himmelmann. 2017. Renewal: A figure of speech or a process sui generis? *Language* 93(2). 381–413. doi:10.1353/lan.2017.0018.
- Ringe, Donald A. 2017. *A linguistic history of English: From Proto-Indo-European to Proto-Germanic*, vol. 1. Oxford: Oxford University Press 2nd edn. doi:10.1093/acprof:oso/9780199284139.001.0001.
- Russell, Bertrand. 1905. On denoting. *Mind* 14. 479–493.
- SanGregory, Erin. 2018. *Differential subject marking in Wakhi*. Graduate Institute of Applied Linguistics MA thesis.
- Sauvageot, Aurelien. 1929. *L'emploi de l'article en gotique*. Paris: Champion.
- Schaarschmidt, Gunter. 1984. Theme-rheme structure and the article in Sorbian. In *Working Papers of the Linguistics Circle of the University of Victoria*, vol. 4 1, 75–90. Victoria: University of Victoria.
- Schumacher, Stefan. 2009. Lehnbeziehungen zwischen Protoalbanisch und balkanischem Latein bzw. Romanisch. In Oliver Jens Schmitt & Eva Anne Frantz (eds.), *Albanische Geschichte: Stand und Perspektiven der Forschung*, 37–60. München: Oldenbourg.
- Schwarz, Florian. 2013. Two kinds of definites cross-linguistically. *Language and Linguistics Compass* 7(10). 534–559. doi:10.1111/lnc3.12048.
- Schwyzler, Eduard. 1936. Die nominale Determination in den indogermanischen Sprachen. *Zeitschrift für vergleichende Sprachforschung auf dem Gebiete der Indogermanischen Sprachen* 63(3/4). 145–167.
- Selig, Maria. 1992. *Die Entwicklung der Nominaldetermination im Spätlatein: Romanischer Sprachwandel und lateinische Schriftlichkeit*. Tübingen: Narr.
- Sommerer, Lotte. 2012. Investigating the emergence of the definite article in Old English: About categorization, gradualness and constructions. *Folia Linguistica* 46. 175–214. doi:10.1515/flih.2012.007.
- Sommerer, Lotte. 2015. The influence of constructions in grammaticalization: Revisiting category emergence and the development of the definite article in English. In Jóhanna Barðdal, Spike Gildea, Elena Smirnova & Lotte Sommerer (eds.), *Diachronic Construction Grammar*, 107–138. Amsterdam: John Benjamins. doi:10.1075/cal.18.04som.
- Szczepaniak, Renata. 2011. *Grammatikalisierung im Deutschen*. Tübingen: Narr 2nd edn.

- Todd, Terry Lynn. 2008. *A grammar of Dimili: Also known as Zaza*. Kandern: Forum Linguistik 3rd edn.
- Trager, George L. 1932. *The use of the Latin demonstratives (especially ille and ipse) up to 600 A.D., as the source of the Romance article*. New York: Publications of the Institute of French Studies.
- Van de Velde, Freek. 2010. The emergence of the determiner in the Dutch NP. *Linguistics* 48(2). 263–299. doi:10.1515/LING.2010.009.
- Van de Velde, Freek & Joop van der Horst. 2013. Homoplasmy in diachronic grammar. *Language Sciences* 36. 66–77. doi:10.1016/j.langsci.2012.03.020.
- de la Villa Polo, Jesús. 2011. Numerals. In Philip Baldi & Pierluigi Cuzzolin (eds.), *New perspectives on historical Latin syntax: Quantification, numerals, possession, anaphora*, vol. 3, 175–238. Berlin: de Gruyter. doi:10.1515/9783110215465.175.
- Vincent, Nigel. 1997. The emergence of the D-system in Romance. In Ans van Kemenade & Nigel Vincent (eds.), *Parameters of morphosyntactic change*, 149–169. Cambridge: Cambridge University Press.
- Vincent, Nigel. 1999. The evolution of C-structure: Prepositions and PPs from Indo-European to Romance. *Linguistics* 37. 1111–1153. doi:10.1515/ling.37.6.1111.
- Vincent, Nigel. 2018. Definiteness in diachrony: Romance vs. Germanic. Specificity, definiteness and article systems across languages, 40th Annual Meeting of the DGfS.
- Wali, Kashi, Omkar Nath Koul & Omakāra Aina Kaula. 1997. *Kashmiri: A cognitive-descriptive grammar*. London: Routledge.
- Wandruszka, Ulrich. 1980. ‚Post-‘ oder ‚Prädetermination‘ in den romanischen Sprachen? *Romanistisches Jahrbuch* 31. 56–72. doi:10.1515/9783110244885.56.
- Wemke, Matthias, Kathrin Kunkel-Razum & Werner Scholze-Stubenrecht (eds.). 1995. *Die Grammatik*. Mannheim: Dudenverlag 8th edn.
- Wendtland, Antje. 2011. *Die Entwicklung von Demonstrativpronomina zu Artikeln im Soghdischen*. Wiesbaden: Harrassowitz. doi:10.2307/j.ctvbqs61b.
- Widmer, Manuel, Sandra Auderset, Johanna Nichols, Paul Widmer & Balthasar Bickel. 2017. NP recursion over time: Evidence from Indo-European. *Language* 93(4). 799–826. doi:10.1353/lan.2017.0058.
- Widmer, Paul. 2018. Indogermanische Stammbäume: Datentypen, Methoden. In Elisabeth Rieken, Ulrich Geupel & Theresa Maria Roth (eds.), *100 Jahre Entzifferung des Hethitischen: Morphosyntaktische Kategorien in Sprachgeschichte und Forschung Akten der Arbeitstagung der Indogermanischen Gesellschaft vom 21. bis 23. September 2015 in Marburg*, 373–388. Wiesbaden: Reichert.
- Wood, Johanna L. 2003. *Definiteness and number: Determiner phrase and number phrase in the history of English*: Arizona State University dissertation.

A Appendix

The data and code used for this paper are archived at [10.5281/zenodo.5520086](https://doi.org/10.5281/zenodo.5520086).

Résumé

Les phénomènes de grammaticalisation sont caractérisés par de fortes asymétries directionnelles (e.g., Kuteva et al. 2019) : les noms de parties du corps, par exemple, deviennent des adpositions spatiales, les minimiseurs scalaires des adverbes négatifs, et les pronoms sujets des marqueurs de l'accord. À l'inverse, les changements en direction opposée sont rares voire inattestés (Garrett 2012: 52). Le fait que ces asymétries soient fortement cross-linguistiques a conduit certains chercheurs à réifier ces processus de grammaticalisation et à en faire des mécanismes linguistiques universels (Heath 1998: 729). Une telle conception implique notamment que la morphosyntaxe d'une langue donnée n'ait que peu ou même aucune influence sur les processus de grammaticalisation. L'objectif principal de cet article est de montrer le rôle crucial que la morphosyntaxe préexistante peut avoir dans les phénomènes de grammaticalisation. L'étude de cas sur laquelle s'appuie cette thèse est la grammaticalisation des articles définis et indéfinis à travers l'histoire des langues Indo-Européennes : on observe que les articles indéfinis se développent plus rapidement dans les langues où un article défini a déjà émergé que dans celles qui n'ont pas d'article défini. Ces deux changements sont donc corrélés. Les résultats de cette recherche suggèrent ainsi que l'étude du rapport entre le système linguistique préexistant et les processus de grammaticalisation qui s'y produisent pourrait grandement nous renseigner sur les contextes et les causes de ces processus (cf. Reinöhl & Himmelmann 2017: 381).

Zusammenfassung

Grammatikalisierungspfade sind durch starke Asymmetrien gekennzeichnet (z.B., Kuteva et al. 2019). So entwickeln sich beispielsweise Körperteilnomina zu räumlichen Adpositionen; Minimierer entwickeln sich zu Negationsmarkern; und Subjektpronomen werden zu Übereinstimmungsmarkern. Veränderungen in die entgegengesetzte Richtung sind entweder selten oder nicht belegt (Garrett 2012: 52). Solche sprachenübergreifenden Asymmetrien haben einige Wissenschaftler dazu veranlasst, Grammatikalisierungspfade als universelle mechanistische Kräfte zu revidieren (Heath 1998: 729). Eine Folge einer solchen Sichtweise ist, dass die Morphosyntax einer Sprache wenig oder sogar keine Relevanz für die Grammatikalisierung hat. Das Hauptziel dieser Arbeit ist es, die entscheidende Rolle aufzuzeigen, die bereits existierende Morphosyntax bei der Grammatikalisierung spielen kann. Die empirische Grundlage für diese Behauptung ist die Grammatikalisierung bestimmter und unbestimmter Artikel in der Geschichte des Indogermanischen: unbestimmte Artikel entwickelten sich schneller unter Sprachen, in denen ein bestimmter Artikel bereits entstanden war, als solche, die keinen bestimmten Artikel hatten. Die beiden Änderungen sind somit korreliert. Die Ergebnisse dieser Fallstudie legen nahe, dass es viel mehr darüber zu erfahren gibt, wann und warum eine Grammatikalisierung stattfindet, indem man ihre Beziehung zum bereits existierenden Sprachsystem untersucht (vgl. Reinöhl & Himmelmann 2017: 381).

Author's address

David Goldstein
Department of Linguistics
University of California, Los Angeles
3125 Campbell Hall
Box 951543
Los Angeles, CA 90095-1543

dgoldstein@humnet.ucla.edu