


# Linguistic Typology

---

## at the Crossroads



ISSN 2785-0943

Volume 3 – issue 2 – 2023

Issue DOI: <https://doi.org/10.6092/issn.2785-0943/v3-n2-2023>

This journal provides immediate and free open access. There is no embargo on the journal's publications. Submission and acceptance dates, along with publication dates, are made available on the PDF format for each paper. The authors of published articles remain the copyright holders and grant third parties the right to use, reproduce, and share the article according to the [Creative Commons Attribution 4.0 International](#) license agreement. The reviewing process is double-blind. Ethical policies and indexing information are publicly available on the journal website:

<https://typologyatcrossroads.unibo.it>

### Editors

*Nicola Grandi* (University of Bologna, Editor in chief)

*Caterina Mauri* (University of Bologna, Editor in chief)

*Francesca Di Garbo* (University of Aix-Marseille)

*Andrea Sansò* (University of Insubria)

### Publisher

*Department of Classical Philology and Italian Studies* (University of Bologna)

*Department of Modern Languages, Literatures and Cultures* (University of Bologna)

The journal is hosted and maintained by [AlmaDL](#)



# Linguistic Typology

---

## at the Crossroads



### Editorial board

*Mira Ariel* (Tel Aviv University)  
*Sonia Cristofaro* (Sorbonne Université)  
*Chiara Gianollo* (University of Bologna)  
*Matti Miestamo* (University of Helsinki)  
*Marianne Mithun* (University of California Santa Barbara)

### Scientific Board

*Giorgio Francesco Arcodia* (Università Ca' Foscari, Venice)  
*Peter Arkadiev* (Johannes-Gutenberg University of Mainz)  
*Gilles Authier* (École Pratique des Hautes Études, Paris)  
*Luisa Brucale* (University of Palermo)  
*Holger Diessel* (University of Jena)  
*Eitan Grossman* (The Hebrew University of Jerusalem)  
*Corinna Handschuh* (Universität Regensburg)  
*Guglielmo Inglese* (University of Turin)  
*Elisabetta Magni* (University of Bologna)  
*Francesca Masini* (University of Bologna)  
*Susanne Maria Michaelis* (MPI EVA – Leipzig)  
*Emanuele Miola* (University of Bologna)  
*Anna Riccio* (University of Foggia)  
*Eva van Lier* (University of Amsterdam)

### Responsible Editor

*Caterina Mauri*, University of Bologna

Department of Modern Languages, Literatures and Cultures, via Cartoleria 5, 40124 Bologna. Email: [caterina.mauri@unibo.it](mailto:caterina.mauri@unibo.it)

### Production editors

*Silvia Ballarè* (University of Bologna)  
*Alessandra Barotto* (University of Insubria)  
*Simone Mattioli* (University of Bologna)  
*Eleonora Zucchini* (University of Bologna)

### Assistant production editors

*Antonio Bianco* (University of Pavia)  
*Valentina Di Falco* (University of Bologna)  
*Sara Gemelli* (University of Pavia)  
*Maria Cristina Lo Baido* (University of Cagliari)  
*Nicola Perugini* (University of Bologna)  
*Antonia Russo* (University of Bergamo)



# Linguistic Typology

---

## at the Crossroads



### CONTENTS

#### Types of clitics in the world's languages

*Martin Haspelmath*----- 1-59

#### Spanish as an argument-indexing language. A view from the analysis of Colombian Andean Spanish

*Sergio Ibáñez Cerda, Armando Mora Bustos, Alejandra I. Ortiz Villega*-----60-99

#### Using a parallel corpus to study patterns of word order variation: determiners and quantifiers within the noun phrase in European languages

*Luigi Talamo*-----100-131

#### The Dimensions of Morphosyntactic Variation: Whorf, Greenberg and Nichols were right

*Siva Kalyan, Mark Donohue* -----132-190

#### Towards a typology of continuative expressions

*Anastasia Panova*-----191-244



# Types of clitics in the world's languages

MARTIN HASPELMATH

MAX PLANCK INSTITUTE FOR EVOLUTIONARY ANTHROPOLOGY

Submitted: 19/12/2022    Revised version: 02/08/2023

Accepted: 02/08/2023    Published: 27/12/2023



Articles are published under a Creative Commons Attribution 4.0 International License (The authors remain the copyright holders and grant third parties the right to use, reproduce, and share the article).

## Abstract

This paper offers and discusses a simple definition of the term *clitic* from a comparative perspective: A clitic is a bound morph that is neither an affix nor a root. It gives examples of several semantic and positional types of clitics from a wide range of languages, and it discusses some typical phonological effects associated with clitics. In the proposed definition, the crucial contrast between affixes and clitics is that affixes are class-selective (occurring always on nouns, on verbs, or on adjectives), while clitics do not exhibit word-class selectivity. In the stereotypical view of clitics, they are “prosodically deficient” in some way, but the phonological effects are quite diverse and cannot serve as a basis for a definition. As clitics are defined as kinds of minimal forms (or morphs), they cannot be nonsegmental, and they cannot interrupt another minimal form (so that there cannot be endoclitics by definition). Finally, I note that the object person indexes of the Romance languages, which have very often been called *clitics*, are actually affixes in the modern languages, although they must go back to earlier clitics.

**Keywords:** clitic; affix; prosodic deficiency; word-class selectivity; object indexing.

## 1. Overview

This paper gives an overview of clitics in human languages, with an emphasis on clear conceptual distinctions and straightforward terminology. In addition to exemplifying a range of clitics from a wide variety of languages, I will discuss some of the earlier conceptual and terminological distinctions, and I will say how the choices made here

relate to the earlier literature. This paper thus has a clear methodological focus and does not claim to make an empirical or explanatory contribution. I begin with the definition of *clitic* as a comparative concept in (1), which is simple and clear.

(1) **Clitic**

A clitic is a bound morph that is neither an affix nor a root.

As a first illustration of clitics, consider the forms in boldface in (2)-(5), which are typical examples of clitics.

(2) English (Indo-European, Germanic)<sup>1</sup>

*my friend = 's house*

(3) Russian (Indo-European, Balto-Slavic)

*Pročita-la = **li** Anna knigu?*

read-PST = PQ Anna book

'Did Anna read a book?'

(4) Persian (Indo-European, Indo-Iranic; Samvelian & Tseng 2010: 215)

*Ru-ye miz = **aš** gozâšt-im.*

on-EZ table = 3SG.OBJ put.PST-1PL.SBJ

'We put it on the table.'

(5) Tagalog (Austronesian, Malayo-Polynesian; Kaufman 2010: 10)

a. *Na-túto = **siya** nang = wika = ng Intsik.*  
 AV-learn = 3SG.NOM GEN = language = LNK Chinese  
 'She learned Chinese.'

b. *Hindí = **siya** na-túto nang = wika = ng Intsik.*  
 NEG = 3SG.NOM AV-learn GEN = language = LNK Chinese  
 'She didn't learn Chinese.'

---

<sup>1</sup> The genealogical classification of the languages mentioned in this paper has been retrieved on Glottolog.

Traditionally, clitics have often been defined as prosodically deficient elements, and/or as forms that are somehow intermediate between affixes and independent words. Below in Section 5, I will explain why the definition in (1) is preferable, even though it does not conform fully to some linguists' intuitions about the nature of clitics.

To understand the definition adopted here, we need to understand the concepts of *bound morph* and of *affix*. Briefly, a morph is a minimal form (Haspelmath 2020), a bound form is a form that cannot occur in isolation (Bloomfield 1933: 160), and an affix is a bound morph that is not a root and that always occurs on roots of the same class (i.e. always on nouns, on verbs, or on adjectives; see Haspelmath 2021).

The clitics in the initial examples given above are evidently not roots (i.e. contentful morphs denoting an object, an action or a property; Haspelmath 2023), and they are not affixes either because they may occur adjacent to different classes of forms, as illustrated in (6) and (7) for English (eng; Indo-European, Germanic) and Russian (rus; Indo-European, Balto-Slavic) (and also in (5a) and (5b) above for Tagalog - tgl; Austronesian, Malayo-Polynesian).

(6) English (Indo-European, Germanic)

- a. *my friend's house* (adjacent to noun)
- b. *the lady I met yesterday's offer* (adjacent to adverb)
- c. *the boy I like's new bike* (adjacent to verb)

(7) Russian (Indo-European, Balto-Slavic)

- a. *Pročita-la li Anna knigu?* (= 3; adjacent to verb)  
read-PST PQ Anna book  
'Did Anna read a book?'
- b. *Knigu li Anna pročita-la?* (adjacent to noun)  
book PQ Anna read-PST  
'Did Anna read a BOOK?'
- c. *Včera li Anna čita-la?* (adjacent to adverb)  
yesterday PQ Anna read-PST  
'Did Anna read YESTERDAY?'

This property of clitics is also called NONSELECTIVITY, contrasting with the WORD-CLASS SELECTIVITY of affixes.<sup>2</sup> An affix such as a Latin genitive suffix must always occur on a noun (contrasting with English 's, which can also occur on adverbs and verbs), and an affix such as a German person-number suffix must always occur on a verb (contrasting with Tagalog *siya*, which can occur on negation markers as well).

It is thus their nonselectivity that picks out clitics in the definition that I use here (see the further discussion in Section 6). Some linguists might prefer a characterization of clitics that makes reference to their phonological properties, but it seems impossible to define clitics phonologically (this is discussed in Section 7 below).

In the next three sections (Sections 2-4), we will see examples of various types of clitics from a wide range of languages, before we move on to a discussion of the definition of the term *clitic* (Section 5). Then I will discuss the lack of word-class selectivity (Section 6), before examining the phonological properties of clitics (Section 7). One consequence of the present definition is that clitics are concrete forms, so that *nonsegmental clitics* or *endoclitics* cannot exist (Section 8). Finally, I will say a few things about Romance *object clitics*, which have played a big role in the literature, but which turn out to be affixes rather than clitics (Section 9).

## 2. Semantic types of clitics

### 2.1. Content words and function words

Linguists often distinguish between CONTENT WORDS and FUNCTION WORDS. Content words are nouns, verbs, and adjectives, and function words are most other types of words.<sup>3</sup> They cannot be easily characterized positively, but there is widespread agreement that the most important classes of function words are relators (adpositions, subordinators), linkers (complementizers, coordinators), articles, tense-aspect auxiliaries, and various kinds of discourse markers. What most function words share (also with affixes) is that the information they convey is discursively secondary (Boye

---

<sup>2</sup> Instead of *nonselectivity*, the literature often uses the term *promiscuous attachment* or *promiscuity* (from Zwicky 1987: 136). I used the latter term in the past, but it seems better to replace it with a term that does not have unwanted associations. (Another possibility that I considered was *indiscriminacy*.)

<sup>3</sup> Some kinds of words, such as numerals and interjections, do not readily fit into this classification.

& Harder 2012). The current section will illustrate various kinds of clitics, most of which are function words.

## 2.2. Person indexes (= bound person forms)

Perhaps the best-known types of clitics are person indexes (Haspelmath 2013), and especially the object indexes of the Romance languages have been discussed extensively (in the wake of Kayne 1975). For example, Spanish *te* ‘you’ is a weak object person form and contrasts with a strong form *ti* ‘YOU’.

(8) Spanish (Indo-European, Italic)

*Te*                    *quiero.*

you.ACC          love.1SG

‘I love you.’ (contrasting with *quiero a ti* ‘I love YOU’)

Below in Section 9 we will see that the Romance object person forms are not really clitics, even though they are usually called *clitics*. But many other languages have subject and/or object person clitics, e.g. Serbo-Croatian (hbs; Indo-European, Balto-Slavic), where subject forms (for past tense constructions) like *smo* and object forms like *mu* and *je* must occur in the second position.

(9) Serbo-Croatian (Indo-European, Balto-Slavic; Bošković 2016: 28)

a. *Zašto*    *smo*                    *mu*                    *je*                    *predstavili*    *juče?*  
why    1PL.SBJ                    him.DAT                    her.ACC                    introduced    yesterday

‘Why did we introduce her to him yesterday?’

b. *Predstavili*    *smo*                    *mu*                    *je*                    *juče.*  
introduced    1PL.SBJ                    him.DAT                    her.ACC                    yesterday

‘We introduced her to him yesterday.’

German (deu; Indo-European, Germanic) has a few clitic subject and object forms in the colloquial language, illustrated in (10).<sup>4</sup>

---

<sup>4</sup> In this connection, English colloquial object forms are sometimes mentioned as well (e.g. *hit 'em* ‘hit them’). But these English forms always follow the verb directly, so they are affixes rather than clitics



## (10) Colloquial German (Indo-European, Germanic)

*Willst de se heute haben?*

want.2SG you them today have

‘Would you like to have them today?’ (Standard: *Willst du sie heute haben?*)

In Lak (lbe; Nakh-Daghestanian, Daghestanian), a subject person index usually follows the verb, but when the focus is on an argument (the subject in (11b), the object in (11c)), it follows this argument.

## (11) Lak (Nakh-Daghestanian, Daghestanian; Kazenin 2002: 293)

a. *Na q̄atri d-ullali-ṣa = ra.*

I house(G4) G4-build.DUR-PTCP = 1SG

‘I am building a house.’

b. *Na = ra q̄atri d-ullali-ṣa.*

I = 1SG house(G4) G4-build.DUR-PTCP

‘The one who is building a house is me.’

c. *Na q̄atri = ra d-ullali-ṣa.*

I house(G4) = 1SG G4-build.DUR-PTCP

‘What I am building is a house.’

Clitic person indexes were also seen above in (4) (Persian) and (5) (Tagalog), and more examples are given below: Halkomelem (hur; Salishan, Coast Salish) (14); Bulgarian (bul; Indo-European, Balto-Slavic) (33), (52), (71b); Ancient Greek (grc; Indo-European, Graeco-Phrygian) (42); Wambaya (wmb; Mirndi, Ngurlun) (43); Polish (pol; Indo-European, Balto-Slavic) (58a); Kugu Nganhcara (wua; Pama-Nyungan, Paman) (72); Udi (udi; Nakh-Daghestanian, Daghestanian) (84).

### 2.3. Tense-aspect forms

In many languages, tense-aspect meanings are expressed by verb-like auxiliaries, and these are commonly bound non-affixal forms, i.e. clitics. For example, the French

---

(if one wants to treat them as distinct from the full forms). By contrast, German shortened forms like *se* can occur both postverbally and in the first position of the Middle Field (e.g. *wenn de se heute haben willst* ‘if you want them today’), so they are not class-selective and are thus clitics.

auxiliary *avoir* ‘have’ is used to form a past tense, and it is a (nonaffixed) bound form rather than a free form, as it must always cooccur with a verb. In (12), for example, the second verb cannot be omitted.<sup>5</sup> This contrasts with English, where the perfect auxiliary *have* can occur without the verb and is thus not a bound form.

(12) French (Indo-European, Italic)

*J’ai*    *changé*    *et*    *mon*    *mari*       *a*    *changé*    *aussi*.  
I have    changed    and    my    husband    has    changed    too  
‘I have changed and my husband has  $\emptyset$ , too.’ (\**J’ai changé et mon mari a  $\emptyset$  aussi.*)

Two more examples of tense-aspect clitics come from Garrwa (wrk; Garrwan) and, Halkomelem.

(13) Garrwa (Garrwan; Mushin 2012: 206–207)

- a. *Jungku yal*    =*i*    *bangkulu-na*.  
stay    3PL    =PST    prison-LOC  
‘They stayed in the prison.’
- b. *Najba*    =*yi*    *bula*.  
see    =PST    3DU.NOM  
‘They two saw him.’

(14) Halkomelem (Salishan, Coast Salish; Gerds & Werle 2014: 251)

*Nem’*    =*?ə*    =*č*             =*ce?* *q<sup>w</sup>ał-ət*    *łə*    *sti:č?*  
go    =QM    =2SG.SBJ    =FUT    wait-TR    DET    bus  
‘Are you going to wait for the bus?’

More tense-aspect clitics are illustrated below: Wambaya (43), English (59), Italian (ita; Indo-European, Italic) (63), Bulgarian (71b).

---

<sup>5</sup> Note that the French auxiliary *avoir* is not prefixal as it need not occur immediately before the verb (e.g. *il a probablement changé* ‘he has probably changed’).

## 2.4. Articles

Definite and indefinite articles accompany nouns and cannot occur on their own, so they are bound forms. In some languages, they always occur directly on nouns and are thus affixes, e.g. in Swedish (Indo-European, Germanic) (*kung-en* [king-DEF] ‘the king’). In other languages, they occur in a peripheral position and noun modifiers may intervene, e.g. in Italian, Basque (eus; Isolate, Eurasia) and Haitian Creole (hat; French-based creole, North America). Such articles are clitics, regardless of their spelling (in Basque, the article is written as if it were a suffix: *etxea*, *etxe berria*).

(15) Italian (Indo-European, Italic)

- a. *il*        *libro*  
       the        book
- b. *l'*        *altro libro*  
       the        other book

(16) Basque (Isolate, Eurasia)

- a. *etxe*        = **a**  
       house        = ART  
       ‘the house’
- b. *etxe*        *berri* = **a**  
       house        new    = ART  
       ‘the new house’

(17) Haitian (French-based creole, North America; Fattier 2013)

- M*        *wè*    *ti*        *nèg*    *ki*        *frekan*        = **an**  
 1SG        see    little    man    REL    impertinent    = DEF  
 ‘I saw the boy who is impertinent.’

Cases like Haitian Creole, where articles occur even outside of relative clauses, are uncommon, but clitic articles are widespread. English *the* and *a(n)* are subminimal (lacking a full vowel), so that they are recognized as clitics even by authors who do not rely on the nonselectivity criterion (e.g. Dixon 2007).

Examples of a few more clitic articles are given below: Welsh (cym; Indo-European, Celtic) (37), Bulgarian (45, 53), Italian (63).

## 2.5. Question and negation particles

Polar question markers are typically clitics, because they can often be associated with a variable focus of the question (as seen in (7a-c) for Russian, and in (54) below for Turkish - tur; Turkic, Common Turkic). In languages where a question particle must occur in a peripheral position, it is still often a clitic because the initial or final expression of the question is not always of the same class. For example, in Mauwake (mhl; Nuclear Trans-New Guinea, Madang), the polar question clitic =*i* may occur on verbs or on nominals.

(18) Mauwake (Nuclear Trans-New Guinea, Madang; Berghäll 2015: 226)

- a. *Sira nain piipua-inan =i?*  
habit that leave-FUT.2SG = QM  
'Will you give up that habit?'
- b. *Nobonob ikiw-e-man nain, owowa eliwa =i?*  
Nobonob go-PST-2PL that village good = QM  
'You went to Nobonob, is it a good village?'

In Mapudungun (arn; Araucanian, Mapudungun), the question marker =*am* may occur in polar questions or constituent questions, and it can occur after different kinds of words.

(19) Mapudungun (Araucanian, Mapudungun; Zúñiga 2014: 165)

- Nepe-le-y ñi püñeñ =am?*  
wake.up-RES-IND my child = QM  
'Is my child awake?'

Languages with polar question markers as verbal affixes are not uncommon either, even though they are less frequent than question particles (Dryer 2005 finds 600 languages with question particles, and 179 languages with "interrogative verb

morphology”). In these languages, the polar question marker by definition always occurs on the verb.<sup>6</sup>

Clitic negation particles are quite common, too, and are illustrated below in (52) for Bulgarian and in (58) for Polish.

## 2.6. Adpositions

In the general literature on clitics, adpositions are not very prominent, though Dixon (2007) notes that several English prepositions (*at, for, to*, etc.) are very much like English auxiliaries and person forms in that they exhibit full forms and shortened forms. But crucially in the present context, adpositions are bound forms which indicate a nominal's semantic role and which do not always occur on the noun, so they are clitics by definition. Two examples are given in (20)-(21).

(20) Korean (Koreanic; Chae 2020: 133)

*Wuli-nun siktang = eyse achim pap = ul mek-ess-ta.*  
 we-TOP restaurant = in morning meal = ACC eat-PST-DECL  
 ‘We ate breakfast in a restaurant.’

(21) Ts'ixa (Khoe-Kwadi, Khoe; Fehn 2016: 108)

*Maá. | àm tsá gére táùn = m ʔò kúũ?*  
 when 2SG.M FUT town = SG.M ALL go  
 ‘When will you go to town?’

A few more examples of adpositional clitics are given below: Sri Lanka Malay (sci; Malay-based creole, Eurasia) (28), Russian (55), Fwe (fwe; Atlantic-Congo, Volta-Congo) (70), as well as in (47). I do not exemplify the nonselectivity of all these markers here, but all of them occur not only next to nouns, but also next to noun modifiers, as seen in (47a) and (70b) (see also fn. 19 for Japanese postposed flags).

---

<sup>6</sup> Dryer does not say so explicitly, but it appears that his *particles* are always nonselective (occurring adjacent to different types of words), while his *affixes* (or other morphology) are always verb-specific. He notes the nonselectivity requirement only for particles that are called *clitics* by the language describers: “Interrogative clitics, which attach to some word, but which exhibit freedom as to the category of word they attach to, are treated here as question particles.” (Dryer 2005: 470)

## 2.7. Subordinators

Like adpositions, subordinators are generally bound forms, and they are often nonselective as well, so that they are clitics rather than suffixes. Their clitic nature is particularly salient when there is phonological reduction, as with English *that* [ðət], French *que/qu'* (reduced before a vowel), or in the Chadic language Makary Kotoko (mpi; Afro-Asiatic, Chadic) *gí/g-* (reduced before a vowel):

(22) Makary Kotoko (Afro-Asiatic, Chadic; Allison 2020: 268)

*Ā*            *gə ən*    *g=*            *ú*    *sī*        *klayaská.*  
2SG.M.COMPL say 1SG    COMP =        1SG take    young.woman  
'He told me to take a young woman (as wife).'

In languages with predominant verb-final order, subordinators are often verbal suffixes (attaching only to verbs), but Turkish *=(y)ken* 'when' is a clitic as it attaches both to verbs and to nouns (Erdal 2000: 42).

(23) Turkish (Turkic, Common Turkic; Göksel & Kerslake 2005: 416)

- a. *Orman-da dolaş-ır =ken bir tilki gör-dü-m.*  
forest-LOC walk-AOR =when a fox see-PST-1SG  
'While walking in the forest, I saw a fox.'
- b. *Ahmet o kitab-ı öğrenci =yken oku-muş.*  
Ahmet that book-ACC student =when read-PRF  
'Ahmet read that book as ('when') a student, it seems.'

A few more examples of subordinator clitics are given in (48) below.

## 2.8. Coordinators

Coordinators meaning *and* or *or* are typically clitics, because they do not occur on their own and combine with forms of different word classes.

- (24) Amharic (Afro-Asiatic, Semitic; Demeke & Meyer 2008: 616)

*Sɛw-u = mm      nəgus-u = mm      dɛnɛgget-u.*  
 man-DEF = CONJ    king-DEF = CONJ    be.surprised-3PL

‘The people as well as the king were both surprised.’

- (25) Tsimshian (or Sm’algyax; Tsimshian; Stebbins 2003: 395)

*ʔgu      Hayda      ʔyuuta gwaʔa    daʔal    aam    wila    Smʔalgyax-t.*  
 small    Haida      man    this    but    good    FOC    Smʔalgyax-3.SBJ

‘This young man is Haida but he speaks good Smʔalgyax.’

A few more examples of coordinator clitics are given in (49) below, as well as in (31b) (English *and* = ), (56) (Latin = *que*), and (83) (Andi = *lo* = ).

## 2.9. Information-structural and discourse markers

Topic and focus markers are often clitics, as illustrated in (26)-(28), where they follow an adverb or a postposition.

- (26) Karbi (Sino-Tibetan, Kuki-Chin-Naga; Konnerth 2020: 466)

*Pini = ke                  etūm                  àn                  chō-ràp-pèt-sināng.*  
 today = TOP              1PL.INCL              rice              eat-together-all-HORT

‘Today, let us eat together.’

- (27) Bunaq (Timor-Alor-Pantar; Schapper 2022: 174)

*Neto    Hulul    gene = na    zol.*  
 1SG    Hulul    LOC = FOC    originate

‘It is Hulul where I come from.’

- (28) Sri Lanka Malay (Malay-based creole, Eurasia; Nordhoff 2009: 275)

*TV = ka = jo                  anà-kuthumung.*  
 TV = LOC = EMPH              PST-see

‘It was on TV that we saw it.’

Further examples of clitic discourse markers are English *however* in (39) and German *já* in (78).

### 3. Positional types of clitics

With respect to their position, we can distinguish several subtypes of clitics: at the most general level, we can distinguish enclitics and proclitics as well as ambiclitics (Section 3.1) and interclitics (Section 3.2). Using further kinds of criteria, we can also define second-position clitics (Section 3.3), clustering clitics (Section 3.4), and epiphrasal clitics (which occur at the edge of a clause, a nominal, or an adverbial; Section 3.5). These classes are not necessarily mutually exclusive.

#### 3.1. Enclitics and proclitics

The two best-known types are enclitics and proclitics, defined as in (29)-(30). (More transparent terms would be *postclitic* and *preclitic*, but two terms *enclitic* and *proclitic* are fairly old and are based on Greek prefixes, not on Latin *post-/pre-*.)<sup>7</sup>

##### (29) Enclitic

An enclitic is a clitic that can occur at the end of a free form but not at the beginning.

##### (30) Proclitic

A proclitic is a clitic that can occur at the beginning of a free form but not at the end.

For example, English Genitive 's can occur at the end of an elliptical answer, as in (31a), and the English coordinator *and* can occur at the beginning of the elliptical expression *and her dog*, as in (31b).

- (31) a.       A: *Is this your bike?* B: *No, (it's) **my friend's**.*  
      b.       A: *Who is coming?* B: *My friend Lee.* C: ***And her dog!***

The opposite situations are quite impossible (*my friend's bike* cannot under any circumstances be shortened to *\*'s bike*, and *Lee and her dog* cannot be shortened to

---

<sup>7</sup> The simple (prefixless) term *clitic* is actually fairly new (going back to Nida 1946; Stockwell et al. 1965). In earlier grammatical descriptions of the classical languages, we mostly find *enclitic*, but *proclitic* was already used in the 19th century.



\**Lee and*), so the reduced free forms in (31a-b) are the basis for classifying 's as an enclitic and *and* as a proclitic.

The nonclitic word preceding an enclitic, and the nonclitic word following a proclitic, will be called its ANCHOR in this paper. According to another very common terminology, the element with which a clitic can occur in such contexts is its HOST, and an equals sign serves as a boundary symbol linking a clitic to its host, as in (32). In many or most cases, *anchor* and *host* refer to the same word.

- (32) a.        *my friend = s bik*  
           b.        *Lee and = her = dog*

It is often said that “a clitic attaches to its host”, and the equals sign is generally taken as signaling this kind of *attachment*, but it is typically unclear what exactly this means. (In this paper, I do not attach particular significance to the notation with the equals sign. It can always be replaced by a space.)

Most often, linguists say that clitics form a prosodic unit (such as a phonological word) with their hosts, and they generally attribute this to their phonological “deficiency” (Halpern 1998: Section 1):

Clitics which form a prosodic unit with a host on their left are enclitics, while those forming a unit to their right are proclitics.

However, as will be discussed further in Section 7 below, it is often unclear how to identify the relevant prosodic units. Consider the Bulgarian object person index *ja* ‘her’, which occurs postverbally when the verb is clause-initial (as in (33a)), but preverbally when there is another preverbal constituent (as in (33b)). It cannot occur preverbally in initial position (see (33c)).

(33) Bulgarian (Indo-European, Balto-Slavic; Avgustinova 1994: 31)

- a.    *Vidjax ja.*  
       I.saw her  
       ‘I saw her.’

- b. *Otnovo ja vidjax.*  
 again her I.saw.  
 ‘I saw her again.’
- c. \**Ja vidjax.*  
 her I.saw  
 (‘I saw her.’)

According to the prosodic-unit criterion, should we say that *ja* in (33b) is proclitic to the verb, or that it is enclitic to the adverb? This is unclear and may not even be decidable, so it is better to define *enclitic* and *proclitic* with respect to their peripheral occurrence in free forms. According to this criterion, *ja* is an enclitic, because (33a) is possible as a free form, but (33c) is not. Thus, *otnovo* is the anchor in (33b), but it may not be the prosodic host.

A clear case of divergence between the criterion of occurrence in a free form and the prosodic criterion comes from Czech (ces; Indo-European, Balto-Slavic). Here too, the cognate person clitics are enclitic as that they cannot (in the formal standard variety) occur at the beginning of a free form. But according to Toman (1996), they lean prosodically to the following word when they occur after a long nominal phrase, as in (34), that forms a separate prosodic constituent.

(34) Czech (Indo-European, Balto-Slavic; Toman 1996: 506)

*Knihy, které tady vidíte, se dnes platí zlatem.*  
 books which here see.PRS.2PL REFL today pay.PRS.PL gold.INS  
 ‘The books you can see here are paid for with gold today.’

There are sometimes clear segmental effects depending on a neighbouring form, e.g. regressive voicing assimilation, as in English (34a) vs. (34b), or otherwise alternating forms, like Tagalog *ng* (after a vowel) vs. *na* (after a consonant), as in (35a) vs. (35b).

- (34) a. *my friend = [z] car*  
 b. *my bike = [s] brakes*

- (35) a. *wika ng Ingles* ‘English language’ (cf. also (5a))  
 b. *Ingles na wika* ‘English language’



Most clitics are enclitics or proclitics, but there are two other possibilities, ambiclitics and interclitics (Section 3.2). Ambiclitics are clitics that may look like a proclitic or like an enclitic, e.g. English *however*.

**(38) Ambiclitic**

An ambiclitic is a clitic that can occur at the end of a free form or at the beginning.

It is not usual to qualify *however* as a clitic, but as it does not occur on its own and is neither a root nor an affix, it is a clitic on the definition of this paper. It can occur initially, medially, or finally in a free form.

- (39) a. *However, our ambitious proposal failed.*  
b. *Our ambitious proposal, however, failed.*  
c. *Our ambitious proposal failed, however.*

Another example is the German adposition *entgegen* ‘against’, which can be used prepositionally or postpositionally (*entgegen meinem Rat* ‘against my advice’ or *meinem Rat entgegen*, van Gijn & Zúñiga 2014: 150).

**3.2. Interclitics**

Some languages have clitics that must occur between two other forms. These are called *interclitics* here, defined as in (40).

**(40) Interclitic**

An interclitic is a clitic that can occur neither at the end of a free form nor at the beginning.

An example is the Taglog linker =*ng/na*=, which was already illustrated in (35a-b) above. Further examples are in (41a-b). This morph (with variant *ng* after a vowel, and *na* after a consonant) occurs between an attributive adjective and a noun (these two elements may occur in either order). According to the positional criterion, it is an interclitic, not an enclitic or proclitic, despite the “backward-leaning” phonological behaviour of =*ng*=.

## (41) Tagalog (Austronesian, Malayo-Polynesian)

- a. *Malaki* = **ng** = *bahay*  
 big = LNK = house  
 'big house'
- b. *bahay* = **na** = *malaki*  
 house = LNK = big  
 'big house'

An even better-known example of an interclitic is the Persian Ezâfe form = (y)e =, as in *lebâs = e = zibâ = ye = Maryam* [dress =EZ= beautiful =EZ= Maryam] 'Maryam's beautiful dress' (Samvelian 2007: 608). This form occurs before an adnominal modifier, but only when a noun or another modifier precedes. It never occurs at the beginning or end of a free form.<sup>8</sup> Another example of an interclitic is the shortened English copula ('s, 're), illustrated in (36a-c) above.

Interclitics could be said to have two anchors, and the same might be said about ambiclitics when they occur in medial position (as in 39b). But it is unclear how to treat ambiclitics in general, so it may be best to restrict the term *anchor* to enclitics and proclitics for simplicity.

**3.3. Second-position clitics**

A number of languages have clitics which must occur in the second position in a clause (see Bošković 2016 for a survey). We saw a Tagalog example in (5a-b), a Russian example in (7a-c), and a Serbo-Croatian example in (9a-b). Clitics of this type were first identified for Ancient Greek, illustrated in (42).

## (42) Ancient Greek (Indo-European, Graeco-Phrygian; Sappho 118.3, Wackernagel 2020: 60)

<i>Aithiopíai</i>	<i>me</i>	<i>korai</i>	<i>Latoûs</i>	<i>anéthēken</i>
Ethiopian.DAT	me.ACC	girl.DAT	Leto.GEN	dedicated

<sup>8</sup> Samvelian (2007) treats the marker = (y)e = as a suffix, but it occurs both after nouns and adjectives, so it is not class-selective and hence a clitic.

*Arista.*

Arista.NOM

‘Aristas dedicated me to Leto’s Ethiopian daughter.’

Wackernagel’s (1892) paper made second-position clitics famous by pointing out that they occur in a number of ancient Indo-European languages (see Walkden 2020 for some background). Clitics of this type have also been found in various other parts of the world, including Uto-Aztec languages, Panoan languages, and a number of Australian languages such as Wambaya.

(43) Wambaya (Mirndi, Ngurlun; Nordlinger 1998: 139, 140)

- a. *Darranggu-nu ngiyi = ng = a irrijabi.*  
 stick-LOC 3SG.NONMASC.ERG = 1.OBJ = NONFUT scratch  
 ‘The stick scratched me.’
- b. *Dagama gini = ng = a ngirra.*  
 hit 3SG.MASC.ERG = 1.OBJ = NONFUT us.EXCL.ACC  
 ‘He hit us.’
- c. *Guyala ngurr = uji ngajbi irra.*  
 NEG 1PL.INCL.ERG = IRR.PRS see them.ACC  
 ‘We have never seen them.’

Nordlinger (1998) calls the second-position clitic clusters *auxiliary* and writes them as one word, but they are not different in nature from the Tagalog or Serbo-Croatian clusters.

Second-position clitics are usually enclitics, but it may be that some of them are interclitics (requiring another form to follow). I define this type as in (44).

(44) **Second-position clitic**

A second-position clitic is a clitic that must occur (possibly as part of a clitic cluster) directly after the first word or nominal or adverbial expression of a clause, or after the first word of a nominal.

Most second-position clitics occur after the first nominal or adverbial of a clause, or after an initial verb or particle (as in (43b-c)), and it is rare to find such clitics after

the first word when this is part of a nominal. An example of such a *nominal-internal* clausal clitic is (42) from Lesbian Greek.

For the definition in (44), one might want to use the more general formulation “must occur after the first constituent”, but the terms *nominal (expression)* and *adverbial (expression)* are very clear (and *word* is fairly clear, too), while *constituent* is more abstract and may not be so clear. In the literature, there are rich discussions of the precise conditions under which second-positions occur in particular languages, often involving prosodic conditions (see, e.g., Bošković 2016). However, the vast majority of what have been called second-position clitics fall under the comparative concept in (44).

Second-position clitics within nominals can be illustrated by the Bulgarian definite article in (45).

(45) Bulgarian (Indo-European, Balto-Slavic; Halpern 1995: 153)

- a. *kniga* = **ta**  
 book = DEF  
 ‘the book’
- b. *xubava* = **ta** *kniga*  
 nice = DEF book  
 ‘the nice book’
- c. *moja* = **ta**      *xubava*      *kniga*  
 my = DEF      nice      book  
 ‘my nice book’

Not only articles, but also adpositions may occur in second position, though this is rare. Dryer (2005: 211) calls such clitics *inpositions* and gives an example from Yawuru (ywr; Nyulnyulan, Eastern Nyulnyulan).

(46) Yawuru (Nyulnyulan, Eastern Nyulnyulan; Hosokawa 1991: 81, 383–384)

- a. *dyungku* = **gun**  
 fire = LOC  
 ‘in the fire’

- b. *bika* = **gun**     *larrkadi*  
 shade = LOC     bottle.tree  
 ‘in the shade of a bottle-tree’
- c. *nyamba* = **gun**     *maya*  
 this     = LOC     house  
 ‘in this house’

### 3.4. Epiphrasal clitics

Many languages have clitics that provide information on a phrase’s relationship with its environment and occur peripherally, either in a nominal expression (i.e. adpositions), or in a clause (i.e. subordinators) (as we already saw in Sections 2.6-7). They may be proclitics or enclitics.

#### (47) Examples of adpositional clitics

- |    |           |             |       |  |
|----|-----------|-------------|-------|--|
| a. | English:  | <i>to</i>   | ‘to’  | <i>to our house</i>                      |
| b. | French:   | <i>pour</i> | ‘for’ | <i>pour notre maison</i> ‘for our house’ |
| c. | Hebrew:   | <i>le-</i>  | ‘to’  | <i>le = David</i> ‘to David’             |
| d. | Japanese: | <i>no</i>   | ‘of’  | <i>Hanako = no</i> ‘of Hanako’           |

#### (48) Examples of subordinator clitics

- |    |          |             |         |                                 |                         |
|----|----------|-------------|---------|---------------------------------|-------------------------|
| a. | German:  | <i>als</i>  | ‘when’  | <i>als wir träumten</i>         | ‘when we were dreaming’ |
| b. | Persian: | <i>ke</i>   | ‘that’  | <i>ke âmadi</i>                 | ‘that you came’         |
| c. | Arabic:  | <i>iðaa</i> | ‘if’    | <i>iðaa kun-ta hunaaka</i>      | ‘if you-are there’      |
| d. | Chinese: | <i>de</i>   | ‘which’ | [ <i>lái de</i> ] <i>nánhai</i> | ‘the boy [who came]’    |

Another class of clitics that commonly occur peripherally to a phrase is coordinator clitics (as already seen in Section 2.8). These are often interclitics.

#### (49) Examples of coordinator clitics (Fortescue 1984: 120)

- |    |          |           |       |   |
|----|----------|-----------|-------|---|
| a. | Spanish: | <i>y</i>  | ‘and’ | <i>guerra y paz</i> ‘war and peace’       |
| b. | Lezgian: | <i>ni</i> | ‘and’ | <i>buba = ni died</i> ‘father and mother’ |



- c. Russian: *ili* 'or' *zdes' ili tam* 'here or there'  
 d. Greenlandic: *= lu* 'and' *ingilluni = lu* 'and she sat down'

Many or most languages also have focus clitics, especially particles meaning 'only' or 'also', which typically occur epiphrasally, as in (50a-c).

(50) Focus clitics

- a. Polish: *tylko* 'only' *tylko dzisiaj* 'only today'  
 b. M. Greek: *ke* 'too' *ke i mitéra mu* 'my mother, too'  
 c. Hungarian: *is* 'too' *én is* 'me too'  
 d. Japanese: *mo* 'too' *watashi mo* 'me too'

However, especially clitics that mean 'also' can occur in a *floating* position, not immediately adjacent to their focus. Example (51) shows the stressed clitic *auch*, which can occur in a preverbal position, not adjacent to its focus. Thus, *AUCH* is a clitic, but it is not epiphrasal.

(51) German<sup>9</sup> (Indo-European, Germanic)

*Meine Schwester ist heute = AUCH gekommen.*  
 my sister has today = also come  
 'My sister came today, too (= 'also my sister' or 'also today').'

A non-peripheral position is not common for the other semantic classes illustrated above, though we sometimes find second-position coordinators (e.g. in Latin, ex. (56) below), and occasionally second-position adpositions (e.g. Yawuru, ex. (46) above).

Epiphrasal clitics are very common, but they are not prominent in the general-theoretical literature on clitics. This is probably because they present no particular problem of analysis, and not because they would not fall under the usual clitic concept.

---

<sup>9</sup> German stressed *AUCH* is an enclitic, because it can occur at the end of a free form (e.g. *heute = AUCH* 'today, too'). There is also a fully synonymous unstressed *auch* (used in a more formal register), which is a proclitic, because it can occur at the beginning of a free form (*auch = HEUTE* 'also today'). If these two instances of *auch* were treated as the same form, this form would be an ambiclitic.

### 3.5. Clustered clitics

Clitics sometimes occur in clusters with rigid internal ordering. Such clusters may occur in second position (as we already saw for Tagalog, Serbo-Croatian, and Wambaya, Section 3.3), but they can also be proclitic, as in the Bulgarian example in (52). Here the clitics occur in a rigid position: *da* – NEG – AUX – DAT – ACC (see Avgustinova 1994: 32 for details).<sup>10</sup>

(52) Bulgarian (Indo-European, Balto-Slavic; Spencer & Luís 2012: 125)

<i>Da</i>	<i>ne</i>	<i>si</i>	<i>mu</i>	<i>go</i>	<i>dala</i>	<i>poveče!</i>
that	not	2SG.AUX	3SG.DAT	3SG.ACC	give	more

‘Don’t give it to him any more, or else!’

For syntactic elements, this kind of rigid ordering is not expected, and this has led some linguists to think of clustered clitics more in *morphological* terms. More generally, Spencer & Luís (2012: 126) note that:

elements that are traditionally called clitics may exhibit a good many features normally associated with affixes ... when they combine into clusters: a fixed order, idiosyncratic alternations in ordering, haplology, idiosyncratic allomorphy, and accidental gaps, not to mention multiple exponence and cumulation.

This is another reason for caution in attributing the properties of types of linguistic forms to larger architectures (“morphology vs. syntax”, “grammar vs. lexicon”, etc.). At present, our understanding of the reasons for these behaviours is quite limited.

## 4. Phonological types of clitics

In this section, I describe and illustrate three types of clitics based on their phonological properties: welded clitics (Section 4.1), stress-affecting clitics (Section 4.2), and shortened clitics (Section 4.3).

---

<sup>10</sup> Above in (33), we saw that single object clitics are enclitics, but (52) shows that when they occur in a cluster that begins with *da*=, the entire cluster is proclitic. So perhaps we should say that object clitics are ambiclitics, because they can occur both at the end of a free form and (as part of a cluster) at the beginning of a free form.

#### 4.1. Welded clitics

A welded clitic is a clitic that interacts with a neighboring form in a segmental way, either by causing segmental change or by undergoing segmental change. For example, the Bulgarian enclitic definite article (*-ət/-ta/-to* in the singular, see (45)) may cause a segmental change in the preceding noun, as seen in (53); the Turkish question particle *mU* shows vowel harmony, harmonizing with the last vowel of the preceding word, as seen in (54); and several Russian prepositions (e.g. *v(o)*, *k(o)*, *o(b)*) have somewhat different shape variants depending on various properties of the next word, as seen in (55).

(53) Bulgarian (Indo-European, Balto-Slavic): definite article

<i>kniga</i>	[book]	<i>kniga-ta</i>	[book-DEF]
<i>vol</i>	[ox]	<i>vol-ət</i>	[ox-DEF]
<i>grək</i>	[Greek]	<i>gərək-ət</i>	[Greek-DEF] (with stem change)

(54) Turkish (Turkic, Common Turkic): polar question particle

<i>geldi mi?</i>	‘did she come?’
<i>öldü mü?</i>	‘did he die?’
<i>Alı mı?</i>	‘Ali?’
<i>dün mü?</i>	‘yesterday?’

(55) Russian (Indo-European, Balto-Slavic): prepositions *v(o)* ‘in’, *s(o)* ‘with’

<i>v nužde</i>	‘in need’	<i>vo vrede</i>	‘in harm’ (* <i>v vrede</i> )
<i>s radost’ju</i>	‘with joy’	<i>so straxom</i>	‘with fear’ (* <i>s straxom</i> )

In English, the difference between the two variants of the indefinite article *a(n)* (e.g. *a tree* vs. *an old tree*) is a striking case of welding in a proclitic.

For welded clitics, we might distinguish between BACKWARD-WELDED clitics (whose segmental shape interacts with the shape of a preceding form) and FORWARD-WELDED clitics (which interact with a following form). In general, backward-welded clitics are enclitics, e.g. the Turkish polar question particle *mi/mü/mu/ml*, and forward-welded forms are proclitics, e.g. the Russian prepositions *v(o)* and *s(o)*. But proclitics may also be backward-welded, as seen in the Welsh definite article *yr/’r* in (37) above, and enclitics may be forward-welded

(e.g. Kukama = *pura*, discussed by Zingler 2020: 266). Quite generally, it seems that backward-welding is more common than forward-welding (Himmelmann 2014). The interclitics that we saw in Section 3.2 are backward-welded (Tagalog *na/ng*, Persian *e/ye*).

Welding does not seem to happen very often with clitics, and it has in fact been suggested that “morphophonological idiosyncrasies” are symptomatic of affixes, but not of clitics (Zwicky & Pullum 1983: 504; Nevis 2000: 389).<sup>11</sup> If phonological interaction plays no role in the definition (as in the present proposal), then the lack (or scarcity) of segmental interactions between clitics and adjacent words becomes an interesting testable prediction that we can make.

#### 4.2. Stress-affecting clitics

Most clitics do not interact suprasegmentally (with respect to their stress or tone properties) with adjacent words, and we can call them *suprasegmentally inert*. For example, while most Turkish words have final stress, the question clitic (e.g. *geldi mi?* ‘did she come?’ in (54) above) is not part of the stress domain and thus does not carry stress. But some clitics are suprasegmentally active in that they are relevant for the stress or tone properties of their anchor words (or perhaps for other adjacent words). Here we will briefly consider clitics which affect the stress of their anchor word.

We can distinguish two types of stress-affecting clitics. First, *STRESS-SHIFTING CLITICS* are clitics which induce a shift of the stress pattern of their anchor word. For example, the Latin conjunctive clitic = *que* induces a stress shift to the final syllable of the anchor word that it annexes to (see Plank 2005).

(56) Latin (Indo-European, Italic)

- a. *ménsa* ‘the table’
- b. *mensá = que* ‘and the table’

---

<sup>11</sup> Conceivably, this could be because clitics are combined “postlexically” with their anchors, while affixes are combined “lexically” and thus undergo “lexical phonological” processes (e.g. Anderson 2005). But on such a view, it is puzzling that phenomena such as Bulgarian *grək ~ gərək-ət* as in (53) are attested at all (see also Halpern 1995 on “lexical clitics”, and Spencer & Luís 2012: Section 4.4.3).

A similar effect is found in Modern Greek (ell; Indo-European, Graeco-Phrygian), where enclitics (such as *mas* ‘our’, *mu* ‘my’) are unstressed but induce an additional stress on anchor words that are stressed on the antepenultimate syllable, as illustrated in (57b). Words that are stressed on the penultimate or ultimate syllable are unaffected, as seen in (57c-d).

(57) Modern Greek (Indo-European, Graeco-Phrygian; van Oostendorp 2012: 1166)

- |    |                        |                 |                      |
|----|------------------------|-----------------|----------------------|
| a. | <i>o jítonas</i>       | ‘the neighbour’ |                      |
| b. | <i>o jìtonáz = mas</i> | ‘our neighbour’ |                      |
| c. | <i>i stafíða = mu</i>  | ‘my raisin’     | ( <i>i stafíða</i> ) |
| d. | <i>i ayorá = mas</i>   | ‘our market’    | ( <i>i ayorá</i> )   |

Stress-shifting clitics seem to be rare, but they have been prominent in the literature, because Greek and Latin are such important languages in Western culture, and the term *enclitic* was originally used for stress-shifting Greek clitics.

The second type is STRESS-INTEGRATED CLITICS. These are clitics which are part of the anchor word’s stress domain and carry stress when the general stress rule would assign stress to them. An example is the Polish negator *nie*, which is stressed when it occurs with a monosyllabic verb form.

(58) Polish (Indo-European, Balto-Slavic; Rubach & Booij 1985: 317)

- |    |                   |                      |
|----|-------------------|----------------------|
| a. | <b><i>nie</i></b> | <i>wiedziáta = m</i> |
|    | NEG               | knew = 1SG           |
|    |                   | ‘I did not know’     |
| b. | <b><i>nié</i></b> | <i>wie-m</i>         |
|    | NEG               | know-1SG             |
|    |                   | ‘I do not know’      |

In this regard, the negator *nie* contrasts with the polar question particle *czy*, which is a clitic, too (*czy wiém?* ‘do I know?’), but which behaves like most clitics in that it is

outside the stress domain.<sup>12</sup> Another clitic that is inside the stress domain is Mapudungun = *ám* (seen in example (19) above), which receives stress as consonant-final words have final stress (Zúñiga 2014: 165).<sup>13</sup>

#### 4.3. Shortened clitics (vs. length-invariant clitics)

Some clitics are closely related to formally similar counterparts and can be regarded as abbreviated variants of them. This type is very well known from English, illustrated in (59).

##### (59) English (Indo-European, Germanic)

FULL	SHORTENED
<i>will</i>	<i>'ll</i>
<i>would</i>	<i>'d</i>
<i>is</i>	<i>'s</i>
<i>are</i>	<i>'re</i>

##### (60) German (Indo-European, Germanic) (see (10) above)

FULL	SHORTENED	
<i>sie</i>	<i>se</i>	'she, her, they, them'
<i>du</i>	<i>de</i>	'you'
<i>es</i>	<i>s</i>	'it'
<i>er</i>	[ <i>e</i> ]	'he'

---

<sup>12</sup> Rubach & Booij (1985: 317) suggest that *nie* is a prefix, but since it occurs nonselectively (e.g. *nie dzisiaj* 'not today'), it is a clitic on the current definition.

<sup>13</sup> Very rarely, a clitic may be part of the stress domain of a preceding word that is not its anchor: In Chamicuro (ccc; Arawakan, Southern Maipuran), definite articles may be part of the stress domain of the preceding verb, noun or demonstrative (e.g. *aná? = na čmežóna* [this = DEF man] 'this man', Parker 1999: 554). This "backward-leaning" behaviour is similar to backward-welded proclitics which were mentioned briefly in Section 4.1 above.

This clitic type was called *simple clitics* by Zwicky (1977), and this term has become very well known in the literature.<sup>14</sup> However, it is not well-defined, so I prefer the new term SHORTENED CLITIC, defined as in (61).

**(61) Shortened clitic**

A shortened clitic is a clitic that has the same semanticosyntactic function as another form from which it appears to have been abbreviated and by which it can be replaced in the same position.

This notion does not seem to be particularly important in the world's languages, because such clitics are not common (see also Zingler 2020: 337–338). They have been prominent in the literature primarily because English has several such pairs where both a full form and a shortened form occur in the standard spelling.

The great majority of clitics are length-invariant, i.e. they do not occur in pairs such as (59)-(60). Of course, all languages have fast-speech phenomena, and variant forms of function words are extremely common. But there is no good reason to associate the shortened forms with cliticness, because the full counterparts of shortened clitics are very often bound forms and thus clitics, too (as with the English forms in (59)).<sup>15</sup> And the existence of pairs of full forms and shortened forms does not seem to be characteristic of clitics as opposed to affixes, because affixes often have full and reduced variants, too (e.g. German genitive suffix *-es* or *-s*, Italian 3PL suffix *-on* or *-ono*). Thus, the very notion of “simple clitics” seems to be primarily based on the peculiarities of the English spelling.<sup>16</sup>

---

<sup>14</sup> It is sometimes said that the full forms are free forms (e.g. “Simple clitics are unaccented variants of free morphemes”, Anderson 2005: 10), but this not the case for these English forms. The full forms are non-deficient in that they contain a full vowel (not just a single consonant, or just a schwa), but they are bound forms, too, like the reduced forms.

<sup>15</sup> It seems that person forms are different from function words in this regard: Independent person forms can occur on their own and are not clitics, but when shortened person forms arise from these (like German *se* and *de*), these are typically bound and are thus clitics.

<sup>16</sup> Some authors do not include the requirement that “simple clitics” must have full-form or free-form counterparts, e.g. Halpern (1998: Section 1): “An unstressed word which is otherwise unexceptional is known as a simple clitic, after Zwicky (1977)”. Given the general unclarity surrounding Zwicky's terms, it is surprising how popular they were for a few decades.

## 5. Defining clitic

The present definition of *clitic* (repeated below from Section 1 above) is somewhat unusual in that it has no direct antecedent in the literature. However, we will see in this section that it accords well with the way the term has been used in the past.

### (1) Clitic

A clitic is a bound morph that is neither an affix nor a root.

Everyone agrees that clitics are bound forms (incapable of occurring in isolation), and nobody would suggest that a root (a morph denoting an object, an action or a property) can be a clitic.<sup>17</sup> It is also clear that a clitic stands in contrast with an affix, as there are a large number of works that aim to distinguish between affixes and clitics in a language (often following the lead of Zwicky & Pullum's famous 1983 paper). There are thus mainly two points where one might want to opt for a modification of the definition in (1), and one might question (62a) or (62b) (or both):

(62)

- a. clitics are monomorphic
- b. clitics are defined as not exhibiting word-class selectivity (= as non-affixal)

That a clitic is a single morph is not something that has often been said, but it is easy to see that this is the case for the great majority of elements that have been called clitics. Almost all question or negation particles, discourse particles, short adpositions, subordinators and coordinators are monomorphic, and so are many person indexes. It is true that auxiliaries and articles are not uncommonly multimorph, as illustrated in (63) by the Italian article *le* and the auxiliary *hanno*, which can be analyzed into smaller constituent morphs (*l-* + *-e*, *ha-* + *-nno*).

(63) Italian (Indo-European, Italic)

*L-e*      *donn-e*      *ha-nno*      *lavorato*.  
DEF-F.PL   woman-PL   AUX.PRF-3PL   worked  
'The women have worked.'

---

<sup>17</sup> Exceptionally, Chae (2020: 105) discusses "clitic nouns" in Korean (kor; Koreanic), but this way of talking is very unusual. The elements he discusses are normally treated as derivational suffixes.



However, the constituent morphs of *le* and *hanno* can be treated as individual clitics, so that *l=e=* and *ha=nno=* can be seen as clitic clusters. It is true that this is a nontraditional way of describing these forms, but there does not seem to be a good alternative. If a clitic could consist of a sequence of morphs (i.e. if there could be such a thing as “a composite clitic”), then all clitic clusters could be composite clitics. Thus, it is best to specify that clitics must be monomorphic by definition.

The other somewhat nontraditional component of the definition is (62b), the lack of word-class selectivity. This criterion (also called *promiscuity* for short) will be discussed further below in Section 6.

In addition to boundness (= non-independence) and nonselectivity, quite a few authors mention phonological criteria, especially *phonological dependence* or *deficiency*, as in (64).

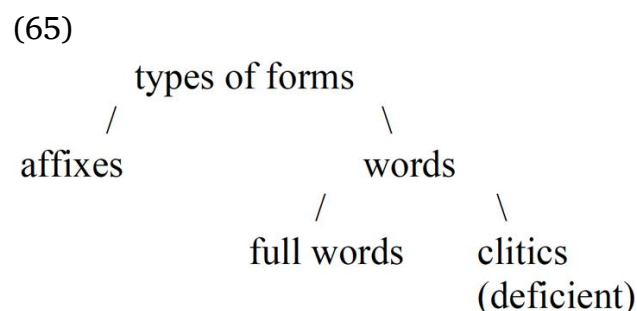
(64)

- a. “[clitics:] grammatical elements which themselves bear no stress and which make up a phonological word with a host item (that bears stress) which the clitic either precedes (it is then a proclitic) or follows (an enclitic)” (Dixon 2007: 574)
- b. “The best way to define the special status of clitics is that in terms of prosodic deficiency: they are words in the morpho-syntactic sense, but not in the phonological sense.” (Booij 2012: 290)
- c. “The most prominent property of clitics is their deficiency. Most often this deficiency is attributed to the phonological status of clitics: clitics are defective in their phonological representation and therefore have to prosodically combine with an adjacent non-clitic word.” (Ionova 2020: 22)
- d. “Clitics are function words that lack independent stress.” (Pescarini 2021: Section 1.1)

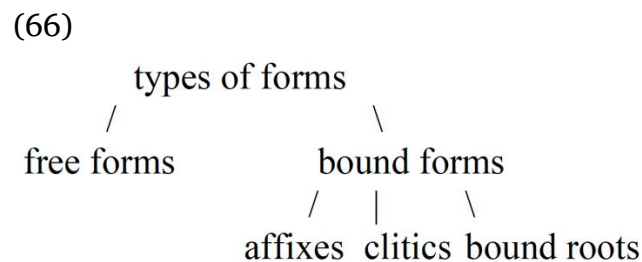
But is such an additional phonological criterion necessary? As Anderson (2011: 4) notes, phonological deficiency or dependence does not distinguish clitics from affixes:

With relatively few exceptions, the affixes found within words as formal markers of derivational and inflectional structure also lack an autonomous organization into prosodic constituents at or above the level of the [p-word].

Thus, the one criterion that clearly distinguishes clitics from affixes is the lack of word-class selectivity (or *promiscuity* of attachment; Section 6). Even though linguists often express the intuition that clitics are “phonologically attached” to their host (resulting in the Greek term *enklitikón* ‘leaning’), this criterion would be significant only if we already knew that clitics are words. If we could initially divide the types of (non-phrasal) forms into affixes and words, as in (65), then the prosodic deficiency of clitics would be relevant.



But in fact, clitics and affixes are very similar in that they are prosodically deficient, and there is no simple phonological criterion distinguishing affixes from words. Thus, the subdivision in (66) below is much more straightforward, as it is based on the simple criterion of boundness (non-occurrence in isolation).



Note that in addition to being bound forms, affixes and clitics must be defined as nonroot forms, because roots may be bound, too. For example, English requires an article or a plural marker with count noun roots (e.g. A: *What do you want to buy?* B: *A book/Book-s/\*Book*), and it requires an object with many transitive verbs (e.g. A:

*What will you do with it?* B: *Replace it/\*Replace*). Thus, many more roots are bound forms than most English-speaking linguists seem to realize.<sup>18</sup>

By specifying that a clitic is not a root in the definition in (1), we distinguish it from simple nouns, verbs and adjectives without making reference to phonological criteria. As the phonological criteria are complex and difficult to apply (see Section 7), this is a clear advantage of this definition.

But the most important way in which the definition in (1) is superior to many other views is that it relies on a small set of clear criteria that could easily be used in textbooks. By contrast, the earlier literature has made reference to a wide range of diagnostics, and the heterogeneity of the criteria is reflected in quotations such as (67).

(67)

- a. “a serious problem which prevails in much of the work on clitics... is that there is no criterial definition, but rather a list of tendencies, general characteristics, and typical features...” (Klavans 1985: 116)
- b. “the various elements which are called clitics form a heterogeneous bunch, at least superficially, and exactly what is meant by “clitic” varies from study to study” (Halpern 1998: Section 1)
- c. “It is extremely difficult to come up with an explicit set of characteristics that may be used to identify clitics cross-linguistically, because the parameters involved vary from case to case” (Stebbins 2003: 385)

Perhaps the most problematic aspect of the earlier approaches (in the tradition of Zwicky & Pullum 1983) is not that they are complex and that the criteria are heterogeneous, but that they may not always point in the same direction and that their application is often subjective (Haspelmath 2015: Section 3). Subjectiveness is a frequent problem in linguistics when “test batteries” are applied and there is no clear rule for how to proceed when the diagnostics do not all point in the same

---

<sup>18</sup> Traditionally, the term “bound root” has been applied especially to roots that occur only in compounds, e.g. Mandarin Chinese *-gōng* ‘worker’, which only occurs in compounds such as *mù-gōng* [wood-worker] ‘carpenter’ and *diàn-gōng* [electricity-worker] ‘electrician’ (Arcodia 2012: 91). English roots such as *book* are often treated as “free morphemes” in textbooks, but by the criterion of independent occurrence, they are not free.

direction (Croft 2010: Section 2.10; Tallman 2020). This problem affects not only the definition of *clitic*, but also the definition of the subtype *special clitic*, which is often said to involve *special syntax*: “A special clitic is ... a “little word” whose syntax is not assimilable to that of full words that might seem to be syntactically parallel” (Anderson 2005: 79). But if there are no limits on the ways in which special clitics might differ from full forms, then this definition cannot be applied objectively.

Taylor (1995: 181) explicitly argues for “graded membership” in the categories *word*, *clitic* and *affix*, but he provides no generally applicable method for measuring degrees of wordhood, cliticness or affixhood. He may well be right that a modular view of grammar (where morphology and syntax are two different modules) is inappropriate, but this does not entail the conclusion that categories like *word* and *clitic* must exist and must have a prototype structure. Maybe these concepts are primarily tools used by linguists and play no role in speakers’ mental grammars. In any event, unless we know exactly what someone means by an affix, a clitic, or word, it is very difficult to evaluate their statements.

## 6. Lack of word-class selectivity

A crucial component of the definition in (1) is the lack of word-class selectivity of clitics, because it is in this way that clitics differ from affixes, which are bound nonroot morphs as well. That clitics contrast with affixes in this way is fairly widely recognized, as is shown by the quotations in (68).

(68)

- a. “Clitics can exhibit a low degree of selection with respect to their hosts, while affixes exhibit a high degree of selection with respect to their stems.” (Zwicky & Pullum 1983: 503)
- b. “It is largely because of their freedom to attach to practically any part of speech that clitics are recognized as a special linguistic unit.” (Taylor 1995: 180)
- c. “[clitic:] a unit that is not a word in a prototypical sense, but with fewer selectional restrictions than a grammatical affix” (Hildebrandt 2015: Section 1)
- d. “an element is only considered a clitic if it has a non-selective distribution but is dependent on a host domain with respect to one or more parameters of phonological wordhood” (Zingler 2022: 9)

In the literature, we also often read that clitics differ from (standard) affixes in that they are associated with phrases rather than words; in fact, clitics are often called “phrasal affixes” (e.g. Anderson 1992: Ch. 8). And as we saw in Section 3.4 above, some clitics (called *epiphrasal clitics*) are clearly associated with nominal (or adverbial) expressions, or with clauses, and they occur strictly on the periphery of such phrases (e.g. English prepositions, or words for ‘also’ in many languages). These elements are clitics and not affixes because the phrases that they occur with do not always have the same kind of word at the periphery; for example, English prepositions are epiphrasal clitics and not prefixes, because they can occur adjacent to nominal modifiers, not only to nouns (as in (47a) above: *to our house*).

The definition of clitics as *nonroot bound forms* that are not affixes has the advantage that it does not rely on a notion of *phrase*, but only on the nature of adjacent forms. While nominal (and adverbial) phrases may be easy to identify in most languages, this is much less straightforward with verb phrases and all kinds of other phrases. Moreover, not only clitics, but also many affixes have a *phrasal* distribution in a certain sense: Tense affixes and case affixes occur on verbs and nouns, respectively, but semantically and functionally they combine with (verb and noun) phrases.

Linguists often treat case affixes and adpositions together (as *flags*, Haspelmath 2019), and also tense affixes and tense clitics (as *tense markers*), because they do not differ except in their position with respect to the noun or verb. And sometimes we find minimal pairs which appear to show that clitics are not really different in nature from affixes. For example, in Egyptian Arabic (arz; Afro-Asiatic, Semitic) *prepositions* such as *mafa-* ‘with’ are prefixes, as illustrated in (69).

(69) Egyptian Arabic (Afro-Asiatic, Semitic; Gary & Gamal-Eldin 1982: 63, 86)

- a. *mafa- xaal-i*  
 with- uncle-1SG  
 ‘with my uncle’
- b. *mafa- l-bint di*  
 with- DEF-girl this  
 ‘with this girl’

They always occur immediately adjacent to the noun because all nominal modifiers follow the noun, including demonstratives. By contrast, their Standard Arabic counterparts are proclitics, because they occur before the demonstratives and these can occur on their own (and are thus not prefixes), e.g. *maʕa = haad̥aa l-kitaab* ‘with this book’.

Similarly to Standard Arabic, the Bantu language Fwe has a preposition-like connective proclitic (*bo =* or *ye =* in (70)) that can occur not only adjacent to nouns, but also to prenominal demonstratives which are not prefixes (Gunnink 2022: 189).<sup>19</sup>

(70) Fwe (Atlantic-Congo, Volta-Congo; Gunnink 2022: 186, 189)

- a. *bàntù bò = kúmùnzì*  
 people CONN = village  
 ‘the people of the village’
- b. *Èmísì yè = cí cìshámù mùshámù.*  
 roots CONN = this tree COP.medicine  
 ‘The roots of this tree are medicine.’

Moving on to tense markers, an instructive pair of languages is Greek and Bulgarian. The Modern Greek future-tense marker *tha-* seen in (71a) is a prefix, while the very similar Bulgarian *šte =* in (71b) is a clitic. Its preverbal position is rigid, but it precedes the clitic *go*, and so both must be proclitics. In Modern Greek, the spelling may suggest that *tha-* is a clitic, too, but the Greek object indexes behave differently from Bulgarian: They are always adjacent to the verb and are thus affixes, which means that the future-tense marker *tha-* is a prefix (see also Joseph 2002).

(71)

- a. Modern Greek (Indo-European, Graeco-Phrygian)  
*tha- to- páro*  
 FUT it take.1SG  
 ‘I will take it’

---

<sup>19</sup> In Japanese, the reason for treating postpositions as clitics rather than case suffixes is similarly subtle: as Nakamura (2018: 249) notes, the phrasal clitic =*dake* ‘only’ may precede the clitic postpositions, e.g. *Hanako = dake = ga* ‘only Hanako (NOM)’. See also Chae (2020: 39–40, 140) for a discussion of the similar situation in Korean.

## b. Bulgarian (Indo-European, Balto-Slavic)

**šte** = **go** = *vzema*  
 FUT it take.1SG  
 'I will take it'

The term *nonselectivity*, or the older term *promiscuity*, may suggest that clitics are completely indifferent as to the words they are adjacent to, but this is not the case. Clitics contrast with affixes, which must be class-selective (occur always with roots of the same class), so any deviation from full class selectivity means that the element in question must be a clitic rather than an affix. In Standard Arabic, prepositions can only occur with nouns and demonstratives, so they are fairly choosy, but they are not affixes. In some of the quotations in (68), the authors assume degrees of selectivity, but a definition of a term like *clitic* must be clear-cut.<sup>20</sup>

It should be noted that bound nonroot forms are affixes also when they are “mobile” in that they may occur on either side of the root (Haspelmath 2021: 19). This means that some class-selective elements which have often been regarded as clitics because of their phonological properties are not clitics. For example, the person-number indexes in Kugu Nganhcara (boldfaced in (72)) are verbal affixes because they always occur next to the verb.

(72) Kugu Nganhcara (Pama-Nyungan, Paman; Klavans 1985: 104; Smith & Johnson 1985: 103–104)

- a. *Nhila pama-ng nhingu pukpe-wu ku?a wa: (=)-ngu.*  
 DET.NOMman-ERG DET.DAT child-DAT dog give 3SG.DAT  
 'The man gave the dog to the child.'
- b. *Nhila pama-ng nhingu pukpe-wu ku?a (=)ngu- wa:*  
 DET.NOMman-ERG DET.DAT child-DAT dog 3SG.DAT give  
 'The man gave the dog to the child.'

<sup>20</sup> Peter Arkadiev (p.c.) has expressed the intuition that clitics are perhaps better characterized as being completely nonselective, whereas affixes may be somewhat flexible with respect to word classes (e.g. number suffixes attaching both to nouns and adjectives). This would be a logical alternative, but I think that it is much easier to determine that a bound morph is fully selective (i.e. occurs only on one class) than to determine that it is fully nonselective. As a result, the definition of *clitic* is fairly broad here, including all bound nonroot morphs that are not fully selective.

- c. *Nhila pama-ng nhingku kuʔa (=)ngku- wa:*  
 DET.NOMman-ERG you.DAT dog 2SG.DAT give  
 ‘The man gave a dog to you.’

They may occur postverbally (as in 72a) or preverbally (as in 72b-c), but they are always verb-adjacent and thus count as affixes (specifically, they are ambifixes; see Arkadiev & Lander 2020). The literature has often regarded them as clitics (Klavans 1985: Section 2.2; Cysouw 2005), and indeed they are phonologically peculiar in that they have a phonotactic shape that excludes them from initial occurrence. In this sense, one might say that a verbal affix such as *ngku-* (in 72c) leans onto the element preceding it. But few clitics are restricted in this way, and alternations between preverbal and postverbal position of person indexes may be conditioned by a variety of factors (see (33) above for Bulgarian, and (87) below for Portuguese - *por*; Indo-European, Italic).

A language that is similar to Kugu Nganhcara in that its person indexes are most often directly postverbal or preverbal is Persian, illustrated in (73) (example (73b) is repeated from (4) above).

(73) Persian (Indo-European, Indo-Iranic; Samvelian & Tseng 2010: 215)

- a. *Ru-ye miz gozâšt-im = aš*  
 on-EZ table put.PST-1PL.SBJ = 3SG.OBJ  
 ‘We put it on the table.’
- b. *Ru-ye miz = aš gozâšt-im.*  
 on-EZ table = 3SG.OBJ put.PST-1PL.SBJ  
 ‘We put it on the table.’

That these Persian forms are clitics (and not ambifixes as in Kugu Nganhcara) can be seen in examples like (74b), where the preverbal element is extracted and fronted, and the object clitic *= aš* must be fronted with it.<sup>21</sup>

---

<sup>21</sup> Here one might want to object that the verb is not *kon* ‘do’, but *bâz* ‘open’. Indeed, deciding which part of the complex predicate *bâz kardan* [open do] ‘to open’ is the verb requires that we say something in addition, but this is beyond the scope of the present discussion.



(74) Persian (Indo-European, Indo-Iranic; Samvelian & Tseng 2010: 221)

- a. *Mi-xâh-i fardâ bâz = aš bo-kon-i.*  
 IMPF-want-2SG tomorrow open = 3SG.OBJ SBJV-do-2SG  
 ‘You want to open it tomorrow.’
- b. *Bâz = aš agar mi-xâh-i fardâ bo-kon-i..*  
 open = 3SG.OBJ if IMPF-want-2SG tomorrow SBJV-do-2SG  
 ‘If you want to open it tomorrow...’

Another question is how we treat person indexes that occur both on nouns and verbs. For example, van Gijn & Zúñiga (2014: 154) note that “many languages of the Americas, (part of) the verbal person markers are isomorphic with the nominal (possessive) person markers. For some language analysts, this is reason to regard them as clitics”, and they give the example in (75).

(75) Plains Cree (Algic, Algonquian-Blackfoot; Wolfart 1996: 412, 420)

- a. *ni-wâpam-â-w*  
 1-see-DIR-3  
 ‘I see him.’
- b. *ni-sîsîp-im*  
 1-duck-POSS  
 ‘my duck’

If the argument indexes and the person indexes were thought to be the same affixes (i.e. to have the same meaning), then they would indeed be clitics. But if they are semantically different (which seems more reasonable to say), then they are homophonous sets of affixes.

A reviewer observes that derivational affixes in European languages sometimes occur on bases of different classes, e.g. English *dis-* in *dis-honest* (adjective), *dis-order* (noun), and *dis-avow* (verb). Again, if these were treated as having the same meaning, they would be clitics. This would be an unintuitive result, but such unexpected effects at the fringes are often unavoidable.

## 7. Phonological “deficiency” and “dependence”

As I noted in Section 5, it is often said in the literature that clitics are “phonologically deficient”, or “prosodically dependent” on an adjacent form, and introductory works typically mention the etymology of the original Greek term *enklitikón* (‘element that leans on another element’). However, there seem to be no proposals for characterizing the phonological properties of clitics in such a way that they can be applied uniformly to all languages, and for this reason, phonological properties play no role in the definition in (1). It should also be noted that the notion *bound* that is part of the definition is a purely syntactic notion; there is no such thing as *phonological boundness*. In this section, I discuss a number of ways in which “deficiency” or “dependence” have been characterized, and I explain briefly why these notions are not suitable for defining clitics in a general way.

### 7.1. Unpronounceability

Clitics do not have independent stress, and it is for this reason that they are typically thought to be in need of a “host” (Bonet 2019):

One consequence of clitics being prosodically defective is that they cannot be the sole element of an utterance, for instance as an answer to some question; they need to always appear with a host.

But what exactly does “prosodically deficient (or defective)” mean? If it just means that a clitic “cannot be the sole element of an utterance” (= that it is a bound form), then Bonet’s statement is tautological.

Now it is sometimes suggested that such elements are “unpronounceable” by themselves: “In order to be pronounced, a formative (word, affix, etc.) needs to be part of an accentual unit” (Halpern 1998: 101). However, stress (or accent) is an abstract property that is very much dependent on the surrounding material. If a monosyllabic morph is an utterance by itself (e.g. English *here!*, or *yes*), the question of stress does not arise because a syllable can be unstressed only in relation to an adjacent syllable. “Lack of stress” is thus not a property that can lead to unpronounceability.

We sometimes observe that clitics are deficient in that they are subminimal, i.e. they have fewer segments or moras than a minimal free form needs to have. Some languages have nonsyllabic clitics, as seen in many cases above (e.g. Amharic =*mm* in (24), English 's, Italian *l'* in (15b), Tagalog =*ng* in (41a)), and some clitics have reduced or short vowels that are not sufficient for minimal free forms. Clitics with schwa [ə] were seen above in (10) (German *de, se*). These clitics could be said to be “unpronounceable in isolation”, but most clitics are not of this sort – clitics like those in (76) are phonologically perfectly complete and just happen to require the syntactic cooccurrence with another adjacent form.

(76) Russian:	<i>li</i>	polar question marker (ex. (2))
Greek:	<i>mas</i>	1st person plural adposessive index (ex. (57b))
Tagalog:	<i>siya</i>	3rd person pronoun (ex. (5))
German:	<i>als</i>	temporal subordinator (ex. (48a))

## 7.2. Stresslessness

Clitics are typically outside the stress domain of their anchor word (Section 4.2) and stressless, and they are sometimes defined as elements that lack stress (e.g. by Dixon 2007; see (64)). But we saw earlier that stress-integrated clitics may carry stress (e.g. Polish *nie wiem* ‘I do not know in Section 4.2), so some authors have added the specification that clitics do not carry *independent* stress. However, even that is not entirely true, as some languages have clitics that are inherently stressed, e.g. English *tóo* and German admonitive *já* (see also Section 8):

(77) *He found the house wonderful, and she liked it, tóo.*

(78) *Komm já rechtzeitig nach Hause!*  
 come ADM in.time to home  
 ‘(I admonish you that) you come home in time!’

Such stressed particles are not usually called *clitics* in the literature, but they fall under the definition in (1) as they are bound morphs and are not class-selective. They are focus and discourse particles and thus fall in the semantic range of forms that are often clitics. Similar reports of stressed clitics are occasionally found in the literature,

e.g. Lowe (2014) on accented clitics in Vedic Sanskrit (san; Indo-European, Indo-Iranic), and Aissen (2017) on a stressed deictic clitic in Tsotsil (tzo; Mayan, Core Mayan).

It should also be noted that not all languages have word stress or word accent, and that it is not even clear how to define *stress* in such a way that the notion can be applied to all languages (Hyman 2014). Stress or stresslessness is thus not suitable as a general criterion for identifying clitics.

### **7.3. Phonological wordhood**

Clitics are often discussed in the context of phonological words (or “prosodic words”; see Hall 1999; Hildebrandt 2015), and it is often said that clitics “do not constitute independent prosodic words, and lean on adjacent lexical heads to form prosodic words” (Elordieta 2014: 19; see also Aikhenvald et al. 2020: 12). The term *host* is usually used for the adjacent form on which a clitic *leans*.<sup>22</sup>

However, there is a wide range of criteria that have been used to identify phonological words, and it has been found that these criteria do not always give the same results even within one language. For example, stress domains and vowel harmony domains are different in Turkish (see (54) above), and in German, the criteria of coordination deletability and vocalic minimality conflict in the case of the diminutive suffix *-chen* (Hall 1999: 18). More generally, languages often show conflicting criteria for phonological domains and wordhood (Bickel et al. 2009; Tallman 2020). It should also be noted that phonological words are widely agreed to be partially isomorphic with morphosyntactic words, but in the absence of any kind of agreement on how to identify morphosyntactic words, it is quite impossible to identify phonological words in general.<sup>23</sup> Thus, the notion of phonological wordhood cannot be said to be well-established, despite its relative popularity since the 1990s.

---

<sup>22</sup> The term *host* was first used by Zwicky (1977). Recall from Section 3.1 above that it is not always clear which form a clitic is associated with prosodically, so in this paper, I use the term *anchor* for the word that is adjacent to a proclitic or an enclitic.

<sup>23</sup> This problem is briefly mentioned by Newell et al. (2017: 2), though without drawing any conclusions from it: “Phonologists can [...] give us some information about word domain. Phonology per se, however, lacks a theory of how the ‘word’ comes to be, and phonologists generally look to morphologists or syntacticians to derive this construct. The latter two groups, however, don’t know, and are often content with the fact that phonologists, at least, can tell them that something is a word, when it is.”

Moreover, even if it were clear how we identify phonological words, this would not be sufficient to identify clitics, because clitics could be related to phonological words in three different ways: (i) they could be integrated into the same phonological word ( $\omega$ ) as their host; (ii) they could form a recursive phonological word that also includes the host as an embedded phonological word; or (iii) they could be adjoined to their host and form a phonological phrase ( $\varphi$ ) with it. These three possibilities are illustrated in (79)-(81). It should be noted that these are hypothetical descriptions, and there is no consensus in the literature about any of these analyses.

(79) Integration: Dutch (Indo-European, Germanic; Booij 1996: 219)

*Jan kocht het boek.*  
 (jan) $_{\omega}$  (kɔx-tət) $_{\omega}$  (buk) $_{\omega}$   
 Jan bought the book  
 'Jan bought the book.'

(80) Recursive phonological word: English (Indo-European, Germanic; Selkirk 1995: 458) (see also fn. 3)

*need 'm*  
 ((nid) $_{\omega}$  əm) $_{\omega}$   
 'need him'

(81) Adjoined to host, forming a phonological phrase: Spanish (Indo-European, Italic; Elordieta 2014: 31)<sup>24</sup>

*leyé-ndo-te-la*  
 ((lejendo) $_{\omega}$  tela) $_{\varphi}$   
 read-GER-2SG-3SG.F  
 'reading it to you'

Quite similar options exist for affixes, which may be COHERING (integrated) or NON-COHERING (adjoined; e.g. Dixon 2020; Raffelsiefen 2023), so phonological wordhood does not seem to be helpful in distinguishing clitics from affixes. Moreover, the motivations for these prosodic analyses are very diverse and cannot be easily

---

<sup>24</sup> We will see in Section 9 below that the Spanish object indexes are affixes, not clitics, but in the literature on prosodic domains, they have often been treated as clitics.

generalized across languages. Some authors have highlighted the special situation of “backward-leaning” proclitics like Dutch *het* in *het boek* ‘the book’ in (79),<sup>25</sup> and such elements have even been called “enclitics” because of their phonological properties. However, as these phonological properties are not uniform, they cannot be the basis for the demarcation of clitics from affixes, or for the distinction between proclitics and enclitics (as we already saw in Section 3.1 for segmental effects).

Clitics and affixes are already distinguished from other forms by being neither free forms nor roots, so it appears that their phonological properties are not needed to single them out. Intuitively, many linguists feel that phonological dependence is part of the nature of clitics, but as this notion is vague and cannot be applied without many additional assumptions, it is better to rely on boundness and on the distinction between roots and nonroots.<sup>26</sup>

## 8. Clitics as concrete forms (morphs)

In the definition proposed here, a clitic is a morph, i.e. a concrete continuous segmental form. This means that there can be no *zero clitics*, that clitics cannot be tonal or otherwise non-segmental (Section 8.1), and that morphs cannot occur inside other morphs, so that there can be no *intraclitics* or *endoclitics* (Section 8.2).

### 8.1. There are no “tonal clitics” or “accentual clitics”

While suprasegmental effects such as stress and tone often show similarities with grammatical forms, they are not forms themselves. Forms are continuous segment sequences, which excludes the possibility of “tonal morphs” (Haspelmath 2020: Section 4). This also means that there can be no tonal clitics, as has occasionally been suggested (e.g. Van de Velde 2009). There cannot be accentual clitics either, as was sometimes discussed for Tongan (ton; Austronesian, Malayo-Polynesian; e.g. Anderson 2005: 94–101). In this language, definiteness is marked by a stress shift to

---

<sup>25</sup> See Cysouw (2005) on “ditropic clitics” (“backward-leaning” proclitics and “forward-leaning” enclitics), and Anderson (2005: Section 2.2) on “backward-leaning” proclitic determiners in Kwakwala (kwk; Wakashan, Northern Wakashan).

<sup>26</sup> The absence of phonological definitional criteria makes it possible to state generalizations about the phonological properties of clitics, e.g. that they are overwhelmingly stressless. This is not so by definition, but it can be treated as a testable empirical claim.

the final mora of the nominal, regardless of which word occurs at the end. This accentual marking shares the property of nonselectivity with clitics, but it cannot be a “processual special clitic” (Anderson 2005: 95) if a clitic is defined as a kind of morph.<sup>27</sup> Similarly, saying that “stress is a proclitic” in Modern Greek (van Oostendorp 2012) is not compatible with the definition of a clitic as a kind of morph.

## 8.2. There are no “intraclitics” or “endoclitics”

Some authors have suggested that languages may have *intraclitics*, i.e. clitics that occur between two morphs of a word-form, or *endoclitics*, i.e. clitics that occur inside a root, just as infixes are often thought of as affixes that occur inside a root. However, if we adopt the definition of *clitic* in (1) and the definition of *affix* in Haspelmath (2021), this is not possible. An affix cannot occur outside of a clitic (because affixes by definition occur next to roots or affixes), and a root cannot be “broken up” by an “infix” or an “endoclitic” (because roots by definition are segment sequences).

An example of a Russian intraclitic might be the preposition *v* = in the reciprocal construction in (82), and an example of an Andi intraclitic might be the additive marker =*lo* in (83). The supposed morph-internal status is shown by the angle brackets in *-do⟨lo⟩sub*.

(82) Russian (Indo-European, Balto-Slavic; Arkadiev 2016: 331)

*Oni razočarovalis' drug v drug-e.*  
 they were.disappointed each in other-LOC  
 ‘They were disappointed by each other.’

(83) Andi (Nakh-Daghestanian, Daghestanian; Maisak 2021: 21)

*Men ru-⟨lo⟩sub, qwar-⟨lo⟩sub.*  
 you say-PROH⟨ADD⟩ write-PROH⟨ADD⟩  
 ‘Neither talk, nor write!’

---

<sup>27</sup> One could imagine a definition of clitic that includes tonal or processual or other kinds of abstract elements, but I would not know how to do this. Nonsegmental effects are based on certain correspondences between two forms, which are often treated as operations or transformations, and they cannot be readily compared with segmental forms.

Arkadiev (2016) and Maisak (2021) regard these forms as *intraclitics*,<sup>28</sup> but this is not possible in the current conceptual framework. If the preposition *v* ‘in’ is not a prefix but a proclitic (Arkadiev 2016: 327), then the first element *drug* in (82) cannot be a prefix or compound member, but must be something else. Affixes cannot stand outside a clitic, and compounds cannot have a clitic inside them either, by definition. This means that the Andi additive marker *=lo* is an enclitic, and so is the element *=s:ub* that follows it in (83). The gloss is thus [say-PROH1 =ADD= PROH2] rather than [say-PROH<ADD>], i.e. there are two prohibitive markers (a suffix preceding the enclitic *=lo*, and an enclitic following the enclitic).

The best-known case of a supposed root-internal endoclititic has been reported from Udi (another Nakh-Daghestanian language), for which Harris (2000, 2002) provides extensive documentation and discussion. She regards bound person forms like *=z=* in (84) as endoclititics because she treats verbs like *a-...-qʔ-* ‘receive’ as single roots (*aqʔ-*).

(84) Udi (Nakh-Daghestanian, Daghestanian; Harris 2000: 598)

<i>Kayuz-ax</i>	<i>a-z-qʔ-e.</i>
letter-ACC	receive1-1SG-receive2-AOR
‘I received the letter.’	

What is surely unusual about Udi is that there are quite a few short bipartite verbs like *a-qʔ-* ‘receive’ (Harris lists 27 such verbs), but they must be treated as consisting of two different morphs, and thus somewhat analogous to English bipartite verbs like *take part* or *make headway*. The proper gloss of (84) is thus [receive1 = 1SG = receive2-AOR2], showing that the person index is a clitic, preceded by the first morph of the verbal expression and followed by its second morph.

When I say here that the Udi *endoclititics* are not clitics, I am not making a substantive claim. According to the definition in (1), these elements cannot be clitics, because a clitic is a type of form, and root-internal segment sequences cannot be forms. As I noted, this follows from the definition of a FORM (Haspelmath 2020: Section 4), and of a ROOT: a form is a sequence of segments that has a linguistic

---

<sup>28</sup> They actually call them “endoclititics” and do not make the distinction between *intraclitics* (between two morphs) and *endoclititics* (inside a root) that I take from Plungian (2000).



function,<sup>29</sup> and a root is a minimal form that denotes an object, an action or property. While one can imagine that a root could be “broken up” by some material that occurs “inside” it, this would not be in line with the definition of a root as a kind of minimal form. “Breaking up” a root is an abstract operation, similar to operations “deleting” or “transforming” forms, or movement operations, or zero elements. Such abstract operations and elements are often useful for language-particular analysis, but they cannot be used in comparative concepts.<sup>30</sup> For example, if we were to relax the definition of a form as a sequence of segments, then all kinds of non-continuous sets of elements could be said to constitute a single form (e.g. English *take ... part*). The continuity requirement is thus crucial and fundamental to our general concepts of grammar.

### 9. Romance object “clitics” as affixes

So far in this paper, I have hardly touched upon object indexes in the Romance languages, even though these kinds of elements are more prominent in the literature on *clitics* than any other type. I left the discussion of these forms until the end because they are not clitics, but affixes. Consider examples such as (85)-(88) (examples from Spanish were cited above in (8) and (81)).

(85) French (Indo-European, Italic)

*Mon frère la connaît.*  
 my brother her knows  
 ‘My brother knows her.’

(86) Italian (Indo-European, Italic; Monachesi 2005: 55)

*Martina te lo spedirà.*  
 Martina you.DAT it.ACC send.FUT.3SG  
 ‘Martina will send it to you.’

---

<sup>29</sup> Since a form is a sequence of segments, “circumfixes” and “circumclitics” cannot be types of forms. One may talk about a “circumfixing construction” (one that includes a prefix and a suffix), or about a “circumcliticizing construction” (one with a proclitic and an enclitic occurring simultaneously), but these constructions must contain two forms.

<sup>30</sup> The reason for this is that comparative concepts must be defined in the same way in all languages (as noted in fn. 27). This is generally impossible for abstract operations. Languages can be readily compared in terms of their forms, but not in terms of their abstract operations and elements.

(87) (European) Portuguese (Indo-European, Italic; Luís & Kaiser 2016: 215, 217)

- a. *Ontem chamou-me.*  
yesterday she.called-me  
'Yesterday she called me.'
- b. *Porque me chamou?*  
why me she.called  
'Why did she call me?'

(88) Romanian (Indo-European, Italic; Monachesi 2005: 44)

- Mihai nu-l așteaptă.*  
Mihai not-him waits  
'Mihai doesn't wait for him.'

These elements are not clitics according to the definition in (1) because they are bound forms that always occur on the verb, whether preverbally or postverbally.<sup>31</sup> That they are affixes rather than clitics is actually fairly widely accepted in the literature (Miller & Sag 1997; Luís 2004; Monachesi 2005: Section 3.3; Bermúdez-Otero & Payne 2011).

Authors who argued for affixal status of the Romance object indexes have typically adduced Zwicky & Pullum's (1983) diagnostic symptoms, pointing out that they occur in rigid clusters, that they sometimes show arbitrary gaps and idiosyncratic phonological behaviour, and that they tend to disallow wide scope over coordination:

(89) French (Indo-European, Italic; Miller & Sag 1997: 7)

- \*Pierre les voit et écoute.* (OK: *Pierre les voit et les écoute.*)  
Pierre them sees and hears  
'Pierre sees and hears them.'

---

<sup>31</sup> It may be unexpected to see mobile bound morphs treated as affixes, but mobile affixes are not unprecedented (e.g. Bickel et al. 2007: 43; Ryan 2010; Jenks & Rose 2015). If elements which always occur on the same type of root but show some mobility were not treated as affixes, this would have to be specified in the definition of "affix", and this definition would need to become still more complex (see Haspelmath 2021: 19).

In the present context, these additional properties of object indexes play no role, because there is only one criterion of cliticness: nonselectivity. Rigid positions in clusters are of course attested in clitics (see Section 3.5), and so is idiosyncratic phonological behaviour (see Section 4.1). An arbitrary gap is also attested in the English Genitive clitic (which does not occur after plural *-s*: *the girls'(\*s) party*, Zwicky 1987), and wide scope in coordination is sometimes even attested with derivational suffixes, so it can hardly be criterial for the clitic/affix distinction.<sup>32</sup> However, by the criterion of word-class selectivity, the object indexes are affixes, so the present conclusion conforms to that reached by Miller & Sag and those following them.

Of course, the Romance object indexes derive from personal pronouns whose position was freer in earlier times, and in medieval texts, they were not always verb-adjacent. Thus, this is a clear instance of a diachronic development from clitics to affixes, and the peculiar distribution of postverbal and preverbal object indexes in European Portuguese is a remnant of this earlier clitic stage. The situation in Modern Greek is quite similar: As we saw in (71a) above, the object person indexes, which have often been called *clitics*, are actually affixes.

## 10. Conclusion: clitics are not intermediate between words and affixes

We have seen a variety of different types of clitics in this paper, as well as a variety of different properties that are found in clitics. I showed that they can all be subsumed under the simple definition in (1) (a clitic is a bound morph that is neither an affix nor a root), but I did not claim that this definition says anything deep about their nature. It is merely a definition, after all. But it is simple and clear, and it has sharp boundaries rather than merely specifying a canon or a prototype.

The definition may seem to be broader than has often been implied, e.g. by including adpositions and subordinators (Section 2.6-7), which have not often been regarded as clitics. However, it is unclear why they should be excluded, and it may be a historical accident that they did not become prominent in the literature on clitics. The result is that most function words are clitics (Section 2.1), but only those that cannot occur on their own (= that are bound forms). Closed-class function words

---

<sup>32</sup> Erdal (2007: 178) cites the following example from Turkish, where the “professional” suffix *-cı* has scope over two nouns (i.e. allows “suspended affixation”): *kum- ve çakıl-cı geldi* [sand- and gravel-PROF came] ‘the supplier of sand and gravel has come’. Nobody would suggest that it is a clitic.

such as demonstratives, auxiliaries and response words ('yes') can often be used in isolation and are therefore not clitics.

Much work on clitics over the last few decades is motivated by the hope of explaining the behaviour of (certain kinds of) clitics by appealing to certain kinds of architectures or rule types, such as lexical vs. postlexical rules (e.g. Halpern 1995; Anderson 2005). However, no particular proposal has been widely accepted, and it appears that the possibilities of an *architectural* approach have been exhausted. Prominent authors like Zwicky (1994) and Spencer & Luís (2012) have suggested that *clitic* is no more than a name for a problem: a label for a range of linguistic expressions that do not fit readily into other classes, not a name of a theoretical construct, and not a name for a "unified class of phenomena" (Zwicky 1994: xiii).

In this paper, by contrast, I do give a definition which by its nature singles out a unified class of phenomena, and this allows the term *clitic* to be more than a name for a problem: It is a comparative concept that helps us to compare languages with respect to phenomena that we find interesting without talking past each other. But since *clitic* is defined as a comparative concept, there is no claim that it carves out part of the underlying reality of languages: Like other terms for comparative concepts, it is a METHODOLOGICAL TOOL, not a "theoretical construct". To the extent that the term allows us to formulate testable claims about the world's languages, and to the extent that these claims are supported, we will have found valuable cross-linguistic generalizations, but it may still be unclear how we can explain these generalizations.<sup>33</sup>

In addition to the architectural approach, a popular view has been that clitics are in some way intermediate between free words and affixes. Zwicky (1977: 1) initially characterized them as "presenting analytic difficulties because they are neither clearly independent words nor clearly affixes", and Nevis (2000: 389) even suggested that a form is a clitic "to the extent that it deviates from the accepted properties of affixes or words".<sup>34</sup> But just as *clitic* is not more than a comparative concept with some

---

<sup>33</sup> I do not actually expect to find robust generalizations that crucially rely on the clitic vs. affix distinction, but as we need to know what a clitic is in order to distinguish words from non-words (see Haspelmath 2023), this definition is very important for all works that make claims about words. Again, it may be that the most robust generalizations will eventually be shown to involve form classes other than words, but the 'word' concept is so central to linguistics that it is good to have a clear definition of it with sharp boundaries.

<sup>34</sup> In a non-serious mode, Sadock (1995: 260) suggested that a clitic could be defined as "an element whose distribution linguists cannot comfortably consign to a single grammatical component".

usefulness for linguists, the familiar affix vs. word distinction (and the morphology vs. syntax division in grammar) could be largely based on the orthographic word. The supposed “analytic difficulties” of clitics would then reflect the difficulties of deciding how to write them (jointly, or separately, or with a hyphen or other boundary symbol). The definition of *affix* is actually much more complex than the definition of *clitic* (once a definition of *affix* is in place), as can be seen in Haspelmath (2021), and defining *word* is not straightforward either (see Haspelmath 2023). But in whatever way we end up defining these terms, the definitions are unlikely to give us deep insights into their nature.

Clitics could be “intermediate” between free words and affixes if there were a single dimension along which they vary, a kind of “scale of coalescence”, or “tightness of bonding”. It has often been suggested that there is such a continuous scale, with a diachronic counterpart in grammaticalization (Hopper & Traugott 2003: 142):

(90) the coalescence scale

free word > clitic > affix

But no systematic way of quantifying the degrees on a scale of coalescence or of tightness of bonding has been suggested, and linguists have mostly relied on their intuitions of what constitutes *tight* or *loose* attachment. In view of the great variety of phenomena that have been cited as diagnostics, we cannot conclude that there is sufficient evidence that the scale is real. Aikhenvald (2002: 42) says that applying a wide range of criteria “suggests a scalar, or continuum-type approach – that is, some morphemes turn out to be more affix-like and others to be more word-like”. But Börjars & Harries (2008) have rightly emphasized that the different dimensions of variation need not correlate with each other, and van Gijn & Zúñiga (2014: 155) make this very concrete: They examine four such dimensions (phonological integration, rigid position, syntactic weight, and lexical class) for twelve morph types from different languages, and they do not find a clear clustering of the dimensions. There is no reason to think that there is a single scale or continuum.

There are thus many open questions that need to be addressed by future research, but I hope that by providing simple and clear definitions of terms such as *affix* (Haspelmath 2021) and *clitic* (definition (1) in Section 1), this research will be facilitated. The definitions do not answer any theoretical questions, but it should have

become clear that it is possible to have such definitions even without answers to our broader questions.

## Acknowledgements

For useful comments on an earlier version of this paper, I am particularly indebted to Peter Arkadiev, Irenäus Kulik and Tim Zingler, but also to quite a few commentators on Academia.edu. In addition, I thank two anonymous reviewers and the editors.

## Abbreviations

1 = 1 <sup>st</sup> person	DUR = durative	NONFUT = non-future
2 = 2 <sup>nd</sup> person	ERG = ergative	NONMASC = non-masculine
3 = 3 <sup>rd</sup> person	EMPH = emphatic	OBJ = object
ADM = admonitive	EXCL = exclusive	PL = plural
ACC = accusative	EZ = ezafe	POSS = possessive
ADD = additive	F = feminine	PQ = polar question
ALL = allative	FOC = focus	PRF = perfect
AOR = aorist	FUT = future	PROH = prohibitive
ART = article	G4 = gender 4	PRS = present
AUX = auxiliary	GEN = genitive	PST = past
AV = actor voice	GER = gerund	PTCP = participle
COMP = complementizer	HORT = hortative	QM = question marker
COMPL = completive	IMPF = imperfective	REFL = reflexive
CONJ = conjunction	INCL = inclusive	REL = relative
CONN = connective	IND = indicative	RES = resultative
COP = copula	INS = instrumental	SBJ = subject
DAT = dative	IRR = irrealis	SBJV = subjunctive
DECL = declarative	LNK = linker	SG = singular
DEF = definite	LOC = locative	TOP = topic
DET = determiner	M = masculine	TR = transitive
DIR = direct Form	NEG = negation	
DU = dual	NOM = nominative	

## References

- Aikhenvald, Alexandra Y. 2002. Typological parameters of clitics, with special reference to Tariana. In R. M. W. Dixon & Alexandra Y. Aikhenvald (eds.), *Word: A cross-linguistic typology*, 42–78. Cambridge: Cambridge University Press.
- Aikhenvald, Alexandra Y., R. M. W. Dixon & Nathan M. White. 2020. The essence of “word”. In Alexandra Y. Aikhenvald, R. M. W. Dixon & Nathan M. White (eds.), *Phonological word and grammatical word: A cross-linguistic typology*, 1–24. Oxford: Oxford University Press.
- Aissen, Judith. 2017. Special clitics and the right periphery in Tsotsil. In Claire Bowerman, Laurence Horn & Raffaella Zanuttini (eds.), *On looking into words (and beyond): Structures, relations, analyses*, 235–262. Berlin: Language Science Press. (DOI:10.5281/zenodo.495449)
- Allison, Sean. 2020. The notion of “word” in Makary Kotoko. In Alexandra Y. Aikhenvald, R. M. W. Dixon & Nathan M. White (eds.), *Phonological word and grammatical word: A cross-linguistic typology*, 260–284. Oxford: Oxford University Press.
- Anderson, Stephen R. 1992. *A-morphous morphology*. Cambridge: Cambridge University Press.
- Anderson, Stephen R. 2005. *Aspects of the theory of clitics*. New York: Oxford University Press.
- Anderson, Stephen R. 2011. Clitics. In Marc Van Oostendorp, Colin J. Ewen, Elizabeth Hume & Keren Rice (eds.), *The Blackwell companion to phonology*. Wiley Online Library: Wiley. (<https://doi.org/10.1002/9781444335262.wbctp0084>)
- Arcodia, Giorgio F. 2012. *Lexical derivation in Mandarin Chinese*. Taipei: Crane.
- Arkadiev, Peter M. & Yu A. Lander. 2020. Ambifiksy i drugie zveri [Амбификсы и другие звери] (Ambifixes and other beasts). In Andrej A. Kibrik, Ks. P. Semënova, D.V. Sičinava, Sergei G. Tatevosov, A. Ju. Urmančieva (eds.), *VAProsy jazykoznanija: Megasbornik nanostatej*, 35–42. Moskva: Buki Vedi.
- Arkadiev, Peter M. 2016. K voprosu ob èndoklitikax v russkom jazyke [On endoclitics in Russian]. In Anton V. Zimmerling & Ekaterina A. Ljutikova (eds.), *Arxitektura klauzy v parametričeskix modeljax: Sintaksis, informacionnaja struktura, porjadok slov*, 325–331. Moskva: Jazyki slavjanskix kul'tur.
- Avgustinova, Tania. 1994. On Bulgarian verbal clitics. *Journal of Slavic Linguistics* 2(1). 29–47.

- Berghäll, Liisa. 2015. *A grammar of Mauwake*. Berlin: Language Science Press. (<http://langsci-press.org/catalog/book/67>)
- Bermúdez-Otero, Ricardo & John Payne. 2011. There are no special clitics. In Alexandra Galani, Glyn Hicks & George Tsoulas (eds.), *Morphology and its interfaces*, 57–69. Amsterdam: John Benjamins.
- Bickel, Balthasar, Goma Banjade, Martin Gaenszle, Elena Lieven, Netra P. Paudyal, Ichchha P. Rai, Manoj Rai, Novel Kishore Rai & Sabine Stoll. 2007. Free prefix ordering in Chintang. *Language* 83(1). 43–73.
- Bickel, Balthasar, Kristine Hildebrandt & René Schiering. 2009. The distribution of phonological word domains: A probabilistic typology. In Janet Grijzenhout & Barış Kabak (eds.), *Phonological domains: Universals and deviations*, 47–75. Berlin: Mouton de Gruyter.
- Bloomfield, Leonard. 1933. *Language*. New York: H. Holt and Company.
- Bonet, Eulalia. 2019. Clitics and clitic clusters in morphology. *Oxford Research Encyclopedia of Linguistics*. (doi:10.1093/acrefore/9780199384655.013.519)
- Booij, Geert. 1996. Cliticization as prosodic integration: The case of Dutch. *The Linguistic Review* 13(3–4). 219–242. (doi:10.1515/tlir.1996.13.3-4.219)
- Booij, Geert E. 2012. *The grammar of words: An introduction to linguistic morphology (3rd edition)*. Oxford: Oxford University Press.
- Börjars, Kersti & Pauline Harries. 2008. The clitic-affix distinction, historical change, and Scandinavian bound definiteness marking. *Journal of Germanic Linguistics* 20(4). 289–350. (doi:10.1017/S1470542708000068)
- Bošković, Željko. 2016. Second-position clitics cross-linguistically. In Franc Lanko Marušič & Rok Žaucer (eds.), *Formal studies in Slovenian Syntax: In honor of Janez Orešnik*, 23–53. Amsterdam: John Benjamins.
- Boye, Kasper & Peter Harder. 2012. A usage-based theory of grammatical status and grammaticalization. *Language* 88(1). 1–44.
- Chae, Hee-Rahk. 2020. *Korean morphosyntax: Focusing on clitics and their roles in syntax*. Abingdon: Routledge.
- Croft, William. 2010. Ten unwarranted assumptions in syntactic argumentation. In Kasper Boye, & Elisabeth Engberg-Pedersen (eds.), *Language usage and language structure*, 313–350. Berlin: Mouton de Gruyter.
- Cysouw, Michael. 2005. Morphology in the wrong place: A survey of preposed enclitics. In Wolfgang U. Dressler, Dieter Kastovsky, Oskar E. Pfeiffer & Rainer,



- Franz (eds.), *Morphology and its demarcations: Selected papers from the 11th Morphology Meeting, Vienna, February 2004*, 17–37. Amsterdam: John Benjamins.
- Demeke, Girma A. & Ronny Meyer. 2008. The enclitic *-mm* in Amharic: reassessment of a multifunctional morpheme. *Linguistics* 46(3). 607–628. (doi:10.1515/LING.2008.020)
- Dixon, R. M. W. 2007. Clitics in English. *English Studies* 88(5). 574–600. (doi:10.1080/00138380701566102)
- Dixon, R. M. W. 2020. Words within words: Examples from Yidiñ, Jarawara, and Fijian. In Alexandra Y. Aikhenvald, R. M. W. Dixon & Nathan M. White (eds.), *Phonological word and grammatical word: A cross-linguistic typology*, 25–38. Oxford: Oxford University Press.
- Dryer, Matthew S. 2005. Polar questions. In Martin Haspelmath, Matthew S. Dryer, David Gil & Bernard Comrie (eds.), *The world atlas of language structures*, 470–473. Oxford: Oxford University Press.
- Elordieta, Gorra. 2014. The word in phonology. In Iraide Ibarretxe-Antuñano & José-Luis Mendivil-Giró (eds.), *To be or not to be a word: New reflections on the definition of word*, 6–65. Cambridge: Cambridge Scholars Publishing.
- Erdal, Marcel. 2000. Clitics in Turkish. In Aslı Göksel & Celia Kerslake (eds.), *Studies on Turkish and Turkic Languages: Proceedings of the Ninth International Conference on Turkish Linguistics, Lincoln College, Oxford, August 12-14, 1998*, 41–48. Wiesbaden: Harrassowitz.
- Erdal, Marcel. 2007. Group inflexion, morphological ellipsis, affix suspension, clitic sharing. In M.M. Jocelyne Fernandez-Vest (ed.), *Combats pour les langues du monde: Hommage à Claude Hagège*, 177–189. Paris: L'Harmattan.
- Fattier, Dominique. 2013. Haitian Creole structure dataset. In Susanne Maria Michaelis, Philippe Maurer, Martin Haspelmath & Magnus Huber (eds.), *Atlas of Pidgin and Creole Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (<https://apics-online.info/contributions/49>)
- Fehn, Anne-Maria. 2016. *A grammar of Ts'ixa (Kalahari Khoe)*. Köln: Universität zu Köln. (Doctoral Dissertation.) (<http://kups.ub.uni-koeln.de/7062/>)
- Fortescue, Michael. 1984. *West Greenlandic*. London: Croom Helm.
- Gary, Judith Olmsted & Saad Gamal-Eldin. 1982. *Cairene Egyptian Colloquial Arabic*. Amsterdam: North Holland.
- Gerds, Donna B. & Adam Werle. 2014. Halkomelem clitic types. *Morphology* 24(3). 245–281. (doi:10.1007/s11525-014-9241-0)

- Göksel, Aslı & Celia Kerslake. 2005. *Turkish: An essential grammar*. London: Routledge.
- Gunnink, Hilde. 2022. *A grammar of Fwe*. Berlin: Language Science Press.
- Hall, Tracy A. 1999. The phonological word: A review. In Tracy A. Hall & Ursula Kleinhenz (eds.), *Studies on the phonological word*, 1–22. Amsterdam: John Benjamins.
- Halpern, Aaron. 1995. *On the placement and morphology of clitics*. Stanford: CSLI Publications.
- Halpern, Aaron. 1998. Clitics. In Andrew Spencer & Arnold M. Zwicky (eds.), *The handbook of morphology*, 101–122. Oxford: Blackwell.
- Hannahs, S. J. & Maggie Tallerman. 2006. At the interface: Selection of the Welsh definite article. *Linguistics* 44(4). 781–816. (doi:10.1515/LING.2006.025)
- Harris, Alice C. 2000. Where in the word is the Udi clitic? *Language* 76(3). 593–616.
- Harris, Alice C. 2002. *Endoclitics and the origins of Udi morphosyntax*. Oxford: Oxford University Press.
- Haspelmath, Martin. 2013. Argument indexing: A conceptual framework for the syntax of bound person forms. In Dik Bakker & Martin Haspelmath (eds.), *Languages across boundaries: Studies in memory of Anna Siewierska*, 197–226. Berlin: Mouton de Gruyter.
- Haspelmath, Martin. 2015. Defining vs. diagnosing linguistic categories: A case study of clitic phenomena. In Joanna Błaszczak, Dorota Klimek-Jankowska & Krzysztof Migdalski (eds.), *How categorical are categories? New approaches to the old questions of noun, verb, and adjective*, 273–304. Berlin: Mouton de Gruyter.
- Haspelmath, Martin. 2019. Indexing and flagging, and head and dependent marking. *Te Reo* 62(1). 93–115. (doi:10.17617/2.3168042)
- Haspelmath, Martin. 2020. The morph as a minimal linguistic form. *Morphology* 30(2). 117–134. (doi:10.1007/s11525-020-09355-5)
- Haspelmath, Martin. 2021. Bound forms, welded forms, and affixes: Basic concepts for morphological comparison. *Voprosy Jazykoznanija* 2021(1). 7–28. (doi:10.31857/0373-658X.2021.1.7-28)
- Haspelmath, Martin. 2023. Defining the word. *WORD* 69(3). 283–297. (doi:10.1080/00437956.2023.2237272)
- Hildebrandt, Kristine A. 2015. The prosodic word. In John R. Taylor (ed.), *The Oxford handbook of the word*, 221–245. Oxford: Oxford University Press.
- Himmelman, Nikolaus P. 2014. Asymmetries in the prosodic phrasing of function words: Another look at the suffixing preference. *Language* 90(4). 927–960.

- Hopper, Paul J. & Elizabeth Closs Traugott. 2003. *Grammaticalization: 2nd edition*. Cambridge: Cambridge University Press.
- Hosokawa, Komei. 1991. *The Yawuru Language of West Kimberley. A Meaning-Based Description*. Canberra: Australian National University.
- Hyman, Larry M. 2014. Do all languages have word accent? In Harry Van der Hulst (ed.), *Word Stress: Theoretical and typological issues*, 56–82. Cambridge: Cambridge University Press.
- Ionova, A. 2019. *The unbearable lightness of clitics*. Leiden: Leiden University-LOT. (Doctoral Dissertation.) (<https://hdl.handle.net/1887/83258>)
- Jenks, Peter & Rose, Sharon. 2015. Mobile object markers in Moro: The role of tone. *Language* 91(2). 269–307. (doi:10.1353/lan.2015.0022)
- Joseph, Brian D. 2002. Defining “word” in Modern Greek: A response to Philippaki-Warbuton & Spyropoulos 1999. In Geert Booij & Jaap van Marle (eds.), *Yearbook of Morphology 2001*, 87–114. Dordrecht: Springer Netherlands. (doi:10.1007/978-94-017-3726-5\_3)
- Kaufman, Daniel. 2010. *The morphosyntax of Tagalog clitics: a typologically driven approach*. Ithaca: Cornell University. (Doctoral Dissertation.)
- Kayne, Richard S. 1975. *French syntax*. Cambridge: MIT Press.
- Kazenin, Konstantin I. 2002. Focus in Daghestanian and word order typology. *Linguistic Typology* 6(3). 289–316. (doi:10.1515/lity.2003.002)
- Klavans, Judith L. 1985. The independence of syntax and phonology in cliticization. *Language* 61(1). 95–120. (doi:10.2307/413422)
- Konnerth, Linda. 2020. *A grammar of Karbi*. Berlin: Mouton de Gruyter.
- Lowe, John. 2014. Accented clitics in the Ṛgveda. *Transactions of the Philological Society* 112(1). 5–43. (doi:10.1111/1467-968X.12013)
- Luís, Ana R. 2004. *Clitics as morphology*. Colchester: University of Essex. (Doctoral Dissertation.)
- Luís, Ana R. & Georg A. Kaiser. 2016. Clitic pronouns: Phonology, morphology, and syntax. In W. Leo Wetzels, João Costa & Sergio Menuzzi (eds.), *The Handbook of Portuguese Linguistics*, 210–233. Wiley Online Library: Wiley. (doi:10.1002/9781118791844.ch12)
- Maisak, Timur. 2021. Endoclitics in Andi. *Folia Linguistica* 55(1). 1–34. (doi:10.1515/flin-2020-2069)

- Miller, Philip H. & Ivan A. Sag. 1997. French clitic movement without clitics or movement. *Natural Language & Linguistic Theory* 15(3). 573–639. (doi:10.1023/A:1005815413834)
- Monachesi, Paola. 2005. *The verbal complex in Romance: A case study in grammatical interfaces*. Oxford: Oxford University Press.
- Mushin, Ilana. 2012. *A grammar of Garrwa*. Berlin: Mouton de Gruyter.
- Nakamura, Wataru. 2018. Case. In Yōko Hasegawa (ed.), *The Cambridge handbook of Japanese linguistics*, 249–275. Cambridge: Cambridge University Press.
- Nevis, Joel A. 2000. Clitics. In Geert E. Booij, Christian Lehmann & Joachim Mugdan (eds.), *Morphology: An international handbook on inflection and word-formation (Volume 1)*, 388–404. Berlin: Mouton de Gruyter. (<https://doi.org/10.1515/9783110111286.1.5.360>)
- Newell, Heather, Máire Noonan, Glyne Piggott, & Lisa deMena Travis (eds.). 2017. *The structure of words at the interfaces*. Oxford: Oxford University Press.
- Nida, Eugene A. 1946. *Morphology: The descriptive analysis of words (1st edition)*. Ann Arbor: University of Michigan Press.
- Nordhoff, Sebastian. 2009. *A grammar of Upcountry Sri Lanka Malay*. Amsterdam: University of Amsterdam-LOT. (Doctoral Dissertation.)
- Nordlinger, Rachel. 1998. *A grammar of Wambaya, Northern Australia*. Canberra: Pacific Linguistics. (<http://dx.doi.org/10.15144/PL-C140>)
- Parker, Steve. 1999. On the behavior of definite articles in Chamicuro. *Language* 75(3). 552–562. (doi:10.2307/417060)
- Pescarini, Diego. 2021. *Romance object clitics: Microvariation and linguistic change*. Oxford: Oxford University Press.
- Plank, Frans. 2005. The prosodic contribution of clitics: Focus on Latin. *Lingue e Linguaggio* 4(2). 281–292. (doi:10.1418/20726)
- Plungian, Vladimir A. 2000. *Obščaja morfologija: Vvedenie v problematiku*. Moskva: URSS.
- Raffelsiefen, Renate. 2023. Morpho-phonological asymmetries in affixation. In Peter Ackema, Sabrina Bendjaballah, Eulàlia Bonet & Antonio Fábregas (eds.), *The Wiley Blackwell companion to morphology*. Wiley Online Library: Wiley. (<https://doi.org/10.1002/9781119693604.morphcom050>)
- Rubach, Jerzy & Booij, Geert E. 1985. A grid theory of stress in Polish. *Lingua* 66(4). 281–320. (doi:10.1016/0024-3841(85)90032-4)

- Ryan, Kevin M. 2010. Variable affix order: Grammar and learning. *Language* 86(4). 758–791. (doi:10.1353/lan.2010.0032)
- Sadock, Jerrold M. 1995. A multi-hierarchy view of clitics. *Chicago Linguistic Society* 31(2). 258–279.
- Samvelian, Pollet & Jesse Tseng. 2010. Persian object clitics and the syntax-morphology interface. In Stefan Müller (ed.), *Proceedings of the 17th International Conference on Head-Driven Phrase Structure Grammar*. Université Paris Diderot, Paris 7, France, 212–232. Stanford: CSLI.
- Samvelian, Pollet. 2007. A (phrasal) affix analysis of the Persian Ezafe. *Journal of Linguistics* 43(3). 605–645.
- Schapper, Antoinette. 2022. *A grammar of Bunaq*. Berlin: Mouton de Gruyter.
- Selkirk, Elisabeth O. 1995. The prosodic structure of function words. In Jill N. Beckman, Laura Walsh Dickey & Suzanne Urbanczyk (eds.), *Papers in Optimality Theory*, 439–469. Amherst: Graduate Linguistic Student Association (GLSA).
- Smith, Ian & Steve Johnson. 1985. The syntax of clitic cross-referencing pronouns in Kugu Nganhcara. *Anthropological Linguistics* 27(1). 102–111.
- Spencer, Andrew & Ana R. Luís. 2012. *Clitics*. Cambridge: Cambridge University Press.
- Stebbins, Tonya. 2003. On the status of intermediate form classes: Words, clitics, and affixes in Smalgyax (Coast Tsimshian). *Linguistic Typology* 7(3). 383–415. (doi:10.1515/lity.2003.019)
- Stockwell, Robert P., Jean Donald Bowen & John Watson Martin. 1965. *The grammatical structures of English and Spanish*. Chicago: University of Chicago Press.
- Tallman, Adam J. R. 2020. Beyond grammatical and phonological words. *Language and Linguistics Compass* 14(2). e12364. (doi:10.1111/lnc3.12364)
- Taylor, John R. 1995. *Linguistic categorization: Prototypes in linguistic theory (2nd edition)*. Oxford: Clarendon Press.
- Toman, Jindřich. 1996. A note on clitics and prosody. In Aaron Halpern & Arnold M. Zwicky (eds.), *Approaching second: Second position clitics and related phenomena*, 505–510. Stanford: CSLI Publications.
- Van de Velde, Mark. 2009. Eton tonology and morphosyntax: A holistic typological approach. In Patience L. Epps & Alexandre Arkhipov (eds.), *New challenges in typology: Transcending the borders and refining the distinctions*, 35–60. Berlin: Mouton de Gruyter.
- van Gijn, Rik & Fernando Zúñiga. 2014. Word and the Americanist perspective. *Morphology* 24(3). 135–160. (doi:10.1007/s11525-014-9242-z)

- van Oostendorp, Marc. 2012. Stress as a proclitic in Modern Greek. *Lingua* 122(11). 1165–1181. (doi:10.1016/j.lingua.2012.05.006)
- Wackernagel, Jacob. 2020 [1892]. *On a law of Indo-European word order*. Berlin: Language Science Press. (Ed. Walkden, George & Sevdali, Christina & Macleod, Morgan.) (doi:10.5281/zenodo.3978908)
- Walkden, George. 2020. Introduction. In George Walkden, Christina Sevdali & Morgan Macleod (eds.), *Jacon Wackernagel's "On a law of Indo-European word order,"* 3–19. Berlin: Language Science Press. (doi:10.5281/zenodo.3978908)
- Wolfart, H. Christoph. 1996. Sketch of Cree, an Algonquian Language. In Ives Goddard (ed.), *Handbook of North American Indians, Vol. 17: Languages*. Washington: Smithsonian Institute.
- Zingler, Tim. 2020. *Wordhood issues: Typology and grammaticalization*. Albuquerque: University of New Mexico. (Doctoral Dissertation.) ([https://digitalrepository.unm.edu/ling\\_etds/71](https://digitalrepository.unm.edu/ling_etds/71))
- Zingler, Tim. 2022. Clitics, anti-clitics, and weak words: Towards a typology of prosodic and syntagmatic dependence. *Language and Linguistics Compass* 16(5–6). e12453. (doi:10.1111/lnc3.12453)
- Zúñiga, Fernando. 2014. (Anti-)cliticization in Mapudungun. *Morphology* 24(3). 161–175. (doi:10.1007/s11525-014-9244-x)
- Zwicky, Arnold M. & Geoffrey K. Pullum. 1983. Cliticization vs. inflection: English n't. *Language* 59(3). 502–513.
- Zwicky, Arnold M. 1977. *On clitics*. Bloomington: Indiana University Linguistics Club. (doi:10.5281/zenodo.7436775)
- Zwicky, Arnold M. 1987. Suppressing the Zs. *Journal of Linguistics* 23(1). 133–148.
- Zwicky, Arnold M. 1994. What is a clitic? In Joel A. Nevis, Brian D. Joseph, Dieter Wanner & Arnold M. Zwicky (eds.), *Clitics: a comprehensive bibliography 1892-1991*, xii–xx. Amsterdam: John Benjamins.

#### CONTACT

[martin\\_haspelmath@eva.mpg.de](mailto:martin_haspelmath@eva.mpg.de)

# Spanish as an argument-indexing language. A view from the analysis of Colombian Andean Spanish

SERGIO IBÁÑEZ CERDA<sup>1</sup> & ARMANDO MORA BUSTOS<sup>2</sup> &  
ALEJANDRA I. ORTIZ VILLEGAS<sup>3</sup>

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO<sup>1</sup>, UNIVERSIDAD AUTÓNOMA METROPOLITANA  
UNIDAD IZTAPALAPA<sup>2</sup>, INSTITUTO DE EDUCACIÓN MEDIA SUPERIOR CDMX<sup>3</sup>

Submitted: 25/05/2023 Revised version: 11/11/2023

Accepted: 07/12/2023 Published: 27/12/2023



Articles are published under a Creative Commons Attribution 4.0 International License (The authors remain the copyright holders and grant third parties the right to use, reproduce, and share the article).

## Abstract

Spanish is considered a dependent-marking language in which argument realization is accomplished through the coding of lexical or referential phrases (RPs). This counter-proposal suggests that it is an argument-indexing language, one where the argument realization is carried out by means of person forms or indexes attached to the verbal word. To prove this, we show that in Standard Spanish (SS) subject and indirect object RPs are not coded in most cases, and that the verb plus the indexes can function as a complete clause. To further discuss these ideas, we analyze Colombian Andean Spanish (CAS), in which DO arguments are also mostly coded through clitic person forms, so CAS has a three index system. We propose that the argument features load is coded in a distributed fashion: the indexes are the syntactic expression of arguments, while the RPs manifest their semantic and pragmatic content.

**Keywords:** Spanish argument realization system; argument indexing languages; Spanish case flag system; cross-reference languages; conominal status in Spanish.

## 1. Introduction

Spanish (spa, Indo-European, Romance) is widely considered a dependent-marking language, in terms of the typological distinction first introduced by Nichols (1986).

This supposes that Spanish is a type of language in which argument realization is accomplished through lexical referential phrases (RPs),<sup>1</sup> and that the semantic and syntactic relations between the verbal predicate and the arguments is marked on those RPs, which are the dependents in relation to the predicate. The flagging on the RPs is usually done by means of case markers or by adpositional marking (analytical case marking).

In this paper, we follow Haspelmath (2013) in setting aside the dependent-marking vs. head-marking opposition and propose that the relevant distinction to explain argument realization systems is between languages in which the argument projection is accomplished by means of lexical referential phrases (RPs) and languages in which arguments are mainly coded through the presence of person forms or indexes in the verb. We also follow Haspelmath (2013) in assuming that the realization of arguments through person forms cannot be labeled as case-marking on the head, as the indexes themselves are the arguments and not a way of flagging the arguments.

In this context, we will show that the main formal means of argument realization in Spanish is not through the RPs, but through indexes in the verbal word. That is, we propose that Spanish is an argument-indexing language.

Despite the dependent-marking tag, it is widely known that Spanish shows what is called the pro-drop parameter, i.e., clauses without a lexical subject are possible, as in (1):

- (1) *Romp-ió*      *el = vaso.*<sup>2</sup>  
break-3PST    ART = glass  
'(He/She) broke the glass.'

The typical (non-explicit) analysis (Alcina & Blecua 1975; Seco 1989; García Miguel 1991, 1995; Bogard 1992, Alarcos Llorach 1994; Bosque & Demonte 1999; Company

---

<sup>1</sup> We use the term referring or referential phrase (RP) instead of noun phrase (NP). We follow Van Valin's proposal (2008) that the basic semantic and syntactic relations in the clause are those of referring expressions (RPs) and predicating expressions (the predicate, which is the nucleus of the clause). The RP function can be fulfilled by any type of lexical categories; hence, RPs do not need to have a specific type of head; so, although in many languages it is the case that NPs typically play that role, it does not need to be so.

<sup>2</sup> Examples without a source indication are elicited directly by the authors. They are provided in the understanding that they are non-controversial and that they are standard examples of Spanish in general.



1998, 2003; Di Tullio 2005; RAE 2009; among many others) assumes that the inflection on the verb is an agreement marker, and that the subject in this type of clauses is a non-coded RP. Haspelmath (2013, 2019) calls this analysis the virtual-agreement view. Following him, we argue that this analysis is misleading, and that the person features in the inflection are the formal manifestation of the argument.

In a similar fashion, some scholars (Heger 1967; Givón 1976; Silva-Corvalán 1981; Suñer 1988; García Miguel 1991; 1995; Bogard 1992; Company 1998, 2003; Belloro 2004, 2007; Kailuweit 2008; Van Valin 2013) have proposed that the Spanish accusative and dative clitics are also agreement markers, as in (2a), in contrast with (2b), where there are full RPs functioning as the arguments:

(2)

- a. *Se = la = dio.*  
 3DAT = 3ACC = give.3PST  
 ‘(He/She) gave it to him/her.’
- b. *Rogelio dio la = noticia a = Pedro.*  
 Rogelio give.3PST ART = news PREP = Pedro  
 ‘Rogelio gave the news to Pedro.’

Here we propose that the agreement analysis is also misleading when applied to the object clitics. Likewise, we consider that the treatment of these clitics as pronouns that substitute the RPs is also incorrect. We will show that the clitic indexes are the primary means for the realization of the indirect object (IO) argument, and also of the DO argument, but the latter only in some varieties, as in the Colombian Andean Spanish.

As it has also been widely discussed, Spanish shows what is called *clitic doubling*, where a RP appears along a dative person form, as in (3a), or an accusative clitic, as in (3b):

(3)

- a. *Rogelio le = dio la = noticia a = Pedro.*  
 Rogelio 3DAT = give.3PST ART = news 3DAT = Pedro  
 ‘Rogelio gave the news to Pedro.’

- b. *Rogelio la = vio                      a    ella en la    calle.*  
 Rogelio 3ACC = see.3PST    DOM her    in the    street  
 ‘Rogelio saw her on the street.’

Interestingly, many authors (Kany 1945; Gili Gaya 1961; García Miguel 1991, 1995; Bogard 1992; Vázquez Rozas 1995; Company 1998, 2003; Belloro 2004, 2007; Di Tullio 2005; Kailuweit 2008; RAE/ASALE 2009, among others) consider that the clitics double the RPs and not the opposite.<sup>3</sup> They treat the RPs as the arguments and the person forms as agreement markers. Here, we argue that the clitic person forms and the RPs jointly are the manifestation of the verbal arguments. This analysis does not imply a double coding of the same referent, but that a single argument information is distributed and coded simultaneously through two distinct forms, the index and the RP. Following Pensalfini (2004), we propose that the indexes only project the syntactic information, while the RPs stand for the semantic and referential information of the argument, so there is not a double instantiation of the same referent in the core of the clause, which in some frameworks (e.g., RRG, Van Valin 2005) operates as an important projection restriction.

Besides establishing the primary role of the indexes in the argument realization system, we will also show that, in contrastive terms, the Spanish flagging system is much more plain or basic, in the sense that it has less overt distinctions than those allowed by the index set. Subject RPs are always unmarked, so the dependent marking features assumed to be present in Spanish are the accusative and dative prepositions that introduce the object RPs, as in (4):

(4)

- a. *La = Inés      cuid-ó      al = guagua.*  
 ART = Inés    care-3PST    DOM.ART = kid  
 ‘Ines cared for the kid.’
- b. *Gerardo    dio              la = noticia    a = Pedro.*  
 Gerardo    give.3PST    ART = news    3DAT = Pedro  
 ‘Gerardo gave the news to Pedro.’

---

<sup>3</sup> Notable exception are Bogard (1992) and García Salido (2013), which treat the lexical phrase as the copy of the clitic.

Nevertheless, as pointed out by some scholars (Torrego Salcedo 1999; Delbecque 2002; Leonetti 2004; Iemmolo 2010; Melis 2018), the accusative *a* in (4a), is a differential object marker (DOM): it marks the animacy of the reference, and it does not really flag the relation between the RP and the predicate. So, two of the three core arguments in Spanish, subject and DO, are not flagged at all. Dative marking on the IO is the only true case flag (4b).

As said before, we propose that Spanish is better explained as a language in which the argument realization is mainly accomplished through the presence of indexes attached to the verb and not through RPs. However, we consider that this structural nature is not actually absolute or uniform, since the argument realization system is not completely or uniquely based on the indexes; to a certain extent it is a mixed system.

To demonstrate what we consider to be the direction the argument realization system is heading, we analyze some aspects of Colombian Andean Spanish (CAS),<sup>4</sup> a dialect spoken in the Southwest part of Colombia, in the Ipiales-Nariño municipality. CAS has a strong index clitic system that “radicalizes” what is present in a more modest fashion in most Spanish varieties. We argue that CAS, as well as some other dialects, such as Rioplatense Spanish (Barrenechea & Orecchia 1970; Fontana 1994; Colantoni 2002; Di Tullio & Zdrojewski 2006; Belloro 2007, 2009, 2012; Di Tullio & Kailuweit 2011) and the Spanish spoken in Chiapas (in the Southeast of Mexico) (Chapa Barrios 2019), is ahead in showing the nature of the Spanish system of argument realization. To prove this, we elaborate on the following ideas:

- a) In most Spanish varieties, as well as in CAS, subject and object RPs are most frequently not coded. The verbal word functions by itself as the clause and the indexes constitute the basic system for the argument realization. The analysis where the arguments are absent RPs does not do justice to this fact.

---

<sup>4</sup> Colombian Andean Spanish has been identified as a proper variety by different authors (Flórez 1961; Montes 1985; Mora et al. 2004, among others). Specifically, we take Ruiz Vásquez’s (2020) proposal, which considers Andean Spanish as a super-dialect and distinguishes two Colombian sub-varieties: Highlands Colombian Andean Spanish and Lowlands Colombian Andean Spanish, both with different regional dialects. The data we present here corresponds to the Highlands variety, and inside this, to the Nariño dialect, specifically the one spoken in Ipiales-Nariño.

- b) Contrastively, there is not a “strong” case flagging system in Spanish. Dative IOs seem to count as the only true case-marked arguments. We will show that in CAS even these IOs are beginning to lose their *a* marking in some contexts.
- c) In CAS both dative and accusative indexes can “remain” coded in presence of RPs, so CAS has a set of indexes that distinguishes three arguments. It also is a “doubled object” variety. Following Siewierska (2004), we assume that it has a cross-indexing system. This is not really a doubling system, but one where the features of a single argument can be simultaneously distributed through both the indexes and the RPs.

The paper is structured as follows: in Section 2 we deal with some structural aspects of Spanish, whose traditional analyses are misleading: the role of RPs in terms of their frequency coding, the status of the flagging system on the RPs, and the identity of the verbal indexes. In Section 3 we address some important features of CAS that show the role of indexes in argument realization in that variety. The data provided in this section is qualitative and not quantitative. In Section 4, we review some proposals in the literature about the status of RPs in head-marking languages and propose an alternative analysis. Finally, in Section 5 we offer some conclusions.

## **2. A re-thinking of some features of Standard Spanish**

In this section, we briefly go through some important structural characteristics of Standard Spanish.<sup>5</sup> The aspects discussed and questioned here are: a) the role of both subject and object RPs as the main device for the instantiation of the verbal arguments; b) the identity of object clitics as pronouns or as agreement markers; and finally, c) the importance and strength of the flagging system of the RPs. “Traditional” analyses<sup>6</sup> of these topics are misleading, because they have their origin in the

---

<sup>5</sup> We use the term *Standard Spanish* in a loose way to refer to what can also be called *general Spanish*, a version of the language that presumably can be recognized by speakers of most varieties, and that excludes controversial uses; something akin to the basic formal or academic version of Spanish, which is, more or less, an abstract supra-version of the language. So *standard* does not have any socio-cultural implications, and does not have linguistic implications, other than those that are directly implicated in this work in relation to the clitic system, the role of RPs and the status of the flagging system.

<sup>6</sup> *Traditional analyses* here means most past and recent approaches to the structural nature of Spanish, which, to our knowledge, in an indirect non-explicit way assume that the main structural means for the coding of the semantic participants in the clause is by means of referential phrases.

imposition of grammatical views that come from the study of other Indo-European languages, such as Latin (lat; Italic) and Greek (ell; Greek), which do not have object clitics or affixes, though being also pro-drop languages. Most importantly, at least in recent times, these analyses have been reinforced by the direct or indirect influence of a general and prevalent theoretical-conceptual framework that has emerged mostly from the study of languages like German (deu; Indo-European, Germanic), English (eng; Indo-European, Germanic) and Russian (rus; Indo-European, Slavic), which have systems of argument coding based on RPs. In what follows, we review these grammatical features one by one.

### 2.1. *The role of RPs in argument realization*

In the Hispanic Linguistic tradition, Spanish RPs are assumed to be the clear manifestation of the verbal arguments in the clause; they are said to function like arguments in semantic and syntactic terms. Any time one looks for a clear example of a clause in this language, it is common to find an example of a clause with full lexical RPs. Nevertheless, in everyday communication RPs strongly tend to be not coded. The most frequent cases, at least in corpus data, are clauses like (5).

(5)

- a. *Qué bueno que ya = lleg-aste.*  
 that.is good SUB PTL = arrive-2PST  
 ‘Good, (you) have just arrived.’
- b. *Tom-a, agárr-a = lo.*  
 take-2PRS.IMP hold-2PRS.IMP = 3ACC  
 ‘(You) take (it), hold it.’
- c. *Ábr-e = me.*  
 open-2PRS-IMP = 1DAT  
 ‘(You) open (the door) for me.’
- d. *Se = lo = di ayer.*  
 3DAT = 3ACC = give.1PST yesterday  
 ‘(I) gave it to him/her yesterday.’

Here the RPs, or some of them, are “missing”, compared to what is expected in other languages, as English and German. However, RPs are not necessary as their referents can be recovered from the indexes or from the situational context, or from both.

The pervasive idea that the RPs are missing, that they are somewhere but have been not coded, clearly based on the model of languages with obligatory RPs (see Haspelmath 2013), has led to the pro-drop analysis and the assumption that the verbal inflection functions as an agreement marker, which agrees with a structurally present, although not explicitly coded, subject RP. This virtual agreement analysis has also been extended to the object clitics, at least the dative one. As mentioned before, this analysis is not convincing, in the first place, since agreement is a two-term syntactic relation established by the co-presence of a controller and a “controlee” or pivot (Van Valin & LaPolla 1997). In other perspectives (Alcina & Blecua 1975; Seco 1989; Alarcos Llorach 1994; Bosque & Demonte 1999; Fernández Soriano 1999; RAE/ASALE 2009; among many others), the object clitics are considered pronouns: in the absence of the RPs, they substitute them and, hence, they function as the arguments; this has given rise to the “image” of a system of complementary distribution between the person forms and the RPs.

However, in corpus data, the verbal word frequently functions as the whole clause itself. Particularly, subject RPs are frequently absent; García Miguel (2015) mentions that in the ADESSE database,<sup>7</sup> which contains syntactic and semantic analyses of almost 160,000 clauses that make up the texts of the Pan-Hispanic ARTHUS corpus,<sup>8</sup> subject RPs appear in only 36% of the total cases; that is, in most of the analyzed clauses, 64% of the cases, the subject argument is directly recovered from the verbal inflection. Bogard (2010) reports even larger percentages for different varieties: in Mexican Spanish 73.4% of the cases appear without a lexical subject; in Colombian Spanish 69.8% of the clauses are in the same situation, and in Peninsular Spanish subjects are not explicitly coded in 66.8% of the clauses. It is necessary to take note that these data are based on the analysis of written discourse. To our knowledge, there are not studies that give a proper account of this type of phenomenon in oral discourse, particularly in dialogic interactions, but presumably, the percentages of cases without a subject RP are much higher, since oral communication is much more anchored to the situational and discourse contexts.

---

<sup>7</sup> Base de datos Alternancias de diátesis y esquemas sintáctico-semánticos del español, (Universidad de Vigo: <http://adesse.uvigo.es/data/>).

<sup>8</sup> Archivo de Textos Hispánicos de la Universidad de Santiago.

In sum, the subject RPs are most frequently not coded and, hence, the argument is directly instantiated by the verbal index. If one starts the analysis from this fact and not from preconceived ideas, it becomes clear that the subject argument information comes from the verbal index and not from RPs that are not present.

This is also true for the indirect object (IO) argument. Vázquez Rozas (1995) provides percentages that go from 91% cases of IOs coded through the clitic index (with or without RP) to only 9% of cases with a lexical IO and without the bound person form. In the same vein, García Miguel (2015) cites a 74.14% of cases of ditransitive constructions without a lexical IO-RP and the object manifested only through the clitic index. In Aranovich's data (2011), the presence of IO-RPs in ditransitive constructions accounts only for a 17% in written texts, although there are dialect differences: while Latin American variants have 29.30% of IO-RPs, the Peninsular dialect shows only 5.80% of similar cases. This last study does not differentiate between IO-RPs which are "doubled" by the clitic and those without the index.

Interestingly, despite the data, in most works of reference the dative clitic is still treated as a copy and the RP is assumed to be the argument, although it is not present in most of the cases. There are three different types of analyses: a) the clitic is assumed to be just a non-informative, redundant form: the "superfluous dative" (Kany 1945: 116; Gili Gaya 1961: 174; Academia Española 1973: Ch. 3.10.4); b) the clitic is assumed to be an agreement marker (Givón 1976; García Miguel 1991; Bogard 1992; Company 1998, 2003); and c) the index is considered a pronoun that substitutes the RP, as in complementary distribution (Alcina & Blecua 1975; Seco 1989; Alarcos Llorach 1994; Bosque & Demonte 1999; Fernández Soriano 1999; RAE/ASALE 2009, etc.), which supposes that the typical scenario is the presence of the RP and then, where it does not appear, the clitic enters as a substitute. Contrary to this, the most frequent case is the presence of the clitic by itself and only in some of these cases it is doubled by a RP. The fact that both the RP and the index can appear together, which is a more frequent scenario than that of clauses with the IO only realized as a RP, shows that the system does not, in fact, operate in complementary distribution.

The problem with all these approaches is that they start from a misconception of the phenomenon. They take for granted the argument realization as a RP. However, if one takes the actual distribution, what comes out clear is that the IO is systematically coded through the index, and then, in some specific cases, it can be

doubled by a RP, most probably for pragmatic reasons (Belloro 2012).<sup>9</sup> In sum, if we start from the consideration of the empirical facts, the main means for the realization of the subject and the IO arguments is by means of bound indexes in the verbal predicate and not through the presence of RPs.

Things are somewhat different, at least in most Spanish varieties, in the case of the accusative or direct object (DO) argument. This is more frequently coded as a RP. For example, Vázquez Rozas (1995) presents a percentage of 75% of lexical realization of that argument, against only a 25% of coding through the clitic index. As has been noted in the literature (Comrie 1981), this is because the DO is usually focal, in information structural terms. This means that its referent represents new information, and it must be explicitly coded through lexical phrases. So, it is possible to say that most Spanish varieties present a mixed system where the subject and IO arguments are typically realized as indexes in the verbal form and the DO is projected as a full RP. Interestingly, as we show in the next subsection, this DO-RP is usually not flagged at all.

## 2.2. The Spanish case marking system

As mentioned before, Spanish has a mixed system for the argument realization. It has a very strong set of argument indexes and it also has a complementary system of flagging in the RPs, when these do appear coded. This system does not include flags for every possible case distinction, so it is relatively basic in comparison to the robust set of indexes that in some Spanish varieties overtly marks the three major type of arguments, and in comparison to languages with a robust set of case distinctions. In this sense, it is not accurate to characterize Spanish as a dependent marking language. In the first place, as it is well known, full lexical subject arguments are not flagged at all, as the examples in (6) show:

(6)

- a. *Enrique jueg-a fútbol todos los = días.*  
Enrique play-3PRS soccer every ART.PL = day  
'Enrique plays soccer every day.'

---

<sup>9</sup> There is also what Haspelmath (2013) calls the dual-nature view (Bresnan & Mchombo 1987; Siewierska 2004; Van Valin 2005), which considers that when the RP is present, the clitic is an agreement marker, and when the RP is absent, the index is the argument. We will discuss this approach in Section 4.



- b. *La = niña toc-a el = piano por = las = mañanas.*  
 ART = girl play-3PRS ART = piano PREP = ART.PL = mornings  
 ‘The girl plays the piano in the mornings.’
- c. *Guillermo trabaj-a hasta tarde.*  
 Guillermo work-3PRS PREP late  
 ‘William works late.’
- d. *La = tienda qued-a lejos.*  
 ART = store be-3PRS far  
 ‘The store is far.’

As can be seen, both animate, (6a), (6b) and (6c), and inanimate (6d) RPs are not marked; similarly, RPs with proper names, (6a) and (6c), or common names, (6b) and (6d), are not flagged; and equally, both subjects of transitive, (6a) and (6b), and intransitive, (6c) and (6d), predicates are unmarked. Of course, there are many languages where one case marker, usually the nominative or the absolutive (Turkish and Chechen<sup>10</sup>, respectively), lacks formal realization and is assumed to be a null or zero morpheme. But this typically happens in languages where the rest of the paradigm has overt coding.

The unmarked or prototypical case of DO argument with an inanimate referent (Comrie 1981), as in (6a) and (6b) above, is not flagged. In Vázquez Rozas (1995) data, almost 81% of the DO are inanimate and they come without a case flag. So, again, if one starts only from the empirical data, it is the case that the two most important arguments of the Spanish clause are not flagged when instantiated as lexical RPs. This is not to say that the semantic and syntactic identity of the arguments cannot be established. This, of course, proceeds through the two other major mechanisms of argument identification: word order and semantic denotation. But this is not the same as saying that the arguments are case-marked. There are three types of evidence that have been adduced to argue that RPs are indeed flagged in Spanish.<sup>11</sup> The first one comes from examples like (7a) and (7b):

<sup>10</sup> Turkish (tur; Turkic, Oghuz); Chechen (che; Nakh-Daghestanian, Nakh).

<sup>11</sup> The classification of Spanish as a dependent marking language implies the fact that there is a system of case marking on the dependents, and although most works in the bibliography accept the idea that the Latin case distinctions only survived in Spanish through the free and clitic pronoun systems, very often terms as nominative and accusative are used “freely” to refer to the syntactic function of the RPs, as synonymous of subject and direct object. In this same direction, the substitution of RPs by the clitics often results in that the substituted RPs are identified as nominative, accusative or dative.

(7)

- a. *Pepe bes-ó a = Lulú.*  
 Pepe kiss-3PST DOM = Lulu  
 ‘Pepe kissed Lulu.’
- b. *Lulú golpe-ó a = Pepe.*  
 Lulu hit-3PST DOM = Pepe  
 ‘Lulu hit Pepe.’
- c. *\*Luisa quem-ó a = la = casa.*  
 Luisa burn.down-3PST DOM = ART = house  
 ‘Luisa burned down the house.’
- d. *\*Ramón romp-ió a = el = vaso.*  
 Ramón broke-3PST DOM = ART = glass  
 ‘Ramon broke the glass.’

In these clauses, the DO arguments appear introduced by the form *a*. But as can be seen from the ungrammaticality of (7c) and (7d), this only happens with animate RPs and very rarely with inanimate ones. The *a* form, then, is not really a device used for marking the functional relation between the argument and the predicate, as true flags are (Haspelmath 2013); rather, it is a differential object marker (DOM) — a device for signaling that the referent is not of the expected semantic type (inanimate). Hence, in a strict sense, the *a* form is not part of a flagging or case marking system.<sup>12</sup>

The second argument usually posited to show the presence of a case system in Spanish is the existence of two sets of independent pronouns: one for the A argument and one for the P argument:

(8)

- a. *Tú (A) me = salud-aste a = mi (P).*  
 2PRON 1ACC = greet-2PST DOM = 1PRON  
 ‘You greeted me.’
- b. *Yo (A) te = empuj-é a = ti (P).*  
 1PRON 2ACC = push-1PST DOM = 2PRON  
 ‘I push you.’

---

<sup>12</sup> As noted in the relevant bibliography (Torrego Salcedo 1999; Delbecque 2002; Leonetti 2004; Iemmolo 2010; Melis 2018), there is considerable dialect variation in the use of *a* as DOM, but Standard or Formal-Academic Spanish does maintain a clear-cut distinction between animate and inanimate.

- c. **Él** (A) *nos = salud-ó*                    **a = nosotros** (P).  
 3PRON 3PL.ACC = greet-3PST DOM = 1PRON.PL  
 ‘He greeted us.’
- d. **Nosotros** (A) *lo = salud-amos*            **a = él** (P).  
 1PRON.PL            3ACC = greet-1PL.PST DOM = 3PRON  
 ‘We greeted him.’
- e. **Ella** (A) *los = felicitó*                    **a = ustedes** (P).  
 3PRON 3PL.ACC = congratulate DOM = 2PRON.PL  
 ‘She congratulated you.’
- f. **Ustedes** (A) *la = regañ-aron*            **a = ella** (P).  
 2PL.PRON 3ACC = scold-2PL.PST DOM = 3PRON  
 ‘You scolded her.’

As can be seen in (8a) and (8b), the pronouns for the A argument, *tú*, “2sg.nom” and *yo* “1sg.nom”, are clearly different from the respective P pronouns *ti* and *mi*. However, this difference is not that systematic, as it only appears between the first and second singular person units (Fernández Soriano 1999). There is no difference between the plurals of the first – (8d) and (8c) – and second persons – (8f) and (8e) – in their use as A or P arguments. And there is no formal distinction between third persons, neither in the singular nor in the plural. The only difference in all these person forms is the presence of the *a* marker. This one appears, again, as a DOM with P animate arguments. 3<sup>rd</sup> person inanimate referents, therefore, have one syncretic pronoun only for both singular and plural. So, in general terms, we can state that the case distinctions of the free pronouns system are minimal.

The third proof of the supposed existence of the Spanish case-marking system comes from the substitution of the object RPs with the set of the so-called clitic pronouns, which makes evident the difference between accusative and dative RPs and between them and the subject RP:

(9)

- a. *Mercedes dio el = dinero a = su = hermana.*  
 Mercedes give.3PST ART = money PREP = 3POSS = sister  
 ‘Mercedes gave the money to his sister.’
- b. *Ella se = lo = dio.*  
 3PRON 3DAT = 3ACC = give.3PRS  
 ‘She gave it to her.’

As can be seen in (9b), the DO of (9a), *el dinero* ‘the money’ is substituted by the bound person form *lo*, considered as an accusative pronoun, and the IO *a su hermana* ‘to her sister’ is substituted by the clitic *se*,<sup>13</sup> which is labeled as dative. The non-marked subject RP *Mercedes* of (9a) is substituted in (9b) by a syncretic (in terms of case) free pronoun. It is mostly due to this methodological procedure, substitution, that linguists talk about case distinctions in Spanish. As argued above, this procedure is inadequate as it starts from the view that the RPs are the main structural way in which arguments are projected. What the facts indicate is that the system of indexes is the main grammatical means for argument coding and for this, it is independent of the RP flagging system. Again, this is not to say that there is not a way of distinguishing the RPs when they appear coded (word order also plays an important role), but to emphasize that the instantiation and identity of each argument is mostly guaranteed by the bound person forms.

It seems then that the only one true case flag in Spanish is the dative *a* form of the IO. Its presence is mandatory in all semantic contexts: before animates – as in (10a) and (10b) – and inanimate referents, as in (10c), as well as in all syntactic contexts: postverbal – (10a) and (10c) – and in dislocated preverbal positions (10b); and before or after DOs – (10a) vs. (10c).

(10)

- a. *Fidel le = prest-ó \*(a) = Pedro un = poco = de = dinero.*  
 Fidel 3DAT = lend-3PST DAT = Pedro ART = some = PREP = money  
 ‘Fidel lent Pedro some money.’
- b. *\*(A) = Pedro le = prest-ó dinero Herminio.*  
 DAT = Pedro 3DAT = lend-3PST money Herminio  
 ‘Herminio lent Pedro money.’
- c. *Patricia les = pus-o cortinas \*(a) = las = ventanas.*  
 Patricia 3DAT.PL = put-3PST curtains DAT = ART.PL = windows  
 ‘Patricia put curtains on the windows.’

In this scenario, Standard Spanish should be classified as a mixed language with a robust set of verbal indexes and a not so robust flagging system (specifically, for the indirect or dative object).

---

<sup>13</sup> Dative *se* appears in the co-presence of the DO clitic forms and it is in complementary distribution with the more frequent *le* (3sg) and *les* (3pl) forms, which appear when there is no DO clitic attached to the verbal predicate.

### 2.3. Person forms are neither agreement markers nor pronouns

As stated in Haspelmath (2013), there are three ways in which indexes and RPs (or conominals in his terminology) can co-exist: 1) indexes with obligatory conominal, as in German, Russian and English, where the subject RP always appears simultaneously with the presence of a person index in the verbal inflection. This is what must be identified as agreement proper, a two-term syntactic relation; 2) indexes with impossible conominals, where the index stands in the place of the RP and acts as a true pro-nominal. When coded, the RP itself appears without a correspondent index, so, this type of system operates in complementary distribution; and 3) indexes with optional conominals, which is the most frequent kind of system in the world's languages; it is usually labeled in the literature as a cross-reference system (Bloomfield 1933; Hockett 1958; Sierwierska 2004).

In terms of what we have said until now, the Spanish argument realization system is a cross-reference or cross-indexing system. More precisely, it is a system where the most basic and frequent case is the one where the arguments (at least subject and IO) are coded through indexes and then, these can optionally be accompanied by the correspondent RPs. The indexes are the arguments in both scenarios. We discuss the status of the conominals in Section 4.

In this context, we consider that the indexes are not agreement markers. The typical (non-explicit) analysis, common to all the Hispanic Linguistics tradition, assumes that, in particular, the subject in pro-drop clauses is a non-coded RP and that the inflection on the verb is an agreement marker (Heger 1967; Alcina & Blecua 1975; Silva-Corvalán 1981; Suñer 1988; Seco 1989; García Miguel 1991, 1995; Bogard 1992, Alarcos Llorach 1994; Bosque & Demonte 1999; Company 1998, 2003; Belloro 2004, 2007; Di Tullio 2005; Kailuweit 2008; RAE/ASALE 2009; among many others).<sup>14</sup> We argue that this analysis is misleading for the following reasons: a) agreement is a two-term syntactic relation that implies the simultaneous presence of a controller and a controlee; b) in Spanish there is a strong tendency for subject RPs to be absent or not-coded; and c) there is not always an anaphoric antecedent in the discursive context; the verbal inflection can be pointing out to a referent in the situational context and not to a discursive antecedent, as in *Está sola* '(she) is alone' (the speaker is looking

---

<sup>14</sup> This non-explicit analysis comes straight from the fact that in all the cited works it is assumed that the subject agrees with the verbal inflection, whether there is a lexical subject or not. This implies that the agreement controller can be coded or not.

at a woman). In this example neither the verbal inflection nor the gender of the adjective *sola* can be said to be controlled, as a referent in the world cannot be a linguistic controller. There is referential matching, but there is no syntactic control.

This analysis can be extended to the case of the clitic indexes, at least to the dative one for most varieties. It cannot be an agreement marker since it is the formal instantiation of the IO argument. As seen before, IO RPs are usually not coded. So, it cannot be the case that absent RPs are the controllers of the verbal indexes. At the same time, the indexes are not pronouns substituting the RPs (Van Valin 2013). They do not substitute anything as they are typically present (subject person forms are obligatory, as well as the dative ones in many Spanish varieties). Also, they do not necessarily have to be linked to an antecedent in the discourse context. Besides, first and second person forms, the most common ones, never substitute anything, since they are deictic forms.

Another indication that the bound person forms are not pronouns comes from the fact that they can be doubled by free pronouns, as in (11a), the same way they can be doubled by RPs.<sup>15</sup> In contrast, free pronouns cannot be doubled by RPs, as shown by the ungrammaticality of (11b). This means that the RPs and pronouns behaves similarly and differently from indexes, which are not pronouns nor nominals.

(11)

- a. *La = vi*                      *a = ella.*  
       3ACC = see.1PST    DOM = 3PRON  
       ‘I saw her.’
- b. \**Vi*                      *a = ella*                      *María.*  
       see.1PST            DOM = 3PRON *María*  
       ‘I saw her (Mary).’

---

<sup>15</sup> The traditional analysis starts from the consideration that it is the clitic which obligatorily has to appear doubling the free pronoun. In our analysis, it is the pronoun which doubles the clitic, just like the RPs do. In topicality contexts, the clitic must appear, and by definition, pronouns always constitute topical information. So, since the referent in turn is topical, the clitic must be coded *La vi en el cine* ‘I saw her at the movies’. Consequently, given certain pragmatic needs (to emphasize or to contrast), the speaker can add the pronoun *La vi a ella en el cine* ‘I saw her at the movies’. So, in contexts where the pronoun can appear, the clitic is always present. This gives the impression that it is the pronoun which requires to be doubled by the clitic.

One last proof that the Spanish verbal indexes are not pronouns comes from the fact that they can signal indefinite (12a) and generic elements (12b), as well as propositions (12c). As Van Valin (2013) notes, true pronouns should only be able to cross-reference definite RPs, since pronouns are themselves definite (Austin & Bresnan 1996).

(12)

- a. *¿Lo = vi-ste, a = un = señor que = pas-ó por = ahí?*  
 3ACC = see-2PST DOM = ART = man REL = pass-3PST by = over.there  
 ‘Did you see him, a man who passed by.’
- b. *Quier-en a = alguien = que sí pued-a hacer el = trabajo.*  
 want-3PL.PRS DOM = someone = REL AFF can-3PRS do ART = job  
*Lo = contratar-ían de = inmediato.*  
 3ACC = hire-3PL.FUT PREP = immediately  
 ‘They want someone who can actually do the job. They would hire him immediately.’
- c. *Consider-o = que no deb-erías ir. Realmente lo = creo.*  
 consider-1PST = SUB NEG should-2FUT go really 3ACC = think  
 ‘I think you shouldn’t go. I really think so.’

In Section 3, we will present some further arguments on the nature of the indexes in Colombian Andean Spanish (CAS), which cannot be considered as pronouns or as agreement markers. For example, in oral communication the dative clitics can lose their plural number feature. Similarly, accusative clitics can lose both gender and number features. So, they do not instantiate referents (they are not pronouns) and do not agree with the correspondent RPs.

### 3. The argument realization system in CAS

In this section, we analyze some aspects of the clitic system of Colombian Andean Spanish (CAS), specifically from the variety spoken in Ipiales-Nariño, a city in southwestern Colombia, near the border with Ecuador. The data come from two main sources: a sample of occurrences in natural discourse, i.e., in real communicative interactions between men and women of different ages, in different types of formal and informal contexts, and from metalinguistic interviews with a group of informants, made with the aim of verifying issues related to the morphosyntax of the indexes. The

informants are around 80 speakers that belong to an extended network of family and friends in Ipiales, most of them are middle class, with high and middle levels of schooling. The analysis is of qualitative nature, and it is not based on quantitative data.

### 3.1. A three-index robust system

As discussed above, the more frequent and basic way in which subject and IO arguments are realized in Standard Spanish and in other varieties of Spanish is through the bound person forms in the verbal word: the dative clitic, in the case of the IO, and the person features in verbal inflection, in the case of the subject. The situation is different for the DO arguments since the basic way of coding these is through RPs. But this is not the case in CAS: in this dialect, similarly to a few other varieties, such as the Rioplatense Argentinian Spanish (RAS) and the Spanish from Chiapas (México), the basic means for the syntactic realization of DO arguments is also as a clitic index as in (13).

(13)

- a. *Mir-a, la = Flor me = los = dio.* (Showing some candies in her hand)  
 look-2PRS ART = Flor 1DAT = 3PL.ACC = give.3PST  
 ‘Look, Flor gave them to me.’
- b. *Pél-a = las, por = favor-cito.* (Pointing to a sack of potatoes).  
 Peel-2PRS = 3ACC PREP = please-DIM  
 ‘Peel them please.’
- c. *¿Dónde lo = compr-aste el = vestido?*  
 where 3ACC = buy-2pst ART = dress  
 ‘Where did you buy the dress?’
- d. *Pás-a = me = lo el = vaso = de = agua.*  
 pass-2PRS = 1DAT = 3ACC ART = glass = PREP = water  
 ‘Pass me the glass of water.’

(13a) and (13b) show the common cases where the DO clitic appears without a RP and without any discourse antecedent, like in any other Spanish dialect. The referent is directly recovered from the speech situation, so the clitic functions as a deictic form, i.e. as an index. In contrast, examples (13c) and (13d), in which a coreferential RP, a conominal, is present along with the clitic, are more pragmatically restricted in



other varieties of Spanish. There are not, to our knowledge, data-based accounts of how frequent this double accusative construction appears in CAS. But it seems, in terms of the Ipiales speakers' perception, that it is common, or 'the natural way of saying it'.

In general terms, in CAS, the accusative index is coded in the presence or absence of the coreferential RP. As a consequence, the index is the argument and the RP, when it appears, should be considered as a duplication.

It is important to note that the accusative doubling is present in most Spanish varieties (Belloro 2012), and it is a grammatical feature of the language. Its presence, however, depends on how restricted it is in pragmatic terms.<sup>16</sup>

Our elicited data from CAS seem to suggest that the doubled accusative construction is unrestricted in most pragmatic contexts, as much as it is in the other important accusative doubling dialect: the Rioplatense Argentinian Spanish (RAS) (Barrenechea & Orecchia 1970, 1977; Blears 2000; Di Tullio & Zdrojewski 2006; Estigarribia 2006; Belloro 2012; Sánchez & Zdrojewski 2013). Both CAS and RAS clearly contrast with Standard Spanish, in which the double construction is much more pragmatically restricted:

(14) While listening to an LP

- a. ??*Prést-a = me = lo el = disco, est-á muy = bueno.* (Topical DO)  
 lend-2PRS = 1DAT = 3ACC ART = album be-3PRS very = good  
 'Lend me it (the album) is very good.'
- a'. *Prést-a = me = lo el = disco, est-á muy = bueno.* (Topical DO)  
 lend-2PRS = 1DAT = 3ACC ART = album be-3PRS very = good  
 'Lend me it (the album) is very good.'
- b. *Prést-a = me = lo, est-á muy = bueno.*  
 lend-2PRS = 1DAT = 3ACC be-3PRS very = good  
 'Lend me it. It's very good.'
- c. *Prést-a = me el = disco, est-á muy = bueno.*  
 lend-2PRS = 1DAT ART = album be-3PRS very = good  
 'Lend me the album, it is very good.'

On the one hand, example (14a) shows that in Standard Spanish the double accusative is not favored in contexts where the DO referent is topical. On the other hand,

---

<sup>16</sup> See Belloro (2012) for a neat account of the phenomenon in Peninsular, Mexican and Argentinian varieties.

according to all our informants, in CAS, in this same context, the double construction in (14a') is very natural. In both varieties, the examples where the DO is realized only as a clitic (14b) or only as a RP (14c) are well-formed and natural.

In the case of new but anchored DOs, the double accusative, as in (15a), is odd and very unusual in Standard Spanish, but this is not so in CAS, where the construction is natural (15a'). Again, the alternative options with the DO as a clitic (15b) or as a RP (15c) are equally possible both in Standard Spanish and in CAS.

(15)

a. ??*Aprovech-é*                      *y*                      *las = compr-é*      (New, anchored DO)

take.advantage.of-1PST    CONJ      3PL.ACC = buy-1PST

*las = papas*                      *en = el = mercado.*

ART.PL = potatoes    PREP = ART = market

'I took advantage and bought the potatoes in the market.'

a'. *Aprovech-é*                      *y*                      *las = compré*      (New, anchored DO)

take.advantage.of-1PST    CONJ      3PL.ACC = buy-1PST

*las papas*                      *en = el = mercado.*

ART.PL = potatoes    PREP = ART = market

'I took advantage and bought the potatoes in the market.'

b. *Aprovech-é*                      *y*                      *las = compr-é*                      *en = el = mercado.*

take.advantage.of-1PST    CONJ    3PL.ACC = buy-1PST    PREP = ART = market

'I took advantage and bought them in the market.'

c. *Aprovech-é*                      *y*                      *compr-é*      *papás*                      *en = el = mercado.*

take.advantage.of-1PST    CONJ    buy-1PST    potatoes    PREP = ART = market

'I took advantage and bought potatoes in the market.'

When the DO is new and non-anchored in Standard Spanish, as in (16a), or it has a generic or *irrealis* interpretation, as in (17a), the accusative doubling is ungrammatical. This is not the case for CAS, as (16a') and (17a') show. As expected, in the cases of a new DO (16b) or a generic DO (17b), the RP is obligatorily needed, both in Standard Spanish and in CAS. On the other side, (16c) and (17c) confirm that a lexical DO makes these constructions viable.

(16)

a. \***La** = *vi*            *una = bicicleta*    *que = est-aba*    (New, non-anchored DO)

3ACC = see.1PST    ART = bicycle    REL = be-3COP

*en = la = puerta.*

PREP = ART = door

‘I saw a bicycle that was at the door.’

a’. **La** = *vi*            *una = bicicleta*    *que = est-aba*    (New, non-anchored DO)

3ACC = see.1PST    ART = bicycle    REL = be-3COP

*en = la = puerta.*

PREP = ART = door

‘I saw a bicycle that was at the door.’

b. \***La** = *vi*

3ACC = see.1PST

‘I saw her.’

c. *Vi*            *una = bicicleta*    *que = est-aba*    *en = la = puerta.*

see.1PST    ART = bicycle    REL = be-3COP    PREP = ART = door

‘I saw a bicycle that was at the door.’

(17)

a. \***Lo** = *contrat-arían*    *a = alguien*    *que = sí*    *pud-iera* (Generic DO)

3ACC = hire-3PL.COND    DOM = someone    REL = AFF    could-3.PSB

*hacer ese = trabajo.*

do    DEM = job

‘They would hire someone who could do that job.’

a’. **Lo** = *contrat-arían*    *a = alguien*    *que = sí*    *pud-iera* (Generic DO)

3ACC = hire-3PL.COND    DOM = someone    REL = AFF    could-3.PSB

*hacer ese = trabajo.*

do    DEM = job

‘They would hire someone who could do that job.’

b. \***Lo** = *contrat-arían.*

3ACC = hire-3PL.COND

‘They would hire him.’

c. *Contrat-arían*            *a = alguien*            *que = sí*            *pud-iera*

hire-3PL.COND            DOM = someone    REL = AFF    could-3.PSB

*hacer* *ese* = *trabajo*.

do DEM = job

‘They would hire someone who could do that job.’

In summary, the examples above show that in CAS the accusative doubling is possible in all pragmatic contexts, as in the case of topical or situationally anchored referents, and new and generic DOs, whereas the double accusative construction is highly restricted in almost all contexts in Standard Spanish.

Table 1 below summarizes the accessibility of double accusative in the relevant pragmatic contexts in Standard Spanish and CAS.

Standard Spanish system	CAS system
1 V + DO clitic + 0 (situationally anchored)	V + DO clitic + 0 (Situationally anchored)
2 V + DO clitic + 0 (topical)	V + DO clitic + 0 (topical)
3 V + DO clitic + Pron (topical)	V + DO clitic + Pron (topical)
4 *V + 0 + Pron (topical)	*V + 0 + Pron (topical)
<b>5 ??V + DO clitic + RP (Topical)</b>	<b>V + DO clitic + RP (Topical)</b>
6 V + 0 + RP (New)	V + 0 + RP (New)
<b>7 *V + DO clitic + RP (New)</b>	<b>V + DO clitic + RP (New-indefinite)</b>
8 *V + DO clitic + 0 (New, non anchored)	*V + DO clitic + 0 (New, non anchored)
9 V + 0 + RP (Generic)	V + 0 + RP (Generic)
<b>10 *V + DO clitic + RP (Generic)</b>	<b>V + DO clitic + RP (Generic)</b>

Table 1: *Distribution of the accusative doubling construction in Standard Spanish and in CAS.*<sup>17</sup>

As can be seen, the main behavioral differences are found in the following contexts (in bold in Table 1): a) in the case of a topical DO, in which the construction is possible and common in CAS but unusual in Standard Spanish; b) in the presence of a new, non-anchored or indefinite DO, in which the accusative doubling is not possible in

<sup>17</sup> The constructional schemes in Table 1 must be read as follows: subjects are omitted; V stands for verb; DO-clitic stands for the direct object clitic; 0 or RP in third position stand for the absence or presence of a DO-RP; 0 in second position stands for an absent direct object clitic, and Pron stands for free pronoun. The information in brackets is relative to the pragmatic value of the referent of the coded or absent RP. The ordering of the acronyms is not as the actual ordering of the lexical and morphological elements in real clauses. In this way, the schema in 1, for example, represent a construction such as *Las compré*, in which a DO-RP is not coded, and the referent of the clitic (*las*) is recoverable from the situational context, for example *las papas* (‘the potatoes’).

Standard Spanish, but is perfectly natural in CAS; and c) in the case of a generic DO, which allows the double construction in CAS, but not in Standard Spanish.

The pragmatic neutrality of the construction in CAS, especially in the cases of new and generic DO contexts, is very important, since it allows the construction to be very natural in many contexts in colloquial communication (as reported by our informants). If we consider that in most languages the typical DO is inanimate and represents new information (Comrie 1981), what we have in CAS is a “natural” increase of the possibilities for the syntactic realization of the DO as a clitic index in doubled constructions.

In this sense, object arguments in CAS, in almost all contexts, can be encoded by bound person forms in the verbal nucleus, both in the cases of IOs, which is a feature CAS has in common with Standard Spanish, and in the case of DOs, which is a pragmatically restricted feature in Standard Spanish and in most dialects, although it is present in all of them.

The “naturalness” of double accusative in CAS is also supported by the fact it can appear in the context of marked constructions as impersonal ones, as in (18a) and (18b),<sup>18</sup> and in relative clauses, as in (18c) and (18d).

#### (18) CAS

- a. *Y ahí se = los = qued-arán esos = dineros.*  
 CONJ over.there 3DAT = 3PL.ACC = keep-3FUT DEM.PL = money  
 ‘And there they will keep that money.’
- b. *Las = ventas = de = hervido y licor se = lo = har-á*  
 ART.PL = sales = PREP = boiled.fruits CONJ liquor 3DAT = 3ACC = do-3FUT  
*en = la = calle.*  
 PREP = ART = street  
 ‘The sales of boiled and liquor will be done on the street.’
- c. *Un = negocio = de = nadie, una = cosa que = la = tien-en*  
 ART = business = PREP = nobody ART = thing REL = 3ACC = have-3PL.PRS

<sup>18</sup> As pointed out by a reviewer, example (18a) can have an interpretation with a third person plural subject, but the general context of the discourse indicates that the speaker is talking about people in general and there is not a specific referent for whoever is going to keep the money. Third person plural inflection is also a well-known mechanism for impersonal constructions. This impersonal interpretation is strengthened by the presence of the locative deictic form *ahí* ‘there’ which, alternatively with *aquí* ‘here’, usually appears instead of a specific referent in impersonal contexts.

*como abandonada.*

as abandoned

‘A nobody’s business, a thing that they have abandoned.’

d. **Estos carros que = los = mir-amos aquí.**

these cars REL = 3PL.ACC = look-1PL.PRS here

‘These cars that we look at here.’

As seen in the examples, the presence of the clitic is pervasive across different syntactic constructions, as well as in different pragmatic contexts.

Another significant feature, common to both the accusative and the dative clitic, is the fact that they do not necessarily agree with the conominals that double them, as can be seen in (19) for the accusative clitic, and in (20) for the dative one:

(19) CAS

a. **Los = baños y el = cuarto = de = aseo lo = arrend-aron.**

ART.PL = bathrooms CONJ ART = room = PREP = cleaning 3ACC = rent-3PL.PST

‘The restrooms and the room where the cleaning supplies are kept were rented.’

b. **Usted déj-e = me decir = lo la = oportunidad**

2PRON let-2PRS = 1ACC tell = 3ACC ART.FEM = opportunity

**que = nos = brind-a.**

REL = 1PL.DAT = give-2PRS

‘You let me tell you the opportunity you give us.’

c. **Lo = traj\_eron los = bultos a = la = casa.**

3ACC = bring-3PL.PST ART.PL = packages PREP = ART = house

‘They brought the packages to the house.’

(20) CAS

a. **Luego = de = escuchar lo.que le = hab-ían dicho**

after = PREP = hear REL 3DAT = have-3PL.COP tell.PRT

**a = los = ecuatorianos.**

DAT = ART = Ecuadorians

‘After hearing what they had told the Ecuadorians.’

b. **De pronto le = vend-en a = otras personas también que = no**

suddenly 3DAT = sell-3PL.PRS DAT = other people also REL = NEG

*labor-aron nunca.*

work-3PL.PST never

‘Suddenly they sell to other people who have never worked.’

c. *Se = le = est-á dando la = autonomía*

IMP = 3DAT = be-3PRS give.GDO ART = autonomy

*a = los = funcionarios.*

DAT = ART.PL = public.workers

‘Public workers are being given autonomy.’

In (19a) and (19c) a plural RP appears doubling the third person singular masculine form *lo*; in (19b) the same masculine form is cross-referred by a feminine RP. This shows that there is no need for agreement of number or gender features. In fact, in (19b) above, there is a simultaneous absence of both person and gender agreement. In this sense, *lo* functions as neuter person form. Similarly, in the three examples of (20), the third person singular form *le* is doubled by plural RPs.<sup>19</sup>

This process of bleaching of number and gender features is common to most Spanish varieties (Company 1998, 2003), but it seems to be much more advanced in CAS. Even more important is to notice this behavior as an indication that the clitics are not functioning as pronouns nor as agreement markers. As said before, they do not show referential features (beyond person) and they do not agree with the RPs. In Section 4, we argue that the indexes are a purely formal or syntactic manifestation of the verbal arguments, as has been partially proposed for head-marking languages in Pensalfini (2004) and Koenig & Michelson (2012).

In essence, CAS is a RP-doubling language (as opposed to clitic-doubling) or, more accurately, a cross-indexing language. This means that the three major direct arguments receive indexing coding on the verb and can optionally be accompanied by a RP or a conominal.

### 3.2. The case flagging system in CAS

We now have established that CAS, in a similar way to the Rioplatense Argentinian dialect, has a more “robust” system of argument indexes than Standard Spanish, as the three major direct arguments are indexed on the verb. It is also the case that in CAS the flagging system for the RPs is, contrastively, slightly “weaker” than in Standard Spanish.

---

<sup>19</sup> Dative clitics do not have gender features.

We have attested two notorious syntactic behaviors that demonstrate this: 1) the dative marker *a* of the IO can be dropped in some contexts, and 2) dative RPs can be substituted by oblique RPs, when doubling the correspondent clitic.

As mentioned before, the strongest evidence of a flagging system in Spanish is the dative *a* marker of the IO. In Standard Spanish it is obligatory, both in postverbal position (21a) and in dislocated constructions (21c), as the ungrammaticality of (21b) and (21d), respectively, shows.

(21) Standard Spanish

- a. *Alicia le = regal-ó un = disco a = Javier.*  
 Alicia 3DAT = give-3PST ART = record DAT = Javier  
 ‘Alicia gave a record to Javier.’
- b. \**Alicia le = regal-ó un = disco Javier.*  
 Alicia 3DAT = give-3PST ART = record Javier  
 ‘Alicia gave a record (to) Javier.’
- c. *A = Javier le = regal-ó un = disco Alicia.*  
 DAT = Javier 3DAT = give-3PST ART = record Alicia  
 ‘Alicia gave a record to Javier.’
- d. \**Javier le = regal-ó un = disco Alicia.*  
 Javier 3DAT = give-3PST ART = record Alicia  
 ‘Alicia gave a record (to) Javier.’

Examples in (22) show that in CAS the dative marker is not mandatory in these two contexts. This indicates that the argument system does not rely on RP flagging, but on verbal indexes. At the present time, we do not know how frequent this kind of phenomenon is, but at least it does not seem rare to our informants.<sup>20</sup> This again shows the relative fragility of the marking system on RPs and the main role that indexes play.

(22) CAS

- a. *Le = pag-a sus = trabajadores.*  
 3DAT = pay-3PRS 3PL.POSS = workers  
 ‘(He/She) pays his/her workers.’

---

<sup>20</sup> In fact, as one reviewer points out, this is a behavior that can be found in other Spanish varieties.



- b. *También cómpr-a = le guagua.*  
 also buy-2PRS = 3DAT child  
 ‘(you) buy for the child too.’
- c. *Usted le = voy a = operar la = car-ita.*  
 2PRON 2DAT = go.1PRS PREP = operate ART = face-DIM  
 ‘I’m going to operate on your face.’
- d. *Porque ellos les = alcanz-a a = dar más barato.*  
 Because3PL.PRON 3PL.DAT = can-3PRS PREP = give more cheap  
 ‘Because (they) are able to sell (to) them cheaper.’

Another indication of the status of the flagging system in CAS is that the dative marking, which counts as a type of direct case marking, can be substituted by prepositional marking, as in (23a) and (23c):

(23)

- a. CAS  
*Ya no nos = da espacios para nosotros.*  
 AlreadyNEG 1PL.DAT = give.3PRS spaces for 1PL.PRON  
 ‘(He/She) no longer gives us spaces.’
- b. Standard Spanish  
*Ya no nos = da espacios a = nosotros.*  
 already NEG 1PL.DAT = give.3PRS spaces DAT = 1PL.PRON  
 ‘No longer it gives us spaces.’
- c. CAS  
*Para ellos les = va a = salir más costoso.*  
 for 3PL.PRON 3PL.DAT = go.3PRS PREP = become more expensive  
 ‘It will be more expensive for them.’
- d. Standard Spanish  
*A = ellos les = va a = salir más costoso.*  
 DAT = 3PL.PRON 3PL.DAT = go.3PRS PREP = become more expensive  
 ‘It will be more expensive for them.’

As can be seen, the dative clitic can be coreferential with a complement introduced by *para*, which is a preposition with greater semantic content than *a*. This highlights two very important facts: 1) the argument marking system does not necessarily rely on a non-

predicative type of case flagging, and therefore it allows the syntactic projection of semantic arguments introduced by predicative prepositions, and 2) the same argument can be simultaneously projected by two distinct units, with different grammatical statuses. In this way, on the one hand, the argument is morpho-syntactically realized by the verbal index and, on the other hand, the argument is semantically and referentially coded through the RP introduced by the preposition. In Section 4, we argue that the cross-reference constructions of CAS can be considered as cases of distributed coding of the same argument and are not doubled constructions, as they have been so far considered. They are not cases of repetition or double coding of the same referent.

The dative clitic can also be cross-referenced by an oblique RP introduced by the genitive case preposition *de*; see examples (24a) and (24c):

(24)

a. CAS

*Se = le = ha*                      *dado*              *el = cumplimiento*    *adecuado*    *de = esto*  
 IMP = 3DAT = have.3PRS    give.PRT    ART = compliance    proper            PREP = DEM  
 ‘Proper compliance has been given to this.’

b. Standard Spanish

*Se = le = ha*                      *dado*              *el = cumplimiento*              *adecuado*  
 IMP = 3DAT = have.3PRS    give.PRT    ART = compliance              proper  
***a = esto.***  
 DAT = DEM

‘Proper compliance has been given to this.’

c. CAS

*Que le = dé*                      *el = funcionamiento*              *de = la = plaza*  
 that 3DAT = give.3PRS    ART = functioning              PREP = ART = square  
***de = mercado***    *como*    *deb-e*              *de = ser*  
 PREP = market    like              should.3PRS    PREP = be

‘(He/She) should give proper functioning to the market-place as it should be.’

d. Standard Spanish

*Que le = dé*                      *el = funcionamiento*              ***a = la = plaza***  
 that 3DAT = give.3PRS    ART = functioning              DAT = ART = square  
***de = mercado***    *como*    *deb-e*              *de = ser*  
 PREP = market    like              should.3PRS    PREP = be

‘(He/She) should give proper functioning to the market-place as it should be.’

This also indicates the distributed projection of the same argument as holding two distinct identities, one as a syntactic argument through the verbal index, and another identity as a semantic argument through an oblique RP.

### 3.3. Another head-marking characteristic of CAS: Applicative constructions

As Yasugi (2012: 7) states, applicative constructions seem to be a characteristic strategy of head-marking languages, or argument-indexing languages, as we call them here. And indeed, the verbal indexation of the applied participant seems to be an important feature of applicative constructions. It is through such indexation that the applied participant is promoted to object status. In this sense, it is noteworthy to see that CAS has developed an applicative marker through the grammaticalization of the verbal form *dar* ‘give’ in the context of an applicative periphrastic construction (Ibáñez Cerda et al. 2022), as shown in the examples in (25).

(25)

- a. *Adela le = dio cocin-ando un = pastel a = su = mamá.*  
 Adela 3DAT = give.3PST cook-GDO ART = cake DAT = 3POSS = mother  
 ‘Adela cooked a cake instead of her mother.’
- b. *Da = me habla-ndo con = el = patrón.*  
 give.2PRS = 1DAT speak-GDO PREP = ART = boss  
 ‘Talk to the boss instead of me.’

In these clauses, *dar* ‘give’ appears along with another verb, *cocinando* ‘cooking’ in (25a), and *hablando* ‘talking’ in (25b), which is a non-finite form (a gerund), but that functions as the main predicate in semantic terms. The *dar* form is inflected and acts as an auxiliary. The construction is a periphrastic one. As proposed in Ibáñez Cerda et al. (2022), the applicative function of *dar* comes from the fact that, besides having no predicative meaning, it is its presence which allows the coding of a deputative beneficiary (Van Valin & LaPolla 1997), a semantically non-required participant, as an object through the presence of the dative index attached to it. In (25a) the clitic *le* (3sg) is cross-referred by the dative RP *a su mamá*; in (25b) the form *me* (1sg) indexes the speaker as the deputative beneficiary.

The construction is mostly used as an attenuation expression in directive/request speech acts, where the speaker, very politely, asks his addressee to do something instead of him, as in *Dame abriendo la puerta, por favor* ‘Please, open the door for me’.

This functional aspect is behind the fact that the most common applied participant is a deputative beneficiary and not the most typical recipient beneficiary.<sup>21</sup>

#### **4. The Status of the RPs in Spanish**

So far, we have posited that the argument realization system of Standard Spanish, and particularly that of CAS, is a cross-referencing or cross-indexing system. More precisely, at least in CAS, it is a system where the most basic case is the one where the three basic arguments are coded through indexes, and then, these can optionally be accompanied by RPs.

As Haspelmath (2013) states, there are three ways in which cross-reference systems, as in CAS, are traditionally analyzed, in terms of the status of the indexes and of the RPs or conominals. Here we briefly recapitulate each of these types of analysis, including a fourth one coming from Van Valin (2013), and finally we present our own alternative proposal.

1) The virtual agreement view. In this, the indexes are considered agreement markers, while the absent or empty RPs are the controlling arguments. This is the non-explicit analysis on which the whole Hispanic Linguistics tradition has been built up, but in terms of the notorious function of the index system and the equally notorious absence of RPs in everyday communication, it seems that there is no reason for such analysis, other than to emulate perspectives coming from other traditions. As Haspelmath (2013: 222) puts it:

It is very likely that this degree of abstractness was widely accepted only because of the influence of well-known European languages like German, English and (somewhat less clearly) Russian, which have gramm-indexing of the subject on the verb, where the conominal is obligatory. From the perspective of these languages, it

---

<sup>21</sup> This periphrastic construction has also been reported in the Andean Zone or Highlands of Ecuador (Haboud 1994, 1998; Bruil 2008; Creissels 2010), and it is seen as a type of calque from the surrounding and neighboring Quichua languages, which are polysynthetic and head-marking languages, and have applicative constructions. Independently of this possible substrate, what arises from the consideration of Spanish as an argument-indexing language, as we propose here, is its inherent structural inclination for developing such construction.

looks as if something is missing in unconominated cross-indexing patterns, so the notion of ‘pro-drop’ may seem natural (Haspelmath 2013: 222).

But now, knowing that languages with “real” agreement (i.e., where the copresence of the index and the conominal is obligatory) are rare (Siewierska 2004), and that cross-reference languages, such as Spanish, are more common, there is no reason to import an analysis that accounts for the former type, but not for the second one. So, this pro-drop analysis can be discarded.

2) The bound-argument view. From this perspective, the indexes are considered to function as pronouns, or nominal-like participants, and fully instantiate the verbal arguments. In this analysis, when the conominals are present, the bound person forms are still viewed as the arguments. In this case, RPs are considered a kind of adjunct or apposition outside of the core of the clause (Jelinek 1984; Nichols 1986). This type of analysis is common to some generative approaches, such as Jelinek (1984) and Baker (1996). As Siewierska (2001) and Van Valin (2013) argue, there is no solid proof for considering the RPs as adjuncts or appositions since they do not necessarily behave differently from arguments or RPs in other non-indexing languages. Most importantly, they do not behave like adjuncts (Van Valin 2013), which are peripheral, non-semantically required participants, and in that sense, they are opposed to arguments.

This also holds true for Spanish. There is no evidence of the non-argument status of the RPs nor of their placement in the periphery, or any other pragmatically motivated positions. So, it seems that this bound-argument view is not the best analytical route to follow.

3) The dual-nature view. In this analysis, the indexes are regarded as both agreement markers and pronouns depending on the circumstances: When the RPs are present, they are the arguments and the indexes are agreement markers; in a complementary fashion, when the conominals are not present, the bound person forms are pronouns and as such they are the arguments. This type of analysis was first proposed by Bresnan & Mchombo (1987) for Bantu languages. Van Valin (2013: 119) also proposes this dual-nature analysis for pro-drop subject languages like Croatian, which is considered a basic dependent marking language. A similar type of analysis has also been proposed for Spanish by García Miguel (1991, 1995), Belloro (2004, 2007) and Kailuweit (2008).

In our perspective, this analysis also does not suit Standard Spanish nor CAS. As we have proposed here, their grammatical structure is basically argument-indexing. As a consequence, if there is an analysis that does start from this consideration and does not force a dual nature for the indexes as agreement markers in the presence of RPs, then such an analysis would be preferable.

4) The Role and Reference Grammar (RRG) view. Van Valin (2005, 2013) presents an alternative analysis for basic head-marking languages. In this view, the bound person forms are neither agreement markers nor pronouns. They are the arguments in the core of the clause.<sup>22</sup> In RRG there is an important projection principle that restricts the instantiation of the same argument twice in the core. This supposes that the cross-referred RPs, when they are coded, cannot be in the core along with the indexes. As mentioned above, Van Valin (2013) states, in the same vein of Siewierska (2001), that the conominals are not adjuncts and do not behave like them, so they cannot be in the clause periphery.<sup>23</sup> In this context, he proposes that the RPs should be placed in what he terms the Extra-Core Slot (ECS), a clause internal but core external position. In this way, he avoids placing the RPs in the periphery, where adjuncts are, and avoids the RRG constraint that precludes the instantiation of the referent of an argument more than once per core. The problem with this approach is that there is no indication of what the behavioral properties of RPs in the ECS are. Are they different from RPs in the core? How are they different? Cross-referred RPs in Spanish, both in the Standard varieties and in CAS, seem to behave as standard argument RPs in other languages.

---

<sup>22</sup> Or more appropriately, in the core of the word, which, in turn, is integrated as part of the core of the clause (Van Valin 2013).

<sup>23</sup> Van Valin (2013) also shows that conominals can neither be in other pragmatically motivated positions out of the core, like those that are recognized in RRG as part of the layered structure of the clause: the Pre-Detached and Post-Detached positions (PrDP and PoDP) and the Pre-Core and Post-Core Slots (PrCS and PoCS). First, PrDP and PoDP imply dislocated elements with the presence of intonation breaks; besides, WH expressions cannot occur in these positions. Standard RPs which appear cross-referring verbal indexes in head-marking languages are not preceded by intonation breaks - and hence, they are not dislocated -, and can be substituted by WH words, so they must be clause-internal. Second, the other core-external, but clause-internal, positions, the PrCS and PoCS, are ruled out as hosts of the conominals, because, among other reasons, there can only be one element in only one of these positions per clause, and in cross-reference languages two or three RPs, depending on the language, can simultaneously appear doubling the argument indexes on the verb.

None of these proposals is completely accurate for explaining the argument realization system of Spanish. Here, then, we propose a fifth type of analysis for cross-reference systems, which picks up some aspects of Pensalfini (2004), Haspelmath (2013) and Van Valin (2013).

5) A new proposal. We first consider, as in Van Valin (2013), that in CAS, as a clear argument-indexing variety, the bound person forms are the arguments in the core of the clause; they are neither pronouns nor agreement markers. When the optional RPs appear coded, the indexes are still the arguments. Next, following Haspelmath's consideration (2013: 224) that there is nothing against the distributed expression of meaning, we propose that indeed in cross-reference languages arguments are expressed simultaneously in two different forms, the indexes and the RPs. This does not need to imply a double instantiation of the same referent. Following Pensalfini's (2004) and Koenig & Michelson's claim (2015) that all major word classes have two components, a formal and an encyclopedic one, we posit that the indexes in cross-reference constructions are the projection of the formal or syntactic component, whereas the RPs are the instantiation of the semantic and referential (or encyclopedic) identity of the arguments.

In this scenario, we propose that the indexes, as purely syntactic forms, do occupy the core of the clause. Then, as they do not have referential information, there is nothing against the instantiation in the same core of another linguistic form carrying the semantic and referential load. This means that in cross-reference constructions the RPs can occupy the core of the clause without violating the constraint on the instantiation of referents no more than once per core, as some frameworks prevent.

This proposal overcomes all other available: first, it eliminates the need for the "fallacy" of the omnipresent, non-explicit, pro-drop analysis: virtual RPs cannot be the syntactic controlling arguments of the verbal indexes. Second, it eludes positing adjunct status for the RPs, for which there is no evidence at all, as Siewierska (2001) and Van Valin (2013) exhibit. Third, it avoids the double nature analysis, as in Bresnan & Mchombo (1987), which is partially based in the pro-drop analysis. And finally, it refrains from positing the existence of framework-based positions, as the extra-core slot (ECS) of the RRG proposal (Van Valin 2013). As mentioned before, RPs in Spanish, both in the Standard varieties and in CAS seem to behave as arguments in semantic and referential terms. The analysis we propose here neatly captures this fact, and at the same time, gives the indexes the syntactic prominence they have in the argument realization system.

## 5. Conclusions

Most of the Hispanic Linguistic tradition literature, as well as typologically-oriented studies, consider that Spanish is basically a dependent-marking language, and for that, they assume that it is a kind of language in which argument realization is accomplished by means of RPs. Here, we have exhibited a different structural reality: 1) RPs are most frequently not coded, and arguments are instantiated directly by verbal indexes at least in the case of subject and IO arguments; 2) The distinction between arguments basically relies on the set of indexes. In this sense, we have provided proofs that Standard Spanish is basically an argument indexing language. We also have determined that this language has a cross-indexing system, where RPs can optionally accompany the indexes.

To present our proposal more clearly, we have analyzed some facts relative to the clitic system of Colombian Andean Spanish (CAS). In this dialect, DO arguments are also basically coded as clitic indexes in most pragmatic contexts, so CAS has a three-argument system consisting in person forms attached to the verbal word.

Finally, after examining some of the most relevant types of analyses about the status of RPs in cross-indexing systems, we have offered an alternative proposal: In cases where the indexes appear accompanied by the correspondent RPs, both are the simultaneous instantiation of the argument features load; the indexes stand for the syntactic or formal realization of the argument, and RPs manifest its referential and encyclopedic content. As such, both can occupy the core of the clause, without violating any type of restriction about the double coding of referents in the core of the clause. This type of cross-indexing construction, hence, is not a doubled construction, as it has been considered so far.

## Abbreviations

1 = First person	DO = Direct Object	PRON = Pronoun
2 = Second person	DOM = Diferential object marker	PRS = Present
3 = Third person	FEM = Feminine	PRT = Participle
ACC = Accusative	FUT = Future	PSB = Possibility
AFF = Affirmative	GDO = Gerund	PST = Past
ART = Article	IMP = Imperative	PTL = Punctual
COND = Conditional	NEG = Negation	REL = Relative



CONJ = Conjunction	PL = Plural	RP = Referential Phrases
COP = Copula	POSS = Possessive	SUB = Subordinate
DAT = Dative	PODP = Post-Detached Position	V = Verb
DEM = Demonstrative	PREP = Preposition	
DIM = Diminutive	PRDP = Pre-Detached Position	

## References

- Academia Española. 1973. *Esbozo de una nueva gramática de la lengua española*. Madrid: Espasa-Calpe.
- Alarcos Llorach, Emilio. 1994. *Gramática de la lengua española*. Madrid: Espasa-Calpe.
- Alcina, Juan & Juan Manuel Bleca. 1975. *Gramática española*. Barcelona: Ariel.
- Aranovich, Roberto. 2011. *Optional agreement and grammatical functions: a corpus study of dative clitic doubling in Spanish*. Pittsburgh: University of Pittsburgh (Doctoral Dissertation).
- Austin, Peter K. & Joan Bresnan. 1996. Non-configurationality in Australian aboriginal languages. *Nat Lang Linguistic Theory* 14. 215–268.
- Baker, Mark C. 1996. *The Polysynthesis parameter*. New York: Oxford University Press.
- Barrenechea, Ana María & Teresa Orecchia. 1970. La duplicación de objetos directos e indirectos en el español hablado en Buenos Aires. *Romance Philology* 24 (1). 58–83.
- Barrenechea, Ana María & Teresa Orecchia. 1977. La duplicación de objetos directos e indirectos en el español hablado en Buenos Aires. In Juan Miguel Lope Blanch (ed.), *Estudios sobre el español hablado en las principales ciudades de América*, 351–381. México: UNAM.
- Belloro, Valeria. 2004. *A Role and Reference Grammar account of third-person clitic clusters in Spanish*. Buffalo: University at Buffalo. (MA thesis).
- Belloro, Valeria. 2007. *Spanish clitic doubling: A study of the syntax-pragmatics interface*. Buffalo: University at Buffalo. (Doctoral Dissertation.)
- Belloro, Valeria. 2009. Spanish datives: remarks on the information-structure side of the story. In Lilian Guerrero & Sergio Ibáñez & Valeria Belloro (eds.), *Studies in Role and Reference Grammar*. 491–516. México: UNAM.
- Belloro, Valeria. 2012. Pronombres clíticos, dislocaciones y doblados en tres dialectos del español. *Nueva Revista de Filología Hispánica* (NRFH) 60(2). 391–424.
- Bleam, Tonia. 2000. *Leísta Spanish and the syntax of clitic doubling*. University of Delaware. (Doctoral Dissertation).

- Bloomfield, Leonard. 1933. *Language*. New York: Holt.
- Bogard, Sergio. 1992. El estatus del clítico de complemento indirecto en español. In Rebeca Barriga Villanueva & Pedro Martín Butragueño (eds.), *Reflexiones lingüísticas y literarias*, vol. 1 Lingüística. 171–186. México: El Colegio de México.
- Bogard, Sergio. 2010. La frase nominal de objeto directo antepuesta al verbo en español. In Sergio Bogard (ed.), *Semántica, pragmática y prosodia. Reflejos en el orden de palabras en español*. 69–115. México: El Colegio de México.
- Bosque, Ignacio & Violeta Demonte (eds.) 1999. *Gramática descriptiva de la lengua española*. Madrid: Espasa-Calpe.
- Bresnan, Joan & Sam Mchombo. 1987. Topic, pronoun and agreement in Chichewa. *Language* 63. 741–782.
- Bruil, Martine. 2008. Give + gerund in Ecuadorian Spanish: A calque from Quichua or a large process of contact induced change? *Leiden Working Papers in Linguistics* 5 (1). 1–23.
- Chapa Barrios, J. Fernando. 2019. *Duplicación de objeto directo en posición no marcada. El caso del español de Chiapas*. México: UNAM. (Bachelor thesis)
- Colantoni, Laura. 2002. Clitic doblado, null objects and clitic climbing in the Spanish of Corrientes. In Javier Gutiérrez-Rexach (ed.), *From words to discourse: Trends in Spanish semantics and pragmatics*. 321–336. Amsterdam: Elsevier.
- Company, Concepción. 1998. The interplay between form and meaning in language change. Grammaticalization of cannibalistic datives in Spanish. *Studies in Language* 22(3). 529–565.
- Company, Concepción. 2003. Transitivity and grammaticalization of object. The struggle of direct and indirect object in Spanish. In Giuliana Florentino (ed.), *Romance objects. Transitivity in Romance languages*. 217–260. Berlin: Mouton de Gruyter.
- Comrie, Bernard. 1981. *Language universals and linguistic typology*. Chicago: The University of Chicago Press.
- Creissels, Denis. 2010. Benefactive applicative periphrases: a typological approach. In Fernando Zúñiga & Seppo Kittilä (eds.), *Benefactives and malefactives*. 29–70. Amsterdam: John Benjamins.
- Delbecque, Nicole. 2002. A construction grammar approach to transitivity in Spanish. In Kristine Davidse & Béatrice Lamiroy (eds.), *The nominative & accusative and their counterparts*. 81–130. Amsterdam: John Benjamins.

- Di Tullio, Ángela. 2005. *Manual de gramática del español*. Buenos Aires: Waldhuter Editores.
- Di Tullio, Ángela & Rolf Kailuweit (eds.) 2011. *El español rioplatense: lengua, literaturas, expresiones culturales*. Madrid-Frankfurt: Iberoamericana-Vervuert.
- Di Tullio, Ángela & Pablo Zdrojewski 2006. Nota sobre el doblado de clíticos en el español rioplatense: asimetría entre objetos humanos e no humanos. *Filología* 38. 13–44.
- Estigarribia, Bruno. 2006. Why clitic doubling? A functional analysis for Rioplatense Spanish. In Timothy L. Face & Carol A. Klee (eds.), *Selected proceedings of the 8th Hispanic Linguistics symposium*. 123–136. Somerville, MA: Cascadilla Proceedings Project.
- Fernández Soriano, Olga. 1999. El pronombre personal. Formas y distribuciones. Pronombres átonos y tónicos. In Ignacio Bosque & Violeta Demonte (coords.), *Gramática descriptiva de la lengua española. Las construcciones sintácticas fundamentales. Relaciones temporales, aceptuales y modales*, vol. 1. 1209–1273. Madrid: Espasa-Calpe.
- Flórez, Luis 1961. El Atlas Lingüístico-Etnográfico de Colombia ALEC. Nota informativa. *Thesaurus, Boletín del Instituto Caro y Cuervo*, XVI (1), 77–125.
- Fontana, Josep 1994. El desarrollo de la conjugación objetiva en español. In *Revista Argentina de Lingüística* 10(1/2). 85–113.
- García Miguel, José María. 1991. La duplicación de complemento directo e indirecto como concordancia. *Verba* 18. 375–410.
- García Miguel, José María. 1995. *Las relaciones gramaticales entre predicado y participantes*. Santiago de Compostela: Universidade de Santiago de Compostela.
- García Miguel, José María. 2015. Variable coding and object alignment in Spanish. Some corpus-based evidence. *Folia Linguística* 49(1). 205–256.
- García Salido, Marcos. 2013. *La expresión pronominal de sujeto y objetos en español. Estudio con datos conversacionales*, Verba. Anexo 70. Santiago de Compostela: Universidade de Santiago de Compostela.
- Gili Gaya, Samuel. 1961. *Curso superior de sintaxis española*. Barcelona: Spes.
- Givón, Talmy. 1976. Topic, pronoun, and grammatical agreement. In Charles N. Li (ed.), *Subject and topic*. 149–188. New York: Academic Press.
- Haboud, Marleen. 1994. *On language contact and grammaticalization in Ecuadorian Highland Spanish*. Oregon: University of Oregon.

- Haboud, Marleen. 1998. *Quichua y castellano en los Andes ecuatorianos: Los efectos de un contacto prolongado*. Quito: Abya-Yala.
- Haspelmath, Martin. 2013. Argument indexing: A conceptual framework for the syntactic status of bound person forms. In Dik Bakker & Martin Haspelmath (eds.), *Languages across boundaries*. 197–226. Berlin: De Gruyter Mouton.
- Haspelmath, Martin. 2019. Indexing and flagging, and head and dependent marking. *Te Reo*, 62 (1), Issue in Honour of Frntisek Lichtenberk. 93–115.
- Heger, Klaus. 1967. La conjugación objetiva en castellano y en francés. *Thesaurus: boletín del Instituto Caro y Cuervo* 22 (2). 153–175.
- Hockett, Charles. 1958. *A course of modern linguistics*. New York: MacMillan Company.
- Ibáñez Cerda, Sergio & Alejandra I. Ortiz Villegas & Armando Mora Bustos. 2022. Applicative periphrastic constructions in the Colombian Spanish from The Andes. In Pacchiarotti, Sara & Fernando Zúñiga (eds.), *Applicative morphology: Neglected syntactic and non-syntactic functions*. Trends in Linguistics. 97–127. Berlin: De Gruyter Mouton.
- Iemmolo, Giorgio. 2010. Topicality and differential object marking. Evidence from Romance and beyond. *Studies in Language* 34(2). 239–272.
- Jelinek, Eloise. 1984. Empty categories, case, and configurationality. *Natural Language and Linguistic Theory* 2. 39–76.
- Kailuweit, Rolf. 2008. “Floating plurals”, prodrop and agreement – an optimality-based RRG approach. In Robert D. Van Valin (ed.), *Investigations of the syntax-semantics-pragmatics interface*. 179–202. Amsterdam: John Benjamins Publishing Company.
- Kany, Charles. 1945. *Spanish-American syntax*. Chicago: University of Chicago Press.
- Koenig, Jean-Pierre & Karin Michelson. 2012. The (non) universality of syntactic selection and functional application. In Christopher Pinón (ed.), *Empirical issues in syntax and semantics* 9. 185–205. Paris: CNRS.
- Koenig, Jean-Pierre & Karin Michelson. 2015. Invariance in argument realization: the case or Iroquoian. *Language*, vol. 91 (1). 1–47.
- Leonetti, Manuel. 2004. Specificity and differential object marking in Spanish. *Catalan Journal of Linguistics* 3. 75–114.
- Melis, Chantal. 2018. Spanish indexing DOM, topicality, and the case hierarchy. In Ilia A. Seržant & Alena Witzlack-Makarevich (eds.), *Diachrony of differential argument marking*, 97–128. Berlin: Language Science Press.

- Montes Giraldo j.j. (1985). *Estudios sobre el español en Colombia*. Bogota: Instituto Caro y Cuervo.
- Mora Monroy, Siervo C. & Mariano Lozano Ramírez & Ricardo A. Ramírez Caro & María B. Espejo Olaya & Gloria E. Duarte Huertas. 2004. *Caracterización léxica de los dialectos del español de Colombia según el «ALEC»*. Bogotá: Instituto Caro y Cuervo
- Nichols, Johanna. 1986. Head-marking and dependent-marking grammar. *Language* 62(1). 56–119.
- Pensalfini, Rob. 2004. Towards a typology of configurationality. *Natural Language and Linguistic Theory* 22. 359–408.
- RAE/ASALE. 2009. *Nueva gramática de la lengua española*. Madrid: Espasa-Calpe.
- Ruiz Vásquez, Néstor F. (2021). El español de Colombia. Nueva propuesta de división dialectal. *Lenguaje* 48(2), 160–195.
- Sánchez, Liliana & Pablo Zdrojewski. 2013. Restricciones semánticas y pragmáticas al doblado de clíticos en el español de Buenos Aires y de Lima. *Lingüística* 29(2). 271–320.
- Seco, Manuel. 1989. *Gramática esencial del español. Introducción al estudio de la lengua*. Madrid: Espasa-Calpe
- Siewierska, Anna. 2001. On the argument status of cross-referencing forms: some problems. *Revista Canaria de Estudios Ingleses* 42. 215–236.
- Siewierska, Anna. 2004. *Person*. Cambridge: Cambridge University Press.
- Silva-Corvalán, Carmen. 1981. La función pragmática de la duplicación de pronombres clíticos. *Boletín de Filología de la Universidad de Chile* 31(2). 561–570.
- Suñer, Margarita 1988. The role of agreement in clitic-doubled constructions. *Natural Language and Linguistic Theory* 6. 391–434.
- Torrego Salcedo, Esther. 1999. El complemento directo preposicional. In Ignacio Bosque & Violeta Demonte (eds.), *Gramática descriptiva de la lengua española. Las construcciones sintácticas fundamentales. Relaciones temporales, aceptuales y modales* 2. 1779–1805. Madrid: Espasa-Calpe.
- Yasugi, Yoshiho. 2012. A head-marking grammar for applicative constructions. In Wataru Nakamura & Ritsuko Kikusawa (eds.), *Objectivization and subjectivization: A typology of voice systems, Senri ethnological studies* 77. 7–22. Osaka: National Museum of Ethnology.
- Van Valin, Robert D. 2005. *Exploring the syntax and semantic interface*. Cambridge: Cambridge University Press.

Van Valin, Robert D. (ed.) 2008. *Investigations of the syntax-semantics-pragmatics interface*. Amsterdam: John Benjamins.

Van Valin, Robert D. 2013. Head-marking languages and linguistic theory. In Balthasar Bickel & Leonore A. Grenoble & David Peterson & Alan Timberlake (eds.), *Language typology and historical contingency*, 91–123. Amsterdam: John Benjamins.

Van Valin, Robert & Randy LaPolla. 1997. *Syntax: Structure, meaning and function*. Cambridge: Cambridge University Press.

Vázquez Rozas, Victoria. 1995. *El complemento indirecto en español*. Santiago de Compostela: Universidade de Santiago de Compostela.

#### CONTACT

[sergioimx@yahoo.com.mx](mailto:sergioimx@yahoo.com.mx)

[lucioamora@gmail.com](mailto:lucioamora@gmail.com)

[aov\\_26@yahoo.com](mailto:aov_26@yahoo.com)

# Using a parallel corpus to study patterns of word order variation: determiners and quantifiers within the noun phrase in European languages

LUIGI TALAMO

LANGUAGE SCIENCE AND TECHNOLOGY, SAARLAND UNIVERSITY (GERMANY)

Submitted: 19/10/2022 Revised version: 21/06/2023

Accepted: 28/07/2023 Published: 27/12/2023



Articles are published under a Creative Commons Attribution 4.0 International License (The authors remain the copyright holders and grant third parties the right to use, reproduce, and share the article).

## Abstract

Despite the wealth of studies on word order, there have been very few studies on the order of minor word categories such as determiners and quantifiers. This is likely due to the difficulty of formulating valid cross-linguistic definitions for these categories, which also appear problematic from a computational perspective. A solution lies in the formulation of comparative concepts and in their computational implementation by combining different layers of annotation with manually compiled list of lexemes; the proposed methodology is exemplified by a study on the position of these categories with respect to the nominal head, which is conducted on a parallel corpus of 17 European languages and uses Shannon's entropy to quantify word order variation. Whereas the entropy for the article-noun pattern is, as expected, extremely low, the proposed methodology sheds light on the variation of the demonstrative-noun and the quantifier-noun patterns in three languages of the sample.

**Keywords:** word order; determiner; quantifier; entropy; Universal Dependency; European languages

## 1. Introduction

Most of the previous studies on word order have been focused on major constituents like subject, verb and object, or adjectives and nouns. Although the two categories of demonstratives and numerals figure in many classic typologies on word order

(Greenberg 1963; Dryer 2009; Hawkins 1983), the closely related categories of articles and non-numerical quantifiers have received little attention (Ioup 1975; Greenberg 1978; Dryer 1992). Quantitative typological studies (Futrell et al. 2015; Naranjo & Becker 2018; Alzetta et al. 2018; Gerdes et al. 2019; Levshina 2019; Talamo & Verkerk 2022), which exploit computational resources, such as annotated treebanks and parsed corpora, and interpret the frequency through information-theoretic measures, have so far not considered these categories either.

The reason behind the neglect of these categories lies in the objective difficulty of defining determiners and non-numerical quantifiers. A quick look to grammars shows that demonstratives are often conflated together with other nominal modifiers such as articles and non-numerical quantifiers; whereas a category of numerical quantifiers, or ‘numerals’, can be quite easily identified, non-numerical quantifiers are often treated together with adjectives, numerals or even non-nominal modifiers like adverbs and intensifiers.

Previous qualitative studies that explicitly consider one of these categories employ a categorical measure to describe the word order pattern i.e., only one possible word pattern can be assigned to a language. On the other hand, quantitative studies use continuous measure such as frequency to capture the variability of word patterns, but the annotation schemata on which they are based do not offer fine-grained distinctions for determiners and non-numerical quantifiers.

In the present paper I aim to fill this gap by looking at the frequencies of noun-article, noun-demonstrative and noun-quantifier orders in a parallel corpus of 17 European languages, which is automatically parsed using tools from the Universal Dependency (UD) project. The rest of the paper is structured as follows: Sect. 2 briefly reviews the qualitative and quantitative studies on the order of determiners and quantifiers; Sect. 3 describes the methodology, presenting the parallel corpus, the information-theoretic measure used to interpret the frequencies and the implementation of the comparative concepts using annotations from the UD project; Sect. 4 presents the results and gives an in-depth analysis of a selection of word order patterns showing high variability; Sect. 5 concludes.

## **2. The order of determiners and quantifiers within the noun phrase in qualitative and quantitative studies**

The term ‘determiners’ is widely employed as an umbrella term for both articles and demonstratives, which is problematic even for a small sample like the one used in the



present article. As already observed by Dryer (2007: 152, 161-162), there are languages in which articles are used in combination with demonstratives and other types of determiners, like possessives, and there are languages in which articles do not exist. In my sample, Greek (ell; Indo-European, Graeco-Phrygian), Irish (gle; Indo-European, Celtic) and Welsh (cym; Indo-European, Celtic) are languages of the former type, while Bosnian-Croatian-Serbian (BCS<sup>1</sup>; hbs; Indo-European, Balto-Slavic), Bulgarian (bul; Indo-European, Balto-Slavic), Czech (ces; Indo-European, Balto-Slavic), Lithuanian (lit; Indo-European, Balto-Slavic), Polish (pol; Indo-European, Balto-Slavic) and Russian (rus; Indo-European, Balto-Slavic) are languages of the latter type.

The term demonstrative is often used interchangeably for both stand-alone words i.e., demonstrative pronouns, and modifiers; the latter are further divided into nominal demonstratives and adverbial demonstratives, which are usually etymologically connected; cfr. English (eng; Indo-European, Germanic) *this* and *that* vs. *here* and *there* (Diessel & Coventry 2020: 1). I consider here nominal demonstratives, which are sometimes described by grammars as ‘demonstrative adjectives’ or ‘adnominal demonstratives’ (Verkerk, p.c.), and I refer here to them as ‘demonstratives’.

As for non-numerical quantifiers, the term is often kept distinct from the similar category of numerical quantifiers, or ‘numerals’, indicating non-numerical words that express quantity; I refer here to this category as ‘quantifiers’. The category of quantifiers is from time to time lumped with determiners and/or adjectives, as in the following quotation from a recent grammar of Irish:

A variety of words referring to quantities also function as determiners within NPs. [...] They are on the whole rather a mixed bag of elements from a syntactic point of view. Many of these forms are treated as adjectives in traditional grammars, although they cannot be declined or compared like the adjectives. [...] (Stenson 2020: 188)

Studies on the order of articles and demonstratives with respect to the nominal head go back at least to Greenberg (1963), where Universal 18 is formulated as follows: “When the descriptive adjective precedes the noun, the demonstrative, and the

---

<sup>1</sup> I follow Alexander (2006)’s usage of the acronym BCS to indicate the pluricentric language formerly known as Serbo-Croatian.

numeral, with overwhelmingly more than chance frequency, does likewise.” (Greenberg 1963: 68).

language	art & noun	dem & noun	quant & noun
BCS	-	dem-noun	quant-noun
Bulgarian	-	dem-noun	quant-noun
Czech	-	dem-noun	quant-noun
Danish	art-noun	dem-noun	quant-noun
Dutch	art-noun	dem-noun	quant-noun
English	art-noun	dem-noun	quant-noun
French	art-noun	dem-noun	quant-noun
German	art-noun	dem-noun	quant-noun
Greek	art-noun	dem-noun	quant-noun
Irish	art-noun#	noun-dem	quant-noun
Lithuanian	-	dem-noun	quant-noun
Polish	-	dem-noun	quant-noun
Portuguese	art-noun	dem-noun	quant-noun
Romanian	art-noun*	mixed	quant-noun
Russian	-	dem-noun	quant-noun
Spanish	art-noun	dem-noun	quant-noun
Welsh	art-noun#	noun-dem	quant-noun

**Table 1:** The order of articles, demonstratives and quantifiers with respect to the nominal head according to Dryer (2008, 2013a), Siewerska (1998). \*: only indefinite articles; #: only definite articles.

The two categories of dependents also feature in subsequent studies such as Hawkins (1983) and Dryer (1992, 2009); according to Dryer (1992, 2009), articles and demonstratives figure among the categories of dependents that do not support the tendency for which dependents follow heads in VO languages and precede in OV languages. Rather than treating word order correlations as a “tendency towards consistent ordering of heads and dependents” (Dryer 1992: 82) as in the previous Head-Dependent Theory (HDT), Dryer’s Branching-Direction Theory (BDT) postulates that constituents follow the same position of either the verb or the object in the verb-object ordering; in a sample of 675 languages, later expanded to over 1500 languages in his 2009 article, Dryer (1992) finds that articles follow the same position of verb i.e. are verb patterners, while demonstratives follow the same position of object i.e., are object patterners. This explains why, from the perspective of the HDT, articles

and demonstratives behave like heads and dependents, respectively. Furthermore, as discussed by Dryer (1992: 121-122), this challenges the notion of determiners as a unitary category for demonstratives and articles; as argued in the beginning of this section, distinct categories for articles and demonstratives are also supported by cross-linguistic evidence, whereas languages that mutually exclude articles and demonstratives in the same position, like half of the languages of my sample, are actually typologically rare.

As for the order of demonstratives with respect to numerals and nominal heads, Greenberg's Universal 20 states that:

When any or all of the items (demonstrative, numeral, and descriptive adjective) precede the noun, they are always found in that order. If they follow, the order is either the same or its exact opposite. (Greenberg 1963: 68-69)

Using an undisclosed sample of languages, Cinque (2005) finds that only 14 out of the mathematically possible 24 orderings are actually attested and accounts for this in terms of movement from a universal underlying demonstrative < numeral < adjective < noun order; by contrast, Dryer (2018) claims that in a sample of 576 languages the attested orderings can be justified by describing the involved categories in semantic terms, rather than using syntactic categories as in Cinque's previous approach.

Owing to the confusion around the quantifier category, there are unsurprisingly very few studies on this category; Greenberg (1963) cautiously suggests that Universal 18 might be extended to non-numeral quantifiers, quoting Romance languages as an example.

With respect to the European languages object of this study, qualitative data on the orderings of the three categories can be collected from the World Atlas of Language Structure (WALS: Order of Demonstrative and Noun: Dryer 2013a) and two other works explicitly focusing on European languages (Dryer 2008 and Siewerska 1998); Table 1 reports this data, showing very little variability in the order of the articles, determiners and quantifiers. All languages with an article place it in the prenominal position, demonstratives are prenominal everywhere except for the Celtic languages and quantifiers are prenominal without exception. The only variability is represented by Romanian (ron; Indo-European, Italic) demonstratives, which Dryer classifies as 'mixed' according to a rule of thumb that states that "if the frequency of

the two orders is such that the more frequent order is less than twice as common as the other, the language is treated as lacking a dominant order for that pair of elements” (Dryer 2013b: 371).

As discussed elsewhere (Levshina et al. 2023; Talamo & Verkerk 2022), the type of data presented in Table 1, as well as the literature discussed above, suffers from what Wälchli (2009) addresses as ‘data reduction’; continuous data, such as the frequencies invoked by Dryer in the quotation above, are reduced to categorical values. For instance, Table 1 uses three out of the six original values proposed by Dryer (2013a) for demonstrative-noun order: prenominal, postnominal and mixed; such an approach loses useful information, such as minor patterns that are not captured by methods like Dryer’s rule of thumb or the quantity of variation behind a ‘mixed’ value.

Thanks to the availability of a growing body of computational resources like corpora and automatic parsers, the last decade has witnessed a number of quantitative studies using information-theoretic measures to capture word order variability (Futrell et al. 2015; Naranjo & Becker 2018; Alzetta et al. 2018; Gerdes et al. 2019; Levshina 2019).

However, these studies either do not consider any of the categories considered in the present study or conflate the three categories into a single category (‘nominal modifier’: Naranjo & Becker 2018: 94; ‘determiner’: Levshina 2019: 539). From a methodological perspective, these studies are problematic since (i) they do not provide a convincing match between cross-linguistically valid categories (comparative concepts: Haspelmath 2010; 2018) and instances of categories as found in corpora (tokens: Levshina 2019: 534) and (ii) they are based on non-comparable corpora (treebanks) which vary wildly regarding genre and size (Levshina et al. 2023: 32-34).

The first point stems from the fact that all studies, except for Levshina (2019), use only one type of annotation provided by the treebanks, namely, the syntactic relation between a dependent token and its head. As for the second point, treebanks are collections of manually or semi-automatically annotated texts, which are used to train Natural Language Processing tools, most notably, parsers; these linguistic resources are generally free from annotation errors, however their size is too small to incorporate semantic facts in the analysis. For this reason, Levshina (2021) uses a UD-parsed version of Leipzig Corpora to study, among other things, the relationship between the semantic properties of the verbal arguments and the order of subject and object.

Talamo & Verkerk (2022) introduce comparative concepts to study the order of four modifiers with respect to the nominal head; they show the implementation of these comparative concepts using two layers of annotation as provided by the UD framework, the syntactic relation layer and the Universal Parts-of-Speech layer, and manually-compiled list of lemmata, which are used to capture words from closed categories, such as articles, demonstratives and adpositions. Their approach allows to disentangle the category of determiners, showing, among other things, that the noun-demonstrative order is quite variable in two out of the 11 languages of their sample.

### 3. Data and Methodology

#### 3.1 The CIEP+ corpus and the sample of languages

The corpus used in the present study is the Corpus of Indo-European Prose and More (henceforth: CIEP+), which has been developed from 2019 (Talamo & Verkerk 2022: 184-186). As the name suggests, the corpus currently features a collection of original versions and translations of 17 fiction books and 1 diary in 33 Indo-European languages, with a planned expansion to include translations from other linguistic families.

The criteria of selection of novels are quite simple: availability in a large number of languages and translations in a modern and accessible language variety. Both criteria are met by the so-called *best-seller* books, as their high demand means that are translated in several languages, using a variety that can be understood by the great majority of speakers. Talamo and Verkerk (2022) then included modern classics such as Marquez's *Cien Años de Soledad* (1967) and Eco's *Il nome della Rosa* (1980), as well as contemporary books such as the *Harry Potter* saga (1997-2007) and novels from Coelho, Musso and Süskind; in order to include minority languages, Talamo and Verkerk (2022) have selected less recent books such as Carroll's novels (*Alice's Adventures in Wonderland*: 1865; *Through the Looking-Glass*: 1871) and Saint-Exupery's *Le Petit Prince* (1943).

Since several translations are not (yet) available for all languages, I select for my sample 15 languages featuring the whole set of books (roughly 120,000 sentences or 2 million tokens for each language); these languages belong to the following branches:

- Germanic: Danish (dan; Indo-European, Germanic), Dutch (nld; Indo-European, Germanic), English (eng; Indo-European, Germanic), German (deu; Indo-European, Germanic);
- Hellenic: Greek (ell; Indo-European, Graeco-Phrygian);
- Romance: French (fra; Indo-European, Italic), Portuguese (por; Indo-European, Italic), Romanian (ron; Indo-European, Italic), Spanish (spa; Indo-European, Italic);
- Balto-Slavic: Bulgarian (bul; Indo-European, Balto-Slavic), Czech (ces; Indo-European, Balto-Slavic), Bosnian-Croatian-Serbian (henceforth: BCS; hbs; Indo-European, Balto-Slavic), Lithuanian (lit; Indo-European, Balto-Slavic), Polish (pol; Indo-European, Balto-Slavic), Russian (rus; Indo-European, Balto-Slavic).

The sample is completed by two minority languages belonging to the Celtic branch, Irish (gle; Indo-European, Celtic) and Welsh (cym; Indo-European, Celtic), each featuring five books (roughly 13,000 sentences, or 300,000 tokens).

The corpus is automatically parsed using Stanford Stanza<sup>2</sup> (Qi et al. 2020), which provides the traditional Natural Language Processing steps of sentence splitting, tokenization, lemmatization, as well as morphological and syntactic annotations using the Universal Dependency pre-trained models (de Marneffe et al. 2021).<sup>3</sup>

### ***3.2 Determiners and quantifiers in European language: comparative concepts and the Universal Dependency framework***

A challenge for typological studies is represented by the cross-linguistically valid definitions of the categories under scrutiny. These definitions, or ‘comparative concepts’, should rely on extra-linguistic factors, such as the semantics and the pragmatics of the categories, and should be different from language-specific categories, which are instead addressed as ‘descriptive categories’ (Haspelmath 2018; Croft 2016).

The usage of automatically annotated linguistic resources poses a series of additional problems, including the quality of the annotated data (Levshina et al. 2023: 29-32) and the cross-linguistic consistency of the annotation (Talamo & Verkerk 2022: 180-184).

---

<sup>2</sup> Version 1.3.0. <https://stanfordnlp.github.io/stanza/>

<sup>3</sup> Version 2.8. [https://stanfordnlp.github.io/stanza/available\\_models.html](https://stanfordnlp.github.io/stanza/available_models.html)

In what it follows, I exemplify both the theoretical and methodological matter on the categories of determiners and quantifiers, showing how comparative concepts can be implemented using the Universal Dependency (UD: de Marneffe et al. 2021) framework.

### 3.2.1 Comparative concepts

In this section, I discuss four comparative concepts and verify their adequacy for the 17 languages of my sample. Talamo and Verkerk (2022: Appendix C) propose the following two comparative concepts for the category of articles and demonstratives:

Within a noun phrase, an *ARTICLE*<sup>4</sup> is a word that occupies a fixed position and expresses certain features of the nominal head, namely: (in)definiteness and/or specificity; additionally, an article may also signal deictic and/or anaphoric reference of the nominal head it modifies.

Within a noun phrase, a *DEMONSTRATIVE* is a word that may vary its position and functionally characterizes the nominal head for deictic and/or anaphoric reference.

Central to the definition of *ARTICLE* is the notion of definiteness; in some languages and along the demonstrative-article grammaticalization path definiteness is found together with specificity (Himmelmann 2001: 831-832). Furthermore, deictic and anaphoric reference, which play a major role in the definition of *DEMONSTRATIVE* (Diessel & Coventry 2020: 1-2), are sometimes found “when articles encode meanings typically associated with demonstratives such as visibility or distance from a deictic center” (Himmelmann 2001: 837). Himmelmann attributes these ‘deictic articles’ to Salish and Wakashan languages, as well as to Austronesian languages (Himmelmann 2001: 837; see also Lyons 1999: 53-57); although the Indo-European languages of my sample lack ‘deictic articles’ i.e., dedicated markers for deixis and/or anaphora, their article systems are able to encode the opposition between ‘familiar/unique reference’ and ‘non familiar/non-unique reference’.

---

<sup>4</sup> As a typographic and grammatical convention, I write comparative concepts using *SMALL CAPS* and treat them as singular nouns; language-specific categories such as English adjectives or Bulgarian demonstratives are written uncapitalized and are treated as plural nouns.

So far, I have discussed the two comparative concepts for their functions which, to a certain extent, tend to overlap; since the two comparative concepts are of the hybrid type (Haspelmath 2018: 86), they also include a formal aspect. It is precisely this formal aspect that distinguishes the two comparative concepts: an ARTICLE is a word occupying a fixed position, whereas a DEMONSTRATIVE is a word that may vary its position.

A number of languages from my sample do not fit, with different degrees, the ARTICLE comparative concept; in Balto-Slavic languages, (in)definiteness and specificity are coded by words belonging to other categories, such as adjectives or demonstratives (BCS: Alexander 2006: 20-21, Czech: Naughton 2005: 88; Lithuanian: Ramonienė et al. 2019: 49-52; Polish: Bielec 2012: 27; Russian: Timberlake 2004: 118-119), while in Bulgarian these features are coded by suffixes (Bulgarian: Antova & Boytchinova & Benatova 2002: 41-48). Celtic languages and Romanian meet the ARTICLE comparative concept only partially. In Irish and Welsh, a positive value of definiteness and/or specificity is coded by words occupying a fixed position, while indefinite nouns are bare nouns, cfr. Irish *an fear* ‘the man’ vs. *fear* ‘man/a man’ and Welsh *yr alarch* ‘the swan’ vs. *alarch* ‘swan/a swan’ (Stenson 2020: 183-185; King 2003: 28-30); in Romanian, fixed-positions words mark non-specific nouns and suffixes mark definite nouns, cfr. Romanian *un munte* ‘a mountain’ vs. *munte-le* ‘the mountain’ (Gönczöl-Davies 2008: 34-40).

The DEMONSTRATIVE comparative concept is valid for all languages of the sample, despite the different levels of deixis that a language may encode: (i) only one deictic value, as in French *ce* ‘this/that’ (Batchelor & Chebli-Saadi 2011: 609-612; see also Dryer 2007: 162-163, Diessel & Coventry 2020: 2-3); (ii) two deictic values, as in English *this* vs. *that*; (iii) three deictic values, as in Spanish *este* ‘this’ vs. *ese* ‘that, close to the hearer’ vs. *aquel* ‘that, distant from both the speaker and the hearer’ (Butt & Benjamins & Rodríguez 2019: 87-88).

Quantifiers can be analyzed cross-linguistically according to the following definition:

Within a noun phrase, a QUANTIFIER is a word that may vary its position and functionally characterizes the nominal head for one of the following three types of non-numeral quantification: (i) distributive, (ii) proportional and (iii) amount-term.



The three types of quantification are described in Croft (2022: Glossary) and roughly correspond to the semantic classes discussed by Keenan (2012: 1-4). For the sake of convenience, I give here Croft’s description of these three types:

- “distributive quantifier: a form that specifies the members of the set but treats them individually (that is, the predicate applies to the whole set by virtue of applying to the individual members of the set)”. For instance, English *every* in *Every dog has fleas* indicates that each member of the *dog* set has *fleas*;
- “proportional quantifier: a form that specifies the set of instances as a proportion of the whole set of individuals/tokens of the type, or at least the contextually relevant whole set.” For instance, English *few* in *few people were pleasantly surprised* indicates that a small proportion of the *people* set were pleasantly surprised;
- “amount-term quantifier: a form used to indicate an imprecise quantity for noncountable entities.” For instance, English *some* in *pour me some wine* indicates an imprecise quantity of the mass noun *wine*.

Note that the first two types of quantification may be also expressed through numerals; these are excluded in the current study.

Although the consulted grammars use other terms to indicate the three types of quantification – only a reference grammar of Romanian (Dobrovie-Sorin & Giurgea 2013: 43-45) explicitly discusses proportional quantifiers – all sampled languages have words corresponding to the QUANTIFIER comparative concept.

For instance, the difference between distributive QUANTIFIER and proportional QUANTIFIER is described in Danish by Lundskær-Nielsen and Holmes (2010: 234) as an opposition between specific and universal application of the quantifier, which results in two different constructions.

(1) Danish (dan; Indo-European, Germanic; Lundskær-Nielsen and Holmes 2010: 234)

- a. *Alle spillerne spillede dårligt.*  
all players.DEF play.PST poorly  
‘All players played poorly.’
- b. *Al magt til folket!*  
all.M.SG power.M.SG to people  
‘All power to the people!’

In the example (1a), the *al* ‘all’ QUANTIFIER is followed by the definite form of *spillern* ‘players’, coding the distributive meaning – the action of playing poorly is predicated for each individual player; by contrast, in example (1b), the *al* ‘all’ QUANTIFIER agrees for gender and number with *magt* ‘power’ – the entire proportion of power should be given to the people.

Instances of amount-term QUANTIFIER are described in Danish by Lundskær-Nielsen and Holmes as “[they] can only modify non-count nouns to specify quantity or degree” (2010: 248), as in the following example using *lidt* ‘some’:

(2) Danish (dan; Indo-European, Danish; Lundskær-Nielsen and Holmes 2010: 248)

*Må jeg låne lidt sukker?*  
May I borrow some.N.SG sugar.NCOUNT  
‘May I borrow some sugar?’

The amount-term QUANTIFIER applies to a non-countable entity – sugar – and the strategy employed by Danish is the lack of agreement between *lidt* and *sukker* ‘sugar’.

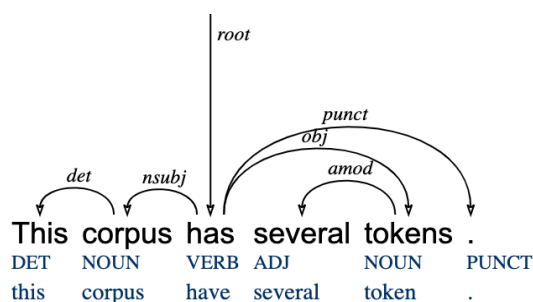
The definition of NOMINAL HEAD involves two comparative concepts, one for the head and the other for the noun; the following comparative concept is based on Croft (definition of the head construction. This definition assumes that word categories are constructions of semantic classes (objects, actions, properties) and information packaging structures (propositional acts: reference, predication and modification; Croft 2001): “Within a noun phrase, a NOMINAL HEAD is the most contentful word that most closely denotes the function of referring as the phrase as a whole.” (Croft 2022: Glossary).

This comparative concept encompasses all instances of HEAD governing a phrase that refers to objects i.e., a referring phrase; all languages of the sample have words corresponding to the definition of NOMINAL HEAD.

### 3.2.2 The UD framework and the List of Lemmata

The UD framework (de Marneffe et al. 2021) consists of several annotation layers spanning different levels of linguistic analysis; the annotation is performed at the token level and within sentence boundaries, with each token getting an incremental identification number (ID) starting from the first token of the sentence. For the purpose of the present study, I will employ two layers of UD annotation: (i) the Universal Parts of Speech (UPOS) layer, which annotates tokens for word categories

using a universal set of 17 tags<sup>5</sup> and (ii) the Relations (Rel) layer, which traces relations between tokens using their ID numbers and a list of 37 syntactic relations.<sup>6</sup> As the name suggests, syntactic relations are conceived of as dependencies, with a token acting as the head and another token acting as the dependent; furthermore, the structure of the annotation is hierarchical, with the sentence predicate acting as the main node (root). This is exemplified in Figure 1, which shows the analysis of the English sentence ‘This corpus has several tokens.’ The main node of the sentence is the ‘has’ token and its immediate dependencies are the two arguments ‘corpus’ and ‘tokens’, which in turn are the head of ‘this’ and ‘several’, respectively. Figure 1 also shows the two layers of UD annotation, in which ‘this’ is annotated as a determiner (UPOS: DET) holding a determination dependency (Rel: det) with ‘corpus’ and ‘several’ is annotated as an adjective (UPOS: ADJ) holding an adjectival modification dependency (Rel: amod) with ‘tokens’.



**Figure 1:** Analysis of the English sentence ‘This corpus has several tokens’ using the UD framework.

While the list of the UPOS tags is closed, the Rel list can be expanded using subtypes of existing relations; for instance, a number of languages uses a subtype of the determiner (det) relation in order to mark the relation between possessive pronouns and their head nouns, labelled ‘det:poss’. Unfortunately, this has led to a proliferation of subtypes, which are quite often specific to a single language or a group of related languages.

Furthermore, the UD framework requires in principle a certain level of consistency between the UPOS and the Relation layer, with determiners (DET) performing determination (det), numerals (NUM) numeral modification (nummod), and so on.

<sup>5</sup> <https://universaldependencies.org/u/pos/all.html>

<sup>6</sup> <https://universaldependencies.org/u/dep/index.html>

Articles, demonstratives and quantifiers are treated as determiners (DET) in the UPOS layer and have a ‘relation determiner’ (det) “between a nominal head and its determiner”.<sup>7</sup>

While the consistency between the UPOS and the Rel layer holds for manually-annotated treebanks, such as the ones available on the project website, it does not for corpora that are automatically parsed using parsers trained on these treebanks. Beside an unavoidable rate of wrong annotations, casual inspection reveals several cases in which the determiner relation is associated with other UPOS tags rather than DET, most notably, adjectives (ADJ) and pronouns (PRON).

In order to reduce the effect of wrong and non-consistent annotations on the quality of data Talamo and Verkerk (2022) propose to add to the UPOS and the Relations annotation layers a third layer, the List of Lemmata (LoL) layer; this layer simply consists of a list of language-specific lemmata, which is compiled using reference grammars and consulting native speakers.<sup>8</sup> The LoL layer is then matched against the lemma annotation layer, which is also provided in the automatic annotation process.

C. Concept	UPOS	Relations	LoL
ARTICLE	DET	det det:predet	articles
DEMONSTRATIVE	DET PRON	det det:predet	demonstratives
QUANTIFIER	DET PRON (ADJ) (NOUN)	det det:predet det:numgov det:nummod (amod) (nmod)	quantifiers
NOMINAL HEAD	NOUN PROP	-	-

**Table 2:** The comparative concepts and their implementation using the UD framework.

Table 1 shows the implementation of the four comparative concepts discussed in the previous section; this implementation is modular i.e., the three layers can be combined or excluded to obtain different results.

<sup>7</sup> <https://universaldependencies.org/u/dep/det.html>

<sup>8</sup> As pointed out by an anonymous reviewer, one may wonder to what extent the UPOS layer is still necessary after the introduction of the the LoL layer. To test this, I computed the entropy by combining the Rel and the LoL layers and keeping the UPOS layer only for the nominal heads; a paired *t-test* shows that the statistical difference between the mean entropy of this combination and of the Rel + UPOS + LoL combination for the three categories is not significant. The mean difference of entropy between the two combinations is .001 for the ARTICLE category, .002 for the DEMONSTRATIVE category and there is no difference for the QUANTIFIER category. As mentioned above, the UPOS layer is however still relevant to capture the nominal heads.

The UPOS tagset does not have specific tags for ARTICLE, DEMONSTRATIVE and QUANTIFIER; all these categories are conflated into the determiner (DET) tag, as described in the UD guidelines for the annotation of determiners;<sup>9</sup> additionally, I have included the PRON tag for DEMONSTRATIVE and QUANTIFIER, as adnominal forms are sometimes mistaken for pronouns by the parser. As for the NOMINAL HEAD, the category is implemented using the NOUN and PROPEN tags.<sup>10</sup>

Along with the determiner relation, I have also included subtypes that are used in at least one language of the sample:

- det:predet, which is used in English to annotate the “relation between the head of an NP and a word that precedes and modifies the meaning of the NP determiner”,<sup>11</sup> as in ‘such a dangerous invention’, where ‘such’ is a predeterminer for ‘a’;
- det:numgov and det:nummod, which are used in BCS, Czech and Polish to mark the difference between quantifiers that do not agree in number with their head (det:numgov) and quantifiers that do agree (det:nummod). For instance, contrast Czech *s několika složkami* ‘with several components’, in which *několika* ‘several’ does not agree for number with *složkami* ‘components’ and Czech *několik let* ‘several years’, in which *několik* agrees for number with *let* ‘years’.

Finally, values given between brackets are used in combination with the LoL layer and only in the implementation of the QUANTIFIER comparative concept; these values include quite broad UPOS tags, adjectives (ADJ) and nouns (NOUN) together with the respective UD Relation, adjectival modification (amod) and nominal modification (nmod).

### 3.2.3 An information-theoretic approach to word-order

Following previous studies on word order (Montemurro & Zanette 2011; Koplenig et al. 2017; Levshina 2019; Talamo & Verkerk 2022), the amount of variability of instances of ARTICLE, DEMONSTRATIVE and QUANTIFIER is captured using information theoretic measures; more specifically, I employ Shannon’s entropy, whose formula is given as follows:

<sup>9</sup> <https://universaldependencies.org/u/pos/all.html#al-u-pos/DET>

<sup>10</sup> <https://universaldependencies.org/u/pos/all.html#al-u-pos/NOUN> and <https://universaldependencies.org/u/pos/all.html#al-u-pos/PROPEN>

<sup>11</sup> <https://universaldependencies.org/en/dep/det-predet.html>

$$H(X) = - \sum_{i=1}^n P(x_i) \log_2 P(x_i)$$

where  $P$  represents the probability of a pattern of word order and  $n$  the possible number of patterns. Since we are concerned here with the order of the nominal head and one of its modifiers,  $n$  is set to 2.

The resulting entropy ranges from 0 i.e., only one of the two possible patterns is attested to 1 i.e., both patterns are attested with the same frequency.

For instance, there are 20 instances of prenominal demonstratives and 978 of postnominal demonstratives in the Greek translation of Marquez's *Cien Años de Soledad*, for a total number of 998 instances of DEMONSTRATIVE. The probability of the DEMONSTRATIVE-NOMINAL HEAD order is 0.02, while the probability of the NOMINAL HEAD-DEMONSTRATIVE ORDER is 0.98; the resulting entropy is obtained by the following equation:

$$H = -(0.02 \times \log_2 0.02 + 0.98 \times \log_2 0.98) = 0.141$$

## 4. Results

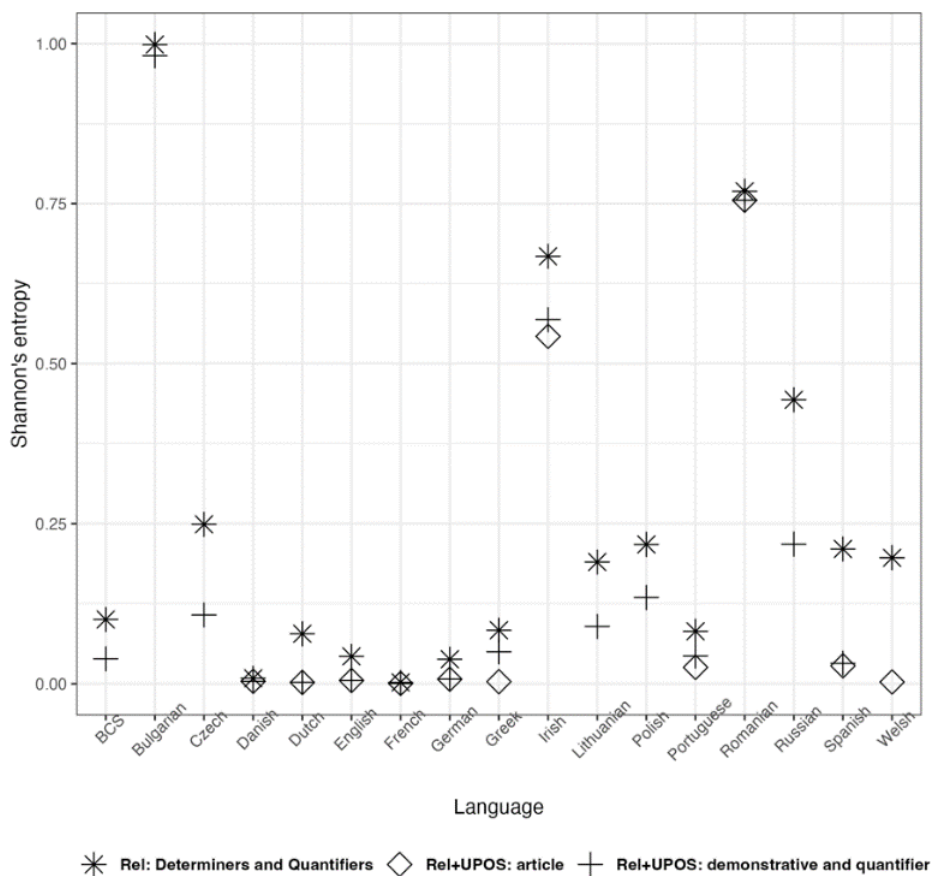
### 4.1. A quantitative overview

Figures 2 and 3 show the entropy of instances of ARTICLE, DEMONSTRATIVE and QUANTIFIER in the 17 languages, as captured by different combinations of annotation layers.

When the Relation layer is used alone, the three categories are indistinguishable from each other<sup>12</sup> and are conflated under the 'Determiners and Quantifiers' category, which is represented by the star shape in Figure 2; this is the methodological approach taken by most of the previous works using UD, as discussed in Sect. 2; this is also the approach capturing the highest level of entropy in all languages, with Bulgarian, Irish and Romanian exceeding the .5 value of entropy.

---

<sup>12</sup> Balto-Slavic languages are an exception here, as they use two Relation subtypes to annotate quantifiers. However, when taken together with the det Relation, the entropy of BCS, Bulgarian, Czech, Lithuanian, Polish and Russian quantifiers is very similar to the entropy of determiners.

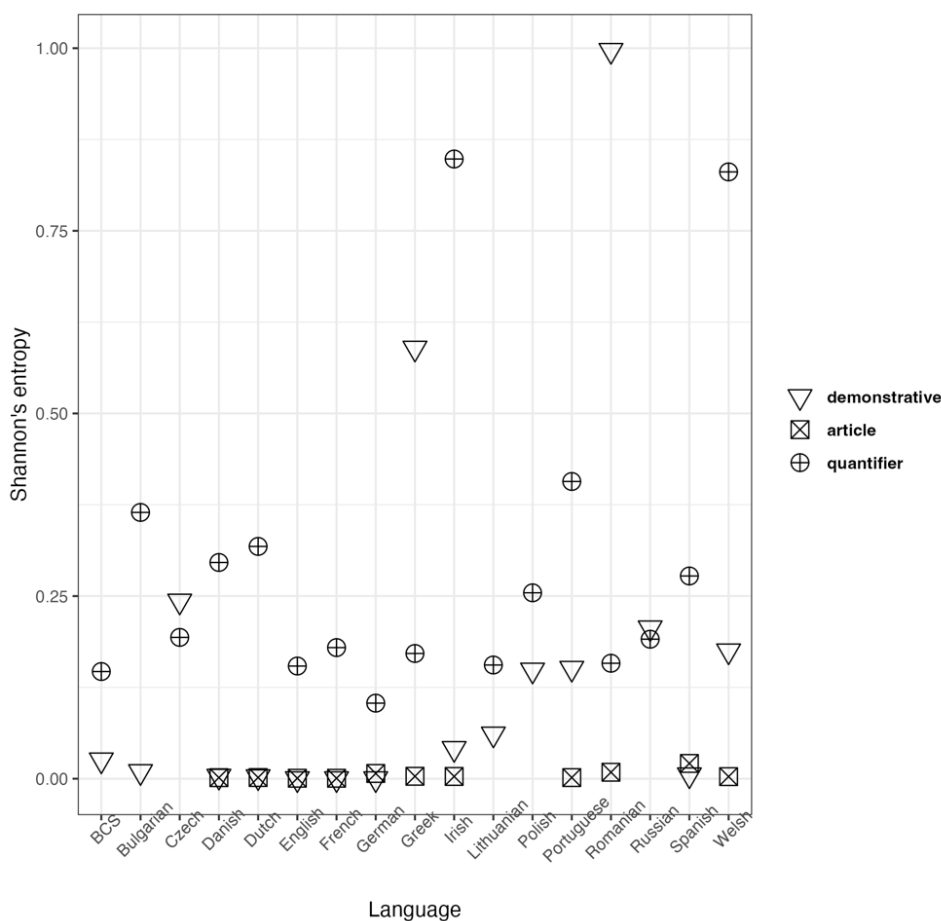


**Figure 2:** The entropy of ‘Determiners and Quantifiers’, as captured by the Relation layer only, the entropy of ARTICLE, as captured by the combination of the Relation and the UPOS layer and the entropy of ‘Demonstratives and Quantifiers’, as captured by the combination of the Relation and the UPOS layer.

The combination with the UPOS layer has the two-fold effect of separating the ARTICLE, which is identified by a diamond shape in Fig. 2, from the ‘demonstratives and quantifiers’ category, which is identified by a plus shape in Fig. 2, and reducing the entropy of all categories. This is particularly clear for languages already below the .5 value, which see their entropy reduced to quasi-null values.

The introduction of the LoL layer, which is combined with the other two layers in Figure 3, unpacks the ‘demonstratives and quantifiers’ category into the DEMONSTRATIVE and QUANTIFIER categories. The high entropy (.981) of the Bulgarian ‘demonstratives and quantifiers’ category is reduced to a quasi-null value (.01) for DEMONSTRATIVE and to .365 for QUANTIFIER, while the moderate entropy (.569) of the

Irish ‘demonstrative and quantifiers’ category raises to .848 for QUANTIFIER and drop to a quasi-null value for ARTICLE (.003).



**Figure 3** - The entropy of ARTICLE, DEMONSTRATIVE and QUANTIFIER, as captured by the combination of the Relations, UPOS and LoL layers.

In sum, there are four languages with entropy values above .500: DEMONSTRATIVE in Greek (.589) and in Romanian (.997), and QUANTIFIER in Irish (.848) and Welsh (.831).

As for Greek demonstratives, a slightly higher entropy is already observed by Talamo and Verkerk (2022) in the same corpus and is justified in terms of information structure. Demonstratives are generally prenominal in Greek, whereas postnominal demonstratives give an emphatic reading to the nominal head (Lascaratou 1998: 164), as in the following example from the Greek translation of Gabriel García Márquez *Cien años de soledad*, in which the rapid aging of Melquíades over a given period of time is emphasized:



(3) Greek (ell; Indo-European, Graeco-Phrygian) Gabriel García Márquez, *Cien años de soledad*, Greek trans. by Maria Palaialogou

Την	εποχή	εκείνη	ο	Μελκιάδες	γερνούσε	φανερά	από	τη
<i>Tin</i>	<i>epochí</i>	<i>ekéini</i>	<i>o</i>	<i>Melkiádes</i>	<i>gernouíse</i>	<i>fanerá</i>	<i>apó</i>	<i>ti</i>
ART.F	time.(F)	that.F	ART.M	Melquiades	age.IPFV.3SG	visibly	from	ART.F
μια	μέρα	στην	άλλη					
<i>mia</i>	<i>mera</i>	<i>stin</i>	<i>alli</i>					
one	day	to.the.F	the.F					

‘At that time Melquíades was visibly aging from one day to the next.’

In the next section, I look more closely to the three other word order patterns with high entropy.

## 4.2. Some patterns of word order with high entropy

### 4.2.1 Variability of the DEMONSTRATIVE in Romanian: information structure or language register?

Instances of DEMONSTRATIVE in Romanian have the highest entropy (.997) across all languages and categories; out of a total frequency of 9086, 4248 demonstratives are prenominal and 4838 are postnominal, meaning that there is almost the same probability for both word order patterns.

According to Giurgea (2013: 160) pre-nominal and post-nominal positions are formally differentiated by the ‘augmented’ form<sup>13</sup> that demonstratives take in post-nominal position: *acest-DEM bărbat-man* ‘this man’ vs. *barbatul-man.DEF acesta-DEM* ‘this man’; the high variability of Romanian demonstratives is evident in prose, where prenominal demonstratives “tend to be used with current discourse topics whereas postnominal demonstratives are preferred for rhematic and contrastive uses” (Giurgea 2013: 163). However, according to the same author, the position of the DEMONSTRATIVE as an information structure marker is lost in the modern-day speaking language and is replaced by an opposition of register: prenominal demonstratives are

<sup>13</sup> Dobrovie-Sorin and Giurgea (2013: 19) account for the difference between non-augmented and augmented forms in terms of phonological constraints.

used in the formal and literary variety, whereas postnominal demonstratives belong to the informal and colloquial Romanian.

Since CIEP+ is a corpus of literary texts, the high variability might be accounted for in terms of information structure; a look at the postnominal demonstratives in Romanian reveals that this strategy mostly codes a cohesion function, namely, anaphoric reference. This is illustrated by an example from *La Jeune Fille et la Nuit*, accompanied by the original sentence in French and the Greek translation; as mentioned above, Greek is the only other language of the sample showing a moderate entropy for the DEMONSTRATIVE, with a function similar to the one described for Romanian.

(4)

a. Romanian (ron; Indo-European, Italic) Guillaume Musso, *La Jeune Fille et la Nuit*, Romanian translation by Constantin Pistea

<i>Știam</i>	<i>foarte</i>	<i>bine</i>	<i>că</i>	<i>imaginea</i>	<i>aceasta</i>	<i>răspundea</i>	<i>aspirației</i>
know.PST.1SG	very	well	that	image.(F)	DEM.F	answer.PST.3SG	aspiration
<i>mele</i>	<i>acea</i>	<i>vreme.</i>					
My	DEM.F	time.(F)					

‘I knew very well that this image answered my aspiration at that time.’

b. Greek (ell; Indo-European, Graeco-Phrygian) Guillaume Musso, *La Jeune Fille et la Nuit*, Greek translation by Maria Gourniezaki

Ἔξερα	ὅτι	αὐτή	ἡ	εἰκόνα	ἀνταποκρινόταν	στις
<i>Íxera</i>	<i>óti</i>	<i>aftí</i>	<i>i</i>	<i>eikóna</i>	<i>antapokrinótan</i>	<i>stis</i>
know.PST.1SG	that	DEM.F	ART.F	image.(F)	answer.PST.3SG	to.the.F
προσδοκίες	εκείνης	της	εποχής.			
<i>prosdokíes</i>	<i>ekeínis</i>	<i>tis</i>	<i>epochís</i>			
expectation	DEM.F	ART.F	time.(F)			

‘I knew that this image met the expectations of that time.’

c. French (fra; Indo-European, Italic) Guillaume Musso, *La Jeune Fille et la Nuit*, original French text

<i>Je</i>	<i>savais</i>	<i>très</i>	<i>bien</i>	<i>que</i>	<i>cette</i>	<i>image</i>	<i>répondait</i>	<i>à</i>
I	know.1SG.PST	very	well	that	DEM.F	image.(F)	anwer.3SG.PST	to

*mon aspiration d' alors.*  
*my aspiration of that.time*

'I knew very well that this image answered my aspiration at the time.'

With respect to the second DEMONSTRATIVE, Romanian and Greek are aligned, in that they both translate with a prenominal distal demonstrative the French expression *d'alors* 'at that time'. By contrast, the French noun phrase *cette image* 'this image', which refers to a previously described image, is translated in Greek using the proximative demonstrative *αυτή aftí* 'this' in its unmarked (prenominal) position, while the Romanian translator uses the proximative demonstrative *acea* 'this' in postnominal position.

The high variability of the DEMONSTRATIVE might be also attributed to the large number of dialogues featured in several texts from CIEP+. Here, it is assumed that dialogues mimic, to a certain extent, the modern-day speaking language. If we look to the distribution of pre-nominal and post-nominal demonstratives across the texts of Romanian CIEP+ (Table 2) we have a partial confirmation of this hypothesis. For instance, the Harry Potter saga is aimed at a young audience, thus featuring a less formal language variety; the seven books from this saga contains 1536 prenominal demonstratives and 2429 postnominal demonstratives. A slight tendency toward a postnominal position of the demonstrative (231 pre-nominal vs. 279 post-nominal demonstratives) is also observed in the Romanian translation of Musso's *La jeune Fille et la Nuit*, which belongs to a literary subgenre - the *novel noir* - that traditionally features a high amount of dialogues.

	Prenominal	Postnominal
<i>Cien años de soledad</i>	729	231
<i>Adventures of Alice in Wonderland</i>	56	74
<i>Het Achterhuis</i>	209	247
<i>O Alquimista</i>	125	103
<i>La jeune fille et la nuit</i>	231	279
<i>Il nome della rosa</i>	721	801
<i>Das Parfum</i>	254	96
<i>Le Petit Prince</i>	30	50
<i>Harry Potter and the Philosopher's Stone</i>	159	179
<i>Harry Potter and the Chamber of Secrets</i>	142	166
<i>Harry Potter and the Prisoner of Azkaban</i>	122	285
<i>Harry Potter and the Goblet of Fire</i>	263	362

	Prenominal	Postnominal
<i>Harry Potter and the Order of the Phoenix</i>	356	574
<i>Harry Potter and the Half-Blood Prince</i>	298	438
<i>Harry Potter and the Deathly Hallows</i>	196	425
<i>Through the Looking Glass</i>	74	37
<i>O Zahir</i>	204	188
Βίος και Πολιτεία του Αλέξη Ζορμπά (Víos kai Politeía tou Aléxi Zorbá)	79	303

**Table 2.** The distribution of the position of Romanian demonstratives across the 18 books of CIEP+.

Finally, data from the largest UD treebank for Romanian (RoReRef: Barbu Mititelu et al. 2016) confirms that formal Romanian has a preference for the prenominal position for DEMONSTRATIVE; the entropy observed for DEMONSTRATIVE in this corpus, which features several genres such as law, medical, academic writing, is lower (.551), with 848 demonstratives in prenominal position and 124 in postnominal position.

#### 4.2.2 Variability of the QUANTIFIER in Celtic languages: artifacts or actual variation?

The entropy of the QUANTIFIER is high for both Irish and Welsh; Irish has a value of .848, with 509 quantifiers in prenominal position and 1343 in postnominal position; Welsh a value of .831, albeit with fewer attested quantifiers i.e., 104 prenominal and 292 postnominal. In order to compare this data with other languages from the sample, it should be kept in mind that Irish and Welsh have only five books from the 18 featured in CIEP+, resulting in approximately one ninth of the total sentences, or one seventh of the total tokens. Furthermore, the performance of the parser for Irish and Welsh is lower with respect to the other languages of the sample;<sup>14</sup> accordingly, I additionally computed the frequency and the entropy of Irish and Welsh QUANTIFIER on the two UD treebanks available for these languages, UD Irish IDT and UD Welsh CCG, which are – at least partially – manually annotated. The entropy of quantifier in the two UD treebanks is higher than the entropy found for CIEP+: .97 for UD Irish IDT and .99 for UD Welsh CCG.

According to Stenson (2020: 188), the position of QUANTIFIER in Irish is lexically determined, as “most precede the noun in the same position as articles and pronominal possessors, but a few follow”. Some quantifiers listed by Stenson are not

<sup>14</sup> See <https://stanfordnlp.github.io/stanza/performance.html> for a comparison between the performance of Stanza’s pretrained models.

considered here, as they are either word combinations such as *go leor* ‘many, much, a lot’ and *ar fad* ‘all’, or are annotated by the parser as heads of nominal phrases, especially in prenominal position (see below).

Lemma	CIEP +		UD Irish IDT	
	Prenominal	Postnominal	Prenominal	Postnominal
<i>beagán</i> ‘a little’	0	15	0	0
<i>céanna</i> ‘same’	0	130	1	75
<i>cuid</i> ‘some, part of’	53	70	81	88
<i>cúpla</i> ‘a couple, a few’	5	5	1	2
<i>éigin</i> ‘some’	0	333	0	30
<i>eile</i> ‘other, another’	2	674	0	285
<i>gach</i> ‘every’	376	3	229	0
<i>gach uile</i> ‘every’	2	0	17	0
<i>mórán</i> ‘many/much’	2	2	1	1
<i>roinnt</i> ‘some, a few’	1	1	4	5
<i>tuilleadh</i> ‘more’	0	6	0	1
<i>uile</i> ‘every’	68	104	16	44

**Table 3.** The distribution of Irish quantifiers at the lemma level and according to their position in CIEP + and in UD Irish IDT.

The distribution of the Irish quantifiers in CIEP + and in UD Irish IDT (Table 3) mostly reflects what Stenson (2020: 189-192) describes in her grammar, with a clear distinction between prenominal and postnominal quantifiers; an exception is represented by *beagán* ‘a little’ and *tuilleadh* ‘more’, which are described as prenominal quantifiers but appears only postnominally, and by some quantifiers appearing in both positions, most notably *cuid* ‘some, part of’ and *uile* ‘all’.

It seems, then, that a certain level of word order variability is also attested at the individual lemma level. However, a closer look to the token of these quantifiers reveals the fictitious nature of this variation, with the possible exception of *cuid*.

Many instances of *beagán* and *tuilleadh* are not captured by the implementation of the QUANTIFIER comparative discussed in Sect. 3.2.2; when they appear in prenominal position, the two Irish quantifiers are annotated both in CIEP + and in the UD treebank as heads of nominal phrases; furthermore, the instances of postnominal quantifiers of *beagán* and *tuilleadh* are words modifying verbs or adjectives. Instances of *uile* in prenominal position are actually the two pronouns *uile dune* ‘everyone’ and *uile rud* ‘everything’, as well as other fixed expressions such as *uile cineál* ‘all kinds’

and *uile bhlas* ‘all flavours’. According to Thurneysen (1990: 229), in Old Irish the position of *uile* is variable and the above-mentioned forms are allegedly relics of previous variability. Finally, *cuid*, along its usage as a prenominal quantifier, is also employed in possessive constructions, following pronominal possessors and preceding possessed objects, usually mass or plural nouns, e.g., *mo.1SG chuid airgid* ‘my money’.

(5) Irish (gle; Indo-European, Celtic) J.K. Rowling, *Harry Potter and the Philosopher’s Stone*, Irish trans. By Máire Nic Mhaoláin

<i>Leag</i>	<i>Mr</i>	<i>Ollivander</i>	<i>méar</i>	<i>fada</i>	<i>bhán</i>	<i>dá</i>	<i>chuid</i>	<i>ar</i>
laid	Mr	Ollivander	finger	long	white	3SG.POSS	CUID	on
<i>an</i>	<i>splanc</i>	<i>thintri</i>	<i>ar</i>	<i>éadan</i>	<i>Harry</i>			
the	flash	lightning	on	face	Harry			

‘Mr. Ollivander laid his white long finger on the flash of lightning on Harry’s face.’

The parser treats *cuid* as the postnominal modifier of the possessed object; for instance, in example (5) *cuid* is parsed as a nominal modifier (nmod) of *méar* ‘finger’; this behavior is perhaps triggered by possessive constructions in which *cuid* is extended to non-pronominal possessors, but with a reversed word order, namely possessed object-*cuid*-possessor, as in example (6). This pattern may originate from a construction which “indicate(s) membership in a specific group” (Stenson 2020: 191) as in *Is inealtóir de chuid Aer Lingus é* ‘He is an engineer from Aer Lingus’.<sup>15</sup>

(6) Irish (gle; Indo-European, Celtic) Saint-De-Exupery, *Le Petit Prince*, Irish trans. By Breandan O Doibhlin

<i>Léiríodh</i>	<i>dom</i>	<i>an</i>	<i>rún</i>	<i>eile</i>	<i>seo</i>	<i>de</i>	<i>chuid</i>	<i>an</i>	<i>phrionsa</i>	<i>bhig.</i>
Show.PASS	me	the	secret	other	DEM	of.it	CUID	the	prince	little

‘I was shown this other secret of the Little Prince.’

As for Welsh, King (2003) describes the position of quantifiers as prenominal, with the *o* preposition preceding the noun in some cases e.g., *chwanag o de* ‘some tea’ but not in others: *sawl anifail* ‘several animals’ (125-126). Data from CIEP + and UD Welsh

<sup>15</sup> In an earlier draft of this paper, following Stenson (2020: 191), I have referred to *cuid* as a quantifier with partitive meaning; an anonymous reviewer suggests that its meaning might be better addressed as a part-whole relation, which is consistent with the group membership meaning discussed here.

CCG seem to contradict this statement, with more quantifiers in postnominal position than in prenominal position.

Lemma	CIEP +		UD Welsh CCG	
	Prenominal	Postnominal	Prenominal	Postnominal
<i>digon</i> ‘enough’	11	34	0	1
<i>gormod</i> ‘too much/many’	0	3	0	0
<i>llawer</i> ‘a lot, much/many’	6	26	2	4
<i>peth</i> ‘some’	28	169	2	8
<i>rhagor</i> ‘more’	3	16	2	2
<i>sawl</i> ‘several’	38	1	13	1
<i>tipyn</i> ‘a (little) bit’	1	6	0	0
<i>ychedig</i> ‘a (little) bit, a few’	17	37	5	3

**Table 4.** The distribution of Welsh quantifier at the lemma level and according to their position in CIEP + and in UD Welsh CCG.

However, these data should be handled carefully; the implementation of the QUANTIFIER category is at the same time too broad and too narrow. It is too broad as the nominal modification (nmod) relation captures several instances in which a word is not constructed as a quantifier; for instance, *peth* is used as the prenominal quantifier ‘some’ only colloquially (King 2003: 128-129), and is largely attested in Welsh CIEP + (169 occurrences) in postnominal position with its original meaning ‘thing’; it is too narrow as, like in Irish, quantifiers are treated as heads of nominal phrases. Furthermore, the Welsh parser, probably because of its small training corpus, performs quite poorly, with several adjectives and/or verbs taken as nominal heads, an issue already encountered for some of the Irish quantifiers; Heinecke and Tyers (2019: 28-29) evaluate a parser trained on their treebank as “comparable with similar sized treebanks”, however concluding that “the current 601 sentences may be a start, but do not cover enough examples to train a robust dependency parser”. The current size of UD Welsh CCG does not also allow for meaningful comparison with the CIEP + data, as the frequency of the Welsh quantifiers is admittedly too low.

Differently from Irish, where there is sound evidence for lexically-based variation with some functionally and diachronically justified exceptions, data for Welsh quantifiers are either too noisy or too small to draw conclusions and the reported high entropy should, for now, be considered an artifact.

## 5. Conclusion

In the present paper I have analyzed the word order variation of articles, demonstratives and quantifiers in 17 European languages; these categories are notoriously hard to define cross-linguistically, and their variation has been poorly investigated in both qualitative and quantitative typological studies on word order.

Following previous quantitative studies, I treat word order variation as a continuous measure rather than a categorical one. However, with respect to previous studies, the methodology of the present paper aims to achieve a better match between typologically-adequate comparative concepts (category-like comparative concepts: Haspelmath 2018) and token-based comparative concepts, here represented by translations from the parallel Corpus of Indo-European Prose (CIEP). Following Talamo and Verkerk (2022), I combine the syntactic and part-of-speech layers of UD annotation with manually-crafted lists of lemmata in order to have a better representation of these categories at the token level.

The proposed methodology allows researchers to disentangle the entropy of the ‘determiners and quantifiers’ category, as captured by the single ‘det’ syntactic relations of the UD framework, into its three different components of ARTICLE, DEMONSTRATIVE and QUANTIFIER. Whereas the category of ARTICLE shows, as expected, no variation, DEMONSTRATIVE shows moderate-to-high values of entropy in Greek and Romanian, and the entropy of QUANTIFIER is high in Celtic languages; a closer look to these word order patterns reveals that the order of demonstratives in Romanian can be accounted for by principles of information structure, as previously shown by Talamo and Verkerk (2022) for Greek. The high entropy of Irish quantifiers is justified on lexical basis, while the high entropy of Welsh quantifiers turns out to be an artifact produced by the computational implementation of the QUANTIFIER category as well as by wrong annotations, which is due to the small training corpus available for Welsh.

The analysis of messy categories such as determiners and quantifiers is a testing ground for typological investigation using computational tools, such as the Stanza parser, UD models and parallel corpora; while these computational tools prove adequate for such a complex task in high-resource languages, the results for low-resource languages such as Welsh are not yet satisfactory enough. However, the development of new NLP tools and the extension of the UD framework to low-resource languages are rapidly evolving, and it will soon be possible to study (formerly) low-resource languages using quantitative typological methods such as the one discussed here.



## Acknowledgements

Earlier versions of this work were presented at the 6<sup>th</sup> edition of the Using Corpora in Contrastive and Translation Studies conference (Bertinoro, September 2021) and at the 4th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (Seattle, July 2022). I thank the audience of both conferences for their comments. I am also indebted to Annemarie Verkerk, who has gone through the manuscript several times and provided precious help, as well as to two anonymous reviewers for improving this paper. All remaining errors are mine. This work is supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Project-ID 232722074 – SFB 1102.

To Dada and Giorgio, in loving memory.

## Abbreviations

1 = 1 <sup>st</sup> person	LOC = locative	PRON = pronoun
3 = 3 <sup>rd</sup> person	M = masculine	PROP N = proper noun
ART = article	N = neuter	PST = past
ADJ = adjective	NCOUNT = non countable	PUNCT = punctuation
DEF = definite	NUM = numerals	SG = singular
DEM = demonstrative	PART = partitive	UPOS = Universal Part of
DET = determiner	PASS = passive	Speech
F = feminine	POSS = possessive	
IMPF = imperfective	PL = plural	

## References

- Alexander, Ronelle. 2006. *Bosnian, Croatian, Serbian, a Grammar*. Madison: The University of Wisconsin Press.
- Alzetta, Chiara & Felice Dell’Orletta & Simonetta Montemagni & Giulia Venturi. 2018. Universal Dependencies and Quantitative Typological Trends. A Case Study on Word Order. In Nicoletta Calzolari et al., *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA). url: <https://www.aclweb.org/anthology/L18-1719>.

- Antova, Evgenia & Ekaterina Boytchinova & Poly Benatova. 2002. *A short grammar of Bulgarian for English speaking learners* (2 ed.). Sofia: ABM Komers.
- Arnaiz, Alfredo R. 1998. The main word order characteristics of Romance. In Siewierska, Anna (ed.), *Constituent Order in the Language of Europe*, 47-74. Berlin: Mouton de Gruyter.
- Barbu Mititelu, Verginica & Elena Irimia & Cenel-Augusto Perez & Radu Ion & Radu Simionescu & Martin Popel. 2016. *UD Romanian RoRefTrees*. [https://github.com/UniversalDependencies/UD\\_Romanian-RRT](https://github.com/UniversalDependencies/UD_Romanian-RRT).
- Batchelor, Ronald E. & Malliga Chebli-Saadi. 2011. *A Reference Grammar of French*. Cambridge: Cambridge University Press.
- Bielec, Dana. 1998. *Polish: An Essential Grammar*. London & New York: Routledge.
- Butt, John & Carmen Benjamin & Antonia Moreira Rodríguez. 2019. *A New Reference Grammar of Modern Spanish* (6 ed.). London & New York: Routledge.
- Cinque, Guglielmo. 2005. Deriving Greenberg's Universal 20 and Its Exceptions. *Linguistic Inquiry* 36(3). 315–332.
- Croft, William. 2001. *Radical Construction Grammar: Syntactic Theory in Typological Perspective*. Oxford: Oxford University Press.
- Croft, William. 2016. Comparative concepts and language-specific categories: Theory and practice. *Linguistic Typology* 20(2). 377–393.
- Croft, William. 2022. *Morphosyntax: constructions of the world's languages*. Cambridge: Cambridge University Press.
- Diessel, Holger & Kenny R. Coventry 2020. Demonstratives in Spatial Language and Social Interaction: An Interdisciplinary Review. *Frontiers in psychology* 11.
- Dobrovie-Sorin, Carmen & Ion Giurgea (eds.). 2013. *A Reference Grammar of Romanian: Volume 1: The Noun Phrase*. Amsterdam / Philadelphia: John Benjamins Publishing Company.
- Dobrovie-Sorin, Carmen & Ion Giurgea. 2013. Introduction: Nominal features and nominal projections. In Carmen Dobrovie-Sorin & Ion Giurgea (eds.), *A Reference Grammar of Romanian: Volume 1: The Noun Phrase*, 1-48. Amsterdam / Philadelphia: John Benjamins Publishing Company.
- Dryer, Matthew S. 1992. The Greenbergian Word Order Correlations. *Language* 68(1). 81-138.
- Dryer, Matthew S. 1998. Aspects of Word Order in the Languages of Europe. In Anna Siewierska (ed.), *Constituent Order in the Languages of Europe*, 283-319. European Science Foundation Language Typology series. Berlin: Mouton de Gruyter.

- Dryer, Matthew S. 2007. Lexical nominalization. In Timothy Shopen (ed.), *Language Typology and Syntactic Description. Grammatical Categories and the Lexicon (Second Edition)*, 151–205. Cambridge: Cambridge University Press.
- Dryer, Matthew S. 2009. The Branching Direction Theory of Word Order Correlations Revisited. In Sergio Scalise & Elisabetta Magni & Antonietta Bisetto (eds.), *Universals of Language Today*, 185-207. Berlin: Springer.
- Dryer, Matthew S. 2013a. Order of Demonstrative and Noun. In Matthew S. Dryer & Martin Haspelmath (eds.), *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. url: <https://wals.info/chapter/88>
- Dryer, Matthew S. 2013b. Determining Dominant Word Order. In Matthew S. Dryer & Martin Haspelmath (eds.), *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. url: <https://wals.info/chapter/s6>
- Dryer, Matthew S. 2018. The order of demonstrative, numeral, adjective, and noun. *Language* 94(4). 798-833.
- Futrell, Richard & Kyle Mahowald & Edward Gibson. 2015. Quantifying Word Order Freedom in Dependency Corpora. In *Proceedings of the Third International Conference on Dependency Linguistics* (Depling 2015), 91–100. Uppsala, Sweden. Uppsala University, Uppsala, Sweden.
- Gerdes, Kim & Sylvain Kahane & Xinying Chen. 2019. Rediscovering Greenberg's Word Order Universals in UD. In *Proceedings of the Third Workshop on Universal Dependencies* (UDW, SyntaxFest 2019). Paris, France: Association for Computational Linguistics, 124–131. doi: 10.18653/v1/W19-8015. url: <https://www.aclweb.org/anthology/W198015>.
- Giurgea, Ion. 2013. The syntax of determiners and other functional categories. In Carmen Dobrovie-Sorin & Ion Giurgea (eds.), *A Reference Grammar of Romanian: Volume 1: The Noun Phrase*, 97-174. Amsterdam / Philadelphia: John Benjamins Publishing Company.
- Gönczöl-Davies, Ramona. 2008. *Romanian: an essential grammar*. London & New York: Routledge.
- Greenberg, Joseph H. 1963. Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements. In Joseph H. Greenberg (ed.), *Universals of Human Language*, 73-113. Cambridge, Mass: MIT Press.

- Greenberg, Joseph H. 1978. Generalizations about Numeral Systems. In Joseph H Greenberg & Charles A. Ferguson & Edith A. Moravcsik (eds.), *Universals of Human Language*, Volume 3: Word Structure, 249–295. Stanford: Stanford University Press.
- Haspelmath, Martin 2010. Comparative concepts and descriptive categories in cross-linguistic studies. *Language* 86(3), 663–687.
- Haspelmath, Martin. 2018. How comparative concepts and descriptive linguistic categories are different. In Daniël Van Olmen & Tanja Mortelmans & Frank Brisard (eds.), *Aspects of Linguistic Variation*, 83-114. Berlin: De Gruyter.
- Hawkins, John A. 1983. *Word Order Universals*. New York: Academic Press.
- Heinecke, Johannes & Francis M. Tyers. 2019. Development of a Universal Dependencies treebank for Welsh. In *Proceedings of the Celtic Language Technology Workshop*. Dublin: European Association for Machine Translation, 21-31. url: <https://www.aclweb.org/anthology/W19-6904>
- Himmelman, Nikolaus P. 2001. Articles. In Martin Haspelmath & Ekkehard König & Wulf Oesterreicher & Wolfgang Raible (eds.), *Language Typology and Language Universals* Vol. 1, 831-841. Berlin: Walter de Gruyter.
- Holmberg, Andres & Jan Rijkhoff. 1998. Word order in the Germanic languages. In, Anna Siewerska (ed.), *Constituent Order in the Language of Europe*, 75-104. Berlin: Mouton de Gruyter.
- Ioup, Georgette. 1975. Some universals for quantifier scope. In John Kimball (ed.), *Syntax and Semantics*, vol. 5, Academic Press, New York.
- Keenan, Edward L. 2012. The Quantifier Questionnaire. In Edward Keenan & David Paperno (eds.), *Handbook of Quantifiers in Natural Language*. Studies in Linguistics and Philosophy, vol 90, 1-20. Dordrecht: Springer.
- King, Gareth. 2003. *Modern Welsh: A Comprehensive Grammar*. London & New York: Routledge.
- Koplenig, Alexander & Peter Paperno & Sascha Wolfer & Carolyn Müller-Spitzer. 2017. The statistical trade-off between word order and word structure - large-scale evidence for the principle of least effort. *PLoS ONE* 12(3)
- Lyons, Christopher. 1999. *Definiteness*. Cambridge: Cambridge University Press.
- Lascaratou, Chryssoula. 1998. Basic characteristics of Modern Greek word order. In Anna Siewerska (ed.), *Constituent Order in the Language of Europe*, 151-171. Berlin: Mouton de Gruyter.
- Levshina, Natalia. 2019. Token-based typology and word order entropy: A study based on Universal Dependencies. *Linguistic Typology* 23(3). 533–572.

- Levshina, Natalia. 2021. Cross-linguistic trade-offs and causal relationships between cues to grammatical subject and object, and the problem of efficiency-related explanations. *Front. Psychol* 12.
- Levshina, Natalia & Savithry Namboodiripad & Marc Allasonnière-Tang & Mathew A. Kramer & Luigi Talamo & Annemarie Verkerk & Sasha Wilmoth & Gabriela Garrido Rodriguez & Timothy Gupton & Evan Kidd & Zoey Liu & Chiara Naccarato & Rachel Nordlinger & Anastasia Panova & Natalia Stoynova. 2023. Why we need a gradient approach to word order. *Linguistics* 61(4). 825-883.
- Lundskær-Nielsen, Tom, & Philip Holmes. 2010. *Danish: A comprehensive grammar*. 2nd edn. Cambridge: Cambridge University Press.
- de Marneffe, Marie-Catherine & Christopher D. Manning & Joakim Nivre & Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics* 47(2). 255-308.
- Montemurro, Marcelo. A., & Damián H. Zanette 2011. Universal Entropy of Word Ordering Across Linguistic Families. *PLoS ONE* 6(5).
- Naranjo, Matías Guzmán, & Laura Becker. 2018. Quantitative word order typology with UD. In Dag Haug & Stephan Oepen & Lilja Øvrelid & Marie Candito & Jan Hajič (eds.), *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018)*, 91-104. Oslo: Linköping Electronic Conference Proceedings.
- Naughton, James. 2005. *Czech: an essential grammar*. London & New York: Routledge.
- Qi, Peng & Yuhao Zhang & Yuhui Zhang & Jason Bolton & Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In Dan Jurafsky & Joyce Chai & Natalie Schluter & Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. url: <https://aclanthology.org/2020.acl-demos.14.pdf>
- Ramonienė, Meilutė & Joana Pribušauskaitė & Jogilė T. Ramonaitė & Loreta Vilkienė. 2019. *Lithuanian: A Comprehensive Grammar*. London & New York: Routledge.
- Siewerska, Anna. (eds.). 1998. *Constituent Order in the Languages of Europe*. Berlin: Mouton de Gruyter.
- Siewierska, Anna & Ludmila Uhlířová. 1998. Word order in the Slavic languages. In Anna Siewerska (ed.), *Constituent Order in the Language of Europe*, 105-149. Berlin: Mouton de Gruyter.
- Stenson, Nancy. 2020. *Modern Irish: A Comprehensive Grammar*. London & New York: Routledge.
- Talamo, Luigi & Annemarie Verkerk. 2022. A new methodology for an old problem: A corpus-based typology of adnominal word order in European languages. *Italian Journal of Linguistics* 34(1). 171-226.

- Tallerman, Maggie. 1998. Word order in Celtic. In Anna Siewerska (ed.), *Constituent Order in the Language of Europe*, 21-45. Berlin: Mouton de Gruyter.
- Thurneysen, Rudolf. 1990. *A Grammar of Old Irish, revised and enlarged edition, translated from the German by Daniel A. Binchy and Osborn Bergin*. Dublin: Dublin Institute for Advanced Studies.
- Timberlake, Alan. 2004. *A reference grammar of Russian*. Cambridge: Cambridge: Cambridge University Press.
- Wälchli, Bernhard. 2009. Data reduction typology and the bimodal distribution bias. *Linguistic Typology* 13(1). 77-94.

**CONTACT**

luigi.talamo@uni-saarland.de

# The Dimensions of Morphosyntactic Variation: Whorf, Greenberg and Nichols were right

SIVA KALYAN<sup>1</sup> & MARK DONOHUE<sup>2</sup>

<sup>1</sup>THE UNIVERSITY OF QUEENSLAND & THE AUSTRALIAN NATIONAL UNIVERSITY,

<sup>2</sup>THE LIVING TONGUES INSTITUTE FOR ENDANGERED LANGUAGES

Submitted: 27/06/2023 Revised version: 23/10/2023

Accepted: 27/11/2023 Published: 27/12/2023



Articles are published under a Creative Commons Attribution 4.0 International License (The authors remain the copyright holders and grant third parties the right to use, reproduce, and share the article).

## Abstract

We examine a database of 3089 languages coded for 351 morphosyntactic features, including almost all of the morphosyntactic features found in *The World Atlas of Language Structures* (Dryer & Haspelmath 2013). We apply Factor Analysis of Mixed Data, and determine that the main dimensions of global morphological variation involve (1) word order in clauses and adpositional phrases, (2) head- versus dependent-marking, and (3) a set of features that show an east-west distribution. We find roughly the same features clustering in similar dimensions when we examine individual macro-areas, thus confirming the universal relevance of these groupings of features, as encapsulated in well-known implicational universals. This study confirms established insights in linguistic typology, extending earlier research to a much larger set of languages, and uncovers a number of areal patterns in the data.

**Keywords:** typology; word order; morphosyntax; head/dependent-marking, computational linguistics; areality.

## 1. Introduction

The goal of much early work in linguistic typology was to categorise languages into distinct overall “types”, under the assumption that once the type of a language was known, a large number of its features could then be predicted – in effect, a holistic

approach to typology (Croft 2003: 31, Humboldt 1836, von der Gabelentz 1901). While the prominence of such taxonomic work has receded in favour of detailed studies of individual features (or clusters of related features), the time is ripe to resuscitate such work in light of the increased amount of linguistic data that has become available. In this paper, we use Factor Analysis of Mixed Data (FAMD; Pagès 2004) to determine the main dimensions of global typological variation in morphosyntax – that is, the features that are most helpful for dividing the world’s languages into different morphosyntactic types. We use a large database of morphosyntactic features, built and substantially expanded from the *WALS* dataset (Dryer & Haspelmath 2013; see Appendix 2), with both features and languages chosen independently of the present study. Upon examining the principal dimensions emerging from this analysis, we find that in each case, they bring together a group of features that has previously been proposed as a basis for the global typological classification of languages: in particular, we find groups of features relating to (mostly clausal) word order (proposed as a basis for classification by Greenberg 1963, refined by Dryer 1992, 2013, and other publications; Dimensions 1 and 4; sections 3.1.1 and 3.1.4), head/dependent marking (Nichols 1986; Dimensions 1 and 2; sections 3.1.1 and 3.1.2), and a set of features that define a global east-west split, one end of which is mainly present in the Old World, and the other end of which is dominant in the “Circum-Pacific” region (Bickel & Nichols 2006, Bugaeva et al. 2021; section 3.1.3).

The remainder of this paper is structured as follows. In section 2, we introduce our dataset (*World\_morphosyntax*), the metadata used as controls, and the technique of FAMD. In section 3, we present our results, for the global set of languages as well as for each macro-area individually; we find that the groupings of features that emerge in the global analysis recur in individual macro-areas. In section 4 we examine some of the negative results, discussing the kinds of features that have the smallest contribution to the global analysis. Finally, in section 5 we summarise our findings, and suggest directions for future research. A number of appendices illustrate the distribution of the individual features that emerge as relevant to defining the four dimensions described in section 3.<sup>1</sup>

---

<sup>1</sup> Appendices are available as supplementary material at:

<https://typologyatcrossroads.unibo.it/article/view/17482/17369>



## 2. Data and methodology

The World\_morphosyntax dataset consists of 3089 language varieties (rows—see Appendix 10), representing 2,693 distinct iso 639-3 codes,<sup>2</sup> coded for 351 morphosyntactic features (columns). It is curated by Mark Donohue (see Appendix 1), and has been developed since 2010. The original database was based on the most robustly coded languages and features from the *World Atlas of Language Structures* (Dryer & Haspelmath 2013). The dataset includes most of the 155 morphosyntactic features in *WALS*, with unordered multivalued features recoded as sets of binary features. (See Appendix 2 for a full listing and description of the features in the World\_morphosyntax dataset.) For instance, *WALS* feature 57A (‘Position of Pronominal Possessive Affixes’) is coded as a single feature in *WALS*, consisting of the features listed in (1).

(1) *WALS* feature 57A ‘Position of Pronominal Possessive Affixes’

- a. Possessive prefixes
- b. Possessive suffixes
- c. Prefixes and suffixes
- d. No possessive affixes

We have recoded this single, categorial, features into three binary features, and added an additional feature, as listed in (2). This recoding captures the variation in *WALS* feature 57A, in M72 and M73; the ‘Prefixes and suffixes’ values of *WALS* 57A is coded with positive values for both of M72 and M73, thus showing commonality with both prefixal languages and suffixal languages, which is not automatically extracted from the *WALS* coding. Positive values for M72 and M73 are unified by M71, which captures the commonality between prefixal and suffixal marking in that both do represent the coding of features of the possessor on the possessum. M70 adds in a typologically-attested variable that is not coded in *WALS*.

---

<sup>2</sup> The most doubled iso codes are cmn (Mandarin varieties), zlm (Malay varieties), adi (Tani languages), each of which has ten or more entries, at least some of which represent different languages by any normal criteria.

(2) Features M70 – M73

- |    |     |                              |     |
|----|-----|------------------------------|-----|
| a. | M70 | Possession: associative tone | +/- |
| b. | M71 | Possessive affixes: any      | +/- |
| c. | M72 | Possessive affixes: prefixes | +/- |
| d. | M73 | Possessive affixes: suffixes | +/- |

In addition to recoding some of the *WALS* features, additional features were added. *WALS* feature 102A codes for the appearance of agreement for A or P arguments. We have added coding for an S argument, as well as a third argument (M238), to account for languages that allow a recipient or dative argument to appear indexed on the verb. Very rarely a fourth or fifth agreement position can be found, and these are also coded, as M243 and M244. Additionally, just as *WALS* codes the position of agreement affixes marking possession on nouns, as prefixal or suffixal, we add in coding for the position of agreement affixes on verbs, as shown in (3).

(3) Coding the position of verbal agreement

- |    |      |                           |     |
|----|------|---------------------------|-----|
| a. | M232 | verb agreement_S prefix   | +/- |
| b. | M233 | verb agreement_S suffix   | +/- |
| c. | M234 | verb agreement_A prefix   | +/- |
| d. | M235 | verb agreement_A suffix   | +/- |
| e. | M236 | verb agreement_P prefix   | +/- |
| f. | M237 | verb agreement_P suffix   | +/- |
| g. | M239 | verb agreement_R/D prefix | +/- |
| h. | M240 | verb agreement_R/D suffix | +/- |
| i. | M245 | verb agreement_tone A     | +/- |
| j. | M246 | verb agreement_tone S     | +/- |
| k. | M247 | verb agreement_tone P     | +/- |

Other *WALS* features that were recoded in order to enhance their matching with a related feature in the database are the features devoted to morphological causatives. These were recoded in line with the features focusing on applicatives (which were also expanded). In *WALS* applicatives are coded for the kinds of bases that allow applicative extensions (intransitive or transitive bases), and the semantic role of the

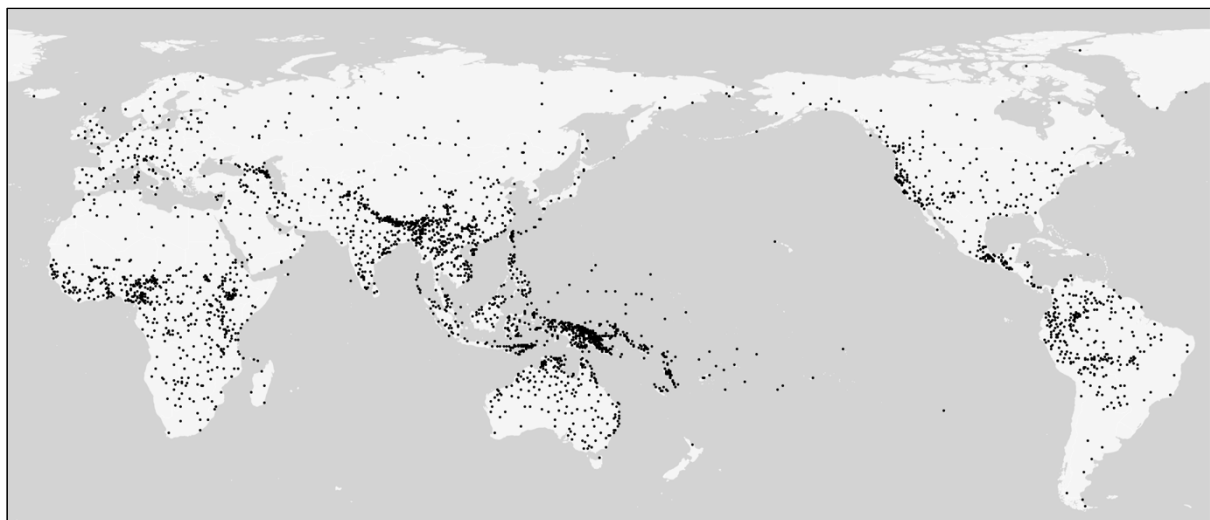
applied object (Benefactive, Instrumental, or Locative) (Polinsky 2013), and (non-periphrastic) causatives are coded according to whether they are morphological or compounds (Song 2013). In the World\_morphosyntax database the range of semantic roles for applicatives was expanded (Benefactive, Instrumental, Locative, Associative/Comitative, Theme, Reason, and Malefactive), and the types of bases allowed were extended to query ditransitive bases, and to distinguish whether agentive or patientive intransitive bases (or both) permit applicatives. Additionally, the possibility of more than one applicative appearing on a single base was coded with two features, as well as the possibility that the ‘applicative’ construction promotes directly to subject (a ‘superapplicative’, as attested in many languages of Taiwan and the Philippines). Matching the detail on applicative constructions present in *WALS* and expanded on in *World\_morphosyntax*, we coded morphological causatives according to whether they are attested in patientive or agentive intransitive bases, or even on ditransitive bases, as well as whether double (or second) applicatives are attested with different bases, what the coding strategy is for the causee of a causative construction with three arguments, and whether there is syncretism between the morpheme used for causatives and applicatives.

Other added features logically extend the scope of *WALS* features (for instance, explicitly coding more semantic roles that can be introduced with applicative constructions, the existence of suppletive negative verbal stems, or morphological processes other than prefixation and suffixation - namely, infixation, and metathesis). Wholly new features centre around the possibilities for nominal incorporation into verbs. Further details on the features in the database can be found in Appendix 2.

On average, the coding of languages reported here from the *World\_morphosyntax* dataset is 86% complete; this compares favourably with *WALS* (18% of 155 features for 2662 languages), as well as more recent datasets such as Grambank (Skirgård et al. 2023; 76% of 195 features for 2430 languages); see Appendix 1. For this study, we excluded known pidgins and creoles, reconstructed proto-languages, and ancient or historical languages. Pidgins and creoles frequently represent lineages that are not original to the area in which they are currently found, and in most cases represent disruptions to the local typological landscape. Ancient or historical languages (i.e. those that are attested only before the era of European colonisation) are by definition not part of any modern linguistic ecology, and so should not be included in an analysis of modern languages. Excluding these, we were left with 3089 languages/varieties,

the locations of which are shown in Map 1.<sup>3</sup> (See Appendix 1 for a full listing of the languages, with their genealogical and areal memberships.)

To analyse the *World\_morphosyntax* dataset, we used Factor Analysis of Mixed Data (FAMD; Pagès 2004), a dimensionality reduction technique that combines Multiple Correspondence Analysis (MCA) and Principal Components Analysis (PCA), as implemented in the *FactoMineR* package for R (Lê et al. 2008; see the Supplementary Materials for our annotated source code). Since MCA is suited for data consisting exclusively of categorical variables, and PCA is suited for data consisting exclusively of continuous variables, we felt that FAMD is the appropriate choice for the *World\_morphosyntax* dataset, which contains 27 ordinal variables (which we treated as continuous) and 324 binary variables. We started by imputing<sup>4</sup> the missing values in the dataset using the regularised iterative FAMD algorithm, as implemented in the “*imputeFAMD*” function in the *missMDA* package for R (Josse & Husson 2016), using 4 components (though the number of components made little difference to the results).



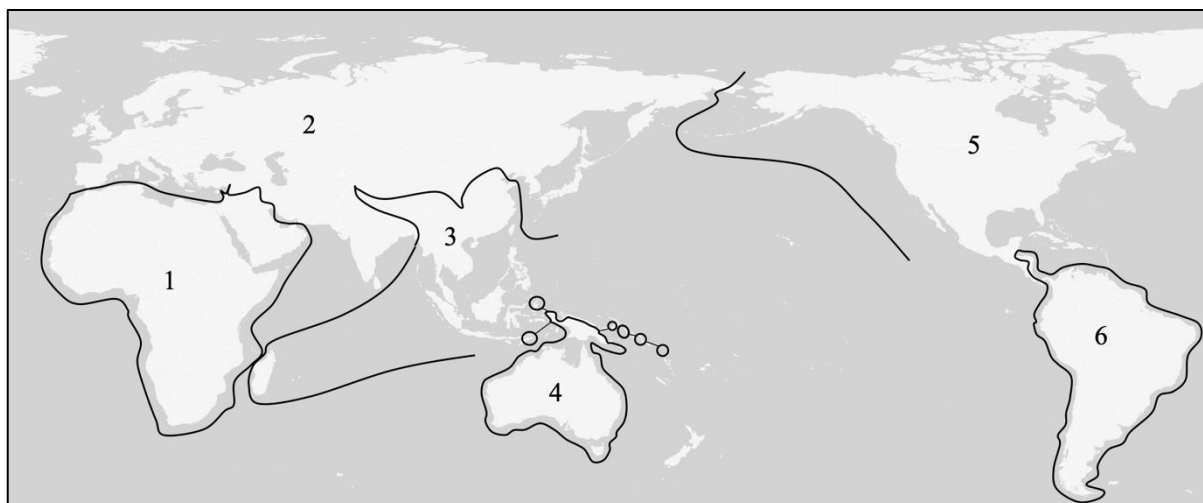
**Map 1:** Languages and language varieties included in the analysis ( $n = 3089$ ).

---

<sup>3</sup> Legend: The map (as well as subsequent maps) shows the world from 60° S to 90° N, and from 30° west extending 360° to the east.

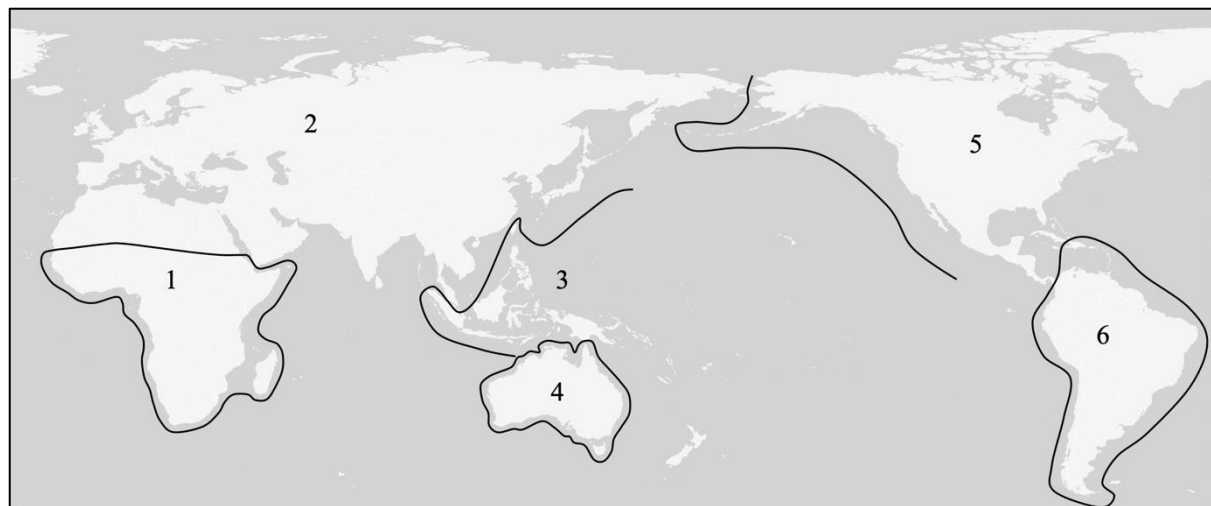
<sup>4</sup> Imputation is a family of techniques for replacing missing values in a dataset with estimates of the most likely values of those data points. It is necessary to perform imputation when applying techniques such as FAMD, since such techniques involve computing a covariance matrix, which requires complete data. The iterative FAMD algorithm for imputation (which we use here) works by first replacing missing values in each column with the column mean; then performing FAMD; then reconstructing the missing values based on the FAMD result; then performing FAMD again; and so on until the imputed values stabilise.

We then applied FAMD to the imputed data, assigning weights to languages in a way that equalises the total weights of different macro-areas, as well as of the different AUTOTYP areas within each macro-area; this was done to increase the likelihood that the dimensions that we find would capture groupings of features that are valid across different macro-areas, and across different areas within each macro-area. Macro-areas were defined according to Hammarström & Donohue (2014), itself a refinement of the macro-areas established by Dryer (1989, 1992), with the exception that languages belonging to the AUTOTYP “North Africa” area were re-assigned from Africa to Eurasia, on the grounds that whereas the Sahara Desert has been a barrier to contact since the end of the African Humid Period (e.g., de Menocal et al. 2000), cross-Mediterranean societies have flourished since antiquity. The two different macro-area divisions are compared in Maps 2 and 3. Most of the changes involve the abandonment of the apparent principle of unifying families into single macro-areas, and the split of Australia and (some of) New Guinea into separate macro-areas.



**Map 2:** Six macro-areas, following Dryer (1989, 1992).

Legend: 1: Africa; 2: Eurasia; 3: Southeast Asia and Oceania; 4: Australia-New Guinea; 5: North America; 6: South America.



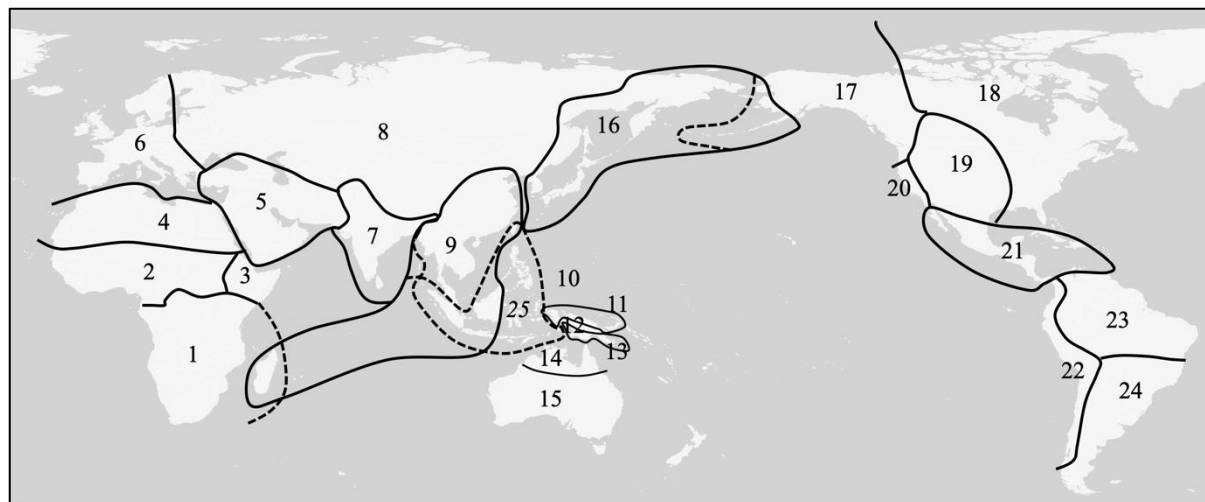
**Map 3:** Six macro-areas, following Hammarström and Donohue (2014), amended here.

Legend: 1: Africa; 2: Eurasia; 3: Pacific; 4: Australia; 5: North America; 6: South America.

The AUTOTYP areas were extrapolated from those described in Bickel et al. (2023), following Bickel (2002) and Nichols et al. (2013) (also <https://www.autotyp.uzh.ch>), with a few major differences: (1) ‘Southeast Asia’ has been split into Mainland Southeast Asia (consisting of the Southeast Asian languages of mainland Eurasia and Hainan) and Island Southeast Asia (consisting of the remaining Southeast Asian languages, as well as languages from ‘Oceania’ west of New Guinea and up to Taiwan), and Madagascar has been moved to Africa, allowing the smaller areas to be embedded unproblematically into macro-areas (as shown in Table 1); (2) The Andaman islands are grouped with “Indic”, based on historical connections; (3) The languages of the Aleutian Islands are included in ‘Alaska-Oregon’, rather than ‘North Coast Asia’, based on geography and cultural connections. The different areas are shown in Map 4, contrasting the original Autotyp areas with the modified set used here.<sup>5</sup> Details of the assignment of individual languages to areas can be found in Appendix 1.

---

<sup>5</sup> Legend for Map 4: 1: S Africa; 2: African Savannah; 3: Greater Abyssinia; 4: N Africa; 5: Greater Mesopotamia; 6: Europe; 7: Indic; 8: Inner Asia; 9: Southeast Asia (mainland); 10: Oceania; 11: N Coast New Guinea; 12: Interior New Guinea; 13: S New Guinea; 14: N Australia; 15: S Australia; 16: N Coast Asia; 17: Alaska-Oregon; 18: E North America; 19: Basin and Plains; 20: California; 21: Mesoamerica; 22: Andean; 23: NE South America; 24: SE South America; 25: Island Southeast Asia.



**Map 4:** The 25 modified AUTOTYP areas compared to the original 24 areas.

We are aware of alternative ways of controlling for area and genealogy (e.g. Guzmán-Naranjo & Becker 2021 on phylogenetic regression and Gaussian processes; Macklin-Cordes & Round 2022 on phylogenetic weighting). However, we opted to stay with areally-weighted FAMD, for the sake of simplicity, and because the patterns we find are strong enough to be visible and consistent regardless of what controls we use (see Appendix 4).

Macro-area (6)	Modified AUTOTYP area (25)	<i>n</i> (languages)
Africa	Africa, African Savannah, Greater Abyssinia	535
Eurasia	N Africa, Greater Mesopotamia, Europe, Inner Asia, Southeast Asia (mainland), N Coast Asia	1024
Pacific	Island Southeast Asia, N Coast New Guinea, Interior New Guinea, S New Guinea, Oceania	760
Australia	N Australia, S Australia	205
North America	Alaska-Oregon, E North America, Basin and Plains, California, Mesoamerica	283
South America	Andean, NE South America, SE South America	282

**Table 1:** Modified AUTOTYP areas arranged by Macro-area.

Another advantage of using areal divisions as a control is that the difference in size between the smallest group and the largest group is less than the difference between the size of the smallest language family or genus (namely 1) and the largest. This means that the area-based controls do not give undue weight to isolates and singleton

genera. An additional advantage of using areas, rather than genealogies, is that we avoid having to make decisions about controversial language families like Nilo-Saharan (Dimmendaal 2011), Trans-New Guinea (Pawley & Hammarström 2018), Transeurasian/Altaic (Clouston 1956, Schönig 2003), Austric (Schmidt 1906, Reid 2005), Hokan and Penutian (Campbell 1997, DeLancey & Golla 1997, Poser 1995), or Dene-Yeniseian (Kari & Potter 2010), or subgroups within families (e.g., Indo-Iranian and the position of Nuristani languages within Indo-European, the existence of Italo-Celtic in the same family, or the internal hierarchy of Tibeto-Burman). A comparison of the results presented here and the (minimally different) results of using genealogically-weighted approaches are discussed in Appendix 4.

### **3. Results**

We examine the results in detail for the world as a whole, and then in summary for each of six macro-areas. Section 3.1 presents the global dimensions of variation, what linguistic features characterise these dimensions, and where languages displaying the highs and lows of these dimensions can be found.<sup>6</sup> In Section 3.2 we examine the dimension plots presented in Figure 3 to show where various areal or genealogical entities can be found, and to what degree they form ‘compact’ clusters in typological space. In Section 3.3 we examine whether, and to what extent, these feature groupings can be considered universal, based on their appearance in the separate analyses of individual macro-areas.

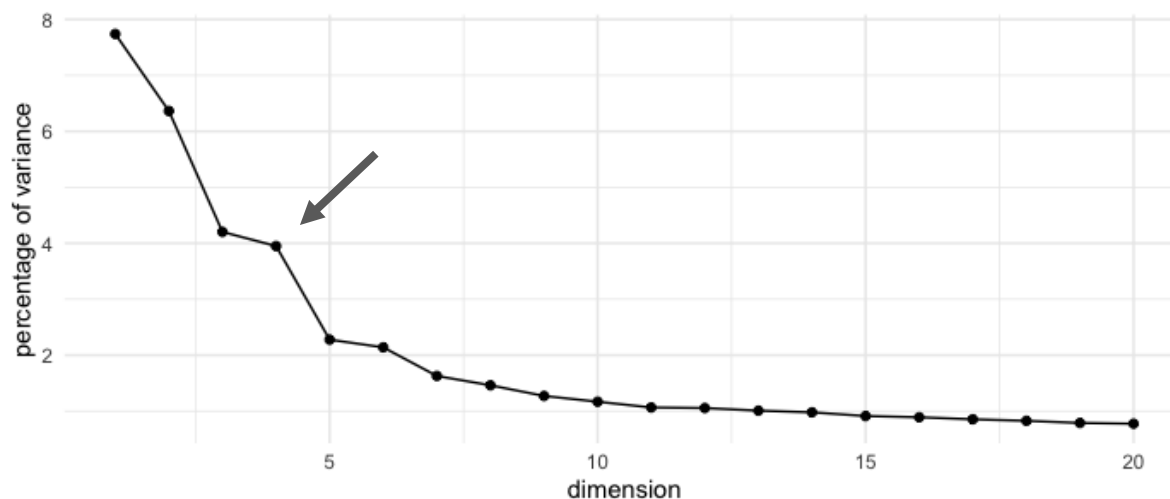
#### **3.1. Overall**

Figure 1 shows the percentage of total variance explained by each of the first 20 dimensions of the FAMD result, with the ‘elbow’ indicated by the arrow. As we can see, there is a sharp drop-off on the scree plot after the first four dimensions; thus, following principles in Cattell (1966), in the following we consider only the first four dimensions.

---

<sup>6</sup> We present four dimensions of variation, for the reasons discussed in Section 2.





**Figure 1:** Scree plot showing variance accounted for by the first 20 dimensions.

The positions of languages according to these four dimensions are plotted in Figure 3. The leftmost column shows Dimension 1 along the  $x$  axis, and Dimensions 2, 3 and 4 on the  $y$  axis in rows 2, 3 and 4, respectively. In the second column Dimension 2 is shown on the  $x$  axis, and Dimension 1 is displayed on the  $y$  axis. In the third column Dimension 3 is plotted on the  $x$  axis, and in the fourth column Dimension 4 is plotted on the  $x$  axis, with the  $y$  axes representing the same dimensions as previously described. The colours of the dots vary according to their positions on the first, second and third dimensions, with these dimensions mapped to red, green and blue colour components, respectively (a technique exemplified in Nerbonne 2009, and other associated works). Combinations of red and green display as yellow, red + blue as purple/magenta, red + green + blue as white. Green + blue is cyan, and the absence of any colouring is black, as shown schematically in Figure 2 (dots can of course also occupy positions inside the cube, where the display colour tends towards grey). Note that in Figure 3 (and later in Map 12) Dimension 4 is not represented in the colours displayed (though see Appendix 8). These four dimensions in total account for 22.2% of the variance in the data (a figure comparable to, for example, Skirgård et al. 2023), as shown in Table 2.<sup>7</sup>

<sup>7</sup> Much of the remaining data can be divided into a) rare features; b) wide-spread common features without strong correlations with other grammatical features; c) geographically restricted features. This is discussed in Section 4. Section 3.3 examines the contribution of other features in determining variation in smaller regions (see also Appendix 5).

Dimension	Variance accounted for?	Section
1	7.7%	3.1.1
2	6.4%	3.1.2
3	4.2%	3.1.3
4	3.9%	3.1.4

Table 2: Variance in the data accounted for by the first four dimensions.

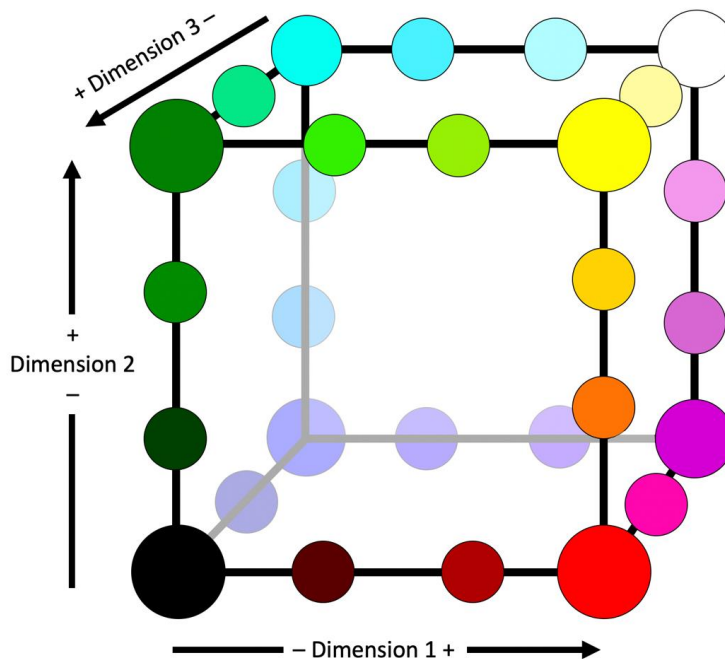
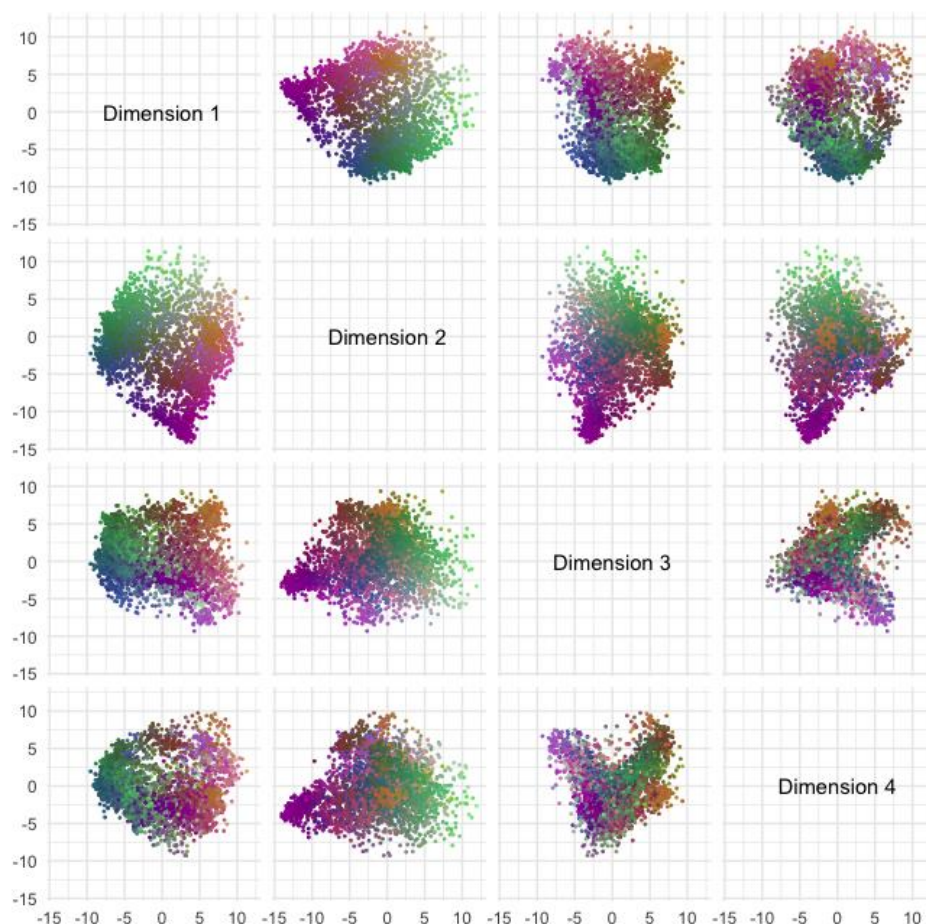


Figure 2: An illustration of a 'Red-Green-Blue' cube.

The interpretation of the different dimensions is presented in 3.1.1 – 3.1.4; in summary, the top end of Dimension 1, shown in red and orange, indicates languages with prepositions, and a tendency towards subject prefixes on verbs, while the bottom end is occupied by SOV languages with case-marking, shown in green and blue. The top end of Dimension 2 correlates with morphologically elaborate verbs, marked in pale green, while the bottom end tends towards isolating languages, with magenta colours. Dimension 3 has languages with gender systems and plural marking on nouns at the top end; languages at this pole are generally brown, while the lower end of this dimension correlates with VOS order and clusivity contrasts, presenting in a mix of colours in Figure 3. The top of Dimension 4 correlates with VSO languages that have prenominal modifiers in the NP, while the lower end correlates with SV order, object prefixes on verbs, and noun-numeral orders. As noted above, the position of a language on Dimension 4 is not indicated by any particular colours in Figure 3, but

Figures A8.1, A8.2 and A8.3, as well as Figures A8.6, A8.8 and A8.9 in Appendix 8 show the effects of having Dimension 4 contributing to the colouring.



**Figure 3:** Languages plotted according to the first four dimensions of variance.

We can see that there are different ‘densities’ of languages in different areas in Figure 3, such as the paucity of languages at approximately (0, 3) in the plot of Dimension 1 vs. Dimension 4 (at the bottom left of Figure 3), and the high concentration of languages at (–6, 2.5) in the plot of Dimension 1 vs. Dimension 2 (at the top left of Figure 3).<sup>8</sup> The dimension plots, based purely on linguistic features, include a number of typologically differentiated or isolated regions that correspond with a high degree of precision to geographically-recognisable areas or genealogically-coherent entities, some of which are discussed below in 3.2 (and see Section 4 for further discussion).

<sup>8</sup> The low-density region corresponds to a mix of languages, including many from modern Iran such as Farsi (pes; west2369, Indo-European, Iranian) and Sorani (ckb; cora1257, Indo-European, Iranian), and the high-density region is occupied by the head-final languages displaying an extreme head-final typology such as is found in Turkish, Daghestanian, and other languages from central Eurasia.

In the following subsections, each dimension is characterised in terms of the features that show the strongest association with it; in the case of binary variables, the strength of this association is measured with an ANOVA test, and for continuous variables, it is measured using Pearson correlation. In both cases, we report an  $r^2$  value. To determine whether the association is positive or negative, we look at the sign of the correlation coefficient (for continuous variables), or (for binary variables) perform a  $t$ -test comparing the dimension values of languages either exhibiting or lacking the feature against the entire set of languages, and note which (if either) of the two  $t$ -tests shows a significant positive value, and which (if either) shows a significant negative value. In Maps 5 – 8 positive values are shown in red/brown, and negative values in blue, according to the scale in Figure 4 (exact values can be found in Appendix 1).



Figure 4: The scale used in Maps 5 – 8.

### 3.1.1 Dimension 1: order of object and verb

The features most strongly associated with the first dimension centre around the order of the verb and its object, as well as a number of further headedness relations such as the position of a marker of subordination, the presence of prepositions, and the presence and position of case markers. Table 3 shows the features that have the strongest associations with Dimension 1.<sup>9</sup>

---

<sup>9</sup> For display purposes a number of related features from our database have been merged in this and subsequent tables for simplicity of presentation. For instance, both ‘SOV’ and ‘OV’ are associated (negatively) with Dimension 1 (since languages with these features on average have a negative value along Dimension 1;  $r^2 = 0.53$  and  $0.54$ , respectively). They are reported in Table 3 simply as SOV. Similarly, ‘Core case (any)’, ‘Dependent marking’, ‘Number of cases’ and ‘Postnominal case’ are all associated (negatively) with Dimension 1 ( $r^2 = 0.43$ ,  $0.49$ ,  $0.49$  and  $0.50$ , respectively), but only two of these features are listed in Table 3. Fuller lists of  $r^2$  values are found in Appendix 1.

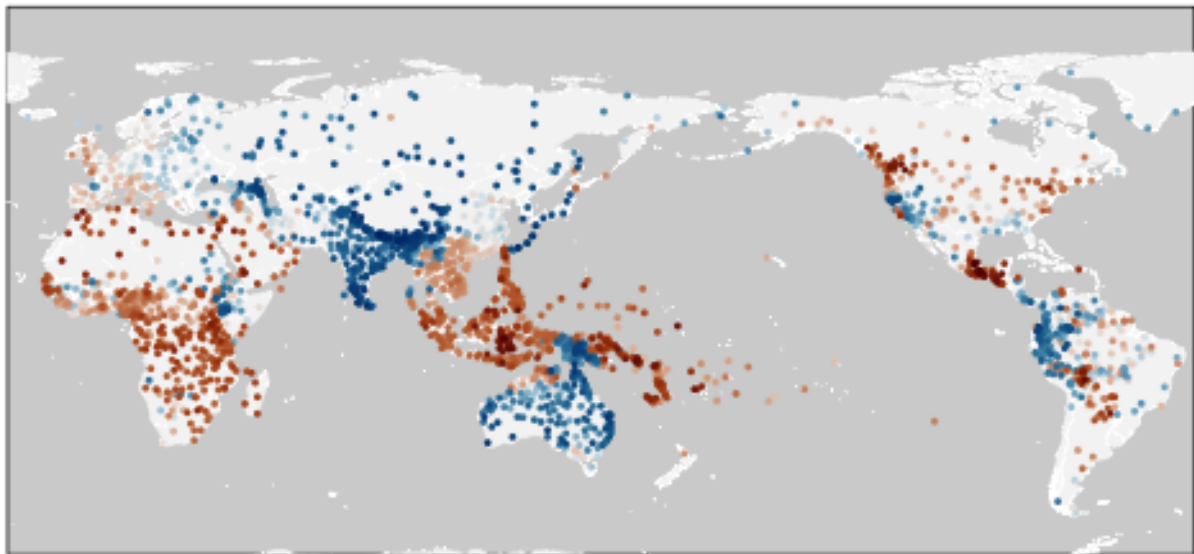
Direction	Feature	$r^2$
<b>High</b>	Prepositions	0.50
	Verb-Object order	0.42
	Initial subordination	0.41
	Nominative agreement by prefix	0.32
	Obliques follow verb	0.31
	Genitive precedes noun	0.31
	Final subordination by suffix	0.39
	Postpositions	0.42
	Obliques precede verb	0.45
	Number of cases	0.49
	Postnominal case	0.50
<b>Low</b>	SOV order	0.53

**Table 3:** Features characterising the extremities of Dimension 1.

These features are strongly reminiscent of (elements of) Greenberg's (1963) discussion of word order universals, and other linguistic features that refer to headedness parameters at the clause level. It is notable that prepositions are more closely associated with the positive (VO) end of this dimension than postpositions are with its negative (OV) end, and that head-final languages are strongly associated with the OV end, while head-initial languages are not as firmly associated with the VO end.

In Map 5 we can see the languages in our sample coded according to their position on Dimension 1, with high values marked in red/brown, and low values in blue, and middling values showing little hue. There are clear areal trends in the distribution of the extremes of this dimension, with large swathes of Eurasia dominated by languages with low (OV-congruent) values, and most of sub-Saharan Africa and Island Southeast Asia showing high (VO-congruent) values. Areas without consistent headedness settings, such as most of western Europe or northern China, are not associated with either extreme. The languages that are highest on Dimension 1 include various Otomanguean languages of the Chinantecan, Zapotecan and Popolocan groups in Central America, as well as Celebic Austronesian languages from central Indonesia, such as Mori (x mz; mori1268), Wolio (w lo; woli1241) and Wotu (w tw; wotu1240). The low end is dominated by South Asian languages, particularly South Dravidian

(Tamil, Tulu and Toda)<sup>10</sup> from the south of the subcontinent, and Bodic Tibeto-Burman (Kurtöp, Ghale and Balti)<sup>11</sup> from the Himalayas.



**Map 5:** Position of languages on Dimension 1 (blue = low, brown = high).

### 3.1.2 Dimension 2: verbal elaboration

The 2<sup>nd</sup> dimension of variation concerns the amount of morphology that can appear on the verb. At one end we have verbs with multiple positions for agreement, valency-increasing morphology for Ps (and, to a lesser extent, As), noun incorporation, and other inflectional material, such as switch-reference marking, Tense/Aspect/Mood, evidentiality, pluractionality, polarity, honorificity, voice marking, etc. (Bickel and Nichols 2013). At the other end, we find languages that lack extensive verbal morphology. The features with the strongest associations involve the lack of subordinating characteristics in “subordinate” clauses of different types, but the absence of the features characteristic of the higher end of this dimension, as well as the tendency for languages low on Dimension 2 to correlate with Dimension 1 (see Figure 3, and see 3.4), means that these languages tend to be more isolating.

---

<sup>10</sup> Tamil (tam; tami1289); Tulu (tcy; tulu1258); Toda (tcx; toda1252).

<sup>11</sup> Kurtöp (xkz; kurt1248); Ghale (ghe; barp1238); Balti (bft; balt1258).

Direction	Feature	$r^2$
<b>High</b>	Total verbal agreement positions	0.37
	Total verbal inflectional synthesis	0.34
	Total Modality affixes	0.30
	Incorporation	0.22
	Applicatives	0.20
	Causatives	0.16
	Possessive prefixes on nouns	0.15
	Total tense distinctions	0.11
<b>Low</b>	SVO order	0.14
	Symmetrical clauses: Purpose	0.17
	Symmetrical clauses: Temporal	0.18
	Symmetrical clauses: Reason	0.21

**Table 4:** Features characterising the extremities of Dimension 2.

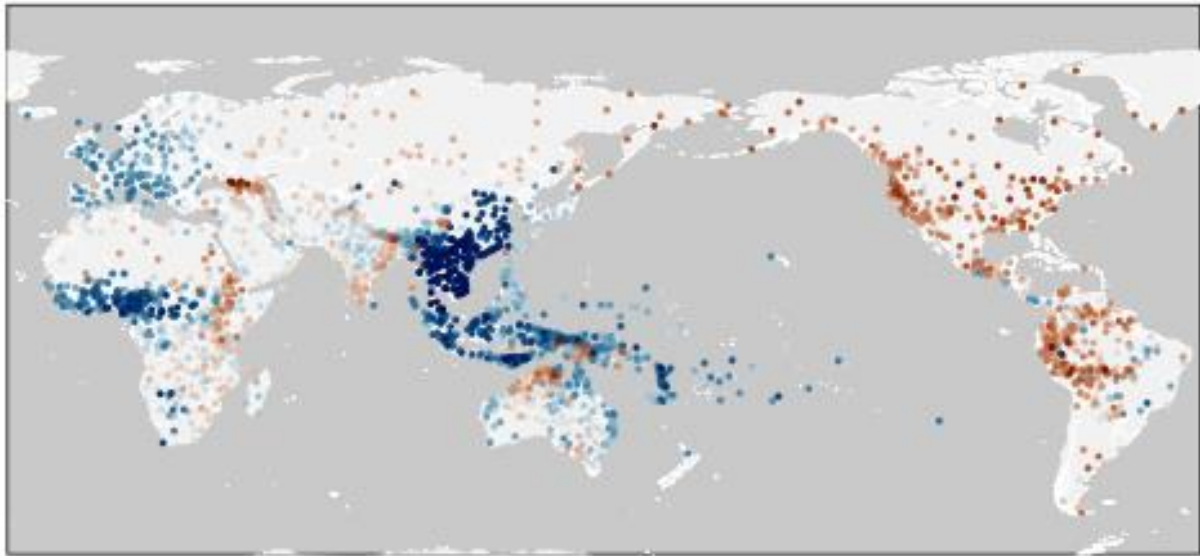
The features here are strongly reminiscent of (and add to) the head-marking end of Nichols' (1986) typology of head-marking vs. dependent-marking languages (itself related to divisions of morphological typology established as early as K.F. Schlegel 1808 and A.W. Schlegel 1818), with languages high on Dimension 2 being more heavily head-marking, and languages low on Dimension 2 showing more isolating/analytic traits. (We have already seen that dependent-marking is associated with Dimension 1, specifically with its lower OV end.)

In Map 6 we can see that the languages of the Americas are almost universally on the head-marking side of this dimension; the opposite extreme, namely absence of head-marking characteristics, dominates in Southeast Asia, to a lesser extent in western Africa, and in small measure in western Europe. The Old World sees clusters of head-marking languages in the Caucasus, in East Africa, in the Munda-Kiranti areas of South Asia, in the north-east of Eurasia on the approach to the Americas; parts of New Guinea, and most of northern Australia, also contain languages that are strongly head-marking, and so high on Dimension 2. Languages from a number of families in Southeast Asia are found at the low end of Dimension 2, including Austronesian (Moken, Cham)<sup>12</sup>, Austroasiatic (Bruu, Vietnamese)<sup>13</sup>, and also Hmong and Thai languages; a number of languages of West Africa, centred on Nigeria (such as Igede

<sup>12</sup> Moken (mwt; make1242, Malayo-Sumbawan); Cham (cjm; east2563, Malayo-Sumbawan).

<sup>13</sup> Bruu (bru; east2332, Katuic); Vietnamese (vie; viet1252, Vietic).

and Yoruba)<sup>14</sup>, are also high on this dimension. The higher end of the scale is occupied by polysynthetic languages from North America (such as Algonquian Arapaho, Cheyenne and Ottawa)<sup>15</sup>, from the North-west Caucasus family (including Abaza, Adyghe and Kabardian)<sup>16</sup>, as well as Chukotko-Kamchatkan Alyutor (alr; alut1245), and a number of western Amazonian languages from the South American (such as Aikanã, Jebero, Matses, Arakmbut and Matsigenka)<sup>17</sup>, and a scattering of languages elsewhere.



**Map 6:** Position of languages on Dimension 2 (blue = low, brown = high).

### 3.1.3 Dimension 3: Western Old World

The third dimension of variation shows the strongest (macro-)areal distribution. Unlike the other three dimensions discussed here, the distribution of Dimension 3 does not identify a number of separate areas throughout the world, but rather a global cline from west to east (as is clearly visible in Map 7, and see below). The features at the high end of this dimension are all morphological; at the low end we see either an absence of extensive nominal morphology, or verb-initial orders. Because of these two

---

<sup>14</sup> Igede (ige; iged1239, Niger-Kongon, Idomoid); Yoruba (yor; yoru1245, Niger-Kongo, Yoruboid).

<sup>15</sup> Algonquian Arapaho (arp; arap1274); Cheyenne (chy; chey1247); Ottawa (otw; otta1242).

<sup>16</sup> Abaza (abq; abaz1241); Adyghe (ady; adyg1241); Kabardian (kbd; kaba1278).

<sup>17</sup> Aikanã (tba; aika1237, Isolate, Aikanã); Jebero (jeb; jebe1250, Cahuapanan); Matses (mcf; mats1244, Panoan, Matses); Arakmbut (hug; arak1258, Harakmbet); Matsigenka (mcb, mach1267, Arawak, Campa).

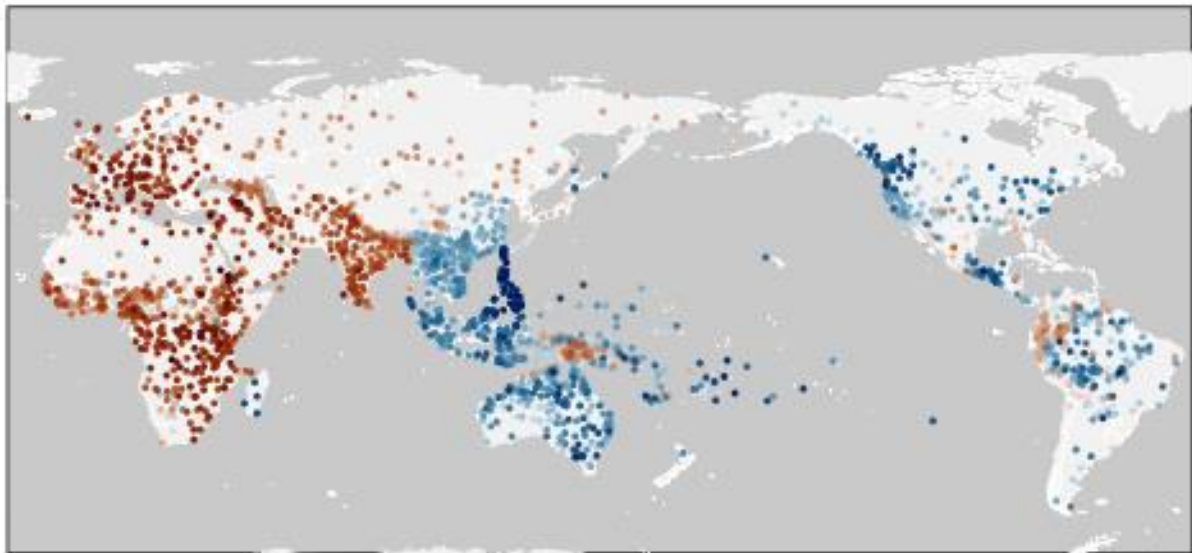


different typologies, the  $r^2$  values of features at the low end of this dimension are not as high as those at the high end.

Direction	Feature	$r^2$
<b>High</b>	Gender	0.28
	Obligatory plural marking on nouns	0.27
	3SG pronominal gender	0.25
	Verb alignment: accusative	0.22
	3PL pronominal gender	0.17
	Suffixal subject agreement on verbs	0.13
	Relative pronouns	0.13
	Ergativity	0.13
<b>Low</b>	VOS order	0.15
	Inclusive/Exclusive contrasts	0.18
	Clause-initial negation	0.18

**Table 5:** Features characterising the extremities of Dimension 3.

As can be seen in Map 7, languages low on this dimension are almost exclusively found in the Circum-Pacific region, an area which “comprises all of the Americas, Oceania (including Australia and New Guinea), and the mainland Asian Pacific Rim”, the last area being the “Pacific-facing coast up to the lower slope of the far side of the major coast range” (i.e., the eastern Himalayas) (Bickel & Nichols 2006: 6). We observe increasingly high values as we go west in the Old World. On the high end we find most of the languages of Europe and the other circum-Mediterranean languages, as well as the Bantu languages, which have significantly higher values than those in the rest of Eurasia and Africa (see 3.2.1 and 3.2.10). In part due to this position, and the relative morphological simplicity of European languages compared to Semitic, Berber or Bantu languages (thus having lower values on Dimension 2), the languages of Europe can be identified as a global outlier (see Figure 3, and 3.2.1).



**Map 7:** Position of languages on Dimension 3 (blue = low, brown = high).

The features that have a strong association with Dimension 3 partially overlap with the list of features often put forward as defining ‘Standard Average European’ (Whorf 1941, and also Haspelmath 2001, van der Auwera 2011, and others). Often-cited ‘Standard Average European’ features that have positive correlations with Dimension 3 include: indefinite articles, have-perfects, relative pronouns (see Table 5), predominantly suffixing morphology (see Table 5), accusative alignment (see Table 5), and negative indefinite pronouns. Features that have negative associations with Dimension 3 include clusivity contrasts (see Table 5), alienability contrasts, identity of ‘and’ and ‘with’, and productive reduplication. In contrast to other studies on Standard Average European, we find that dative subjects have a (weak) positive association with Dimension 3 ( $r^2 = 0.10$ ). (See Appendix 1 for details of the associations of these features with the different dimensions.)

It is clear from Map 7 that Dimension 3 negatively correlates with ‘eastness’ (as displayed on Map 7, such that Iceland is west and Greenland is east). The correlations of the different dimensions with ‘eastness’ in different domains are shown in Table 6. The strong negative correlation of Dimension 4 (3.1.4) with eastness in Eurasia reflects the far western position of the verb-initial Celtic, Berber and Semitic languages; there are very few verb-initial languages in the east of mainland Eurasia.

Dimension 3, however, shows strong correlations across Eurasia, the Old World, and globally.<sup>18</sup>

Dimension	Global	Old World	Eurasia
1	-0.02	-0.11	-0.45
2	0.36	-0.14	-0.22
3	-0.54	-0.61	-0.74
4	-0.21	-0.26	-0.73

**Table 6:** Correlation with eastness (*r*).

The languages at the top end of Dimension 3 are Niger-Kongo Bantu (Ruwund, KinyaRwanda and Runyankore)<sup>19</sup>, Indo-European Romance (Spanish, Romansch, Galician and French)<sup>20</sup> or Afro-Asiatic Semitic (Cypriot Arabic, Mlaḥsô and Fezzan Arabic)<sup>21</sup>, in addition to a number of other European languages (such as Albanian, Czech and Tabarchino)<sup>22</sup>. As can be seen in Figure 3, the lower end of Dimension 3 is quite dispersed typologically, and consequently there is a range of different languages that are maximally different from those of the western Old World, as measured on this dimension. Languages at the bottom of Dimension 3 include verb-initial Texistepec (poq; texi1237, Totozoquean, Chitimacha–Zoque), Kuikuro (kui; kuik1246, Cariban, Nahukwa), Shuswap (shs; shus1248, Salishan, Interior Salish), and many languages of the Philippines and Taiwan (such as Hanunoo, Saaroa and Maranao)<sup>23</sup>, and Polynesia (Samoan and Niuean)<sup>24</sup> in the Pacific. In addition to their verb-initial clauses, these languages also lack gender in nouns or pronouns, accusative alignment, or obligatory plural marking.

<sup>18</sup> Strong negative correlations are also found in South America (-0.41), due to the presence of a large area in the northern Andean region occupied by languages with higher values on Dimension 3, belonging to the Jivaroan, Quechuan, Tucanoan and Boran families, amongst others.

<sup>19</sup> Ruwund (rnd; ruun1238); KinyaRwanda (kin; kiny1244); Runyankore (nyn; nyan1307).

<sup>20</sup> Spanish (spa; stan1288), Romansh (roh; roma1326), Galician (glg; gali1258), French (fra; stan1290).

<sup>21</sup> Cypriot Arabic (acy; cypr1248); Mlaḥsô (lhs; mlah1239); Fezzan Arabic (ayl; liby1240).

<sup>22</sup> Albanian (als; tosk1239, Indo-European, Albanian), Czech (ces; czech1258, Indo-European, Slavic); Tabarchino (lij; ligu1248, Indo-European, Romance).

<sup>23</sup> Hanunoo (hnn; hanu1241, Austronesian, Philippines); Saaroa (sxr; saar1237, Austronesian, Tsouic); Maranao (mrw; mara1404, Austronesian, Philippines).

<sup>24</sup> Samoan (smo; samo1305, Austronesian, Oceanic); Niuean (niu; niue1239, Austronesian, Oceanic).

## 3.1.4. Dimension 4: order of subject (and negator) and verb, and NP orders

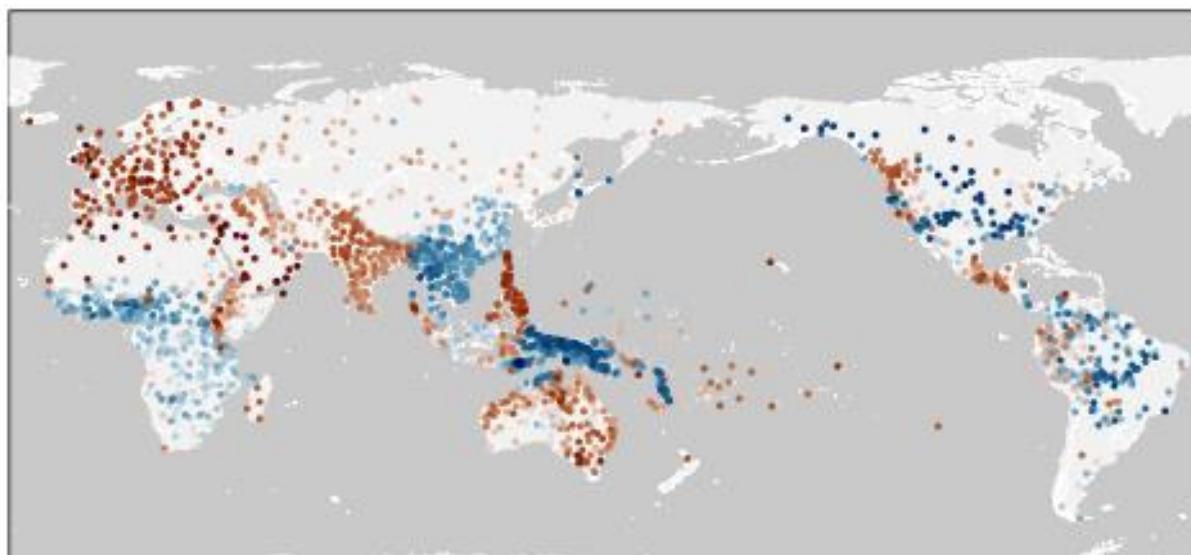
The other major aspect of clausal word order, the order of subjects and predicates, is found to have the strongest associations with both ends of Dimension 4. Clause-initial negation, which is overwhelmingly (but not exclusively) associated with verb-initial languages, also shows strong positive associations with Dimension 4. Unlike the word order correlations evident in Dimension 1, a number of NP-internal correlations are found with Dimension 4, leading to a number of languages which are not verb-initial nonetheless displaying high values on this dimension.

Direction	Feature	$r^2$
High	VSO order	0.22
	Clause-initial negation	0.20
	Numeral precedes noun	0.19
	Clause-initial Wh-question words	0.11
	Adjective precedes noun	0.11
	Genitive precedes noun	0.10
	Relative pronouns	0.10
	Clause-final negation	0.10
	Inalienable possession	0.10
	Object agreement prefix	0.14
Numeral follows noun	0.14	
Low	SV order	0.21

**Table 7:** Features characterising the extremities of Dimension 4.

The order of subject and verb again reflects Greenberg's classification of the world's languages by clausal word order. As with the order of object and verb, seen in Map 5, we can identify a number of contiguous areas which are high or low on this dimension. The relative paucity of VS languages, compared to SV languages, means that it is easiest to consider the distribution of these languages compared to a background of SV languages. The languages at the top of Dimension 4 are mostly Semitic and Berber languages from north Africa and the Middle East, and the Celtic languages of western Europe, though certain south-eastern Australian languages such as Warrnambool (gjm; warr1257, Pama-Nyungan, Kulinic), Wembawemba (xww; wemb1241, Pama-Nyungan, Kulinic) and Muk-Thang (Garnai) (unn; gana1278,

Pama-Nyungan, Gippsland) are also found at this extreme. Languages at the opposite extreme of this dimension are found in North America, including the Athabaskan languages Dena'ina (tfn; tana1289), Kaska (kkz; kask1239) and Slavey (xsl; sout2959), and the Siouan languages Lakhota (lkt; lako1247), Stoney (sto; 1ton1242), Hidatsa (hid; hida1246) and Hocąk (Winnebago) (win; hoch1243), as well as in languages from various families on the fringes of New Guinea, such as Puare (pux; par1240) and Barupu (wra; wara1302) (Skou family); Riantana (ran; rian1263, (Trans New Guinea?), Kolopom), and the Timor-Alor-Pantar languages Tanglapui/Sawila (tpg; sawi1256), Lamma/Western Pantar (lev; lamm1241), Adang (adn; adan1251), Abui (abz; abui1241) and Kamang (woi; kama1365).



**Map 8:** Position of languages on Dimension 4 (blue = low, brown = high).

### ***3.2. Geographically or genealogically recognisable regions***

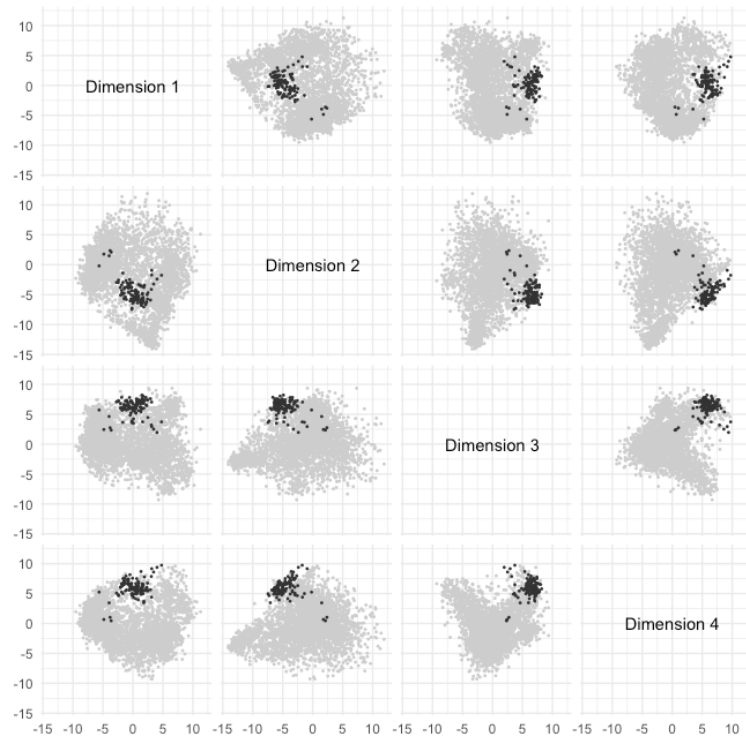
In this section we return to the dimension plots seen in Figure 3 (and compare also with Map 12), and examine recognisable geographic or genealogical regions to determine whether, and to what extent, they correspond to distinct ‘regions’ in typological space. To assess whether a given geographic or genealogical group clusters on one side of a given dimension, we perform a one-sided *t*-test comparing the values of languages within the group along that dimension, and the values of all languages in the dataset along that dimension. Generally, *t* values greater than 20 or less than –20 indicate that a group of languages shows extreme values along a given dimension, and/or forms a tight cluster along that dimension. A *p*-value close to zero

indicates that the means of the two populations being compared are significantly different; however, since  $p$  values are generally lower for larger datasets, the results should be interpreted on the basis of the  $t$  statistic as well as the  $p$  value. We also report the degrees of freedom ( $df$ ) for each analysis.

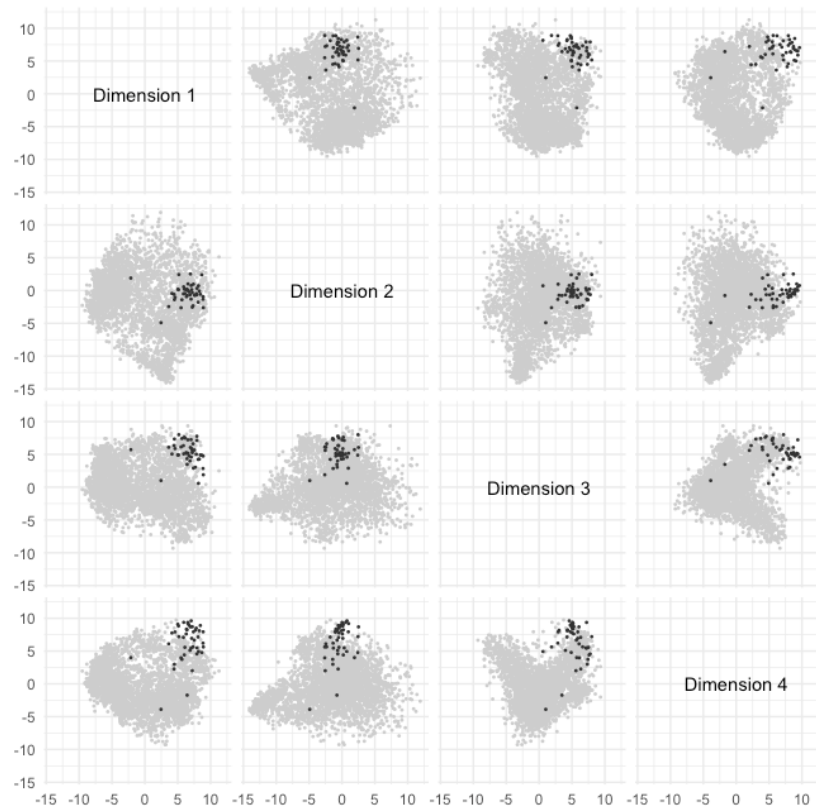
### 3.2.1. *Western Old World: Europe, Arabia and North Africa*

As mentioned in 3.1.3, the languages of (western) Europe almost all occupy a position high on Dimension 3 (Western Old World) ( $t = 40.24$ ,  $df = 178.89$ ,  $p < 0.001$ , according to a one-sided  $t$ -test) and 4 ('order of subject (and negator) and verb, and NP orders'), and moderately low on Dimension 2 ('verbal elaboration') ( $t = -14.39$ ,  $df = 175.57$ ,  $p < 0.001$ ). The region of typological space that can be seen in the combination of these two dimensions is quite separate from the rest of the cloud; exceptions to this separation, found much lower on Dimension 3, are recognised as outliers within Europe: Basque varieties (eus; basq1248), Hungarian (hun; hung1274), Gagauz Turkish (gag; gaga1249), and (to a lesser extent) the Celtic languages. The mixed word-order typology of most of the European languages (with head-initial parameters dominating at the clause level, and head-final parameters predominant within NPs) means they occupy a position in the middle of Dimension 1 ('order of object and verb'), and they can be seen to occupy a distinct, albeit interior, position in the plot of Dimension 1 vs. Dimension 2. In Figure 4 we can see that the languages of Europe occupy a compact region in typological space in each of the dimension plots, including those that do not involve dimensions 2 or 3, though they are not part of the 'fringe' of typological space.

Figure 5 shows the position of the languages of Arabia and North Africa; not as compact as the European languages, they can also be characterised as occupying a fringe positions on the plot of Dimensions 1 and 3 ( $t = 20.49$ ,  $df = 52.25$ ,  $p < 0.001$ ;  $t = 18.38$ ,  $df = 49.89$ ,  $p < 0.001$ ), and are higher on Dimension 1 than the European languages ( $t = 18.42$ ,  $df = 73.60$ ,  $p < 0.001$ ), but not significantly higher on Dimension 4 (two-sided  $t = 0.92$ ,  $df = 53.17$ ,  $p = 0.36$ ). The outliers at the lower end of Dimension 3 for this group of languages are mixed languages in the areas, such as Kumzari (zum; kumz1235, Indo-European (?), Iranian), between Arabia and Persia, and Kwarandzyey (/Korandje) (kcy; kora1291), a Songhai language spoken in the extreme north-east of the Sahara in a Berber linguistic environment.



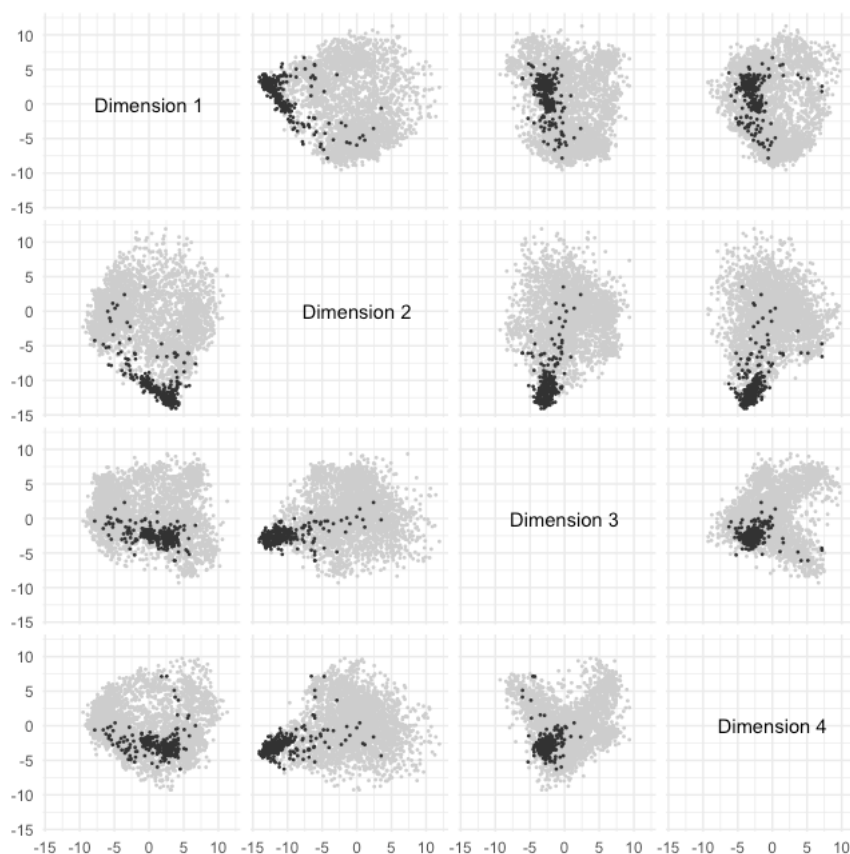
**Figure 5:** Languages of Europe highlighted on the dimension plots.



**Figure 6:** Languages of Arabia and North Africa highlighted on the dimension plots.

### 3.2.2. Mainland Southeast Asia

The languages of Southeast Asia represent a number of typologically convergent language families, all low on Dimensions 2 and 3 ( $t = -41.17$ ,  $df = 366.98$ ,  $p < 0.001$ ;  $t = -30.25$ ,  $df = 821.84$ ,  $p < 0.001$ ). The outliers for this group, typologically, are also outliers geographically. The most divergent languages are the Nungish languages of northern Myanmar and adjacent China, high on Dimension 2, and the Nicobarese languages of the Nicobar Islands, high on Dimension 4 (raising questions about their inclusion in a ‘Mainland Southeast Asia’ area). As with the languages of Europe, the languages of Southeast Asia largely cluster together even in plots that do not involve Dimension 2 or Dimension 3.



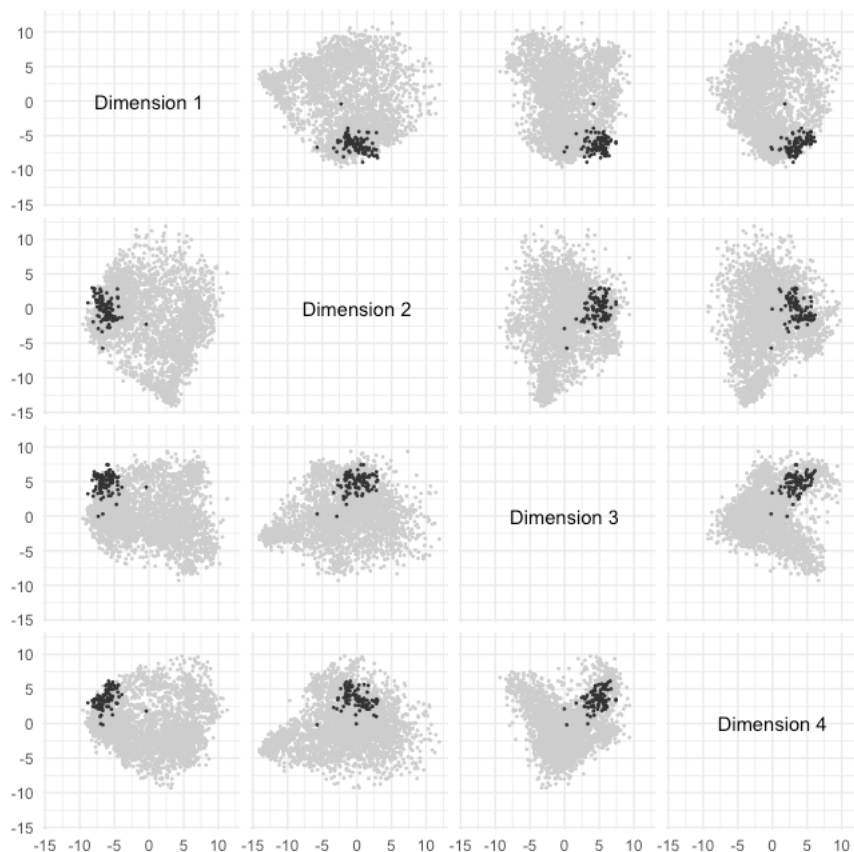
**Figure 7:** Languages of mainland Southeast Asia highlighted on the dimension plots.

### 3.2.3. Core South Asia

The Indic (< Indo-European) and Dravidian languages of South Asia also cluster together, though not at the periphery of any of plots, except for their low position on



Dimension 1 ( $t = -41.99$ ,  $df = 205.58$ ,  $p < 0.001$ ), and relatively high position on Dimension 3 ( $t = 29.02$ ,  $df = 136.18$ ,  $p < 0.001$ ). The typological outliers for this area, low on Dimension 3 or high on Dimension 1, are Vedda (ved; vedd1240, Indo-European, Indic), from Sri Lanka, and Dari (prs; dari1249, Indo-European, Iranian), the eastern variety of Farsi spoken in Afghanistan and not typologically assimilated to the languages of the region.



**Figure 8:** Languages of ‘core South Asia’ highlighted on the dimension plots.

### 3.2.4. Inner Asia

The core Eurasian profile of a radically head-final language (low on Dimension 1:  $t = -25.13$ ,  $df = 191.91$ ,  $p < 0.001$ ) with a modest level of morphological elaboration (moderately greater than zero on Dimension 2:  $t = 4.52$ ,  $df = 165.25$ ,  $p < 0.001$ ) is most strongly realised in Inner Asia, where Mongolic, Tungusic, Turkic and Uralic languages share many typological features. The outliers in this group are recently-

arrived varieties of Mandarin (Dungan, Urumqi and Taz)<sup>25</sup> and Arabic (Afghanistani Arabic and Bukhara Arabic)<sup>26</sup>, which are low on Dimension 2 and high on Dimension 1, respectively.

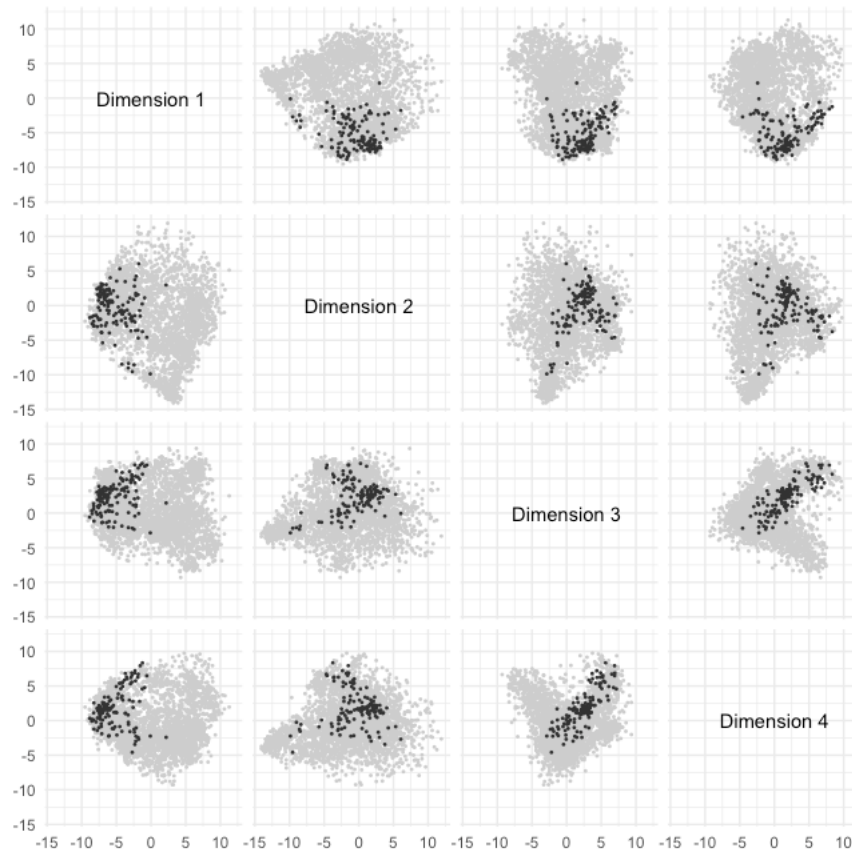


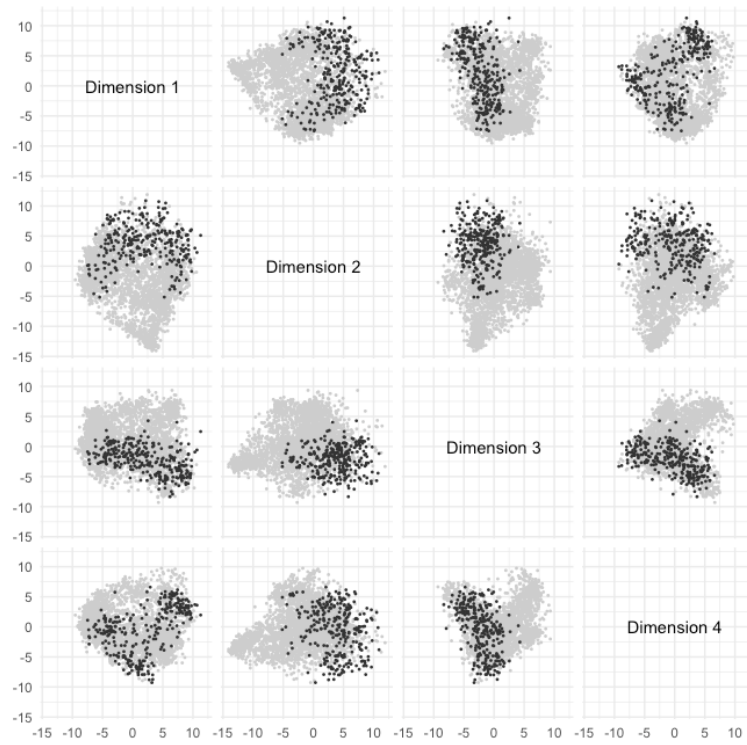
Figure 9: Languages of Inner Asia highlighted on the dimension plots.

### 3.2.5. North America

The languages of North America are widely dispersed, though the average position, and densest grouping, is both high on Dimension 2 ( $t = 24.8$ ,  $df = 406.47$ ,  $p < 0.001$ ) and low on Dimension 3 ( $t = -17.3$ ,  $df = 428.62$ ,  $p < 0.001$ ), indicating a head-marking, morphologically complex language that is maximally different from the languages of western Eurasia. In Dimension 1 and Dimension 4 there is no apparent pattern (two-sided  $t = -0.94$ ,  $df = 322.14$ ,  $p = 0.35$ ), but in Dimension 1 the languages on average have values slightly greater than zero ( $t = 6.60$ ,  $df = 339.84$ ,  $p < 0.001$ ).

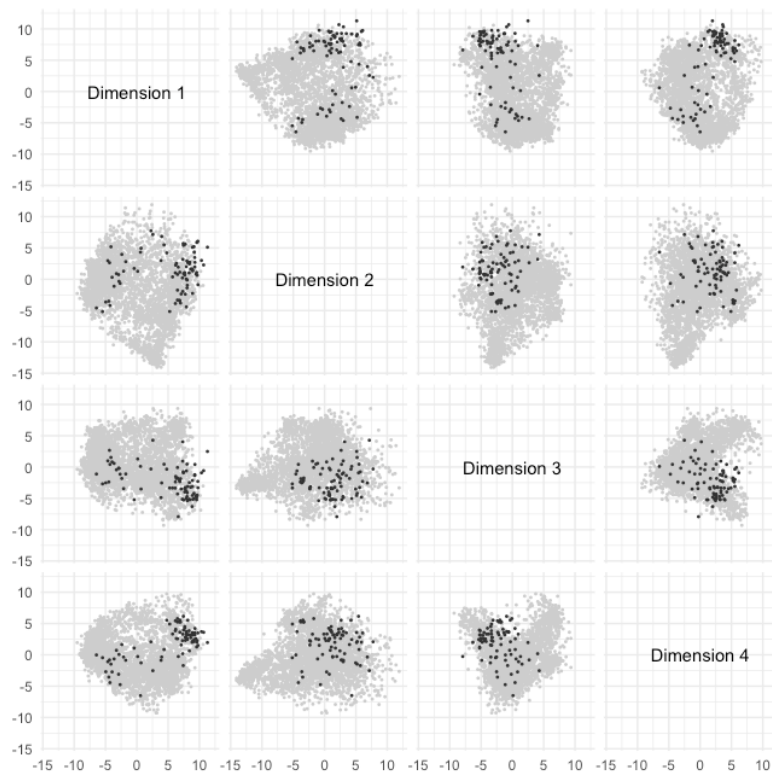
<sup>25</sup> Dungan (dng; dung1253); Urumqi (cmn; wulu1243); Taz (cmn; north3283).

<sup>26</sup> Afghanistani Arabic (abh; taji1248); Bukhara Arabic (auz; uzbe1248).



**Figure 10:** Languages of North America highlighted on the dimension plots.

### 3.2.6. Mesoamerica

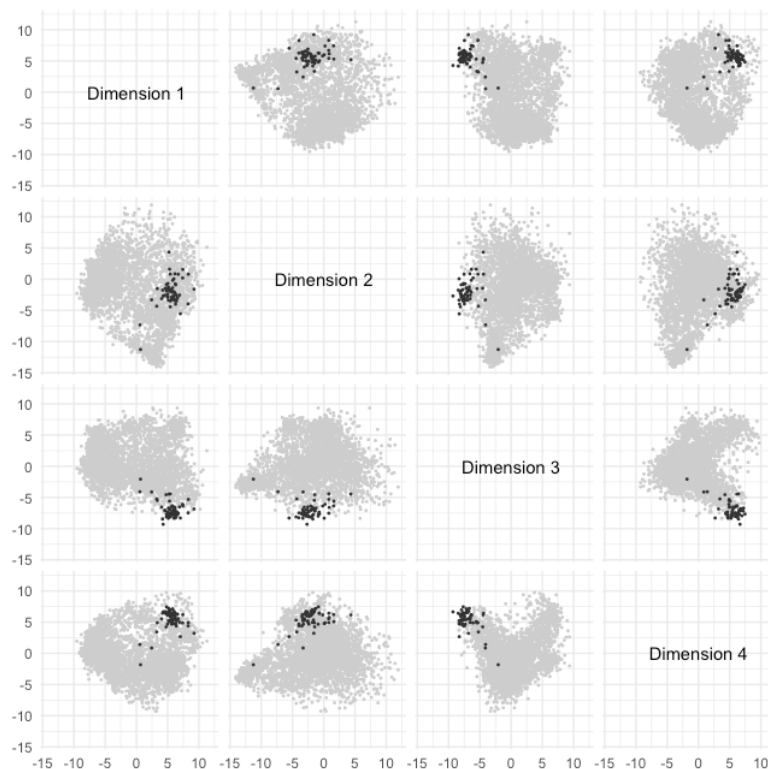


**Figure 11:** Languages of Mesoamerica highlighted on the dimension plots.

Focussing just on the languages of Mesoamerica as a sub-region within North America we find a high degree of typological dispersal, but with a cluster high on Dimension 1 ( $t = 8.63$ ,  $df = 86.28$ ,  $p < 0.001$ ) and middling high on Dimension 4 ( $t = 6.81$ ,  $df = 89.38$ ,  $p < 0.001$ ).

### 3.2.7. *The Philippines and Taiwan*

The ‘Philippine-type languages’ of the Philippines and Taiwan, which, while mostly Austronesian, do not form a valid subgroup within that family, can be found high on Dimension 4 ( $t = 31.15$ ,  $df = 56.38$ ,  $p < 0.001$ ) and low on Dimension 3 ( $t = -7.88$ ,  $df = 64.48$ ,  $p < 0.001$ ), where they form a fringe to typological space.



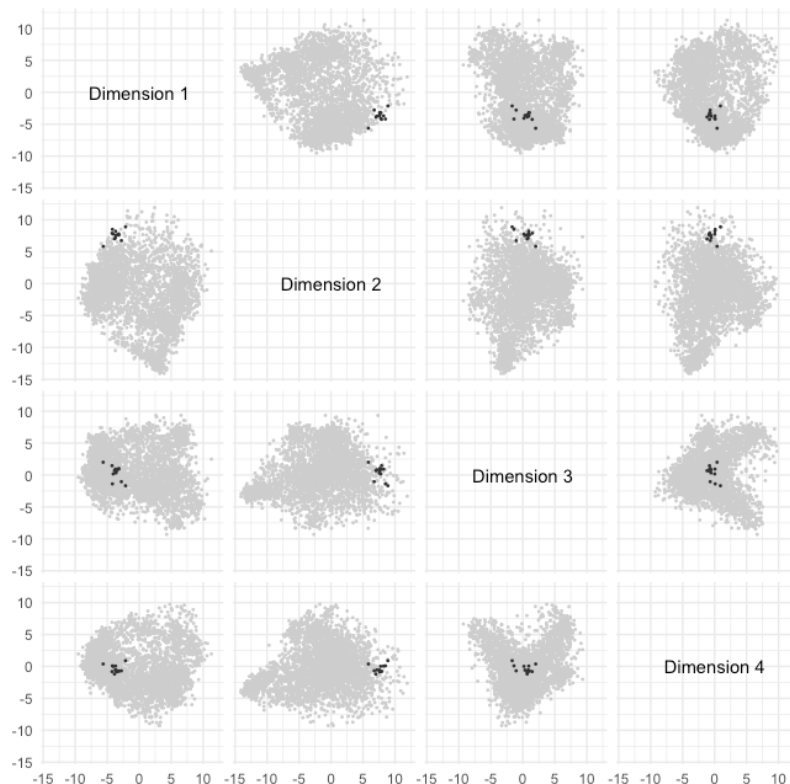
**Figure 12:** Languages of the Philippines and Taiwan highlighted on the dimension plots.

In other dimension plots they also form a tight cluster, with the divergent languages from this region (high or low on Dimension 2, or low on Dimension 1 or 4) being 1) the Austronesian languages of Taiwan (high on Dimension 2), 2) Iraya (iry; iray1237, Austronesian) from Mindoro in the Philippines, and Taiwanese (nan; taib1242, Tibeto-Burman, Sinitic), the intrusive Sinitic language of Taiwan (low on Dimension 2),

and 3) the southern Austronesian languages in this cluster, such as Talaud (tld; tala1285) and Sangir (sxn; sang1336) from northern Indonesia (low on Dimensions 1 and 4).

### 3.2.8. Eskimo-Aleut

The languages of the Eskimo-Aleut family are very low on both Dimension 1 ( $t = -15.51$ ,  $df = 15.79$ ,  $p < 0.001$ ) and very high on Dimension 2 ( $t = 39.45$ ,  $df = 16.29$ ,  $p < 0.001$ ), indicating a morphologically elaborate group of extremely SOV languages. As a small and young family they form a tight cluster, and represent an extreme extension of the North American (or North-east Asian) linguistic type.



**Figure 13:** Languages of the Eskimo-Aleut family highlighted on the dimension plots.

### 3.2.9. North-west Caucasus

The languages of the North-west Caucasus family occupy a position similar to the Eskimo-Aleut languages, but more extreme (Dimension 1:  $t = 5.24$ ,  $df = 8.01$ ,  $p < 0.001$ ; Dimension 2:  $t = 5.47$ ,  $df = 7.88$ ,  $p < 0.001$ ).

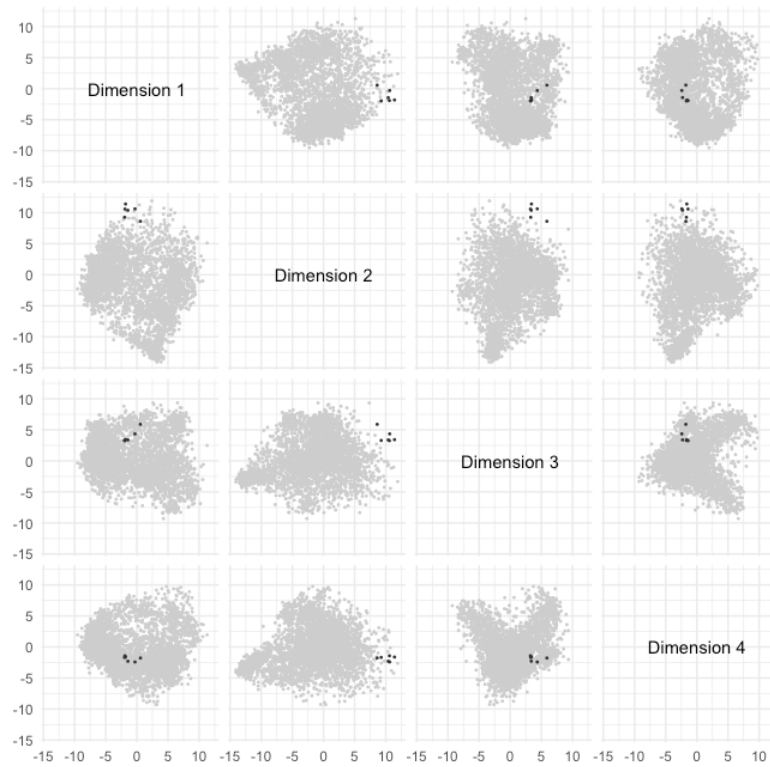


Figure 14: Languages of the Eskimo-Aleut family highlighted on the dimension plots.

### 3.2.10. Narrow Bantu

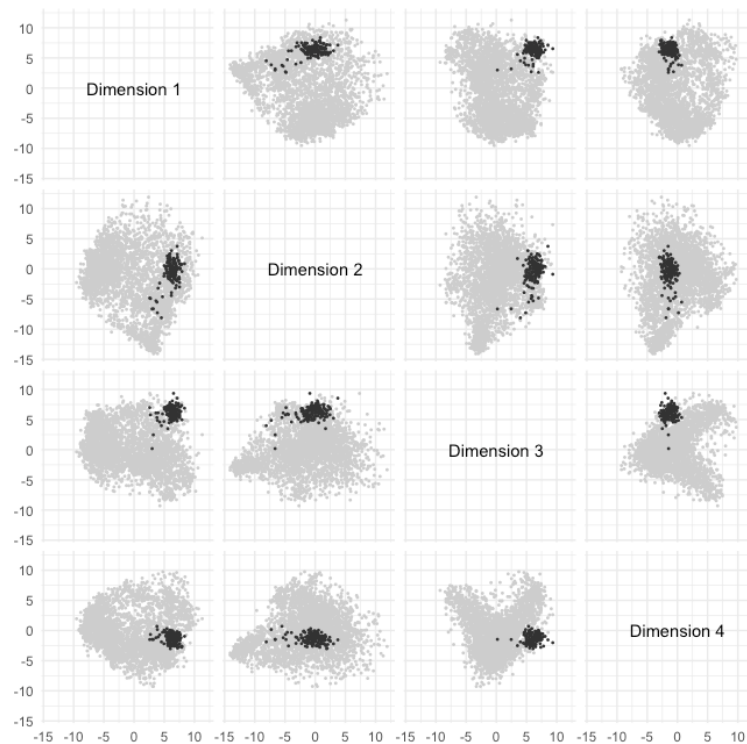
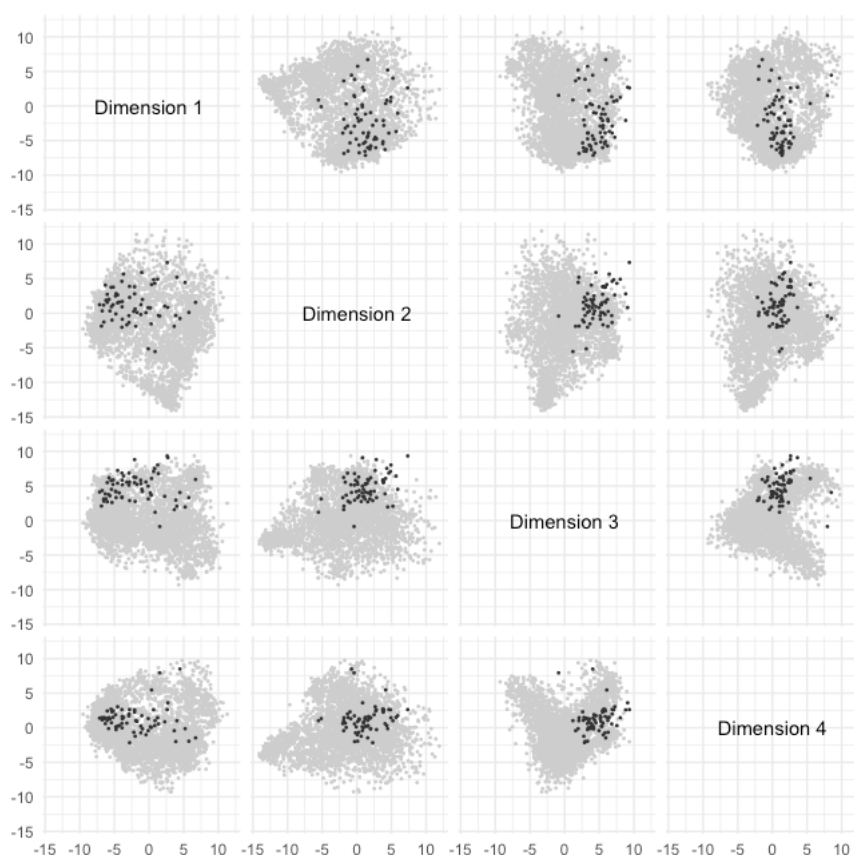


Figure 15: Languages of the Narrow Bantu subgroup highlighted on the dimension plots.

The many languages of the Narrow Bantu subgroup, found in a broad, contiguous range across southern Africa, are highly typologically congruent, being found high on Dimensions 1 and 3 ( $t = 50.14$ ,  $df = 587.47$ ,  $p < 0.001$ ;  $t = 51.67$ ,  $df = 318.28$ ,  $p < 0.001$ ), and low on Dimension 4 ( $t = -12.09$ ,  $df = 518.94$ ,  $p < 0.001$ ). They represent a typological extension away from the rest of the language cloud, seen in the plot of Dimension 3 vs. Dimension 4. Typological outliers of this group (lower on Dimension 2, or lower on Dimension 3) include the peripheral Bantu languages from the north-west of the Bantu expanse, in Cameroon, The Congo, or the Democratic Republic of the Congo, which are more isolating than the ‘modal’ Bantu language.

### 3.2.11. Greater Abyssinia

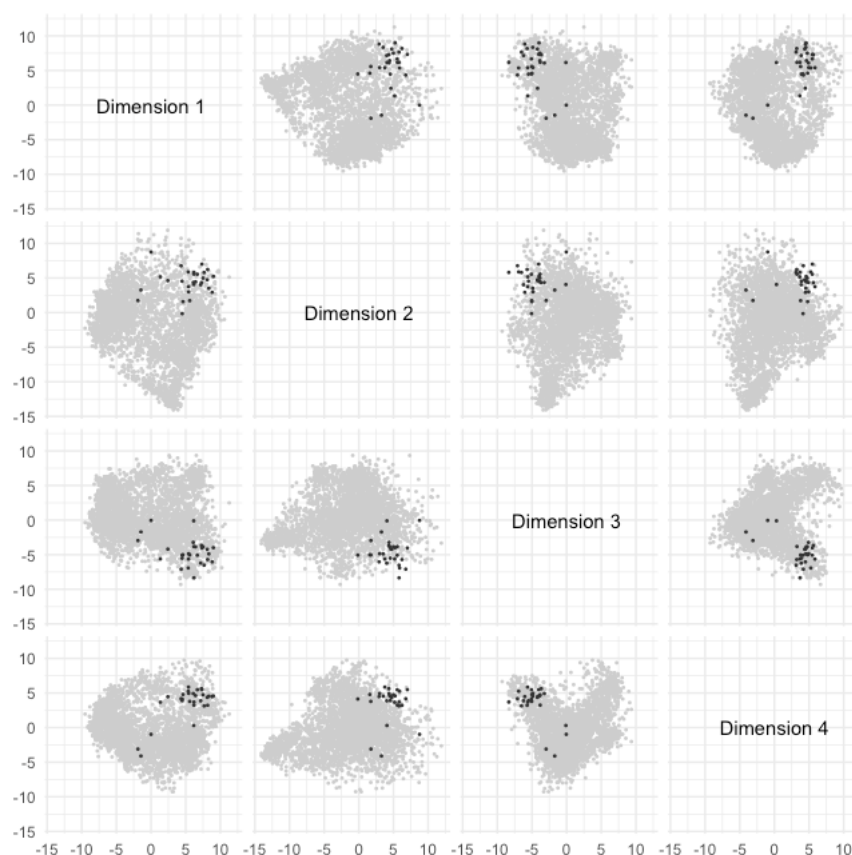
The languages of Greater Abyssinia, centred around the Horn of Africa, are typologically diverse, but are all relatively higher on Dimensions 2 ( $t = 5.25$ ,  $df = 105.57$ ,  $p < 0.001$ ), and lower on Dimension 3 ( $t = -5.77$ ,  $df = 84.27$ ,  $p < 0.001$ ), than the Bantu languages.



**Figure 16:** Languages of Greater Abyssinia highlighted on the dimension plots.

### 3.2.12. Pacific North-west

The languages of the Pacific North-west are typologically convergent languages from a number of families, high on Dimension 2 ( $t = 17.36$ ,  $df = 31.86$ ,  $p < 0.001$ ), low on Dimension 3 ( $t = -14.24$ ,  $df = 30.02$ ,  $p < 0.001$ ), and high on Dimension 4 ( $t = 8.30$ ,  $df = 29.04$ ,  $p < 0.001$ ). The outliers lower on Dimensions 1 or 4 are at the northern or southern edges of the area (Tlingit (tli; tlin1245, Na-Dene, Tlingit) and Haida (hdn; nort2938, Haida), Klamath (kla; klam1254, Klamath-Modoc), Kalapuya (kyl; kala1400, Kalapuyan) and Molala (mbe; mola1238), respectively).



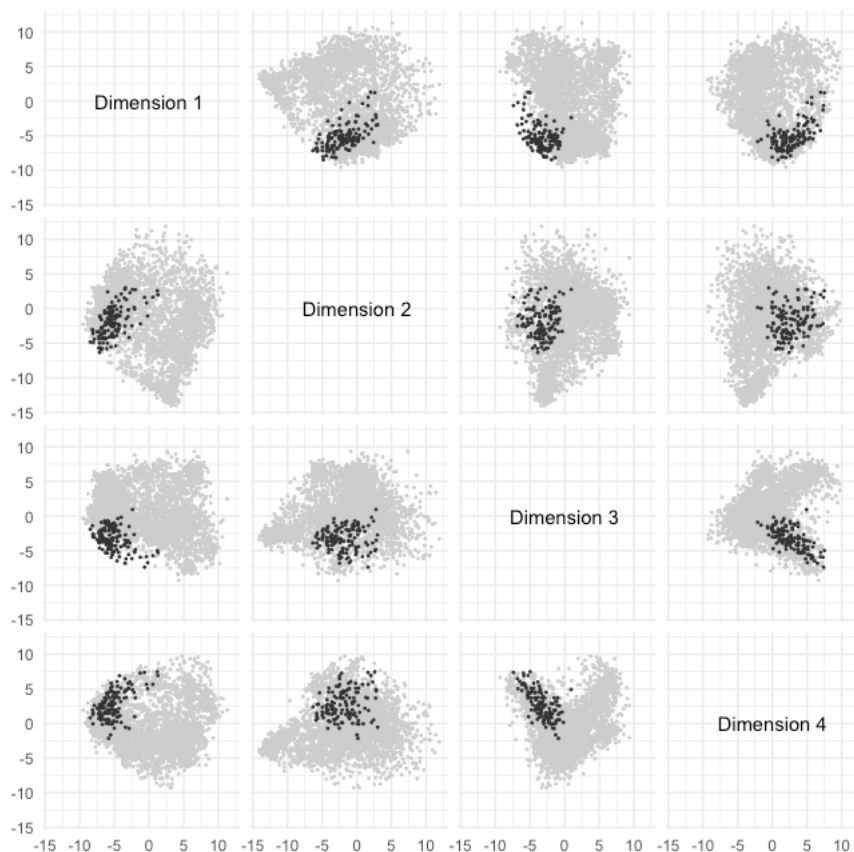
**Figure 17:** Languages of the Pacific North-west highlighted on the dimension plots.

### 3.2.13. Southern and Central Australia

The Pama-Nyungan languages of southern and central Australia occupy a small region of typological space which is low on Dimensions 1, 3, and 4 ( $t = -27.94$ ,  $df = 207.27$ ,  $p < 0.001$ ;  $t = -23.40$ ,  $df = 183.71$ ,  $p < 0.001$ ;  $t = 15.38$ ,  $df = 158.29$ ,  $p < 0.001$ ), and in the middle of Dimension 2 (two-sided  $t = -1.06$ ,  $df = 182.96$ ,  $p = 0.29$ ).



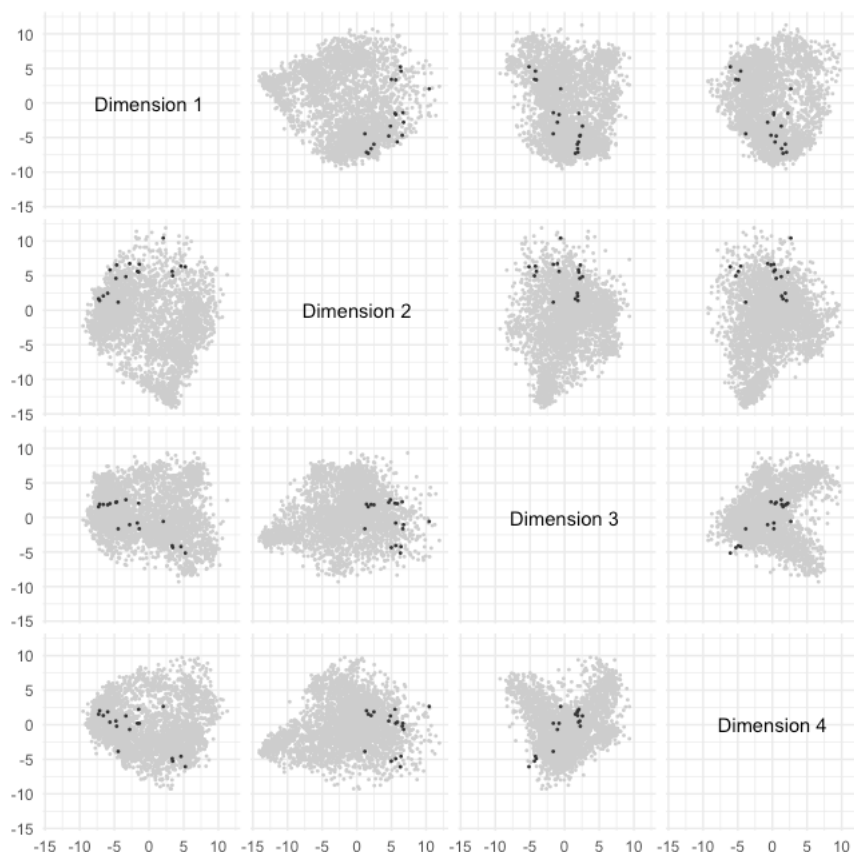
Languages higher on Dimension 2 are those which have some form of agreement, on the very or via clitics, and languages higher on Dimension 4 tend to be found in the south-east of the continent.



**Figure 18:** Languages of southern and central Australia highlighted on the dimension plots.

### 3.2.14. North-east Asia

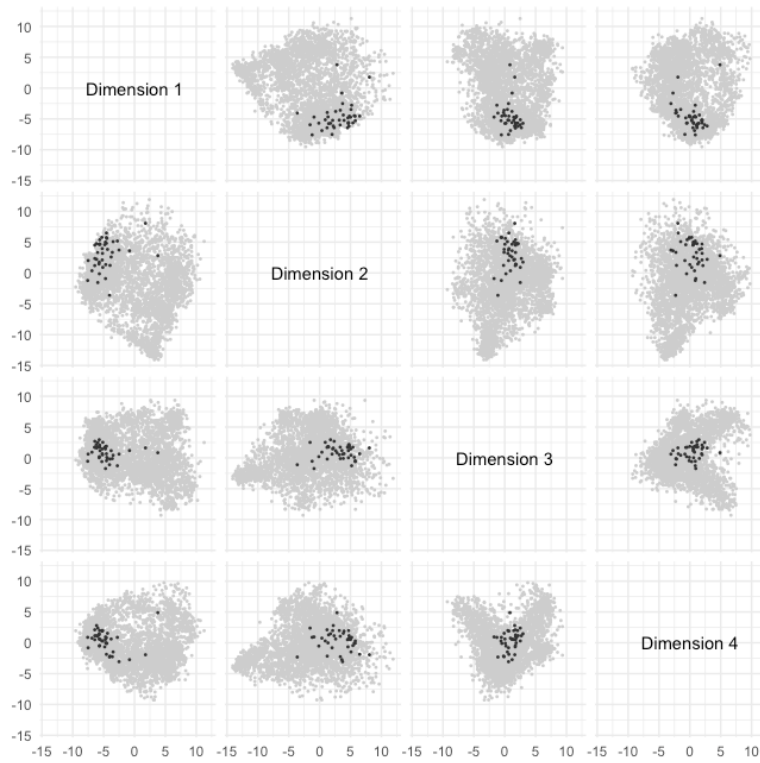
The languages of North-east Asia, comprising the Ainu and Chukotko-Kamchatkan language families, as well as the Tungusic languages north of Hokkaido and the Eskimo-Aleut languages spoken west of the Bering Strait, and the Yukaghir languages. These languages are all high on Dimension 2 ( $t = 11.59$ ,  $df = 17.85$ ,  $p < 0.001$ ), but do not occupy a typologically compact space in terms of the other three dimensions examined here (Dimension 1: two-sided  $t = -2.28$ ,  $df = 17.27$ ,  $p = 0.035$ ; Dimension 3: two-sided  $t = -1.08$ ,  $df = 17.36$ ,  $p = 0.29$ ; Dimension 4: two-sided  $t = -0.54$ ,  $df = 17.28$ ,  $p = 0.59$ ).



**Figure 19:** Languages of North-east Asia highlighted on the dimension plots.

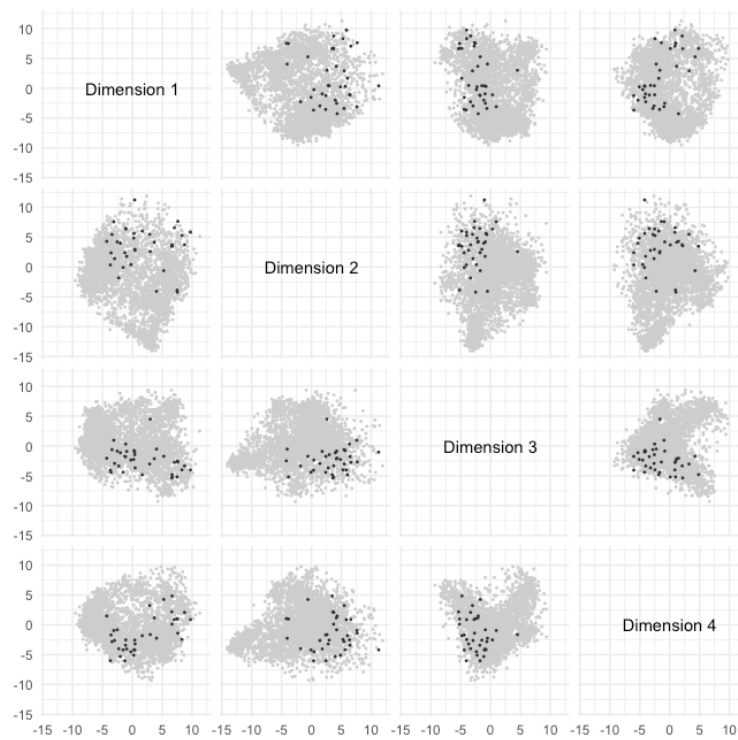
### 3.2.15. Andes

The languages of the Andes are for the most part quite compact; exceptions are the isolates Camsá (kbh; cams1241), Esmeraldeño (atac1235), and Cholon (cht; chol1284), and to a lesser extent the Chibchan language Kuna (kvn; sanb1242), all in the north of the region except for Cholon, in Pre-Andine Peru. The main group of Andean languages, from the Aymaran, Barbacoan, Chocoan, Jivaroan and Quechuan families, are low on Dimension 1 ( $t = -13.75$ ,  $df = 45.60$ ,  $p < 0.001$ ) and high on Dimension 2 ( $t = 12.13$ ,  $df = 44.46$ ,  $p < 0.001$ ), and occupy middle positions on Dimensions 3 and 4 ( $t = 2.80$ ,  $df = 51.54$ ,  $p < 0.01$ ; two-sided  $t = 2.30$ ,  $df = 44.66$ ,  $p = 0.026$ ).



**Figure 20:** Languages of the Andes family highlighted on the dimension plots.

### 3.2.16. *Mamoré–Guaporé*



**Figure 21:** Languages of the Mamoré–Guaporé area family highlighted on the dimension plots.

The languages of the Mamoré–Guaporé area are typologically diverse in terms of the global dimensions of variation. In an analysis restricted to just South America these languages emerge as distinct, reflecting the widespread use of prefixal agreement and possessive affixes in these VO languages. As with other languages of the Americas, these languages are low on Dimension 3 ( $t = -8.18$ ,  $df = 37.66$ ,  $p < 0.001$ ).

### **3.3. Universality? Macro-areas examined**

In this section we examine whether the features that characterise the dimensions described in 3.1 are also relevant within individual macro-areas, paralleling the methodology advocated by Dryer (1989, and subsequent publications) that seeks confirmation for universals by their universal, independent attestation around the world. We have already seen (3.2.5) that the languages of North America occupy a position that does not cover the full extent of the dimensions, but in many dimensions does occupy the fringe, suggesting that the parameters of variation within this macro-area will be different, in at least some respects, from those that pertain to the globe as a whole.<sup>27</sup> In this section we report in outline the results of applying FAMD to individual macro-areas.<sup>28</sup> Tables 8 – 11 show the features that were discussed for each of the global dimensions of variation that were described in 3.1, with each row corresponding to a different macro-area, indicating, for each feature of the global FAMD, which dimensions (if any) of the local macro-area show associations with that feature (if any). For example, the difference between dominant OV vs VO order, a feature associated with Dimension 1 in the global analysis (3.1.1), appears in Dimension 1 in Africa, Eurasia, North America, the Pacific, and South America, but is relegated to Dimension 2 in Australia, where word order is less dominant a variable; in Africa VO order is also a feature with a strong association with Dimension 4. Case marking, also a feature of (global) Dimension 1 in 3.1.1, appears in Dimension 1 in all of the macro-areas except North America, where it is only found in the third dimension. In Eurasia case marking is associated with both the first and second

---

<sup>27</sup> In Appendix 6 similar plots are given to show the distribution of the languages of the other macro-areas in terms of the global variation.

<sup>28</sup> Note that the individual analyses result in different numbers of relevant dimensions, following the methodology described in section 2. For Africa eight dimensions emerged as relevant (though just the first two are sufficient to account for the variance in most of the languages); Eurasia, Australia and South America require four dimensions each, and for the Pacific and North American macro-areas three dimensions are optimal.

dimensions, indicating its importance in that continent. Gender, a correlate of Dimension 3 in the global analysis, appears in Dimension 1 in Africa, Eurasia and Australia, but is not correlated with any of the major dimensions in the Pacific or the Americas because of its rarity in those areas.

Importantly, most of the features we encountered in the global analysis are also relevant in (most of) the individual macro-areas, though their representation is increasingly scattered in the higher dimensions. As mentioned above, the order of object and verb is relevant in all macro-areas, though less prominently in Australia than in other areas; dependent-marking is relevant in five macro-areas, though less so in North America than in others. Agreement appears in the first two dimensions in all macro-areas, and the order of subject and verb is relevant to different degrees in all of the macro-areas; the scarcity of verb-initial languages in Africa and Eurasia lowers the relevance of this feature in the Old World. The different valency-adding devices (causatives and applicatives), which show a similarity to agreement in that they encode argument information on the verb, are generally less prominent in individual macro-areas, but are still relevant. The features associated with Dimension 3 globally are much less well-attested in individual macro-areas; this is to be expected, since, as we saw in 3.1.3, this dimension essentially presents as a cline across the Old World, and so is much less prominent from analyses of variation in the languages of Australia, the Pacific or the New World. Nonetheless, these languages from the western edge of the populated world are typological outliers, as seen earlier in Figure 3, in contrast to most of the languages of the Circum-Pacific region, and so this dimension must be part of a global investigation.

Features	OV / VO	Case	Initial/final subordination	Prepositions	Agreement <sub>prefix</sub>	(Agreement <sub>suffix</sub> )	Obl V / V Obl	Postpositions	Total cases
Africa	1,4	1		1	1,2	1,3	1	1	1
Eurasia	1	1,2	1	1	3		1		2
Pacific	1	1	1	1	2		1	1	
Australia	2	1	4		1	2			
North America	1	3	1	1	3	3	1	1	3
South America	1	1	1		1	2	1	1	

**Table 8:** Features associated with Dimension 1, and their positions in macro-areal analyses.

Features	Total agreement	Verbal synthesis	Modality affixes	Incorporation	Applicatives	Causatives	Possessive prefixes	Total tenses	SVO order	Symmetrical
Africa	1,2,7	2	2	4,5	6	2,4	7	8	1	2
Eurasia	2,3	2	1	4	3,4	2		2	2	2
Pacific	2,3	1	1		3		3	1	1	1,2
Australia	1,3,4	1		1,4	3	3		3		
North America	2,3			2	2	2	3			
South America	1,3				2	2	1	1	1	3,4

**Table 9:** Features associated with Dimension 2, and their positions in macro-areal analyses.

Features	Gender	Obligatory plural marking	Agreement suffix	Relative pronouns	Ergativity	VOS order	Clusivity contrasts	negation	Initial
Africa	1	1	1	4,5	6,8	3,4	3,6		3
Eurasia	1	1	1	1,3	1,2				1
Pacific			1			1	1		2
Australia	1	1	2		1,4	2			2
North America		2			1	1			1
South America		3,4	1				3		2

**Table 10:** Features associated with Dimension 3, and their positions in macro-areal analyses.

In summary, most of the groupings of features we identified in 3.1 can be justified in the context of the analysis of individual macro-areas, suggesting that these associations between features are likely to arise from universal properties of human language.

The following sections briefly discuss the features that appear in the FAMD analyses of individual macro-areas, including those which are not present in the global analysis.<sup>29</sup>

<sup>29</sup> A more detailed explication of the FAMD analysis of the individual macro-areas is shown in Appendix 5.

Features	Initial negation	VSO order	Gen/Adj/Num N	Relative pronouns	Final negation	Inalienability	N Num	SV order
Africa	3	3,6		4,5	2		7	1,3
Eurasia	1	1		1,3		3		1
Pacific	2	2				1		2
Australia	2	2				1		2
North America	1	1			3		1	1
South America	2	2			2	3,1	2	2

**Table 11:** Features associated with Dimension 4, and their positions in macro-areal analyses.

### 3.3.1. Africa

We can see in Table 12 that most of the features that determine variation within Africa are consistent with the parameters of global variation.<sup>30</sup> The differences that can be found involve the strong correlation of postpositions with SOV languages in Africa, which is not found globally, and the widespread use of prefixal plural marking, which is so common amongst the Bantu and other Niger-Congo languages that it plays a large role in the continent as a whole. In the fourth dimension, varieties of Malagasy (bhr; bara1369) are differentiated by the nature of its voice system (here dubbed ‘superapplicative’, following Naylor 1995), and in the fifth and sixth dimensions, which are justified following the same procedures described in Section 2, we find features that identify certain Chadic and South Semitic languages which display infixation, and a small number of mostly East Sudanic languages which have ergative patterns.<sup>31</sup>

<sup>30</sup> Note that, as discussed in Section 2, we consider the languages of northern Africa, north of the Sahara, to be part of the macro-area Eurasia, rather than Africa, for the reasons outlined there. As such Arabic and Berber languages are not included in the analysis of Africa separate from Eurasia.

<sup>31</sup> Dimensions 2 and 3 correspond very closely to Dimension 2 (3.1.2) and Dimension 4 (3.1.4) from the global analysis.

Dimension	Low	High
1	SVO Agreement prefixes	Plural suffixes SOV, Postpositions
2	Negative particle	Verbal agreement Causatives
3	Subject-Predicate	Gender Predicate-Subject
4	'Superapplicative'	Double causatives, VOS, Incorporation Head marking
5	Noun-modifier orders	Incorporation Infixes
6	Gender in 1/2 pronouns	Applicatives Ergativity
7	Incorporation Possessive suffixes	Double negation Third agreement position
8	Possessive classes Third agreement position	VSO order Ergativity

**Table 12:** Relevant features: Africa.

'Low': features showing a negative correlation with the relevant dimension;

'High': features showing a positive correlation with the relevant dimension.

### 3.3.2. Australia

Australia is most at variance with global norms in terms of morphosyntactic variation. As seen above, word order is not a correlate of the first dimension of variation in Australia (though it is represented in the second dimension). The features correlating with the major dimension of variation in Australia correspond to the long-discussed Pama-Nyungan/non-Pama-Nyungan divide, with the north(-west)ern non-Pama-Nyungan languages displaying prefixal agreement on verbs, often with portmanteau subject/object morphemes, clusivity contrasts in bound morphology, and gender systems. Opposing this are the Pama-Nyungan languages that occupy most of the continent, which tend to be more dependent-marking, with ergative case marking, and typically lacking gender contrasts. The second dimension picks out languages, typically in the south-east of the continent, which are verb-initial and which employ pronominal bases to which a productive affix is added (Daniel 2013).



Dimension	Low		High	
1	Ergativity	Suffixing	Gender, Clusivity	Prefixal agreement
2	SOV order		Pronominal bases	VOS order
3	Verb agreement		Causatives	Applicatives
4	Verb agreement	Subordinating suffixes	Incorporation	N Dem order

**Table 13:** Relevant features: Australia.

### 3.3.3. Eurasia

As with Africa, the mapping of Eurasia in Section 4 reveals a number of clearly separated areas. Unique features associated with dimension 1 include the contrast between isolating languages and tense-marking languages. The second dimension has a strong east-west distribution, with high values in the west, where word order is manipulated to form content questions, and relative pronouns are used as subordinators.

Dimension	Low		High	
1	Tense, SOV	Case marking	VO	Isolating
2	Prenominal relative clause		Initial Wh-, subject suffixes, gender	Relative pronouns
3	Accusative pronouns		Applicatives	Prefixal agreement
4	VS			SV
5	Causatives			Ergativity

**Table 14:** Relevant features: Eurasia.

### 3.3.4. Pacific

In the Pacific we again see an OV vs. VO divide along the first dimension, correlating with suffixal subject morphology amongst the ‘OV Papuan’ languages and clusivity contrasts in the VO Austronesian languages. The second dimension introduces prefixal agreement as a major correlate.

Dimension	Low	High
1	SOV Subject agreement suffixes	SVO Clusivity contrasts
2	Prefixal agreement SV	Initial negation, VS Case marking
3	Applicatives, total agreement positions	

Table 15: Relevant features: Pacific.

### 3.3.5. North America

Most of the features that appear in the analysis of North America are also present in the global analysis. Additionally, the second dimension distinguishes between prefixing and suffixing languages.

Dimension	Low	High
1	SOV	Negator-verb Verb-predicate
2	Suffixing Causatives	Prefixal agreement Prefixing
3	Applicatives Total agreement positions	

Table 16: Relevant features: North America.

### 3.3.6. South America

Dimension	Low	High
1	SOV Case marking	Prefixal agreement Prefixal possession
2	Symmetrical	Suffixal possession, object agreement Applicatives
3	Double negation Clusivity in bound morphology	Verb agreement Causatives
4	Modality affixes Tense	Suffixes plural marking Initial question particles
5	Applicatives Dem N order	

Table 17: Relevant features: South America. ‘Low’: features showing negative associations with the relevant dimension; ‘High’: features showing a positive association with the relevant dimension.

South America shows the same VO vs. OV divide in the first dimension, but with less clear areality than is seen in other macro-areas. There is a strong Andean area defined by Dimension 2, abutting a Pre-Andine area defined by Dimension 3.

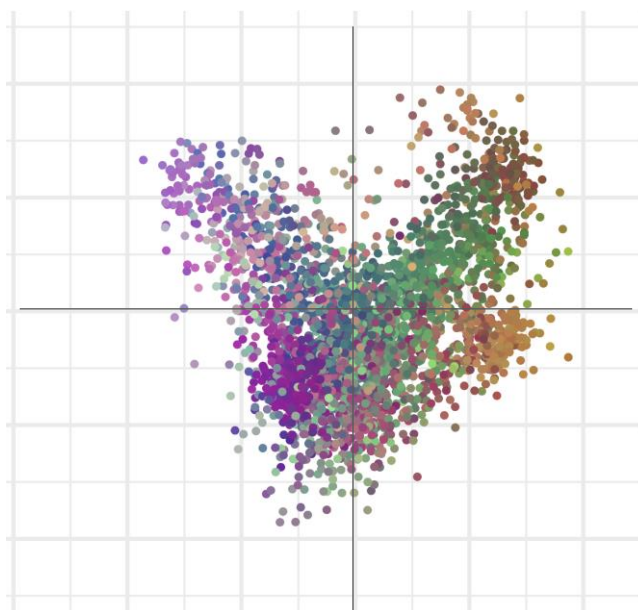
### 3.4. Correlations between dimensions?

By definition the different dimensions are as independent of each other as possible. Table 18 shows the overall correlations that can be found between the different dimensions; none of these correlations are significant, as shown in Table 18.

Dimension	1	2	3
1			
2	<i>0.034</i>		
3	<i>0.021</i>	0.025	
4	<i>0.001</i>	0.023	0.020

**Table 18:** Overall correlations between dimensions ( $r^2$ ; negative correlations shown in italics).

Despite the different dimensions being overall independent, some correspondence is inevitable due to the presence of the same or similar features in more than one dimension.



**Figure 22:** Dimension 3 (x) and Dimension 4 (y).

More interestingly, some dimensions correlate with others in just part of their range. Most dramatically, while there is only minimal correlation between Dimension 3 and Dimension 4 ( $r^2 = 0.020$ ), the plot showing these two dimensions together (Figure 22, from Figure 3) shows clearly that at the lower half of Dimension 3 (on the left) there is a negative correlation with Dimension 4 ( $r^2 = 0.290$ ), while at the upper half of Dimension 3 (on the right) there is a positive correlation with Dimension 4 ( $r^2 = 0.258$ ). This reflects a split in the typology of verb-initial languages depending on whether they are found in the west (Celtic, Semitic) or the east (Philippines, Central America), since these languages are not typologically uniform. There is a weak positive correlation between Dimensions 1 and 4 in the upper half of Dimension 1 ( $r^2 = 0.064$ ), reflecting the rarity of languages that are both OV and VS in their word order. We find a slightly higher positive correlation with Dimension 2 at the bottom tail (that is, for those values more than one standard deviation below the mean) of Dimension 1 ( $r^2 = 0.094$ ); this corresponds to the fact that the very bottom of Dimension 1 is occupied by languages from the Himalayas (which are both low on Dimension 1, and concentrated in central Eurasia), which are less morphologically elaborate than many less ‘extreme’ SOV languages, and moving away from this edge almost inevitably leads to greater morphological elaboration, higher on Dimension 2. The lower half of Dimension 2 shows positive correlations with both Dimension 3 ( $r^2 = 0.205$  in the extreme bottom) and Dimension 4 ( $r^2 = 0.153$  in the lower half). These correlations largely reflect the position of the languages of Europe, high in Dimension 3 and Dimension 4, but in the lower half of Dimension 2. The last correlation we draw attention to is the lower half of Dimension 4, where we find a weak positive correlation with Dimension 3 ( $r^2 = 0.085$ ) (see Figure 22).

#### **4. Features with minimal contribution to global linguistic variation**

In Section 3 we discussed the features that contribute to the dimensions that best describe global (and regional) morphosyntactic variation. This section briefly discusses some of the features that have the least contribution to global variation, either because they are so rare, they have a very limited distribution, or they appear in so many languages with little or no association with other parts of the language (at least, as far as is coded in the database used). Some of these features are listed in Table 19, which is not intended to be exhaustive.

Extremely Rare	Geographically Limited	Widely Ubiquitous
Polar questions formed by verbal reduplication	Genitive subjects	Predicative possession with a 'have' verb, or genitive subject
More than three agreement positions on the verb	Polar questions forms with word order change	Polar questions forms with particles or intonation
Marked absolutive case	Verb agreement by tone	Adnominal demonstrative identical to pronominal demonstratives
Incorporation of transitive subjects into verbs	Philippine-type voice systems	Presence of a perfective in the aspect system

**Table 19:** Different features with minimal contribution to global categories of variation.

Examples of some of the 'extremely rare' features are shown in (4) – (7); in Yao'an Lolo (ycl; lolo1259, Tibeto-Burman, Lolo-Burmese; Merrifield 2010) the only marker of the question is the reduplication of the verb (the only language in our database with this feature). In KinyaRwanda (Kimenyi 1980) we see a verb with five agreement positions filled on the verb; the database contains only 12 languages with more than three agreement positions. The Nias (nia; nias1242, Austronesian, Batak-Barrier Islands; Donohue and Brown 1999, Brown 2005, Donohue 2008) sentences show the alternation of the unmarked *ulö* 'snake' in an A function, and the marked *g-ulö* 'snake' in absolutive functions; sixteen other languages in the database have marked absolutive cases, most (11) of which also mark the ergative role. In Boni (/Aweer) (orm; awee1242, Afro-Asiatic, Cushitic, Omo-Tana; Sasse 1984) we see the rare case of an incorporated A (Sasse notes that while a natural translation involves the passive, the verbform in (7) is clearly transitive); only three other languages are known to us with this feature.

(4) Yao'an Lolo: reduplication on verbs marking polar questions

*Ni pia cir-cir ho ar?*  
 2SG clothes wash-RED REAL PFV  
 'Have you already washed the clothes?'

(5) Kinyarwanda: more than three agreement positions

*Abáana ba-zaa-ha-ki-mu-b-eerek-er-a.*

children they-FUT-there-it-him-them-show-BEN-ASP

‘The children will show it to him for them there.’

(6) Nias: marked absolutive case

a. *I-usu g-ulö asu hö'ö.*

3SG.ERG-bite ABS-snake dog DIST

‘That dog bit the snake.’

b. *I-usu n-asu ulö hö'ö.*

3SG.ERG-bite ABS-dog snake DIST

‘That snake bit the dog.’

c. *Möi ga g-ulö.*

go here ABS-snake

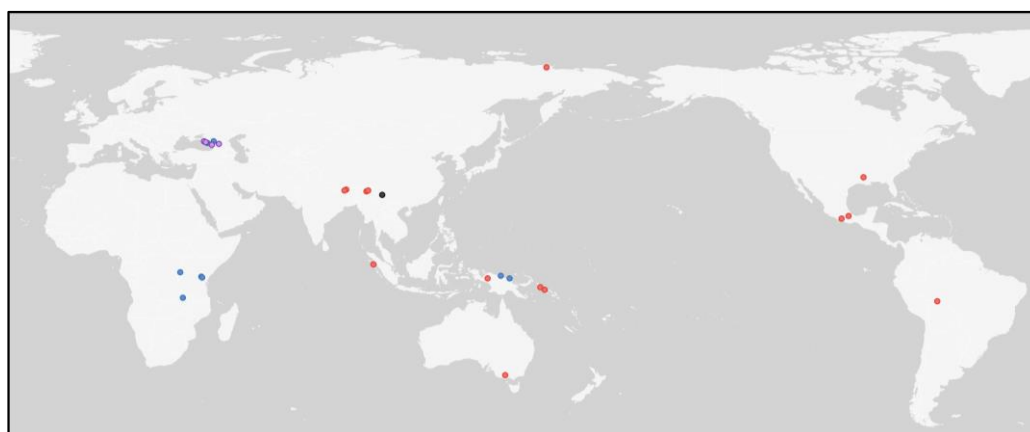
‘The snake came this way.’

(7) Boni: incorporated transitive subject

*Míŋ qweŋra kawáyð'aadéed'i idohóo^d'isa.*

house Boni/GEN usually women^build/IMPERF/3SG.M

‘Boni houses are usually built by women.’

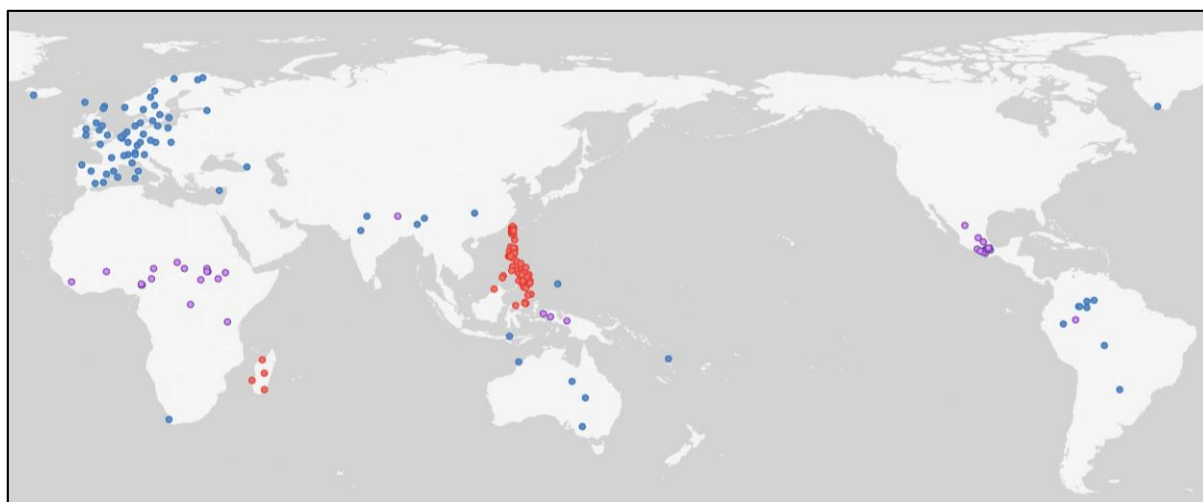


**Map 9:** Languages with extremely rare features.

Blue = more than three agreement positions; Red = marked absolutive case; Purple = both the preceding features; Black = polar questions marked by reduplication.

The distributions of languages displaying the first three of these features are shown in Map 9, where red indicates languages with marked absolutes, blue indicates languages with more than three agreement positions on the verb (compare with Map 1, which shows the total sample examined).

Examples of features that are more common than those shown in Map 9, but which have strong geographic concentrations, are shown in Map 10. While a number of parts of the world have languages in which word order changes in polar questions, the concentration in western and northern Europe is striking. Languages with Philippine-type voice systems are largely restricted to the Philippines and Taiwan, with the outlier group in Madagascar reflecting the migration from Southeast Asia ca 1,500 years ago. Languages which have tone as an exponent of verbal agreement are concentrated in Central America and in Central Africa.



**Map 10:** Features with geographically restricted ranges.

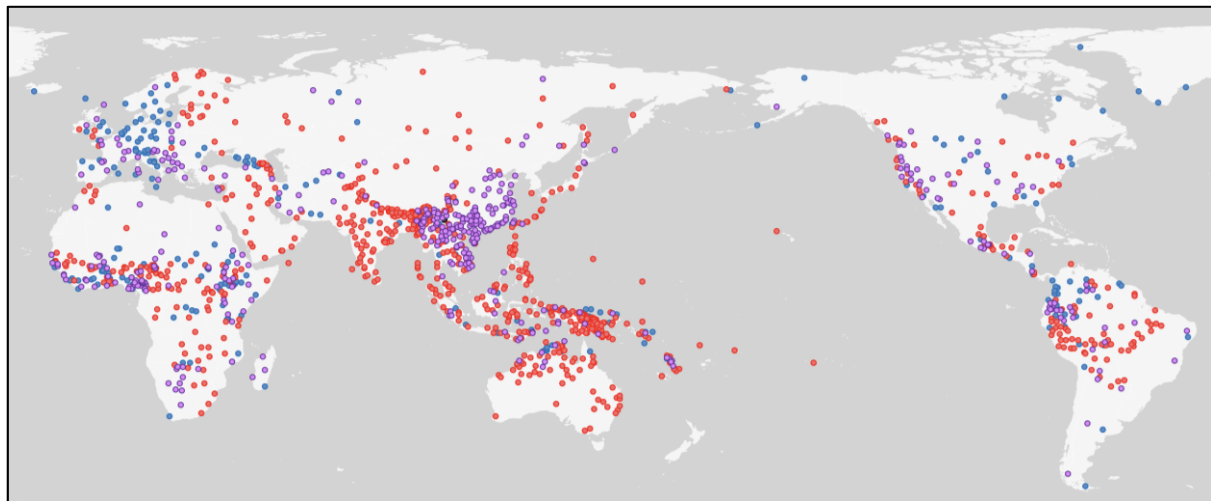
Blue = word order changes in polar questions;

Red = languages with Philippine-type voice systems;

Purple = tone as an exponent of verbal agreement.

Two features listed as widely ubiquitous in Table 19, the existence of a ‘have’ verb in the language, or the use of a particle to mark polar questions, are shown in Map 11. Both of these features are found across the map, though with different frequencies in

different continents. They are so widespread that they have very little probative value in understanding global morphosyntactic variation.<sup>32</sup>



**Map 11:** Features with widely ubiquitous distributions.

Blue = language includes a 'have' verb;

Red = language uses a particle to mark polar questions;

Purple = language has both a 'have' verb, and a polar question particle.

## 5. Conclusions

Without explicitly setting out to do so, our study has quantitatively confirmed many of the insights of 20th-century typological research concerning the main dimensions of morphosyntactic variation, by finding them as emergent properties of a bottom-up investigation of a large body of morphosyntactic data. We have shown that much of the variation between languages, both globally and within macro-areas, can indeed be largely explained by established typological parameters, as described in 3.1: the order of subjects and objects with verbs, dependent-marking settings, and the position of genitives, numerals and adjectives with respect to the nouns that they modify. The features that correlate with Dimensions 2 and 3, head-marking settings, verbal elaboration, and a number of features that are reminiscent of 'Standard Average European', have not all been proposed as factors underlying typological variation, but have been demonstrated here to be as important as more familiar word order

---

<sup>32</sup> A glance at Map 11 raises the suspicion that these might be relevant features at local levels; the distribution of 'have' verbs in South America, for instance, appears to be concentrated in the north-west.



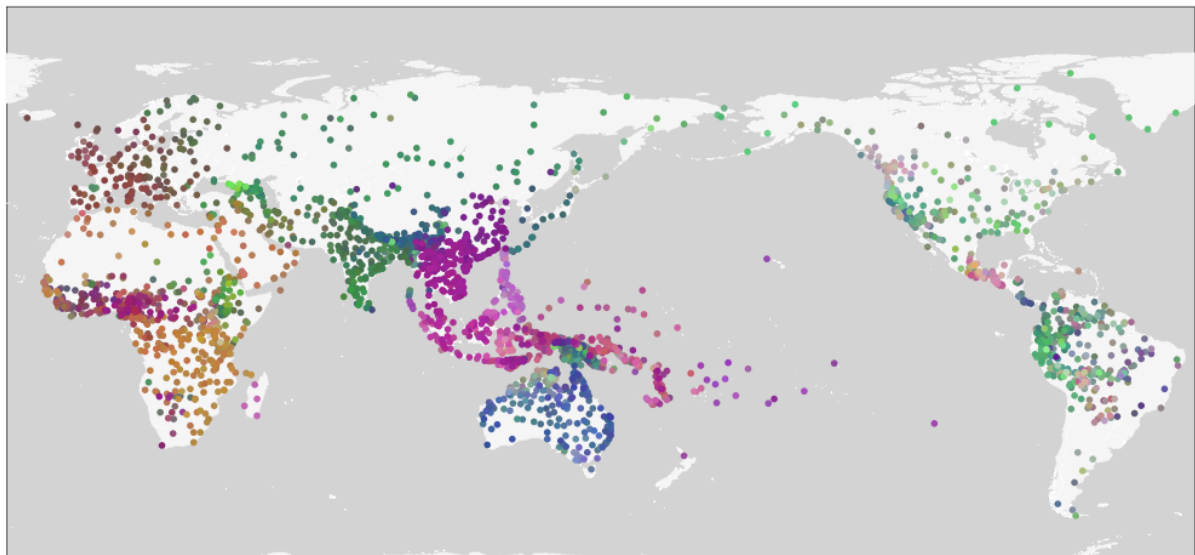
universals. The kinds of features described in 3.1 are summarised in Table 20; we can see that the main correlates of these dimensions are evenly split into those involving word order, and those involving morphology, with different dimensions addressing different kinds of word order or morphology. Secondly we find that the presence of case-marking morphology, or the lack of extensive morphological processes in the language, or else word order parameters more relevant to Noun Phrase-internal elements, are also significant factors in assessing global morphosyntactic variation.

Dimension	Main type of feature	Secondary types of features
1	Word order (clausal; object)	Dependent marking
2	Verbal morphology, head marking	Isolating profile
3	Nominal morphology, gender	Ergativity, Clusivity
4	Word order (clausal; subject)	Word order (Noun Phrase)

**Table 20:** Different features with minimal contribution to global categories of variation.

In Map 12 we see the 3089 languages of the database, coloured according to the position of each language on Dimensions 1–3, with these dimensions mapped to red, green and blue colour components, as described in 3.1. Each language is then represented with a dot of the same colour that was seen in Figure 3, so that in Map 12 the colouring solely represents the position of the individual languages in typological space as defined by the first three dimensions – that is, the colour scheme is an emergent property of the linguistic analysis, and in no way involved any phylogenetic information, and only involved reference to geography in that AUTOTYP areas were used as controls in the analysis.<sup>33</sup>

<sup>33</sup> This means that Dimension 4 is not represented in the colouring in Map 12. Dimension 4 is shown in Map 8, and in Appendix 8 alternatives to Map 12, and Figure 3, are shown with colourings representing different combinations of dimensions, including Dimension 4.



**Map 12:** Emergent areality around the world. Map coloured according to the first three dimensions of variation, as discussed in 3.1, and shown in Figure 3

A number of regions are clearly identified by both geography and typology, as discussed in 3.2. Europe (e.g., Haspelmath 2001) is represented with a distinct brown colour, which is largely due to the languages being moderately low on Dimension 2, and high on Dimension 3 (compare Figure 3 in Section 3.1 and Figure 4 in Section 3.2.1). The isolating languages of Southeast Asia (e.g., Enfield 2005) are shown in magenta, as they are relatively high on Dimension 1, low on dimension 2, and moderately low on Dimension 3 (compare Figure 3 and Figure 6 in 3.2.2). Similar, though not as extreme, colours are found in the Macro-Sudan Belt (Güldemann 2009) in west-central Africa, and in Polynesia (e.g., Krupa 1982). Most of Australia (e.g., Bowern 2006, Dixon 2017) is in dark blue, as the languages there are low on Dimension 1 and moderately low on Dimension 2. The only bright green regions, high on Dimension 2, are found in the Caucasus (Catford 1977) and parts of North America (e.g., Mithun 1999). In North America the Pacific North-west (Mithun 2010) stands out, coloured in grey and having much higher values on Dimension 1; a similar typology is evident in the Oaxaca area in Central America, and the Mamore-Guapore region of South America (Campbell et al. 1986; Crevels & van der Voort 2008; see 3.2.16). The orange colour found in much of Sub-Saharan Africa represents the Niger-Congo Bantu languages, high in Dimensions 1, 2 and 3 (and low in Dimension 4, though this is not apparent from the colouring on the map); see 3.2.10. Separated by the Macro-Sudan belt, the verb-initial languages of North Africa and Arabia appear in dark orange, reflecting their position at the top of Dimension 4, but also high on

Dimension 3 (3.2.1). The languages of Taiwan and the Philippines are also high in Dimension 4, coloured in mauve following their position low on Dimension 3 (3.2.7).<sup>34</sup> The densely occupied space low in Dimension 1 and middling in Dimensions 2 and 3 is coloured in dark green, and spans the Eurasian steppe and South Asia (Janhunen 2023, Emeneau 1956, and many others; see 3.2.3 and 3.2.4); higher on Dimension 2, but otherwise in a similar position, the languages of the Ethiopian linguistic areas (Crass 2009) are coloured in a slightly lighter shade of green, and a darker green is found for the languages of Japan and Korea, representing a position lower on Dimension 2. Many of the languages of the Andes in South America are reminiscent of this pan-Eurasian typology (Constenla Umaña 1991, Adelaar 2009, Michael et al. 2012; see 3.2.15). In addition to these previously-discussed regions in typological space, we can also identify a number of emergent areas on the map, such as South-west China, and North-west Australia, Oaxaca (within Meso-America; 3.2.6), the Kimberleys in Australia, and the South-east Amazon, all clearly identifiable on Map 12.<sup>35</sup>

We mentioned in Section 2 our decision to use nested geographic areas, rather than genealogies, as controls. While most ‘controls’ in recent linguistics studies are based on genealogies, we have based our work on culturally-defined areas, specifically a set of 25 areas slightly modified from the AUTOTYP areas, as described in Section 2.

Our results not only confirm typologists’ intuitions about the features that are most important for typological classification, but also show the efficacy of a bottom-up approach to the detection and mapping of areal patterns in morphosyntax.<sup>36</sup> The success (in terms of interpretable results) of the use of a large set of linguistic features, without any cherry-picking, shows that a holistic (or even ‘super-holistic’) approach to language typology (following, e.g., Ramat 1986, Plank 1998, Comrie 1988, 2001, and others) is a valid way to objectively assess claims about linguistic areality or linguistic universals.

---

<sup>34</sup> Taken with the grey areas discussed in the Americas, and the mauve from the Philippines, the languages of North Africa-Arabia represent a third verb-initial linguistic ‘types’, with a large number of typological features not associated with the verb-initial parameter.

<sup>35</sup> A higher-resolution version of this map can be found in Appendix 7.

<sup>36</sup> It has been suggested that we attempt a similar analysis using the Grambank database. This research has already been performed (Skirgård et al. 2023), and, owing to the different and smaller set of languages coded for a different and smaller set of features, the results are very different, though we note that Skirgård et al. 2023 also appear to have identified word order, head/dependent marking and gender as relevant to their analysis.

## Acknowledgments

We thank two anonymous referees, whose input has greatly improved this paper, and the input from the editorial team that similarly contributed to the clarity of our presentation. We are grateful for their feedback.

## Abbreviations

1 = 1 <sup>st</sup> person	BEN = benefactive	GEN = genitive
2 = 2 <sup>nd</sup> person	Dem = demonstrative	IMPERF = imperfective
3 = 3 <sup>rd</sup> person	DIST = distal	M = masculine
ABS = absolutive	ERG = ergative	PFV = perfective
ASP = aspect	FUT = future	REAL = realis

## References

- Adelaar, Willem F. H. 2009. *The Languages of the Andes*. Cambridge: Cambridge University Press.
- van der Auwera, Johan. 2011. Standard Average European. In Bernd Kortmann & Johan van der Auwera (eds.), *The Languages and Linguistics of Europe: A Comprehensive Guide*, 291–306. Berlin: de Gruyter Mouton.
- Bickel, Balthasar. 2002. The AUTOTYP research program. Invited talk given at the Annual Meeting of the Linguistic Typology Resource Center Utrecht, September 26–28, 2002.
- Bickel, Balthasar & Johanna Nichols. 2006. Oceania, the Pacific Rim, and the Theory of Linguistic Areas. *Proceedings of the 32nd annual meeting of the Berkeley Linguistics Society*, 3–15. Berkeley Linguistics Society and the Linguistic Society of America.
- Bickel, Balthasar & Johanna Nichols. 2013. Inflectional Synthesis of the Verb. In Matthew S. Dryer & Martin Haspelmath (eds.), *The World Atlas of Language Structures Online*, Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at <http://wals.info/chapter/22>.)
- Bickel, Balthasar, Johanna Nichols, Taras Zakharko, Alena Witzlack-Makarevich, Kristine Hildebrandt, Michael Rießler, Lennart Bierkandt, Fernando Zúñiga & John B. Lowe. 2023. The AUTOTYP database (v1.1.1). <https://doi.org/10.5281/zenodo.7976754>.

- Bowern, Claire. 2006. Another Look at Australia as a Linguistic Area. In Yaron Matras, April McMahon & Nigel Vincent (eds.), *Linguistic Areas*, 244–265. London: Palgrave Macmillan.
- Brown, Lea. 2005. Nias. In Alexander Adelaar & Nikolaus Himmelmann (eds.), *The Austronesian Languages of Asia and Madagascar*, 562–589. London: Routledge.
- Bugaeva, Anna, Johanna Nichols & Balthasar Bickel. 2021. Appositive possession in Ainu and around the Pacific. *Linguistic Typology* 26(1). 43–88.
- Campbell, Lyle. 1997. *American Indian languages: The historical linguistics of Native America*. New York: Oxford University Press.
- Campbell, Lyle, Terrence Kaufman & Thomas Smith-Stark. 1986. Meso-America as a linguistic area. *Language* 62(3). 530–558.
- Catford, John C. 1977. Mountain of tongues: the languages of the Caucasus. *Annual Review of Anthropology* 6. 283–314.
- Cattell, Raymond B. 1966. The Scree Test for The Number of Factors. *Multivariate Behavioral Research* 1(2). 245–276.
- Clauson, Gerard. 1956. The case against the Altaic theory. *Central Asiatic Journal* 2. 181–187.
- Comrie, Bernard. 1988. Linguistic Typology. *Annual Review of Anthropology* 17. 145–159.
- Comrie, Bernard. 2001. Different views of language typology. In Martin Haspelmath, Ekkehard König, Wulf Oesterreicher & Wolfgang Raible (eds.), *Language typology and language universals: An international handbook, Vol. 1*, 24–39. Berlin: Walter de Gruyter.
- Constenla Umaña, Adolfo. 1991. *Las lenguas del área intermedia: introducción a su estudio areal*. San José: Editorial de la Universidad de Costa Rica.
- Crass, Joachim. 2009. Ethiopia. In Bernd Heine & Derek Nurse (eds.), *A Linguistic Geography of Africa*, 228–250. Cambridge: Cambridge University Press.
- Crevels, Mily & Hein van der Voort. 2008. The Guaporé-Mamoré region as a linguistic area. In Pieter Muysken (ed.), *From Linguistic Areas to Areal Linguistics*, 151–179. Studies in Language Companion Series. Vol. 90. Amsterdam: John Benjamins.
- Croft, William. 2003. *Typology and Universals*. Cambridge: Cambridge University Press.
- Daniel, Michael. 2013. Plurality in Independent Personal Pronouns. In Matthew S. Dryer & Martin Haspelmath (eds.), *The World Atlas of Language Structures Online*.

- Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at <http://wals.info/chapter/35>.)
- DeLancey, Scott & Victor Golla. 1997. The Penutian hypothesis: Retrospect and prospect. *International Journal of American Linguistics* 63(1). 171–202
- Dimmendaal, Gerrit J. 2011. *Historical Linguistics and the Comparative Study of African Languages*. Amsterdam: John Benjamins.
- Dixon, R. M. W. 2017. The Australian Linguistic Area. In Alexandra Y. Aikhenvald & R. M. W. Dixon (eds.), *The Cambridge Handbook of Linguistic Typology*, 624–650. Cambridge: Cambridge University Press.
- Donohue, Mark. 2008. Semantic alignment systems: what's what, and what's not. In Mark Donohue & Søren Wichmann (eds.), *Semantic alignment: typological and descriptive studies*, 24–75. Oxford: Oxford University Press.
- Donohue, Mark & Lea Brown. 1999. Ergativity: some additions from Indonesia. *Australian Journal of Linguistics* 19(1). 57–76.
- Dryer, Matthew S. 1989. Large linguistic areas and language sampling. *Studies in Language* 13. 257–292.
- Dryer, Matthew S. 1992. The Greenbergian Word Order Correlations. *Language* 68(1). 81–138.
- Dryer, Matthew S. 2013. Against the six-way order typology, again. *Studies in Language* 37. 267–301.
- Dryer, Matthew S. & Martin Haspelmath (eds.) 2013. *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at <http://wals.info>, Accessed on 2023-01-06.)
- Emeneau, Murray. 1956. India as a Linguistic Area. *Language* 32 (1). 3–16.
- Enfield, Nicholas J. 2005. Areal Linguistics and Mainland Southeast Asia. *Annual Review of Anthropology* 34(1). 181–206.
- Gabelenz, Georg von der. 1901. *Die Sprachwissenschaft: Ihre Aufgaben, Methoden und bisherigen Ergebnisse*. Leipzig: Chr. Herm. Tauchnitz.
- Greenberg, Joseph. 1963. Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements. In Joseph Greenberg (ed.), *Universals of Language*, 73–113. London: MIT Press.
- Güldemann, Tom. 2009. The Macro-Sudan belt: towards identifying a linguistic area in northern sub-Saharan Africa. In Bernd Heine & Derek Nurse (eds.), *A Linguistic Geography of Africa*, 151–185. Cambridge: Cambridge University Press.

- Guzmán-Naranjo, Matías & Laura Becker. 2021. Statistical bias control in typology. *Linguistic Typology* 26(3). 605–670.
- Hammarström, Harald & Mark Donohue. 2014. Some principles on Macro-Areas in Typological Comparison. *Language Dynamics and Change* 4(1). 167–187.
- Haspelmath, Martin. 2001. The European linguistic area: Standard Average European. In Martin Haspelmath, Ekkehard König, Wulf Oesterreicher & Wolfgang Raible (eds.), *Language Typology and Language Universals, Vol. 2*, 1492–1510. Berlin: De Gruyter.
- Humboldt, Wilhelm von. 1836. *Über die Verschiedenheit des menschlichen Sprachbaues und ihren Einfluß auf die geistige Entwicklung des Menschengeschlechts*. Berlin: Druckerei der Koniglichen Akademie der Wissenschaften.
- Janhunen, Juha A. 2023. The Unity and Diversity of Altaic. *Annual Review of Linguistics* 9(1). 135–154.
- Josse, Julie & François Husson. 2016. missMDA: A Package for Handling Missing Values in Multivariate Data Analysis. *Journal of Statistical Software* 70(1). 1–31.
- Kari, James & Ben Potter (eds.). 2010. *The Dene-Yeniseian connection*. Anthropological Papers of the University of Alaska, New Series 5 (1-2). Fairbanks: Alaska Native Language Center.
- Kimenyi, Alexandre. 1980. *Relational Grammar of Kinyarwanda*. University of California Publications, Linguistics, 91. Berkeley: University of California Press.
- Krupa, Viktor. 1982. *Polynesian Languages: a survey of research*. London: Routledge and Kegan Paul.
- Lê, Sébastien, Julie Josse & François Husson. 2008. FactoMineR: An R Package for Multivariate Analysis. *Journal of Statistical Software* 25(1). 1–18.
- Macklin-Cordes, Jayden & Erich Round. 2022. Challenges of sampling and how phylogenetic comparative methods help, with a case study of the Pama-Nyungan laminal contrast. *Linguistic Typology* 26(3). 533–572.
- de Menocal, Peter, Joseph Ortiz, Tom Guilderson, Jess Adkins, Michael Sarnthein, Linda Baker & Martha Yarusinsky. 2000. Abrupt onset and termination of the African Humid Period. *Quaternary Science Reviews* 19(1–5). 347–361.
- Merrifield, Judith Thomas. 2010. *Yao'an Lolo Grammar Sketch*. Dallas: Graduate Institute of Applied Linguistics. (MA thesis.)
- Michael, Lev, Will Chang & Tammy Stark. 2012. Exploring phonological areality in the circum-Andean region using a Naive Bayes Classifier. *Language Dynamics and Change* 4(1). 27–86.

- Mithun, Marianne. 1999. *The Languages of Native North America*. Cambridge: Cambridge University Press.
- Mithun, Marianne. 2010. Contact and North American Languages. In Raymond Hickey (ed.), *The Handbook of Language Contact*, 673–694. Oxford: Wiley-Blackwell.
- Naylor, Paz Buenaventura. 1995. Subject, topic, and Tagalog syntax. In David Benett, Theodora Bynon & George Hewitt (eds.), *Subject, Voice and Ergativity*, 161–201. London: School of Oriental and African Studies.
- Nerbonne, John. 2009. Data-Driven Dialectology. *Language and Linguistics Compass* 3(1). 175–198.
- Nichols, Johanna. 1986. Head-marking and dependent-marking grammar. *Language* 62(1). 56–119.
- Nichols, Johanna, Alena Witzlack-Makarevich & Balthasar Bickel. 2013. The AUTOTYP genealogy and geography database: 2013 release. <http://www.spw.uzh.ch/autotyp/>.
- Pagès, Jérôme. 2004. Analyse factorielle de données mixtes. *Revue de Statistique Appliquée* 4. 93–111.
- Pawley, Andrew & Harald Hammarström. 2018. The Trans New Guinea family. In Bill Palmer (ed.), *The Languages and Linguistics of the New Guinea Area: a Comprehensive Guide*, 21–196. *The World of Linguistics, Vol. 4*. Berlin: De Gruyter Mouton.
- Plank, Frans. 1998. The co-variation of phonology with morphology and syntax: A hopeful history. *Linguistic Typology* 2. 195–230.
- Polinsky, Maria. 2013. Applicative Constructions. In Matthew S. Dryer & Martin Haspelmath, (eds.), *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at <http://wals.info/chapter/109>.)
- Poser, William J. 1995. Binary Comparison and the History of Hokan Comparative Studies. *International Journal of American Linguistics* 61(1). 135–144.
- Ramat, Paolo. 1986. Is a Holistic Typology possible? *Folia Linguistica* 20 (1-2). 3–14.
- Reid, Lawrence A. 2005. The current status of Austric: A review and evaluation of the lexical and morphosyntactic evidence. In Laurent Sagart, Roger Blench & Alicia Sanchez-Mazas (eds.), *The peopling of East Asia: putting together archaeology, linguistics and genetics*, 134–162. London: Routledge Curzon.



- Sasse, Hans-Jürgen. 1984. The pragmatics of noun incorporation in eastern Cushitic languages. In Frans Plank (ed.), *Objects: toward a theory of grammatical relations*, 243-268. London: Academic Press.
- Schlegel, August Wilhelm von. 1818. *Observations sur la langue et la littérature provençales*. Paris: Libraire grecque-latine-allemande
- Schlegel, Karl Friedrich von. 1808. *Über die Sprache und Weisheit der Indier: Ein Beitrag zur Begründung der Alterthumskunde, nebst metrischen Übersetzungen indischer Gedichte*. Heidelberg: Mohr and Zimmer.
- Schmidt, Wilhelm. 1906. Die Mon-Khmer-Völker, ein Bindeglied zwischen Völkern Zentralasiens und Austronesiens ('The Mon-Khmer Peoples, a link between the Peoples of Central Asia and Austronesia'). *Archiv für Anthropologie* 5. 59-109.
- Schönig, Claus. 2003. Turko-Mongolic Relations. In Juha Janhunen (ed.), *The Mongolic Languages*, 403-419. London: Routledge.
- Skirgård, Hedvig, et al. 2023. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances* 9(16). 1-15. DOI: 10.1126/sciadv.adg6175
- Song, Jae Jung. 2013. Nonperiphrastic Causative Constructions. In Matthew S. Dryer & Martin Haspelmath (eds.), *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at <http://wals.info/chapter/111>.)
- Whorf, Benjamin Lee. 1941. The Relation of Habitual Thought and Behavior to Language. In Leslie Spier, A. Irving Hallowell & Stanley S. Newman (eds.), *Language, Culture, and Personality: Essays in Memory of Edward Sapir*, 75-93. Menasha, Wisconsin: Sapir Memorial Publication Fund. Reprinted in John B. Carroll (ed.) 1956. *Language, Thought and Reality. Selected Writings of Benjamins Lee Whorf*. Cambridge, Mass: The MIT Press.

**CONTACT**

s.kalyan@uq.edu.au

mhdonohue@gmail.com.

# **Appendices of *The Dimensions of Morphosyntactic Variation: Whorf, Greenberg and Nichols were right***

SIVA KALYAN<sup>1</sup>, MARK DONOHUE<sup>2</sup>

<sup>1</sup>THE UNIVERSITY OF QUEENSLAND & THE AUSTRALIAN NATIONAL UNIVERSITY,

<sup>2</sup>THE LIVING TONGUES INSTITUTE FOR ENDANGERED LANGUAGES

## **Appendices**

Appendix 1.: List of languages with dimension correlates

Appendix 2.: List of features

Appendix 3.: Extended list of feature associations with first four dimensions

Appendix 4.: Comparison of area-controlled, family-controlled and genus-controlled FAMDs

Appendix 5.: Explication of FAMD analyses of macro-areas

Appendix 6.: Macro-areas plotted on the first four dimensions of global variation

Appendix 7.: Larger version of the map combining the first three dimensions

Appendix 8.: Alternative visualisations of dimensions

Appendix 9.: Maps and plots of linguistic features with strong associations with the different dimensions.

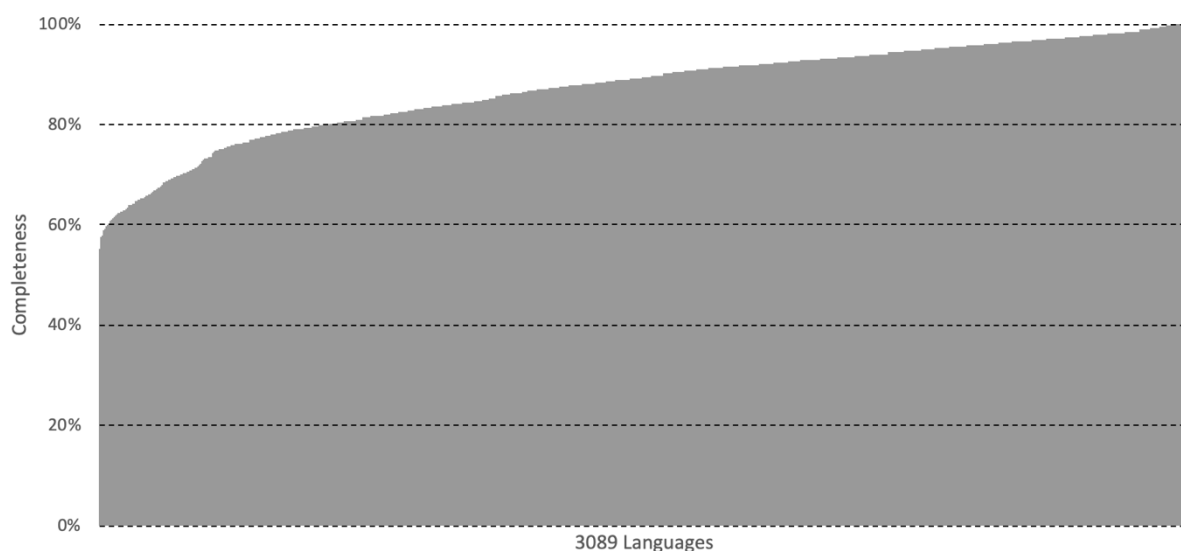
Appendix 10.: Source code

**Appendix 1. List of languages with dimension correlates****Appendix 2. List of features**

Appendix 1 and Appendix 2 can be found in the files WorldFAMD\_Appendix-12.xlsx and WorldFAMD\_Appendix-2.xlsx, available at <https://osf.io/u9qbe/>.

**Appendix 3. Extended list of feature associations with first four dimensions**

Appendix 3 can be found in the file WorldFAMD\_Appendix-3.xlsx, available at <https://osf.io/u9qbe/>.



**Figure A3:** The distribution of data in the database: coding completeness of 3089 languages/ varieties across maximally 351 features.

#### **Appendix 4. Comparison of area-controlled, family-controlled and genus-controlled FAMDs**

In the body of this paper we examined the data controlling for area, balancing the weighting for both macro-area and modified AUTOTYP area (Section 2). This method is easy to implement, avoiding decisions about relatedness of controversial languages. We tested whether a family-controlled or genus-controlled analysis produced different results, and found that essentially the same features emerge as the drivers of variation, though sometimes in different orders. We also examined a FAMD analysis without controls for area or genealogy, similarly finding very little difference. The number of dimensions that are optimal varies across the different analyses, as shown in Table A4.1, which also described the variance accounted for (in each different analysis) by each dimension.<sup>1</sup>

<b>Dimension</b>	<b>Area</b>	<b>Family</b>	<b>Genus</b>	<b>∅ controls</b>
1	7.7%	7.2%	7.8%	9.4%
2	6.4%	6.5%	6.8%	7.0%
3	4.2%	3.6%	3.9%	4.4%
4	3.9%	2.6%	2.8%	3.9%
5		2.2%	2.1%	
6		1.9%	1.7%	
7			1.6%	

**Table A4.1:** Dimensions and variance in the different analyses.

The following tables show the correlates of the different dimensions discussed in the body of the paper, and indicate in which dimension they occur with different controls. For instance, VO order is a correlate of Dimension 1 in the area-controlled analysis. It correlates to Dimension 2 in a family-controlled analysis, to both Dimension 1 and Dimension 2 in the genus-controlled analysis, and to Dimension 1 when no weighting is applied. We can see that there is little discrepancy between the different analyses in terms of which features correlate together, though we note that a family-controlled analysis places all of these features as part of Dimension 2, and that in the genus-controlled and

---

<sup>1</sup> We do not discuss these alternative analyses in great detail here, as they are the topic of a forthcoming paper (in preparation).

unmoderated analyses most of the features listed here occur in more than one dimension (just as, for instance, Gen N order is a correlate of both Dimension 1 and Dimension 4).

	Prepositions	VO	Subordinator-Clause	Prefixal nominative	V Obl	Gen N	Cl Sub	Postpositions	Obl V	Total Case	N Case	SOV
Area	1	1	1	1	1	1,4	1	1	1	1	1	1
Family	2	2	2	3	2	2	2	2	2	2	2	2
Genus	1	1,2	1,2	2,3	1,2	1	1,2	1,2	1	1,3	1	1,2
∅	1,3	1	1,3	2	1	1	1,3	1	1	1	1	1

**Table A4.2:** Correlates of Dimension 1.

The features that correlate with Dimension 2 in the area-controlled analysis tend to be part of the first dimension for the other analyses, but we also find some features correlating in the lower dimensions. Again, all of the features that showed associations with dimensions in the area-controlled analysis appear in the alternative analyses.

	Balanced	SVO	Total tenses	Possessive affixes	Causatives	Applicatives	Incorporatio	Modality affixes	Verbal synthesis	Total agreement
Area	2	2	2	2	2	2	2	2	2	2
Family	1	2	2	1	1,5	1,4	1,4	2	1	1,3
Genus	1	1	1,7	2,3	2,3,5	2,6	2	1	1	2,3
∅	1	1,3	1	2,3	2	2	2,3	1	1	2,4

**Table A4.3:** Correlates of Dimension 2.

The features that correlate with Dimension 3 across the different controls similarly find parallels in the analyses with other controls, but for this set of features we find

some that are not prominent in most of the analyses. Relative Pronouns, for instance, feature in the unweighted analyses, but not in either of the phylogenetically controlled analyses. While clusivity contrasts are relevant for the unweighted analysis and the genus-weighted analysis, it does not feature in the family-weighted analysis.

	Gender	Obligatory plural	Gender in 3sg	Accusative verbs	Gender in 3pl	Relative pronouns	Suffixal nominative	Ergativity	VOS	Clusivity	Clause-initial negation
Area	3	3	3	3	3	3,4	3	3	3	3	3,4
Family	4	4	4	4			(4)	6	1		1
Genus	3,4	4	4	4			4	5	2,3	2	2,3
∅	3	3	3	3	3	3	1	1	2,4	2,3	2

Table A4.4: Correlates of Dimension 3.

The features associated with the fourth dimension are replicated in the other analyses, though, as with the features associated with the third dimension, we see more features that are not prominent in one or more of the alternative analyses, and more variability in terms of which dimension they are found in.

	VSO	Clause-initial negation	Numeral Noun	Initial wh-questions	Adjective Noun	Genitive Noun	Relative Pronoun	Clause-final negation	Inalienability	Prefixal accusative	Noun Numeral	SV
Area	4	3,4	4	4	4	2,4	3,4	4	4	4	4	4
Family	1	1	3	1	(3)	2		1		3	3	
Genus	2,3	2,3	3,7	2	3	(2)	4	3			3,7	2,3,7
∅	2,4	2	4	2,4	1,4	1	3	4	2	2,4	4	2

Table A4.5: Correlates of Dimension 4.

Very few novel features emerged as significant in the higher dimensions of the alternative analyses. Table A4.6 lists the features that are found in at least two analyses; Table A4.7 offers a partial list of features occurring in just one analysis. We see the order of demonstratives and nouns features in the genus-weighted and un-weighted analyses, as does the position of Tense-Aspect-Mood affixes on the verb. Other verbal morphology accounts for the rest of the ‘new’ features, including the appearance of Philippine-type voice systems, and the possibility for double applicative or double causative constructions, but only in the genus-weighted analysis. In that analysis tonal morphology appears, including tonal marking of TAM categories, or marking agreement on the verb.

	Negative affixes on verb	Interrogative verb forms	Noun Demonstrative	TAM suffixes	Philippine- style voice
Area					
Family	2	6			
Genus		1	1	1	3
∅	1		1	1	4

**Table A4.6:** Additional features with strong associations emerging from the alternative analyses.

Dimension	Family	Genus	∅ controls
1			Dem Noun
2		Double applicatives TAM prefixes	
3	Number-sensitive suppletive verb forms Passives		Tense-sensitive suppletive verb forms
4		Tonal morphology	
5		Double causatives	
6			
7		Tonal agreement on verbs	

**Table A4.7:** Additional features with strong associations in one of the alternative analyses.

## Appendix 5. Explication of FAMD analyses of macro-areas

This appendix reports in more detail the results of the analyses that were reported in summary in 3.3. In each case only a selection of the features that are most strongly associated with a particular dimension are shown.

### A5.1. Africa

Features characterising the extremities of Dimensions 1 – 8 in the African FAMD.

Direction	Feature	$r^2$	
High	SOV order	0.68	
	Oblique precedes verb	0.62	
	Postpositions	0.55	
	Number of cases	0.51	
	Genitive precedes noun	0.49	
	Plural suffixes on nouns	0.48	
	Subject agreement suffixes	0.46	
	Prefixing	0.44	
	Object agreement prefixes	0.49	
	Prepositions	0.52	
	Oblique follows verb	0.57	
	Low	SVO order	0.63

**Table 5.1.1:** African Dimension 1.

Direction	Feature	$r^2$
High	Morphological causatives	0.50
	Total agreement positions	0.43
	Accusative verb alignment	0.41
	Negative affix on verbs	0.41
	Negator follows subject	0.32
Low	Negative particle	0.34

**Table 5.1.2:** African Dimension 2.



Direction	Feature	$r^2$
High	Predicate-subject	0.67
	VSO order	0.48
	Clause-initial negation	0.40
	Object verb agreement suffix	0.19
	Clusivity contrasts	0.18
Low	Gender	0.14
	Gender in 3SG pronoun	0.14
	SV order	0.67

Table 5.1.3: African Dimension 3.

Direction	Feature	$r^2$
High	Head marking	0.16
	Relative pronouns	0.16
	Accusative verb alignment	0.13
Low	Double causatives	0.12
	VOS order	0.14
	Incorporation	0.15
	Superapplicatives	0.22

Table 5.1.4: African Dimension 4.

Direction	Feature	$r^2$
High	Infixes	0.23
	Incorporation	0.23
	Demonstrative precedes noun	0.18
	Relative pronouns	0.12
	Superapplicatives	0.11
	Adjective follows noun	0.09
Low	Relative clause follows noun	0.10
	Demonstrative follows noun	0.11

Table 5.1.5: African Dimension 5.

<b>Direction</b>	<b>Feature</b>	<b><math>r^2</math></b>
High	Ergative verb alignment	0.24
	Ergative pronominal alignment	0.14
	Applicatives	0.13
	Clusivity contrasts	0.10
	VSO order	0.08
Low	Gender in first or second person pronouns	0.11

**Table 5.1.6:** African Dimension 6.

<b>Direction</b>	<b>Feature</b>	<b><math>r^2</math></b>
High	Third agreement position on verbs	0.11
	Double negation	0.11
	Numerals precede nouns	0.11
	Adjectives follow nouns	0.09
	Possessive suffixes	0.12
Low	Incorporation	0.14

**Table 5.1.7:** African Dimension 7.

<b>Direction</b>	<b>Feature</b>	<b><math>r^2</math></b>
High	Ergativity	0.22
	VSO order	0.11
	Total tense distinctions	0.06
	Possessive suffixes	0.07
	Third agreement position on verbs	0.11
Low	Possessive classes	0.12

**Table 5.1.8:** African Dimension 8.

Dimension	Variance accounted for?
1	10.8%
2	7.9%
3	4.7%
4	3.0%
5	2.1%
6	1.8%
7	1.6%
8	1.6%

**Table A5.1.9:** Variance in the data accounted for by the first eight dimensions of the Africa analysis.

### A5.2. Australia

Features characterising the extremities of Dimensions 1 – 4 in the Australian FAMD.

Direction	Feature	$r^2$
High	Prefixal agreement for subject and object	0.71
	Fused arguments in verbal agreement	0.56
	Incorporation	0.55
	Gender	0.54
	Bound clusivity contrasts	0.51
Low	Accusative pronominal alignment	0.36
	Dependent marking	0.45
	Suffixing	0.45
	Ergativity	0.50

**Table 5.2.1:** Australian Dimension 1.

Direction	Feature	$r^2$
High	VOS order	0.39
	PNG-affixes forming pronouns	0.29
	Object suffix agreement	0.25
	Demonstrative precedes noun	0.16
Low	Obliques precede verb	0.14
	SOV order	0.29

**Table 5.2.2:** Australian Dimension 2.

Direction	Feature	$r^2$
High	Applicative types	0.55
	Causative~Applicative syncretism	0.32
	Causatives	0.23
	Past tense	0.10
Low	Verb agreement for subject and object	0.08

**Table 5.2.3:** Australian Dimension 3.

Direction	Feature	$r^2$
High	Demonstrative precedes noun	0.12
	Negator follows verb	0.11
	Incorporation	0.11
	Ergative alignment	0.12
	Final subordinating suffix	0.14
Low	Agreement	0.14

**Table 5.2.4:** Australian Dimension 4.

Dimension	Variance accounted for?
1	13.2%
2	4.8%
3	3.7%
4	3.5%

**Table A5.2.5:** Variance in the data accounted for by the first four dimensions of the Australia analysis.

**A5.3. Eurasia**

Features characterising the extremities of Dimensions 1 – 4 in the Eurasian FAMD.

<b>Direction</b>	<b>Feature</b>	<b><math>r^2</math></b>	
High	Genitive follows noun	0.64	
	Prepositions	0.60	
	VO order	0.58	
	Oblique follows verb	0.55	
	Gender in 3SG pronouns	0.53	
	Relative clauses follow nouns	0.52	
	Relative clauses precede nouns	0.55	
	Genitive precedes noun	0.56	
	Final subordination	0.58	
	SOV order	0.58	
	Postpositions	0.59	
	Low	Oblique precedes verb	0.60

**Table 5.3.1:** Eurasian Dimension 1.

<b>Direction</b>	<b>Feature</b>	<b><math>r^2</math></b>
High	Verbal synthesis	0.44
	Tense marking	0.44
	Nominative verb agreement	0.39
	Morphological causatives	0.39
	Number of cases	0.37
	Numeral classifiers	0.36
	Symmetrical subordination	0.55
	Low	Isolating

**Table 5.3.2:** Eurasian Dimension 2.

Direction	Feature	$r^2$
High	Number of applicatives	0.39
	Double applicatives	0.35
	Subject agreement prefixes	0.32
Low	Dative subjects	0.22
	Relative pronouns	0.23
	Accusative pronominal alignment	0.39

**Table 5.3.3:** Eurasian Dimension 3.

Direction	Feature	$r^2$
High	Number of applicatives	0.33
	<i>Have</i> verb	0.28
	Incorporation types	0.24
	Perfect formed with <i>have</i>	0.21
	Double applicatives	0.18
Low	Asymmetrical negative clauses	0.15
	Null copular verb in nominal predicate clauses	0.15
	Possession formed with a locative possessor	0.17

**Table 5.3.4:** Eurasian Dimension 4.

Dimension	Variance accounted for?
1	12.7%
2	10.2%
3	7.0%
4	4.8%

**Table A5.3.5:** Variance in the data accounted for by the first four dimensions of the Eurasian analysis.

**A5.4. Pacific**

Features characterising the extremities of Dimensions 1 – 3 in the Pacific FAMD.

Direction	Feature	$r^2$
High	SOV order	0.77
	Suffixal case	0.70
	Obliques precede verbs	0.69
	Postpositions	0.65
Low	Clusivity contrasts	0.53
	Initial subordination	0.64
	VO order	0.72
	Prepositions	0.73

**Table 5.4.1:** Pacific Dimension 1.

Direction	Feature	$r^2$
High	Dependent marking	0.39
	VS order	0.36
	Initial negation	0.33
Low	Negator follows subject	0.25
	Prefixal nominative agreement	0.33
	SV order	0.37

**Table 5.4.2:** Pacific Dimension 2.

Direction	Feature	$r^2$
High	Total agreement positions	0.54
	Possessive affixes on nouns	0.29
	Applicatives	0.28
	Accusative verb alignment	0.27
Low	Polar question particles	0.06
	Isolating	0.21

**Table 5.4.3:** Pacific Dimension 3.

Dimension	Variance accounted for?
1	14.8%
2	6.1%
3	4.5%

**Table A5.4.4:** Variance in the data accounted for by the first three dimensions of the Pacific analysis.

### A5.5. North America

Features characterising the extremities of Dimensions 1 – 3 in the North American FAMD.

Direction	Feature	$r^2$
High	Prepositions	0.62
	Initial negation	0.58
	VSO order	0.54
	Initial subordination	0.45
	Genitive follows noun	0.45
Low	Final subordination	0.47
	Genitive precedes noun	0.48
	Oblique precedes verb	0.52
	Postpositions	0.63
	SOV order	0.75

**Table 5.5.1:** North America Dimension 1.

Direction	Feature	$r^2$
High	Agreement for two arguments	0.40
	Applicatives (multiple associations with types of applicatives)	0.32
	Incorporation	0.17
Low	Morphological causatives	0.11
	Indirect trivalent verbs	0.12

**Table 5.5.2:** North America Dimension 2.



Direction	Feature	$r^2$
High	Suffixing	0.44
	Dependent marking	0.36
	Number of cases	0.32
	Nominative suffix agreement	0.22
Low	Possessive prefixes	0.16
	Object agreement prefixes	0.34
	Subject agreement prefixes	0.38
	Prefixing	0.44

Table 5.5.3: North America Dimension 3.

Dimension	Variance accounted for?
1	10.4%
2	6.0%
3	5.8%

Table A5.5.4: Variance in the data accounted for by the first three dimensions of the North American analysis.

### A5.6. South America

Features characterising the extremities of Dimensions 1 – 4 in the South American FAMD.

Direction	Feature	$r^2$
High	VO order	0.50
	Possessive prefix	0.40
	Initial subordination	0.39
	Nominative agreement prefix	0.35
	Verb initial order	0.29
Low	Oblique precedes verb	0.34
	Final subordination	0.41
	Postpositions	0.45
	Suffixal core case marking	0.46
	SOV order	0.57

Table 5.6.1: South American Dimension 1.

Direction	Feature	$r^2$
High	Applicatives	0.38
	Morphological causatives	0.28
	Object suffix on verbs	0.26
	VSO order	0.24
Low	Clause-final negation	0.18
	SV order	0.22
	Balanced 'subordinate' clauses	0.28

**Table 5.6.2:** South American Dimension 2.

Direction	Feature	$r^2$
High	Subject and object agreement	0.32
	Clusivity contrasts in bound morphology	0.23
	Polar question particles	0.21
Low	Plural prefixes on nouns	0.15
	Initial Wh question words	0.15
	VS order	0.16

**Table 5.6.3:** South American Dimension 3.

Direction	Feature	$r^2$
High	Plural suffixes on nouns	0.24
	Accusative alignment on verbs	0.22
	Negator precedes verb	0.21
Low	Negative suffix on verbs	0.19
	Symmetrical subordination	0.21
	Active alignment on verbs	0.22

**Table 5.6.4:** South American Dimension 4.

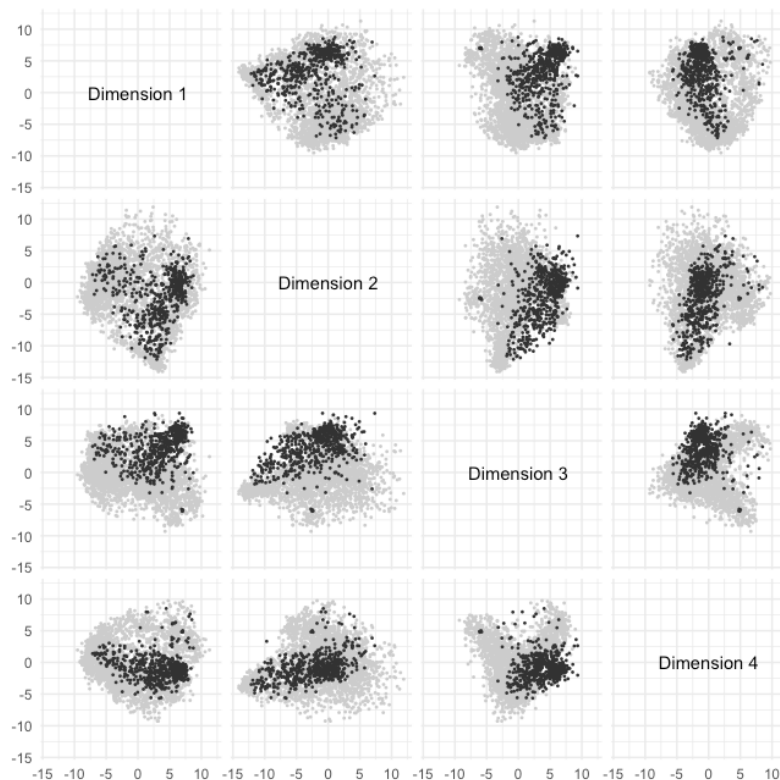
Dimension	Variance accounted for?
1	9.4%
2	6.4%
3	4.7%
4	4.5%

**Table A5.6.5:** Variance in the data accounted for by the first four dimensions of the South American analysis.

## Appendix 6. Macro-areas plotted on the first four dimensions of global variation.

In 3.2.5 we saw the position of North America plotted on the dimension plots that were first introduced in Figure 3 (in 3.1). Here we present the other macro-areas, with brief commentary.

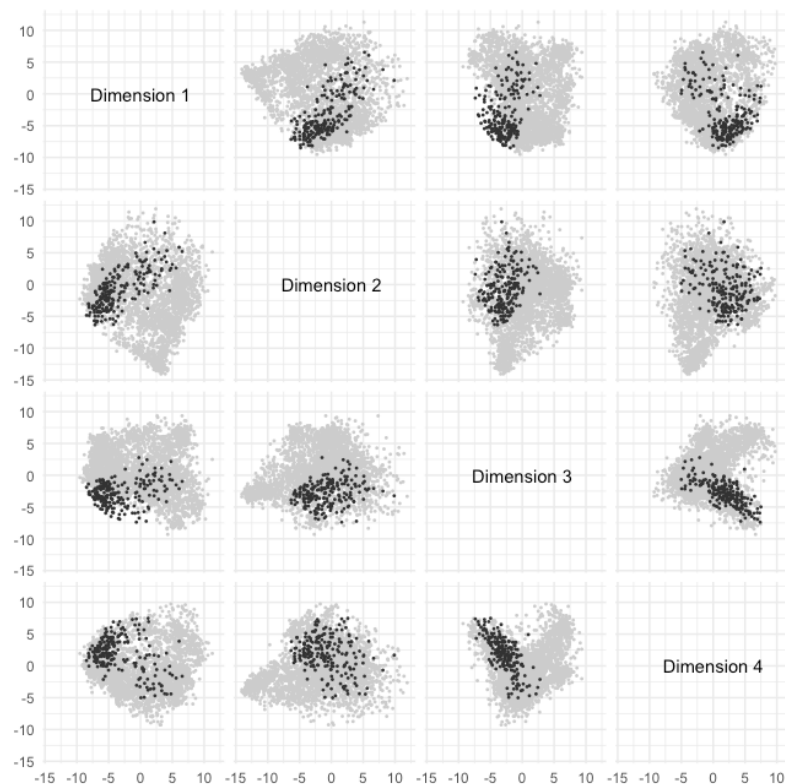
### *Appendix 6.1. Africa.*



**Figure A6.1:** The languages of Africa in morphosyntactic space.

The languages of Africa are generally low on Dimension 3; while the Bantu languages form a tight cluster near the top of Dimension 1, the non-Bantu languages spread across most of the typological space defined by Dimensions 1 and 2, with predominantly Afro-Asiatic languages towards the bottom of Dimension 1. In terms of Dimension 4 the majority of African languages occupy a middle position, due to the verb-initial languages spoken north of the Sahara being grouped with Eurasia (see Section 2, and 3.2.1). We can see that, for the languages of Africa, there is considerable correlation between Dimensions 1 and 2 with Dimension 4.

### ***Appendix 6.2. Australia.***



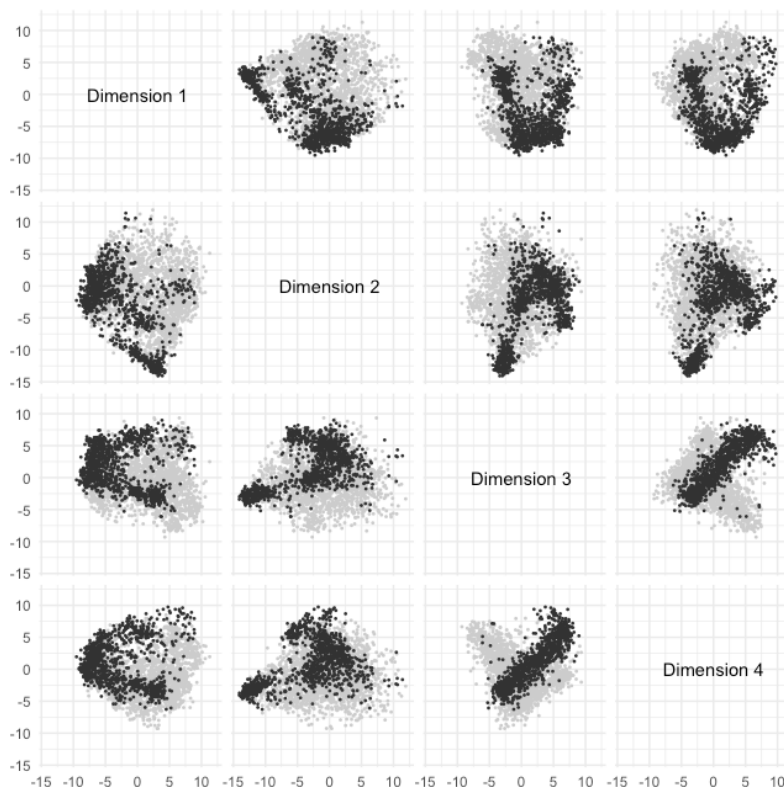
**Figure A6.2:** The languages of Australia in morphosyntactic space.

We have seen the typological position of the Pama-Nyungan languages that occupy most of Australia in 3.2.13. In this section the non-Pama-Nyungan languages of northern Australia are added. We can see that the non-Pama-Nyungan languages dramatically extend the typological range of the languages of Australia, occupying

space higher on Dimensions 1, 2, and lower on Dimension 4. In Australia Dimension 1 strongly correlates with Dimension 2, and Dimension 3 strongly correlates with Dimension 4.

### ***Appendix 6.3. Eurasia.***

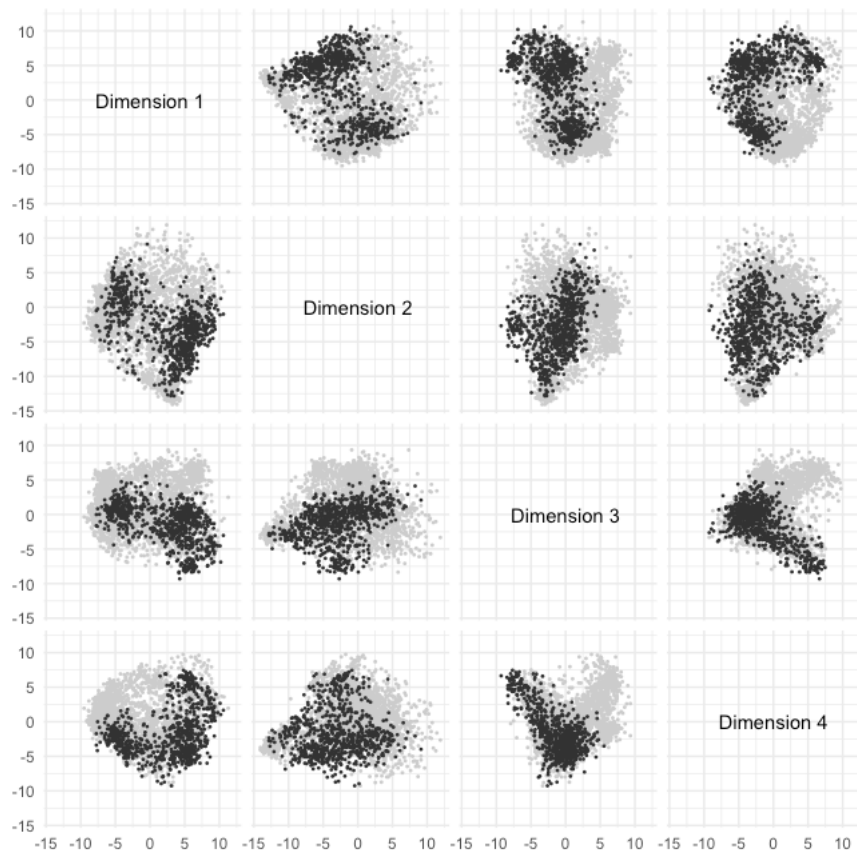
The languages of Eurasia present an oddly fractured profile in certain dimension plots. In any plot involving Dimension 1 or Dimension 2 we can see either an empty interior in a U-shaped pattern of distribution, or else a separate region (occupied by languages from mainland Southeast Asia) with no strong connection to the rest of the languages of Eurasia, typologically. We have seen, in Map 9, that the only concentrations of languages that approximate the isolation of the languages of mainland Southeast Asia are found in West Africa and central Indonesia. In the plot of Dimension 3 vs. Dimension 4 we see a strong correlation between the two dimensions, with the languages of Eurasia avoiding the space low on Dimension 3 and high on Dimension 4, a space occupied by languages of the Circum-Pacific regions.



**Figure A6.3:** The languages of Eurasia in morphosyntactic space.

### **Appendix 6.4. Pacific.**

The languages of the Pacific occupy a large region in typological space, but do not extend to the higher levels of Dimension 2 or Dimension 3. Other regions in which Pacific languages are not found can be detected in Figure A6.4; the space high on Dimension 4 and low on Dimension 1; the space high on Dimension 4 and high on Dimension 3; the space high on Dimension 2 and low on Dimension 3. In the plot of Dimensions 1 and 2 we see an approximate bifurcation, and in the plot of Dimension 3 and Dimension 4 a strong correlation is apparent.



**Figure A6.4:** The languages of the Pacific in morphosyntactic space.

### **Appendix 6.5. North America.**

This plot has already been described in 3.2.5, and is repeated here purely for the benefit of having all of the macro-areal plots in the same place. We note that Dimension 3 and Dimension 4 show a negative correlation, as with the Pacific.

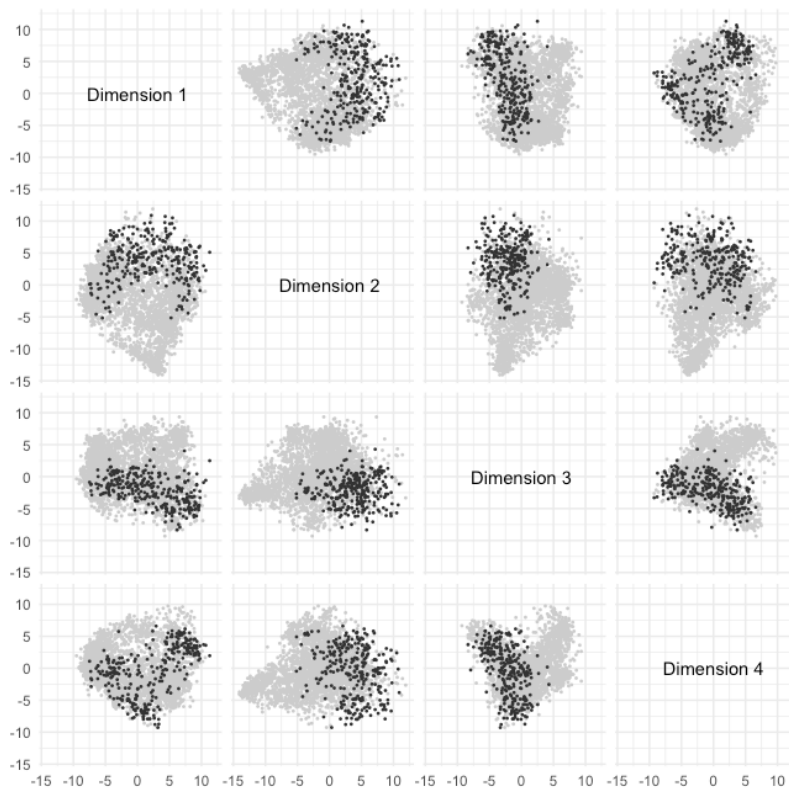


Figure A6.5: The languages of North America in morphosyntactic space.

**Appendix 6.6. South America.**

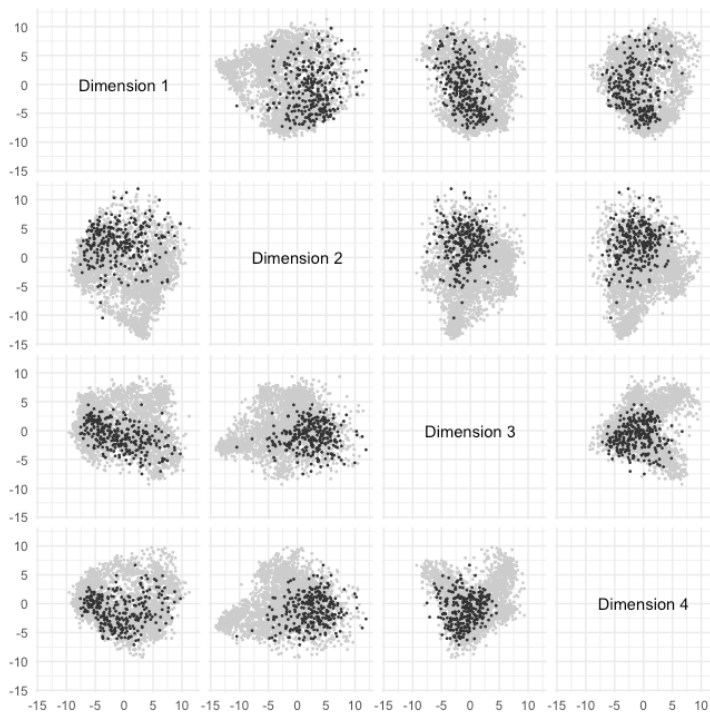


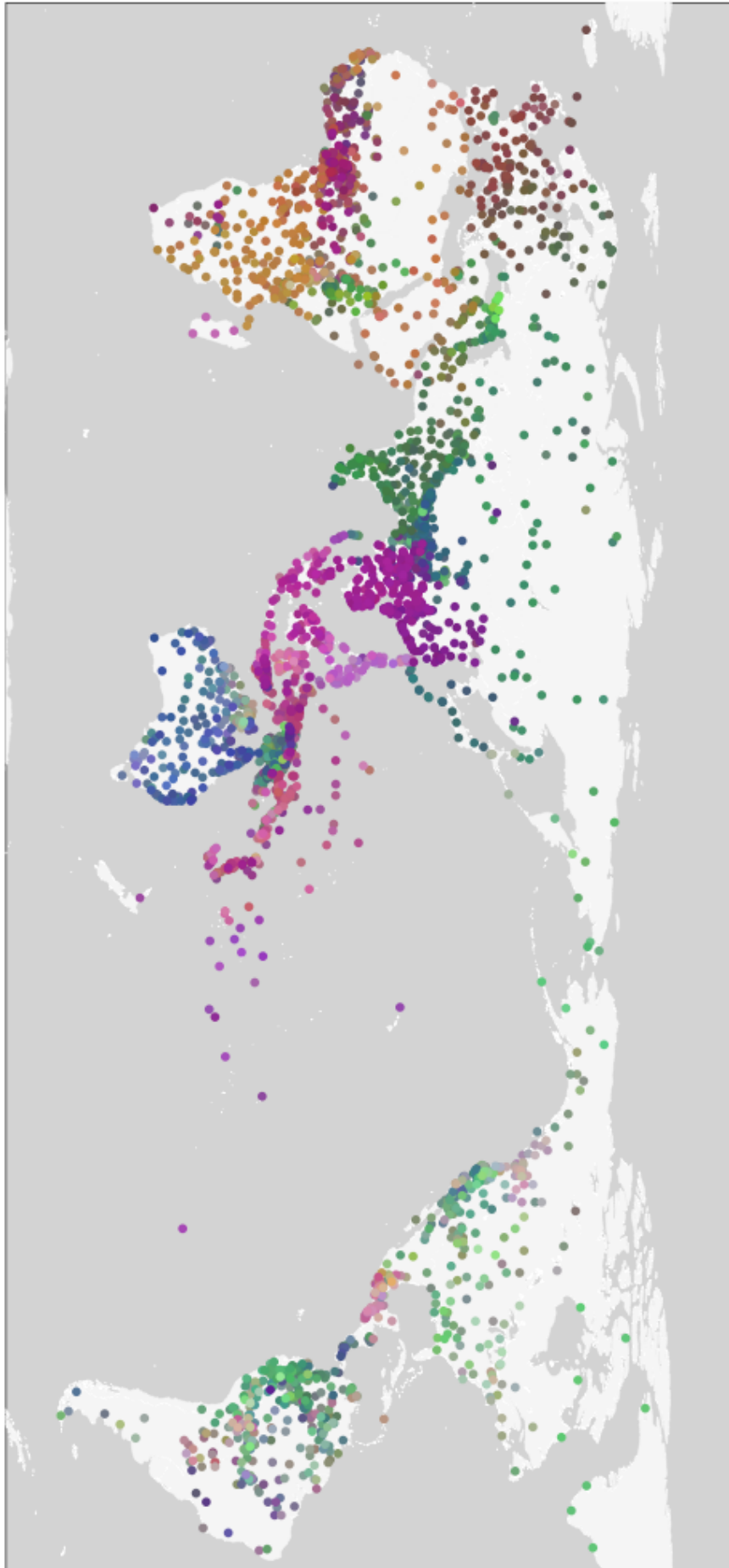
Figure A6.6: The languages of South America in morphosyntactic space.

The languages of South America can be seen as a slightly less peripheral version of the North American languages, slightly higher on Dimensions 3 and 4, and slightly lower on Dimension 1.

### **Appendix 7. Larger version of the map combining the first three dimensions**

An interactive, version of this map, best viewed with Google Chrome, can be found online at <https://skalyan91.github.io/d3-language-maps/globe.html?data=ms-points-2023-noPC-Autotyp>. Large downloadable versions of the map can be found at <https://osf.io/u9qbe/>.





Map A7: Larger version of Map 12.

## Appendix 8. Alternative visualisations of dimensions

The colourings in Figure 3 and Map 9 follow combinations of Red, Blue and Green assigned to the first, second and third dimensions, respectively, while in Maps 5 – 8 we saw maps coloured according to a single dimension. In this appendix we present alternative visualisations colouring according to different combinations of dimensions.

### A8.1. Colours reflecting Dimensions 1, 2 and 4

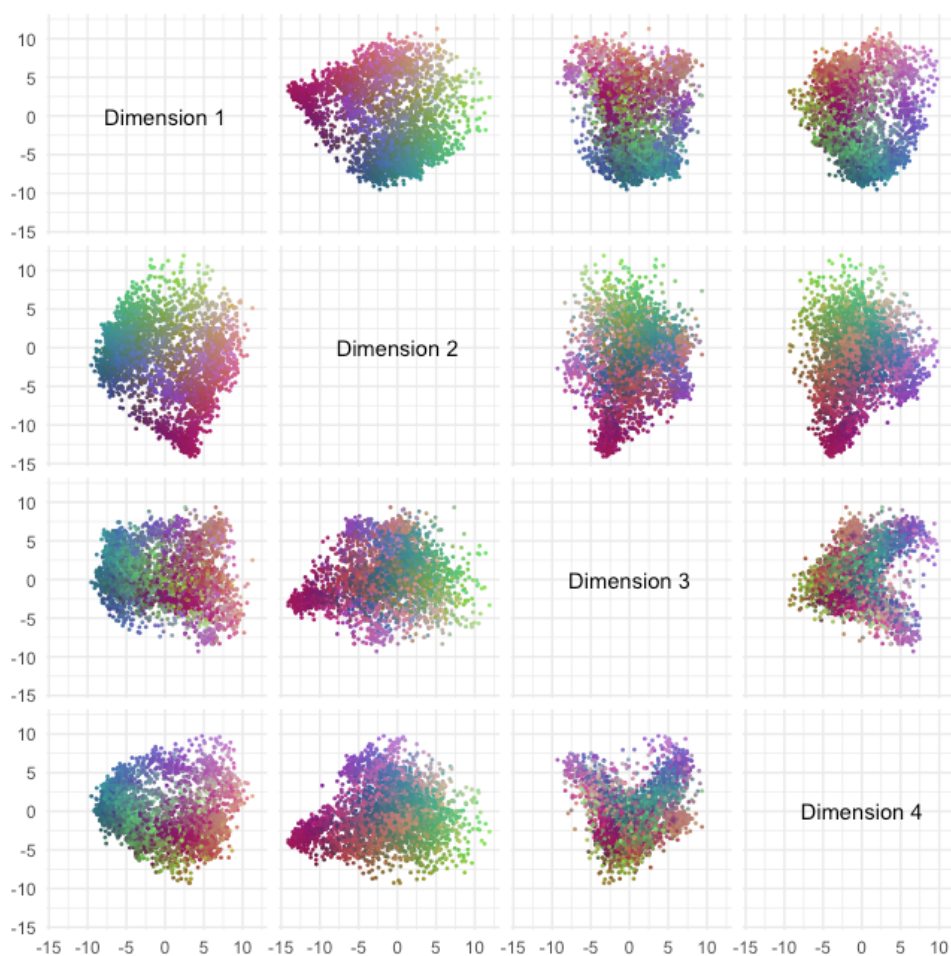
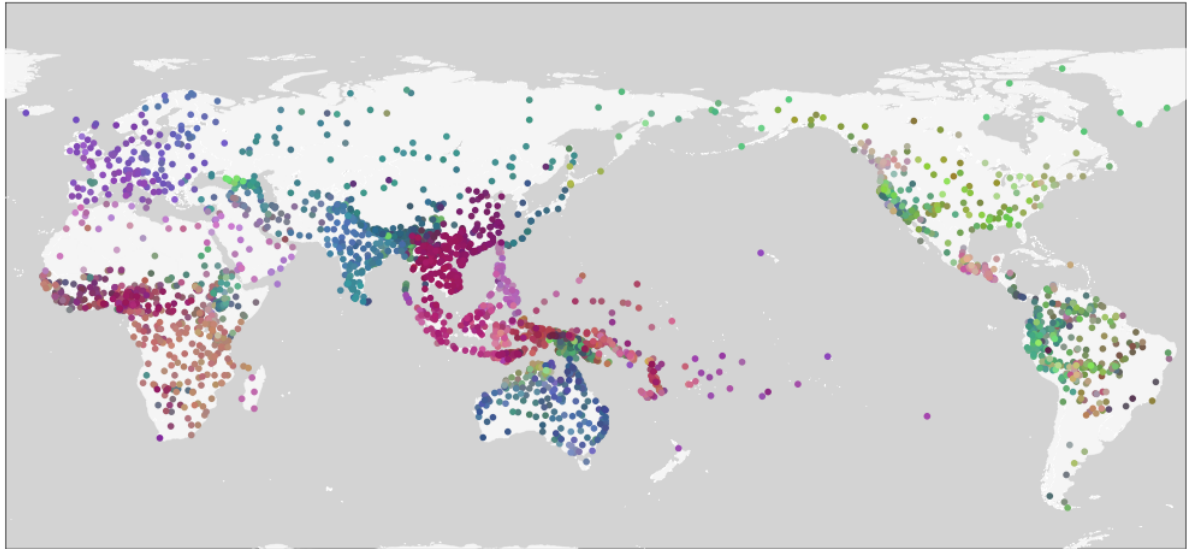
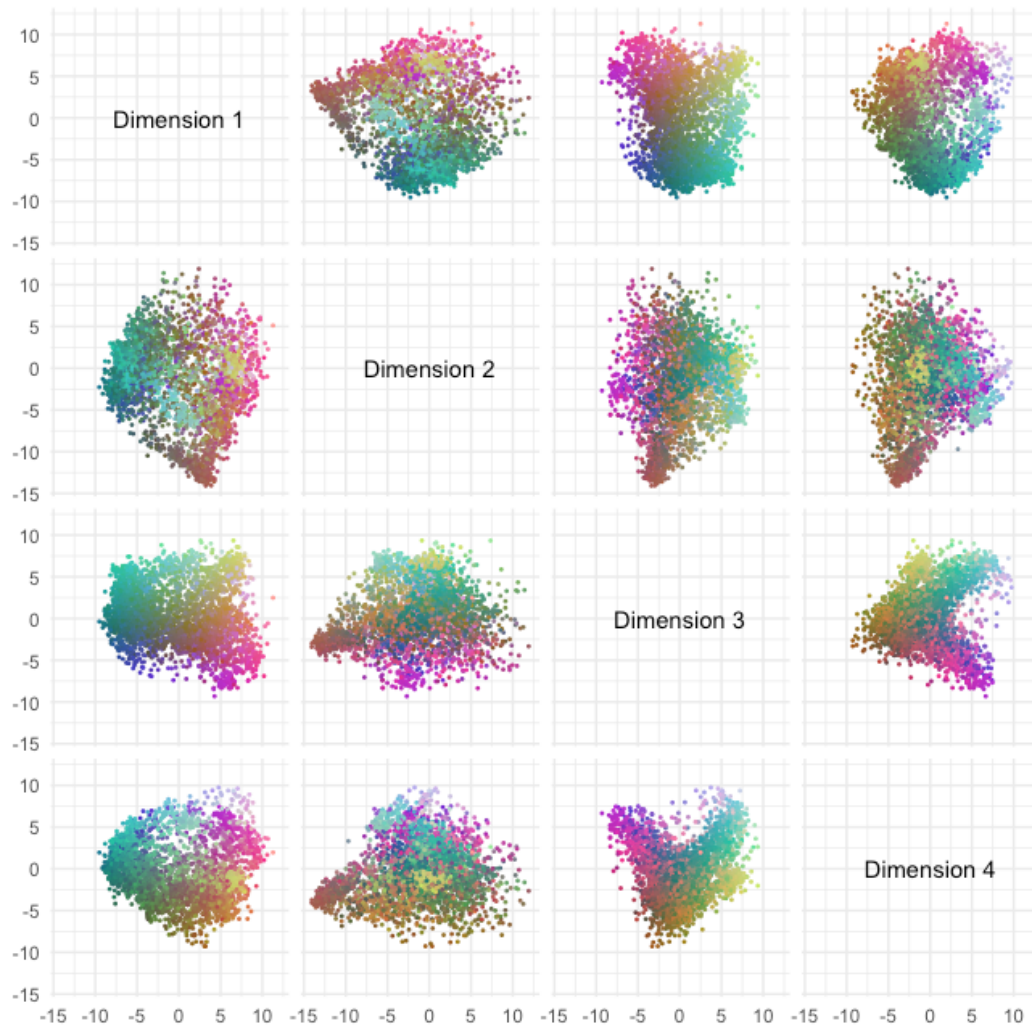


Figure A8.1: Dimension plots coloured for Dimensions 1, 2 and 4.

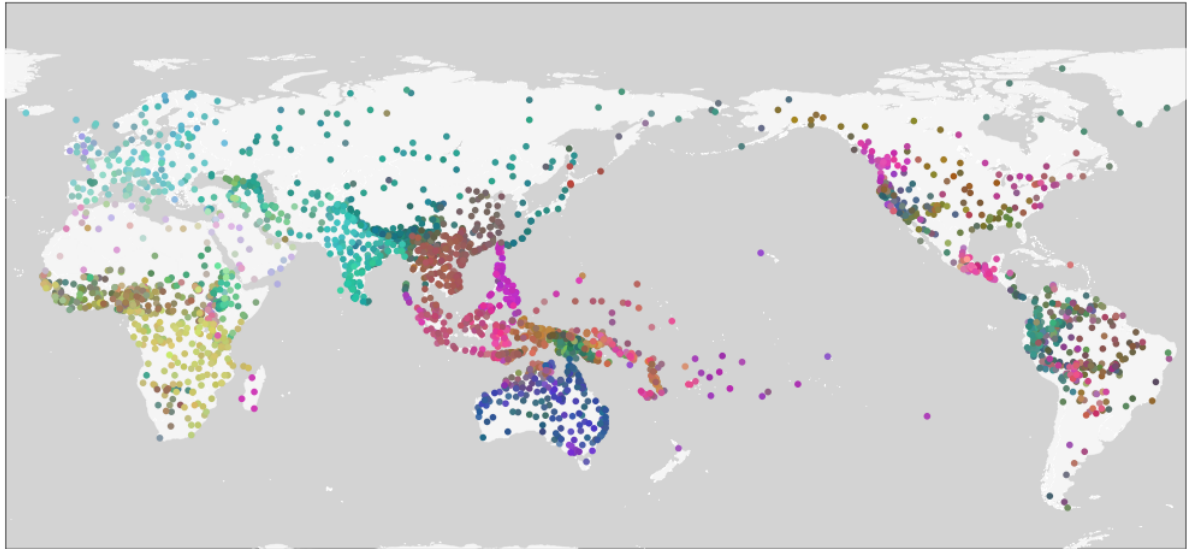


**Map A8.1:** Global map coloured for dimensions 1, 2 and 4.

**A8.2. Colours reflecting Dimensions 1, 3 and 4**

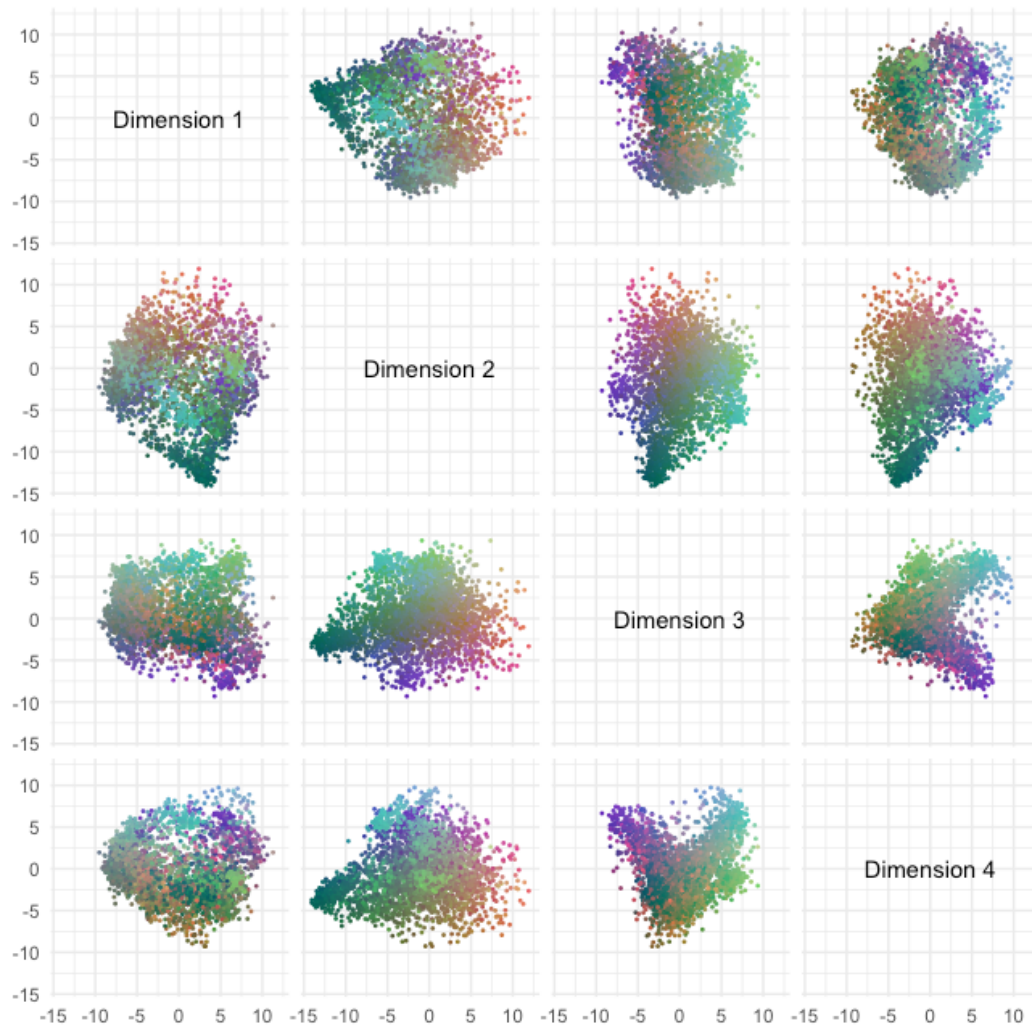


**Figure A8.2:** Dimension plots coloured for Dimensions 1, 3 and 4.

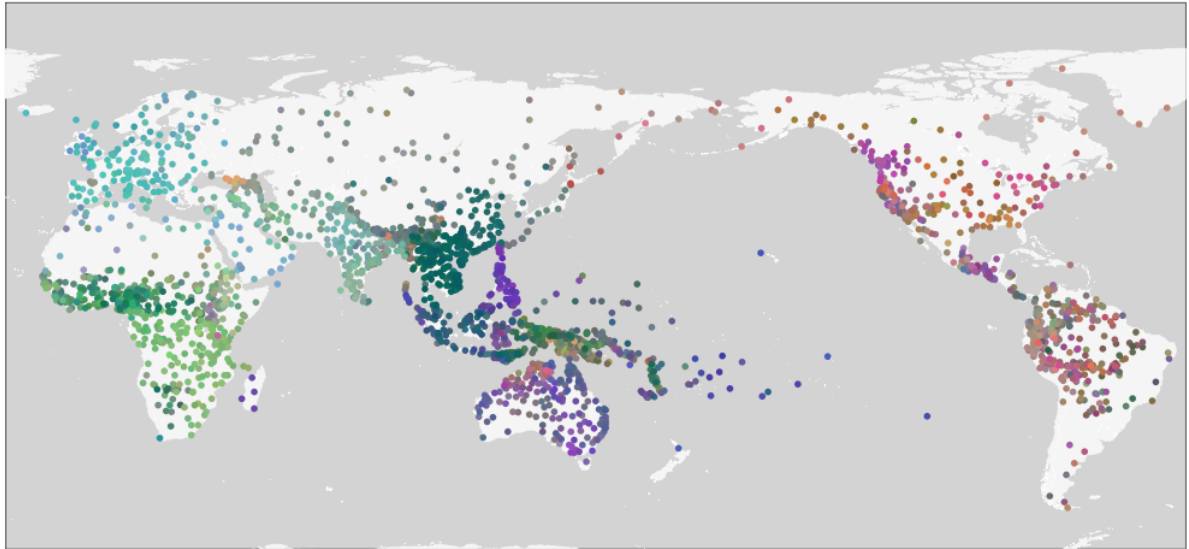


**Map A8.2:** Global map coloured for dimensions 1, 3 and 4.

**A8.3. Colours reflecting Dimensions 2, 3 and 4**

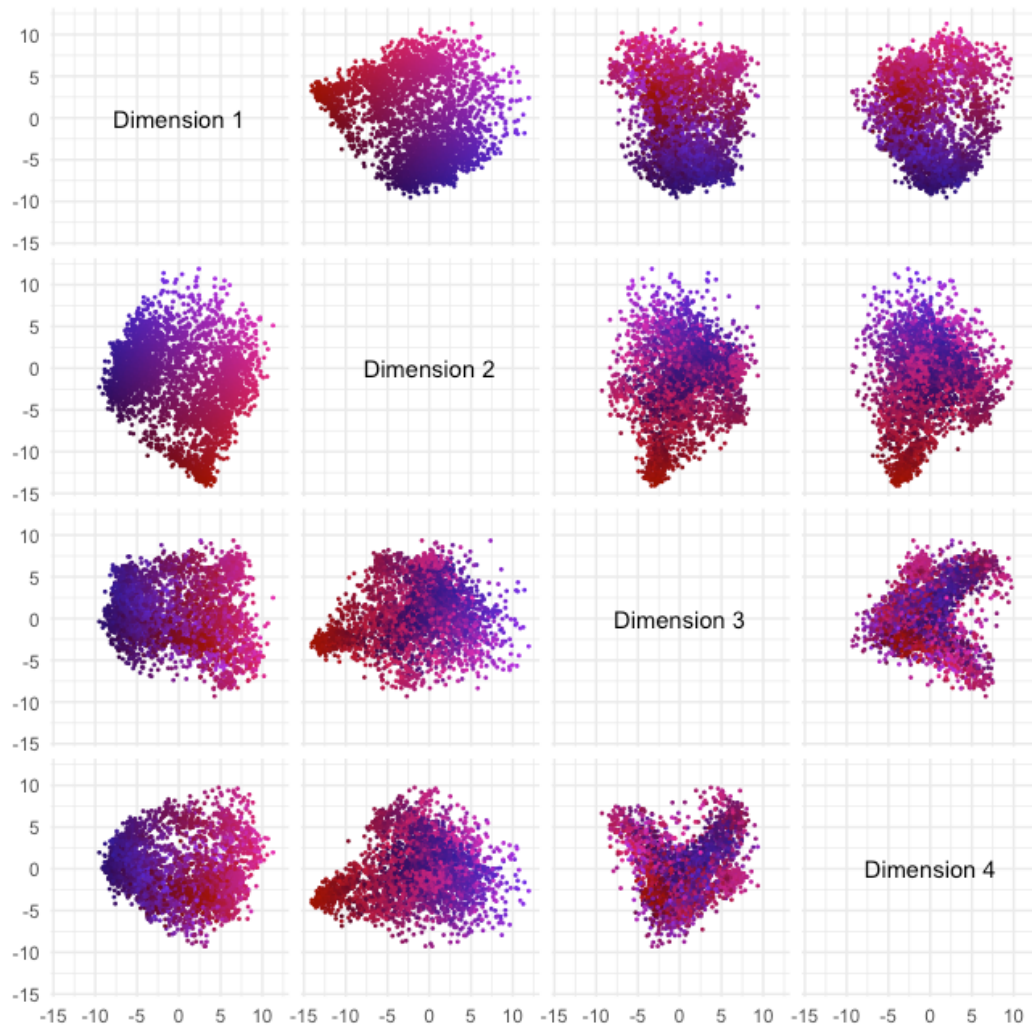


**Figure A8.3:** Dimension plots coloured for Dimensions 2, 3 and 4.



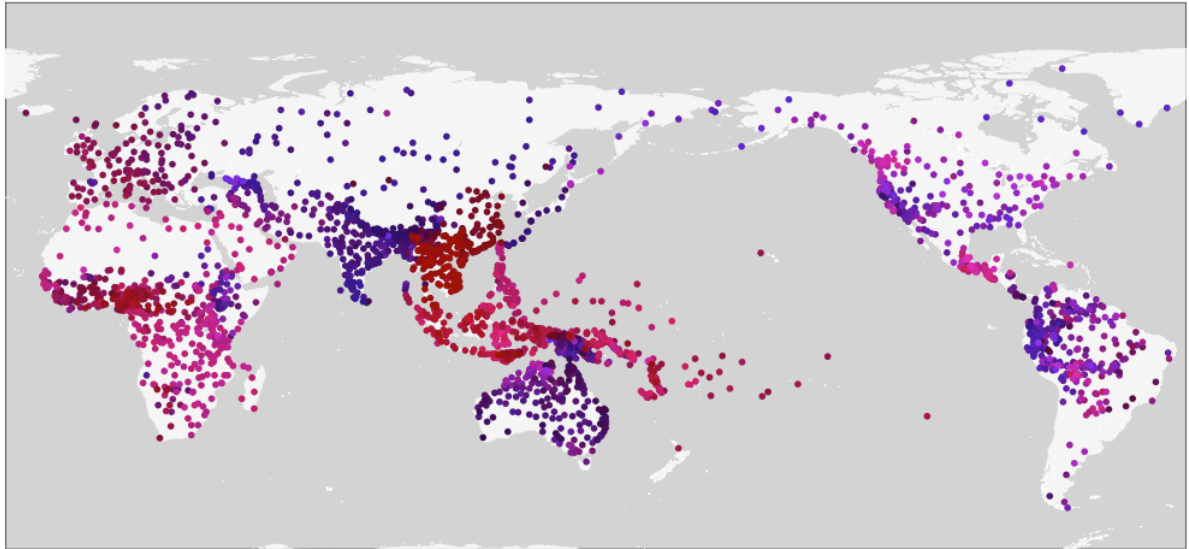
**Map A8.3:** Global map coloured for dimensions 2, 3 and 4.

**A8.4. Colours reflecting Dimensions 1 and 2**



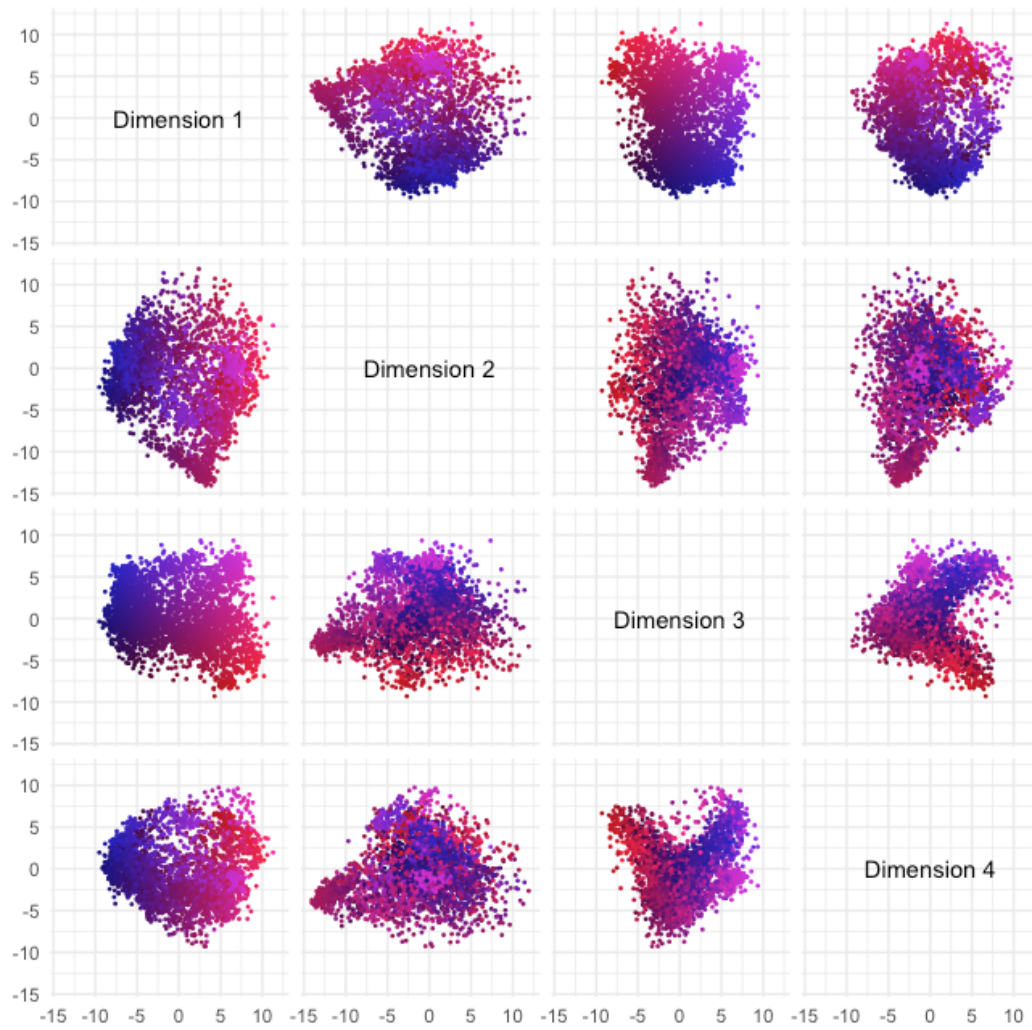
**Figure A8.4:** Dimension plots coloured for Dimensions 1 and 2



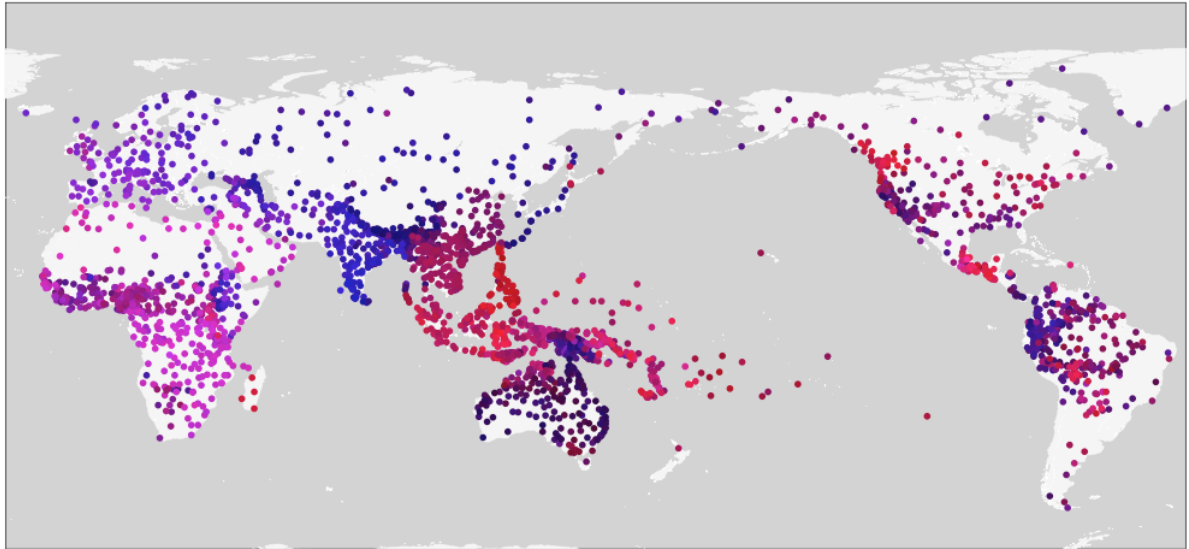


**Map A8.4:** Global map coloured for dimensions 1 and 2.

**A8.5. Colours reflecting Dimensions 1 and 3**

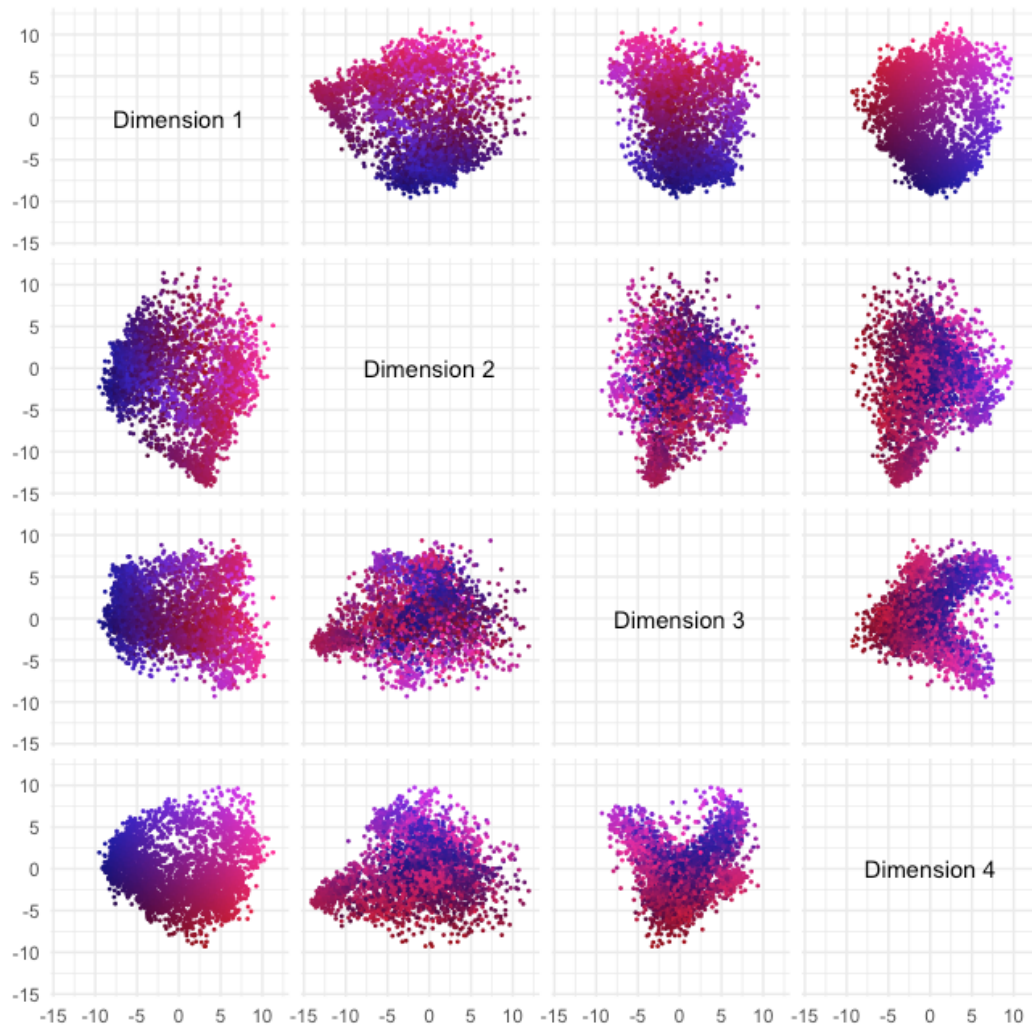


**Figure A8.1:** Dimension plots coloured for Dimensions 1 and 3.

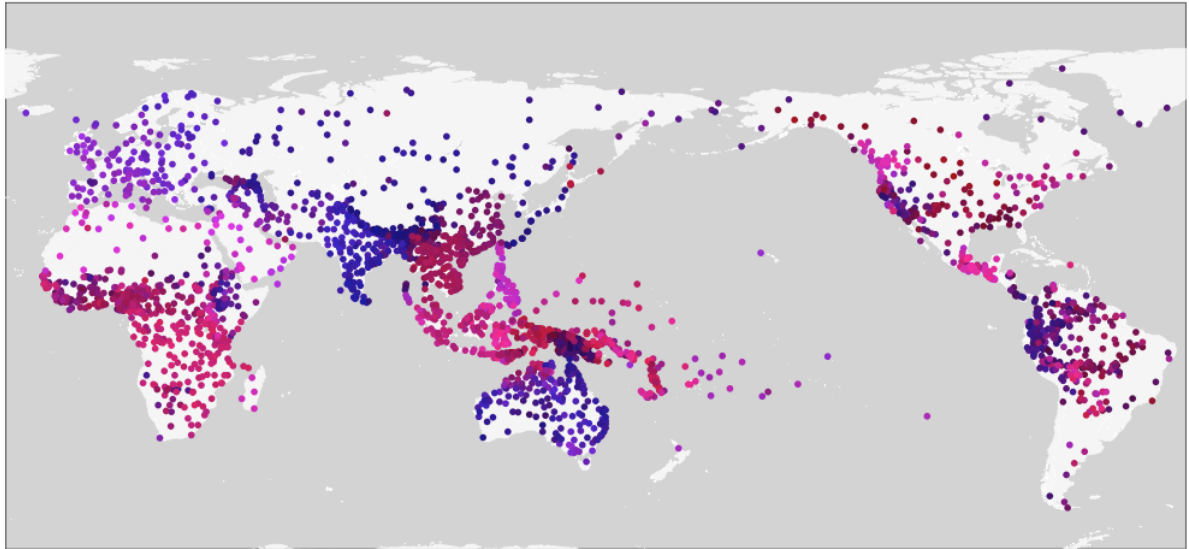


**Map A8.1:** Global map coloured for dimensions 1 and 3.

**A8.6. Colours reflecting Dimensions 1 and 4**

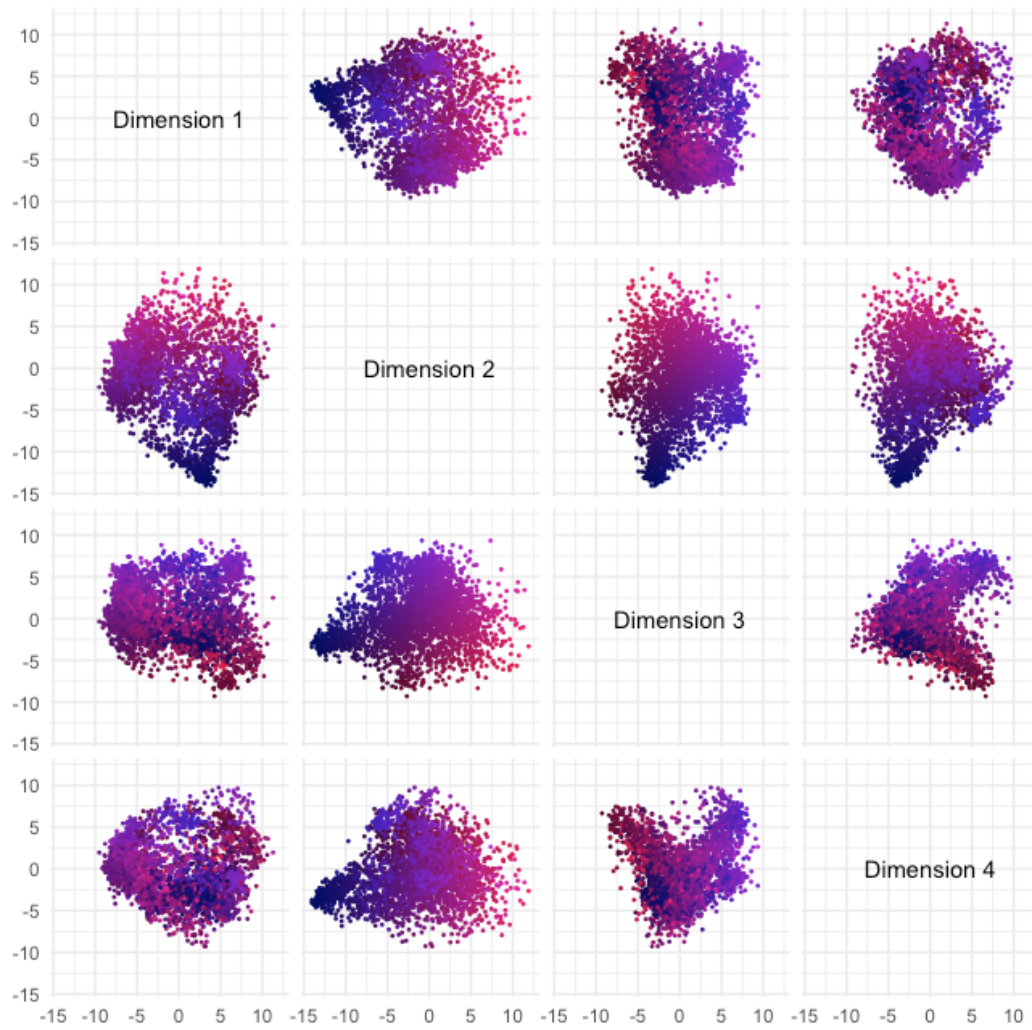


**Figure A8.1:** Dimension plots coloured for Dimensions 1 and 4.

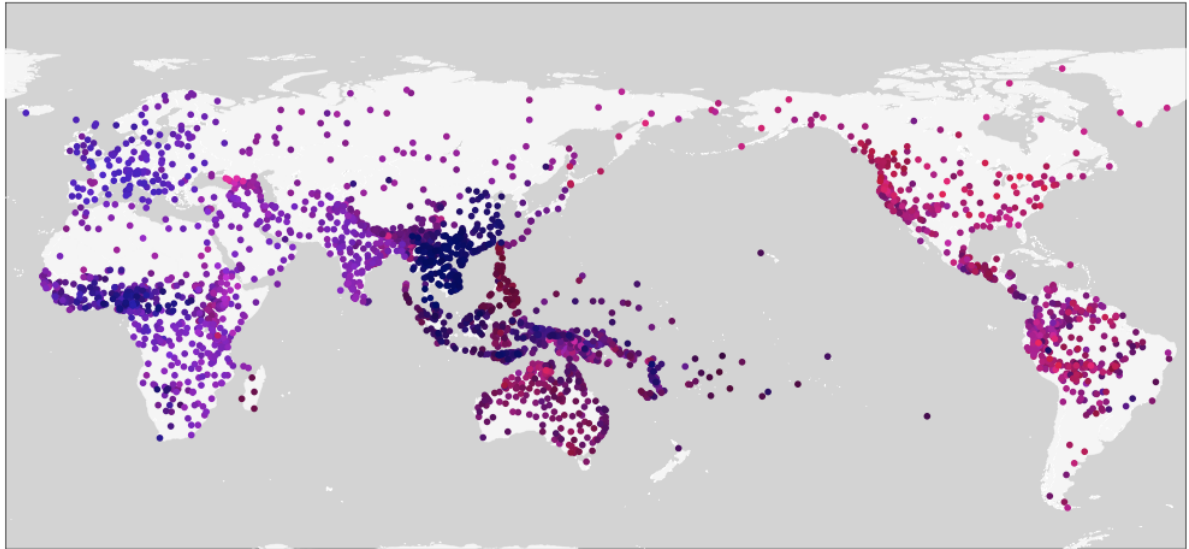


**Map A8.1:** Global map coloured for dimensions 1 and 4.

**A8.7. Colours reflecting Dimensions 2 and 3**

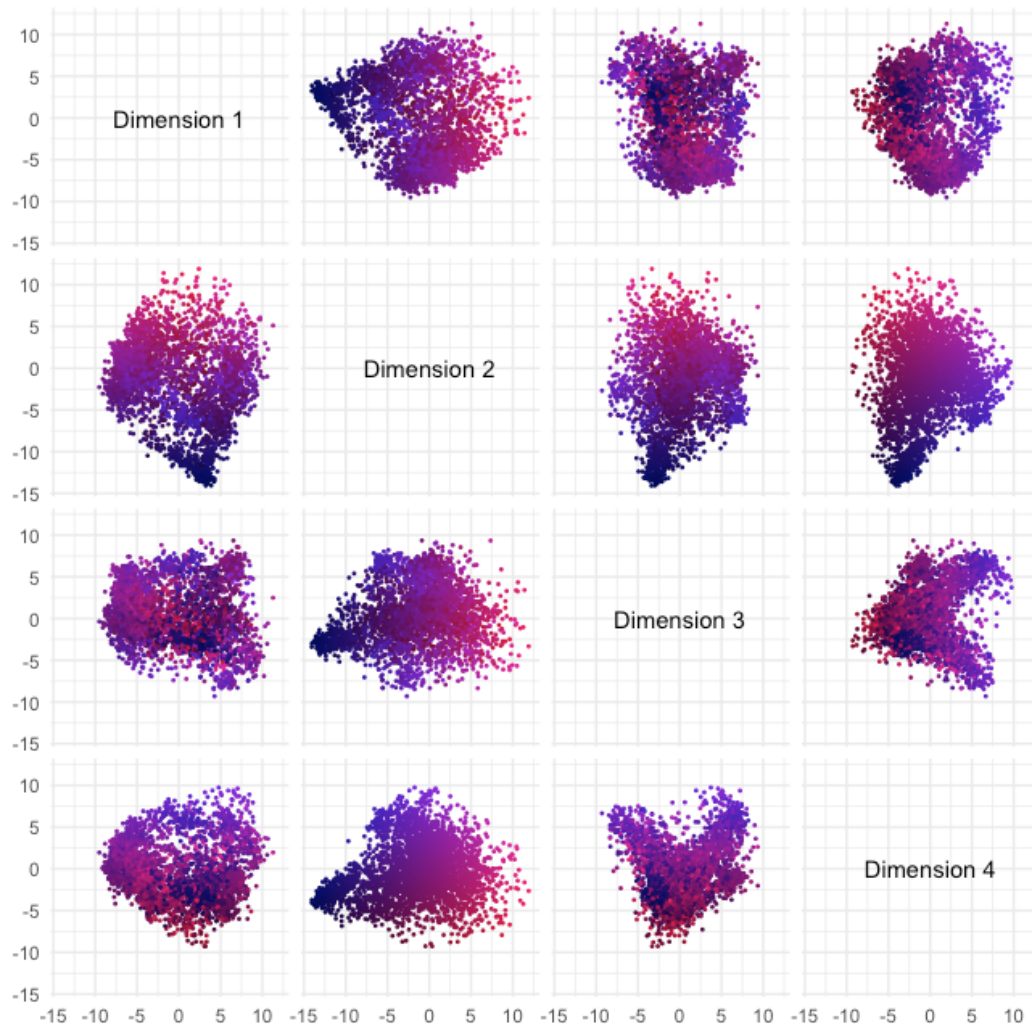


**Figure A8.1:** Dimension plots coloured for Dimensions 2 and 3.



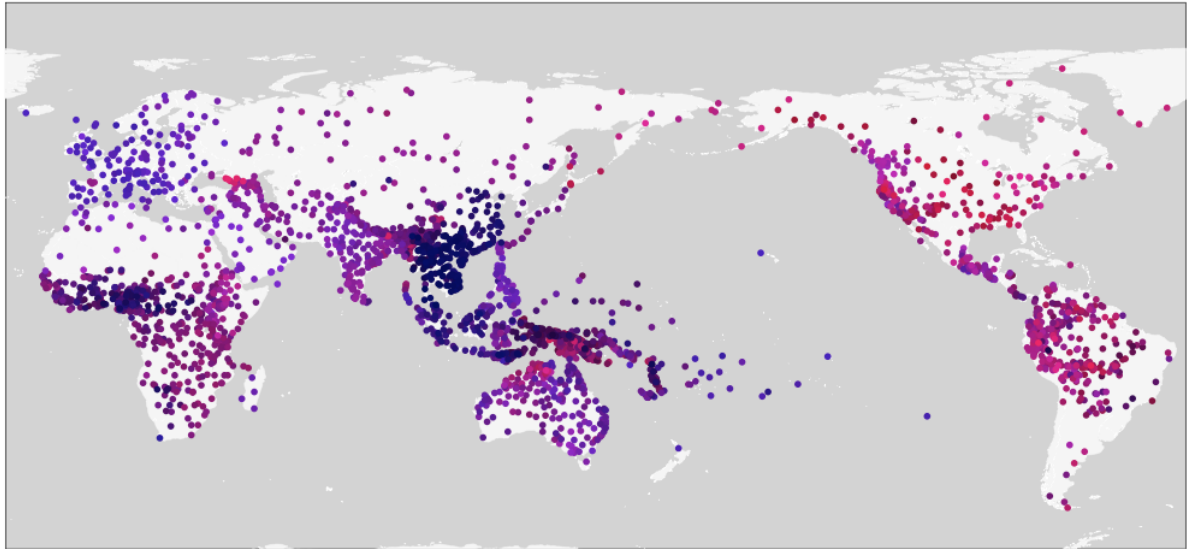
**Map A8.1:** Global map coloured for dimensions 2 and 3.

**A8.8. Colours reflecting Dimensions 2 and 4**



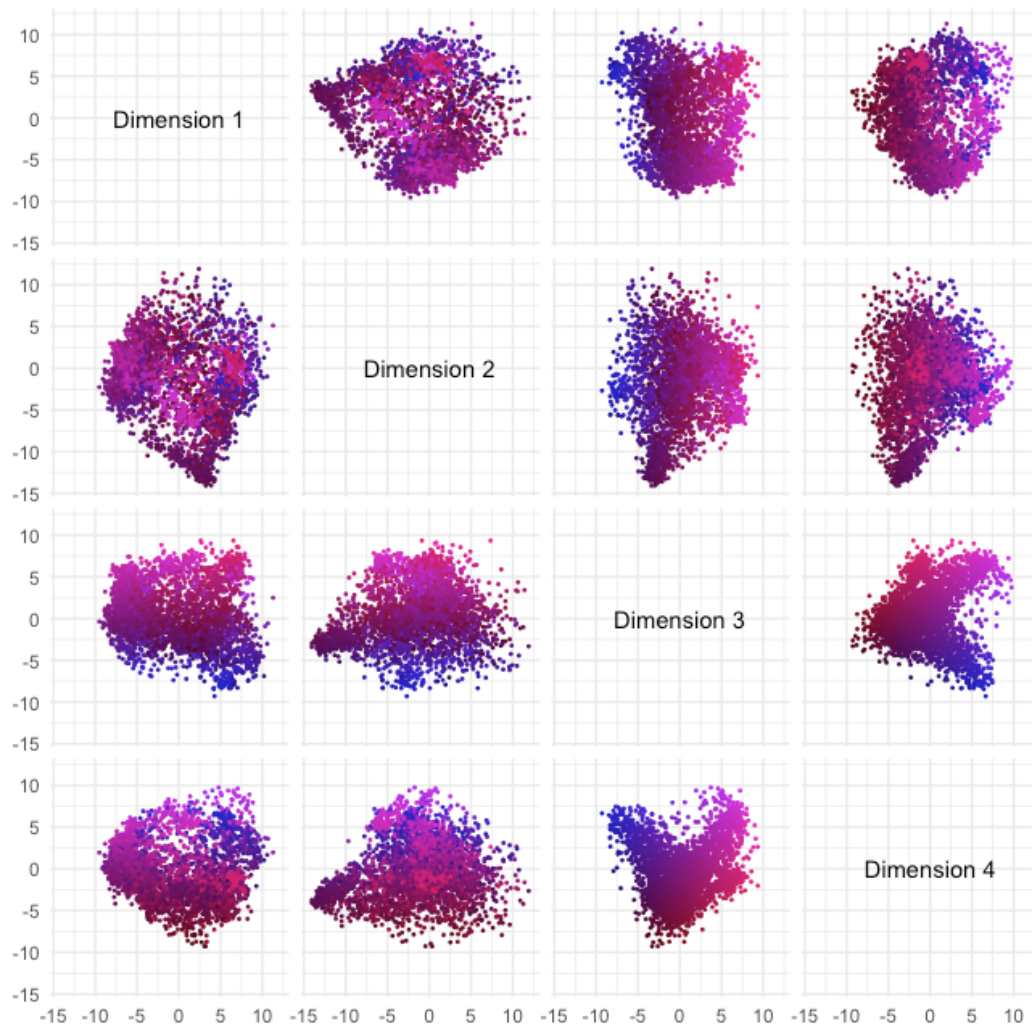
**Figure A8.1:** Dimension plots coloured for Dimensions 2 and 4.



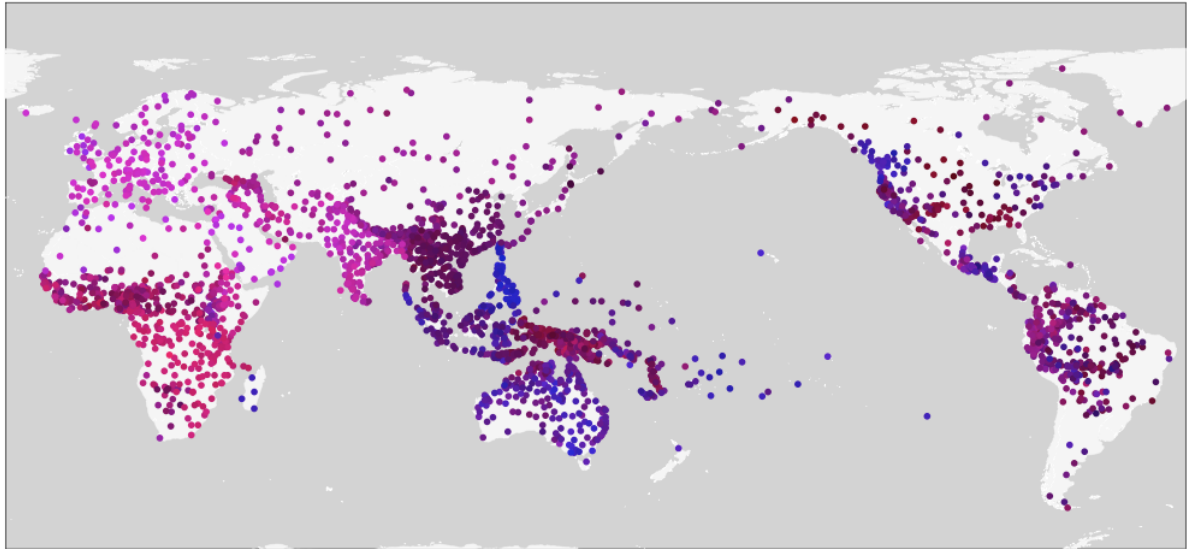


**Map A8.1:** Global map coloured for dimensions 2 and 4.

**A8.9. Colours reflecting Dimensions 3 and 4**



**Figure A8.1:** Dimension plots coloured for Dimensions 3 and 4.



**Map A8.1:** Global map coloured for dimensions 3 and 4.

## Appendix 9. Maps and plots of features contributing to the different dimensions

In 3.1 a number of different features were identified as contributing to the different dimensions that define typological variation. Here we present dimension plots illustrating the distribution of the different features that were reported in 3.1.1, 3.1.2, 3.1.3 and 3.1.4 in typological space, along the two dimensions for which they show the greatest correlations, and maps showing the geographical distribution of these features.

### A9.1 Features associated with Dimension 1

Dimension plots of features with strong associations with Dimension 1.

(all plots show V1 vs V2 except ‘Nominative agreement by prefix’, which plots V1 vs V3).

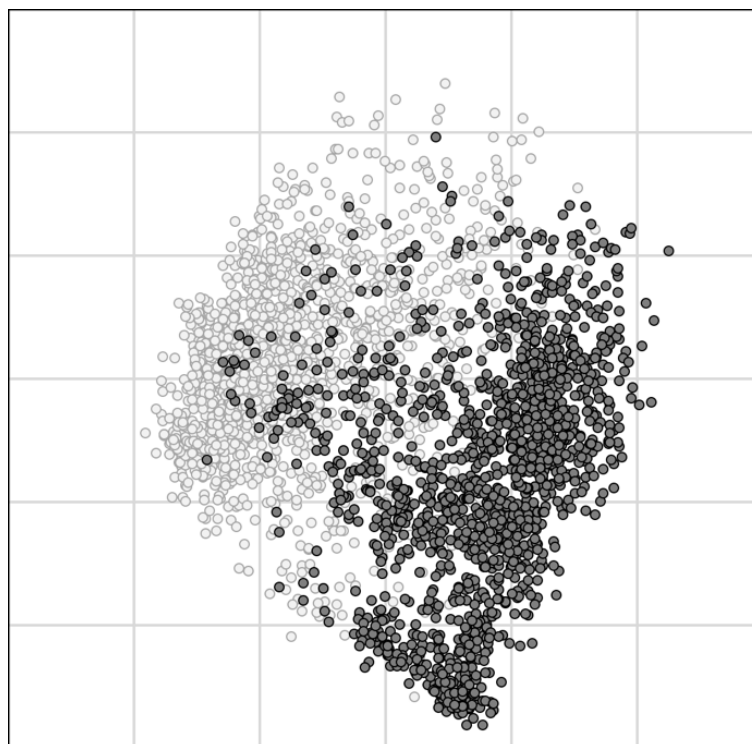
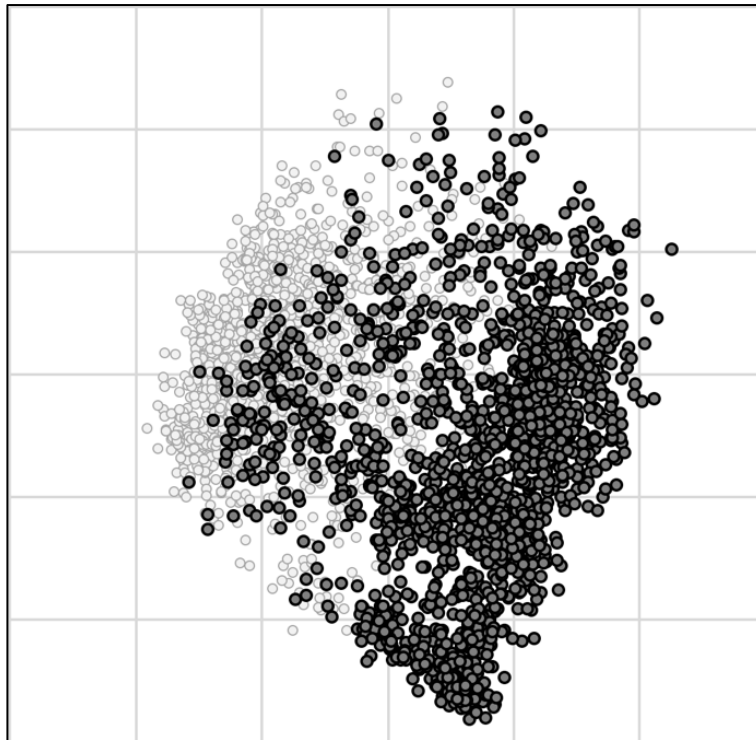
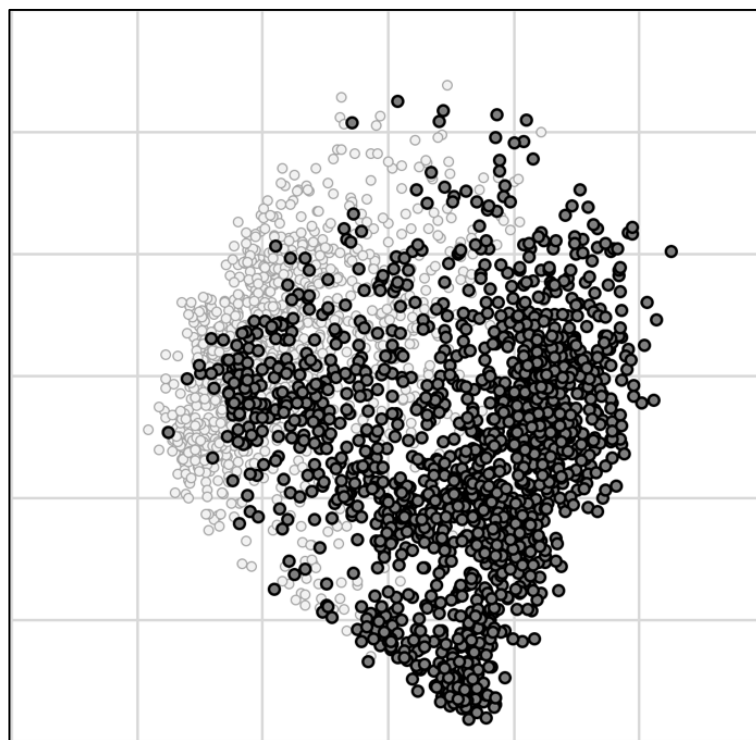


Figure A9.1.1: Prepositions in Dimensions 1 and 2 ( $r^2 = 0.50$ ).



**Figure A9.1.2:** VO order in Dimensions 1 and 2 ( $r^2 = 0.42$ ).



**Figure A9.1.3:** Initial subordination in Dimensions 1 and 2 ( $r^2 = 0.41$ ).

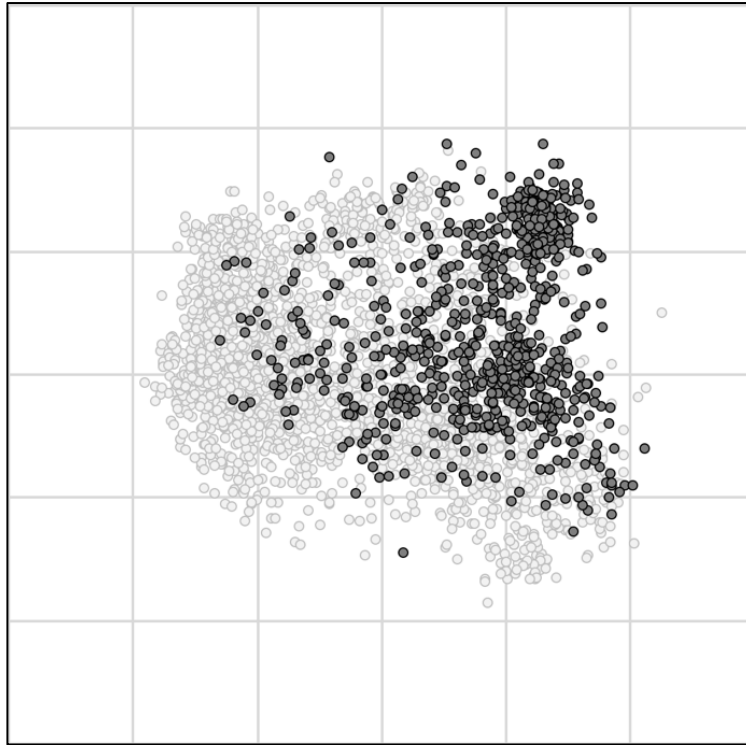


Figure A9.1.4: Nominative agreement by prefix in Dimensions 1 and 3 ( $r^2 = 0.32$ ).

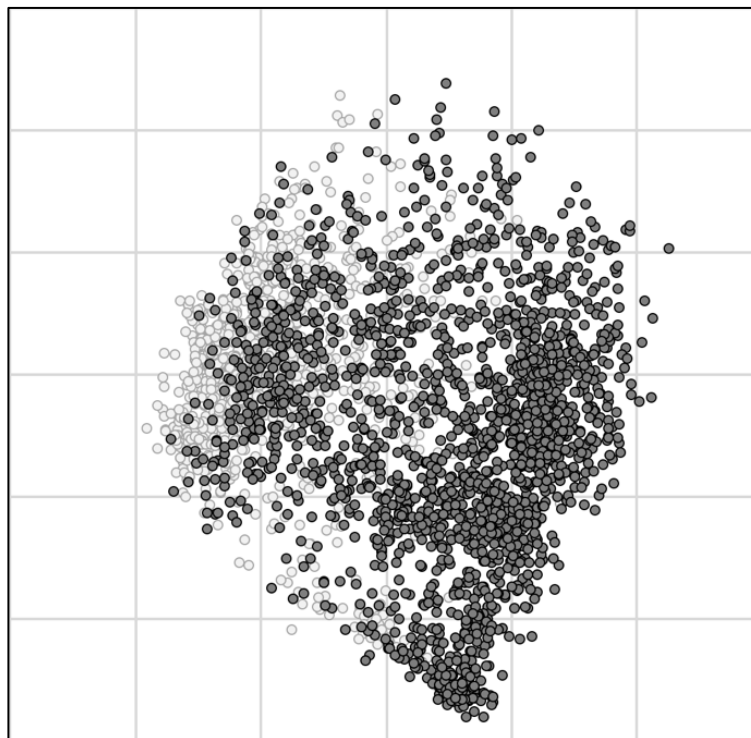


Figure A9.1.5: V Oblique order in Dimensions 1 and 2 ( $r^2 = 0.31$ ).

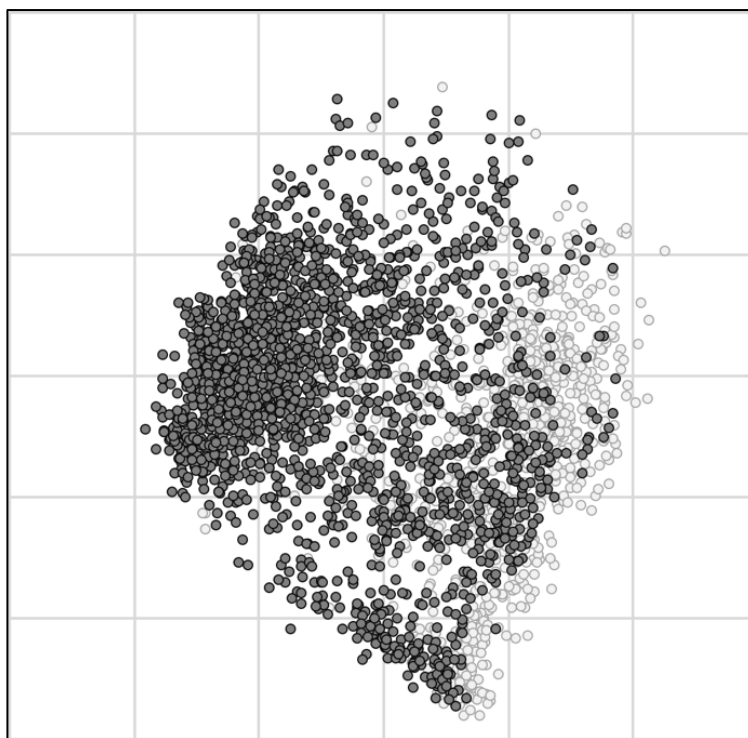


Figure A9.1.6: Genitive N order in Dimensions 1 and 2 ( $r^2 = 0.31$ ).

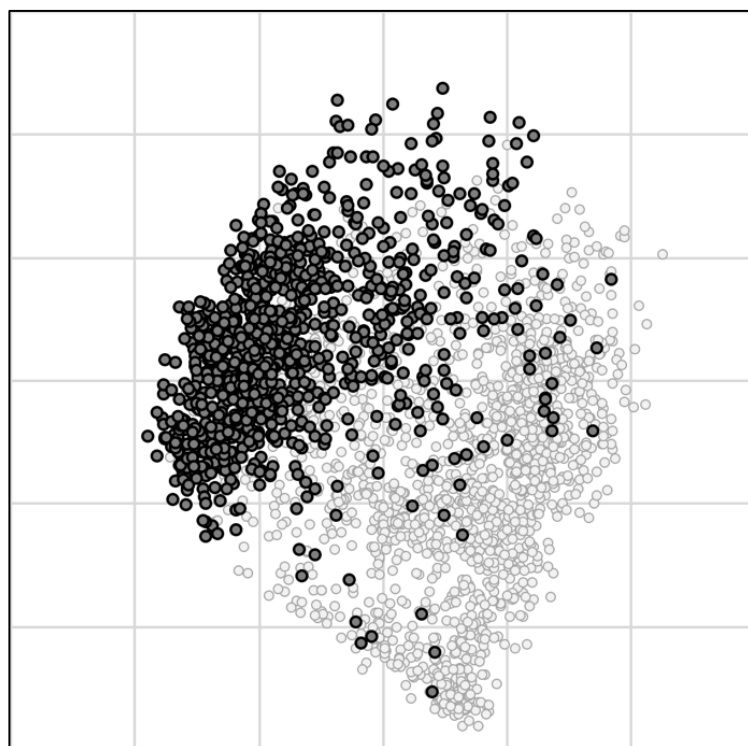


Figure A9.1.7: Subordination by suffix in Dimensions 1 and 2 ( $r^2 = 0.39$ ).

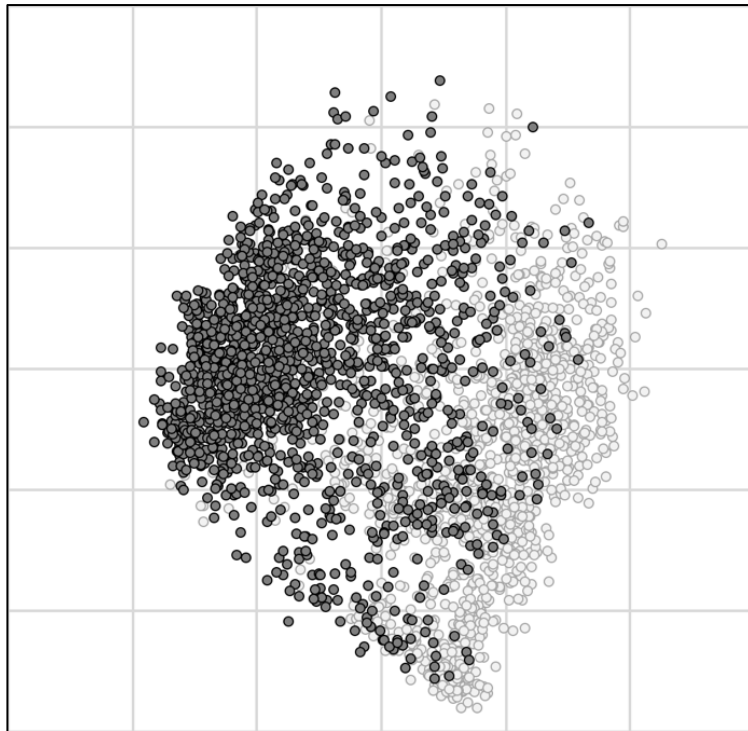


Figure A9.1.8: Postpositions in Dimensions 1 and 2 ( $r^2 = 0.42$ ).

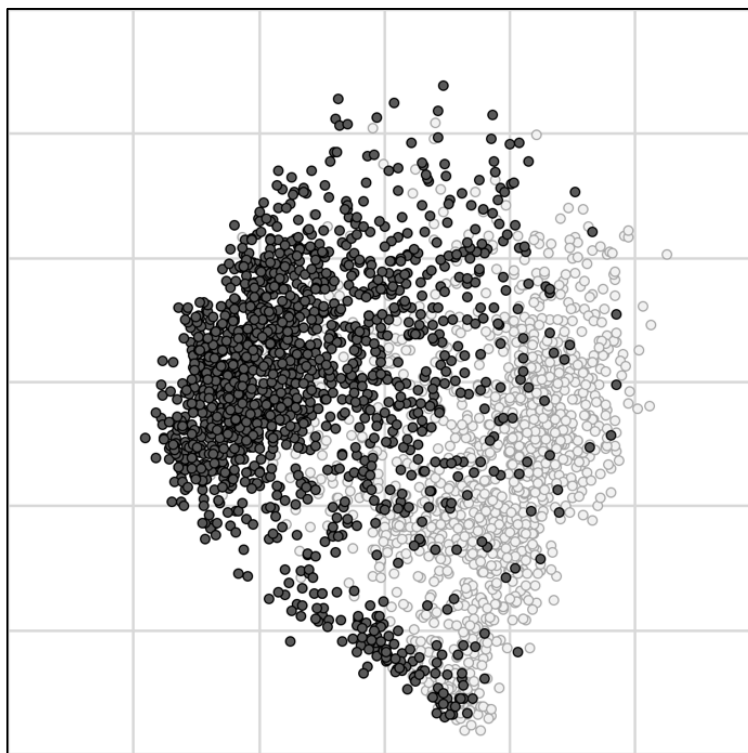
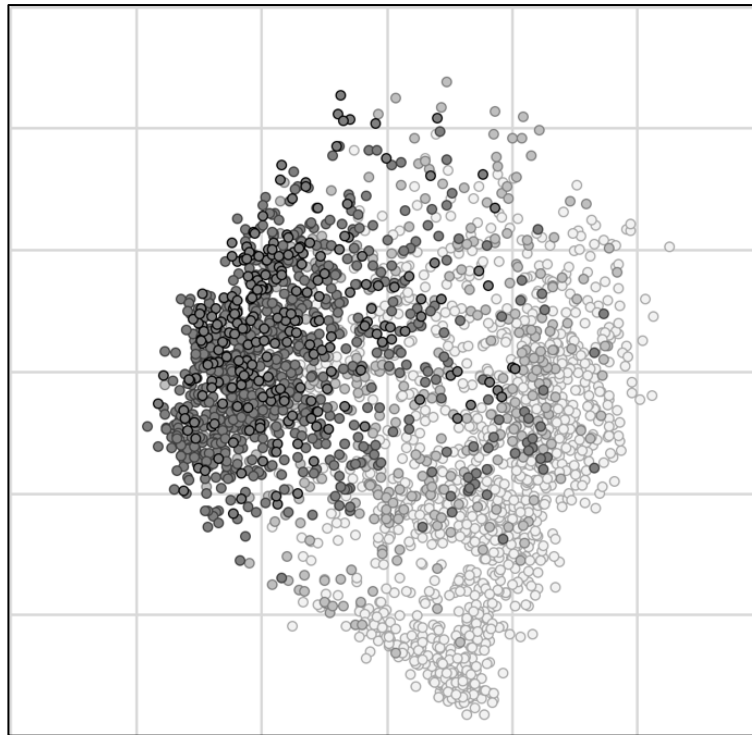
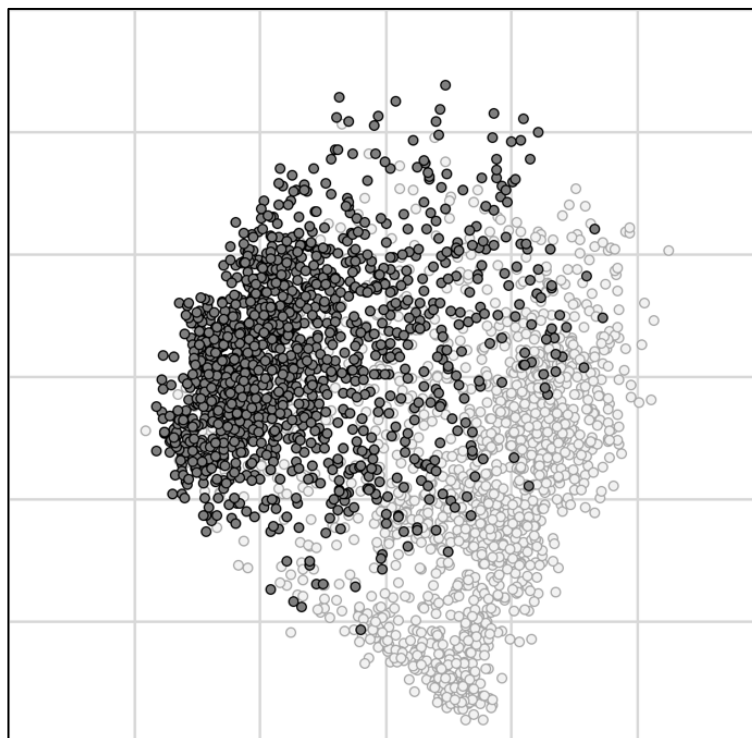


Figure A9.1.9: Oblique V order in Dimensions 1 and 2 ( $r^2 = 0.45$ ).





**Figure A9.1.10:** Number of cases in Dimensions 1 and 2 ( $r^2 = 0.49$ ).



**Figure A9.1.11:** Postnominal case in Dimensions 1 and 2 ( $r^2 = 0.50$ ).

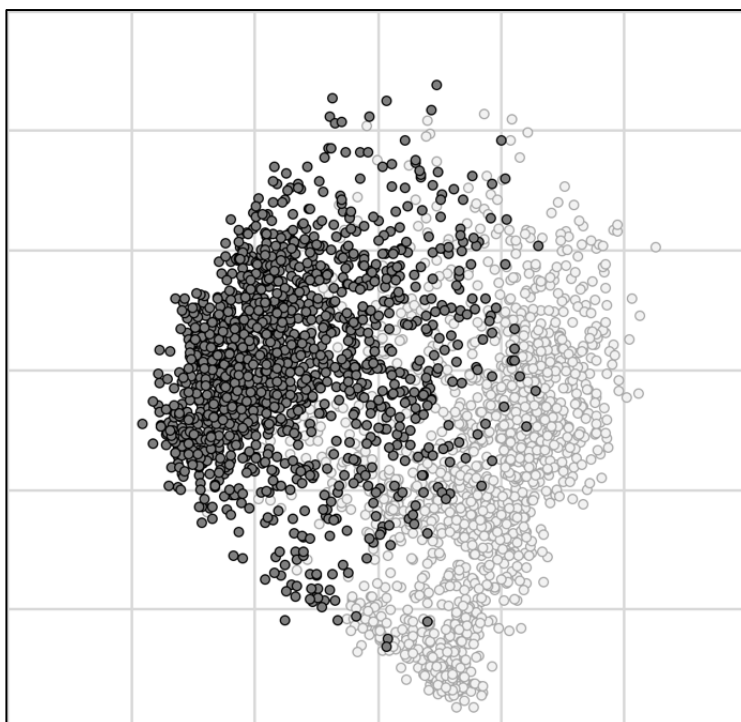
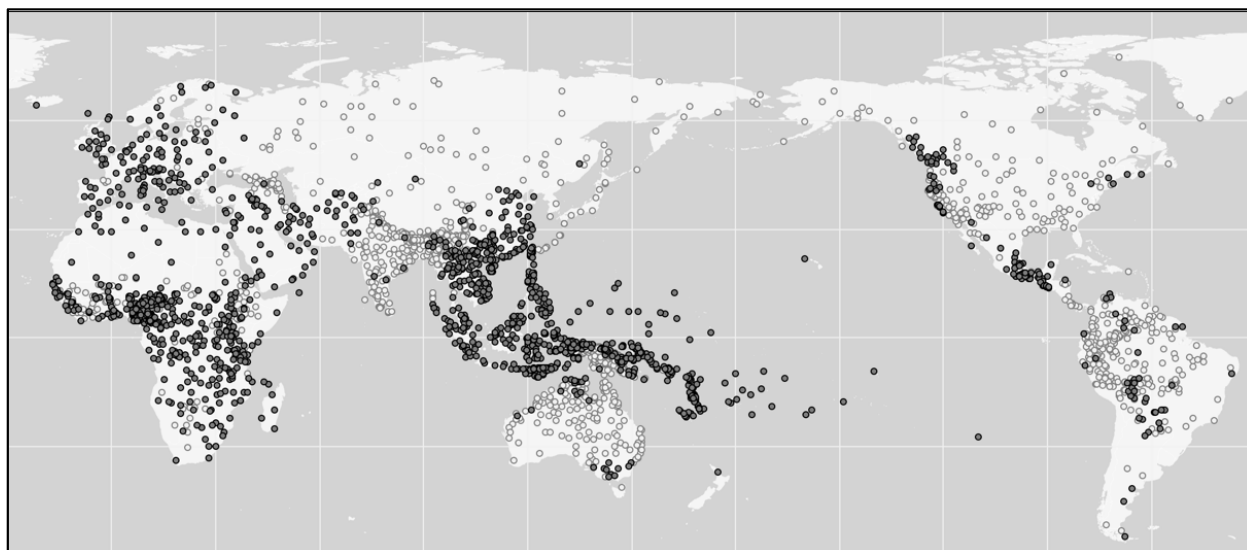
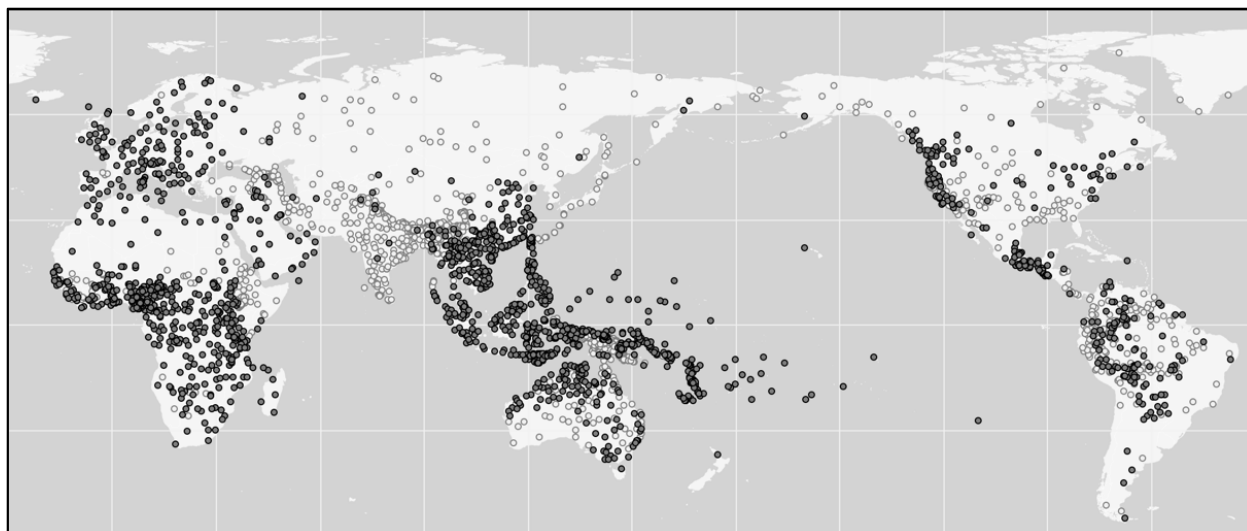


Figure A9.1.12: SOV order in Dimensions 1 and 2 ( $r^2 = 0.53$ ).

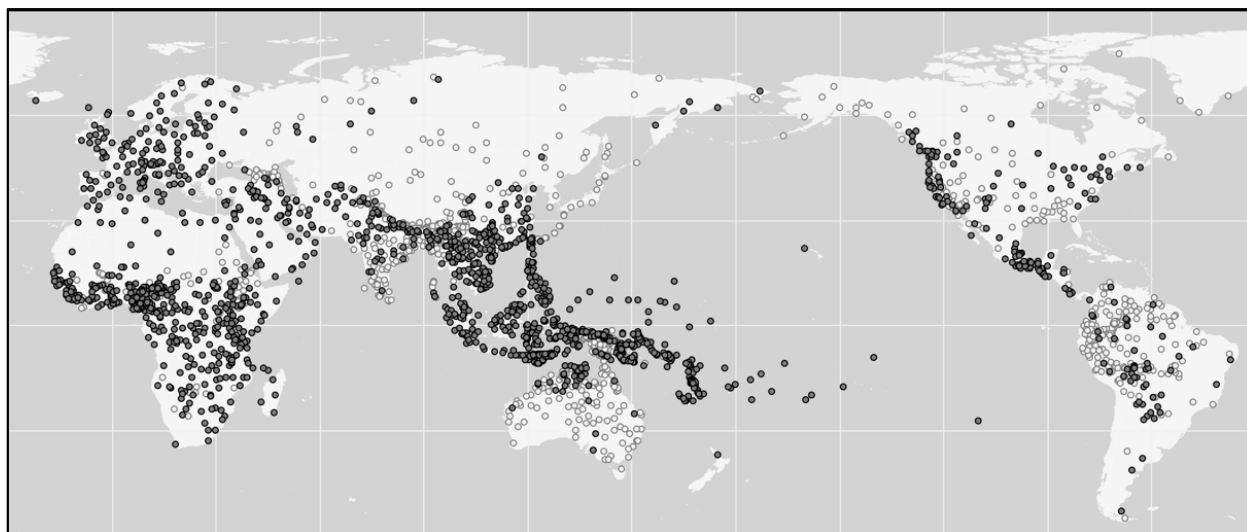
Maps of features with strong associations with Dimension 1.



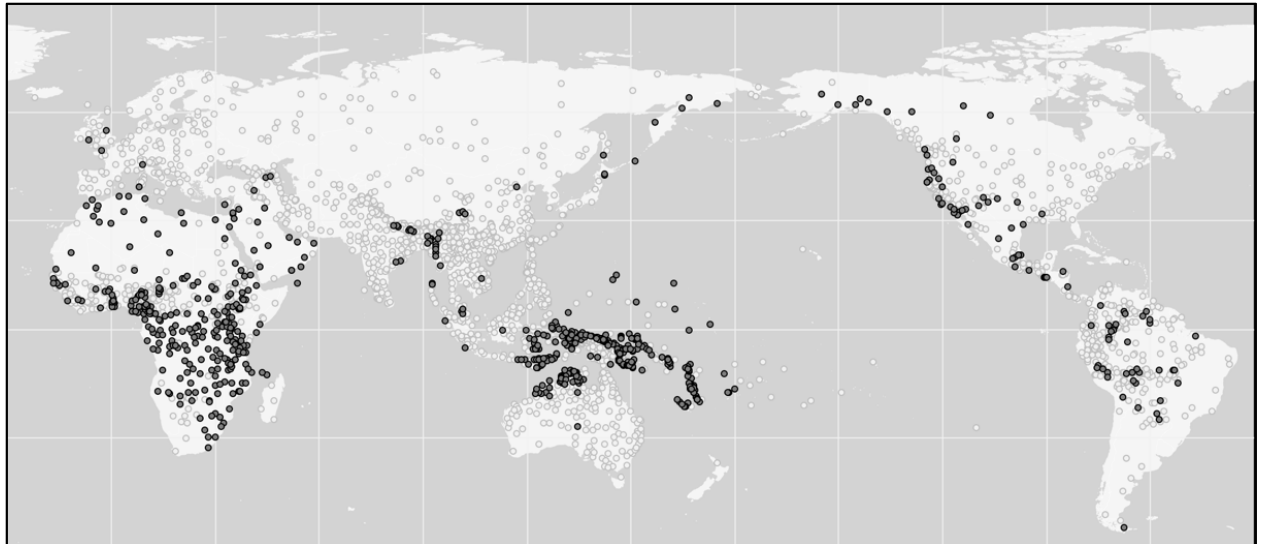
Map A9.1.1: Prepositions.



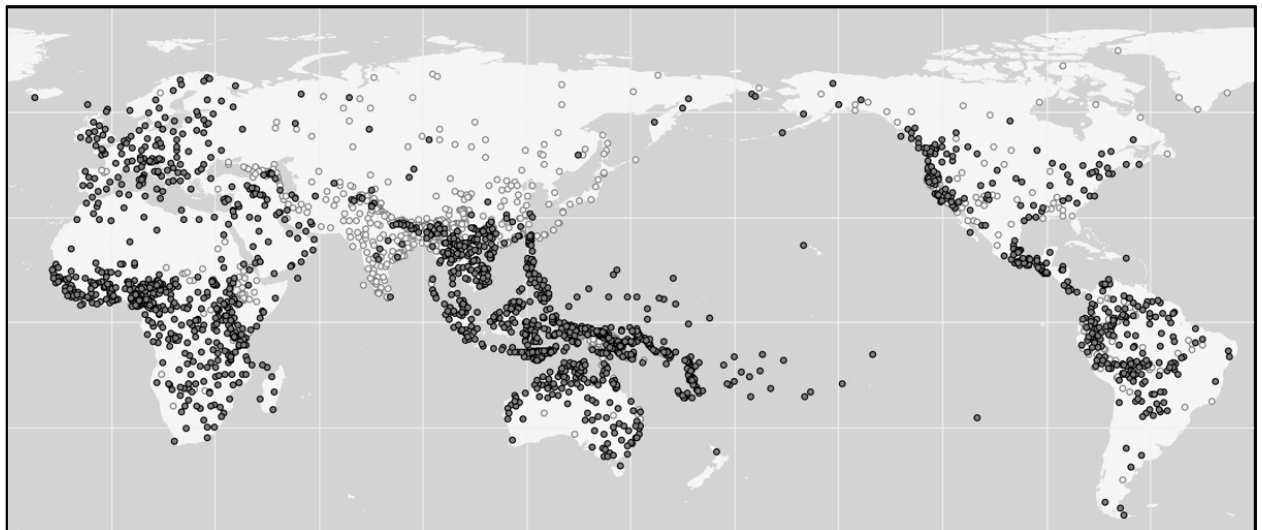
**Map A9.1.2:** VO order.



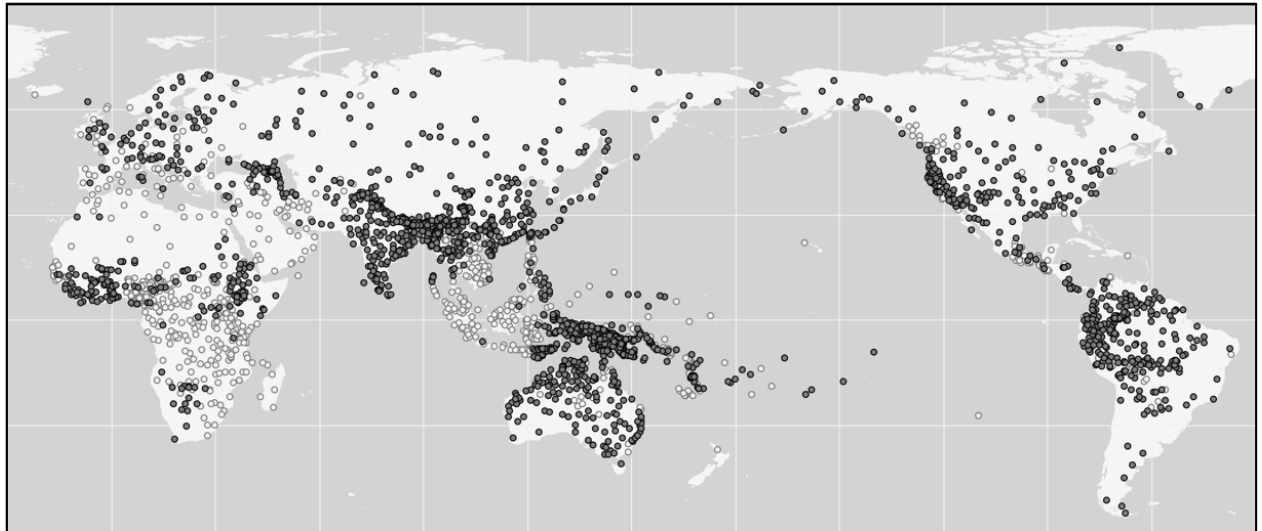
**Map A9.1.3:** Initial subordination.



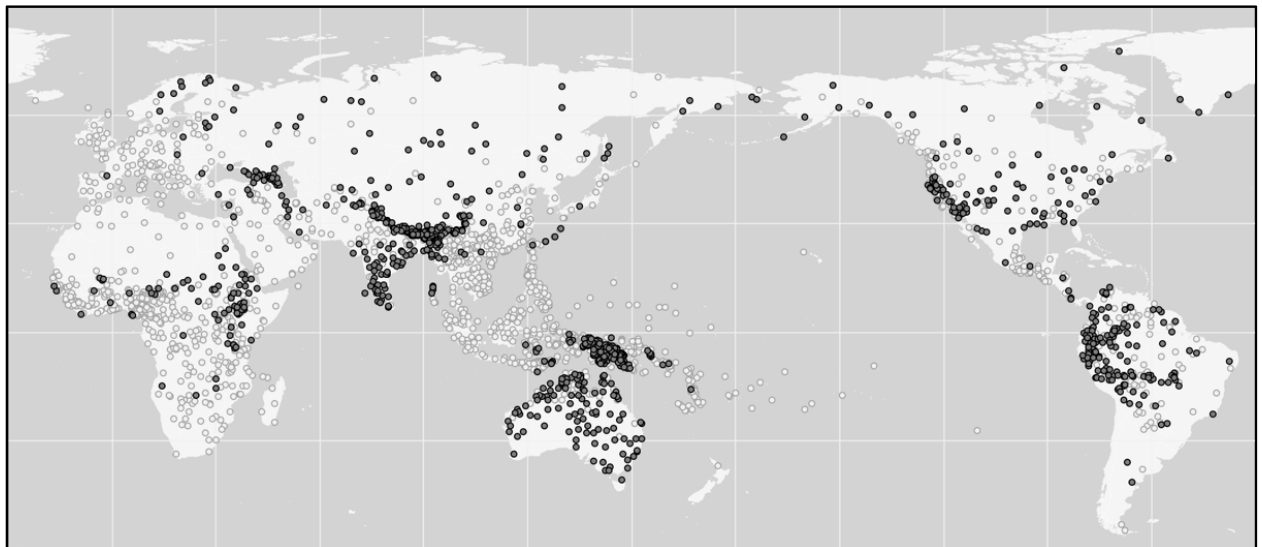
**Map A9.1.4:** Nominative agreement by prefix.



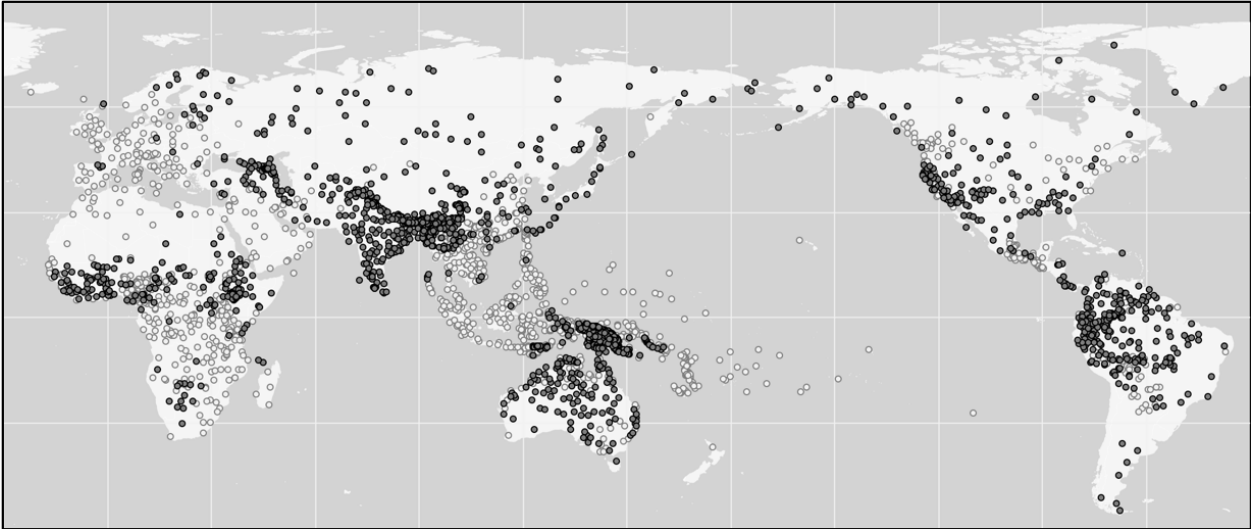
**Map A9.1.5:** V Oblique order.



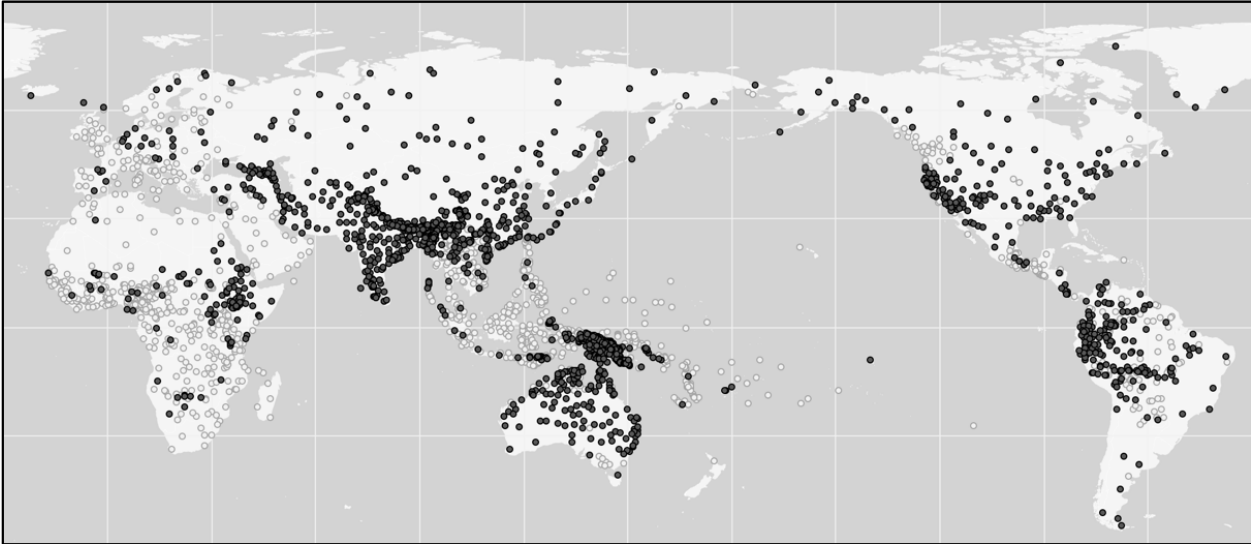
**Map A9.1.6:** Genitive N order.



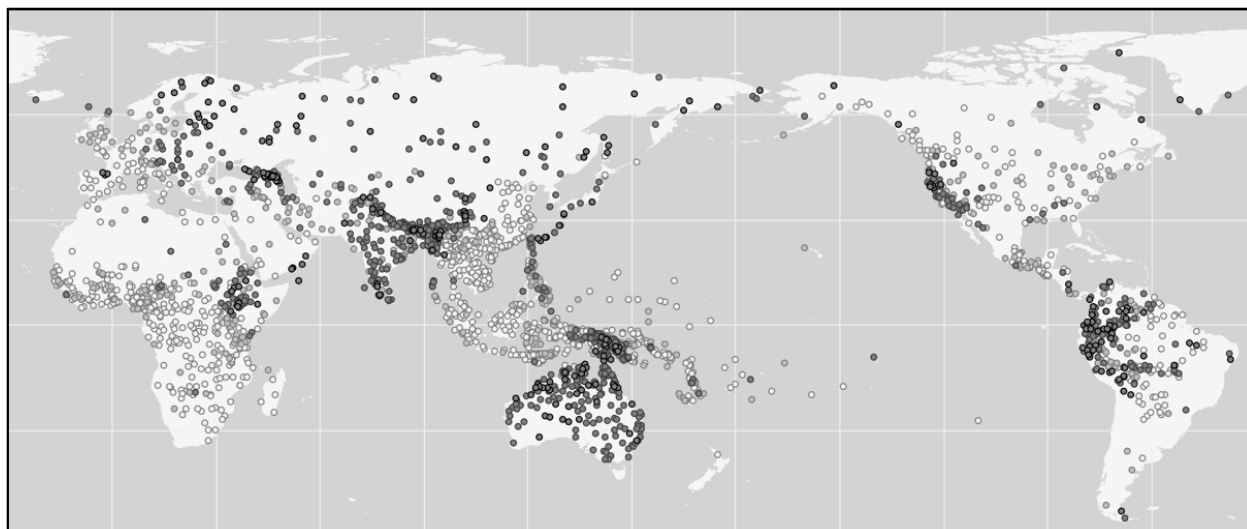
**Map A9.1.7:** Subordination by suffix.



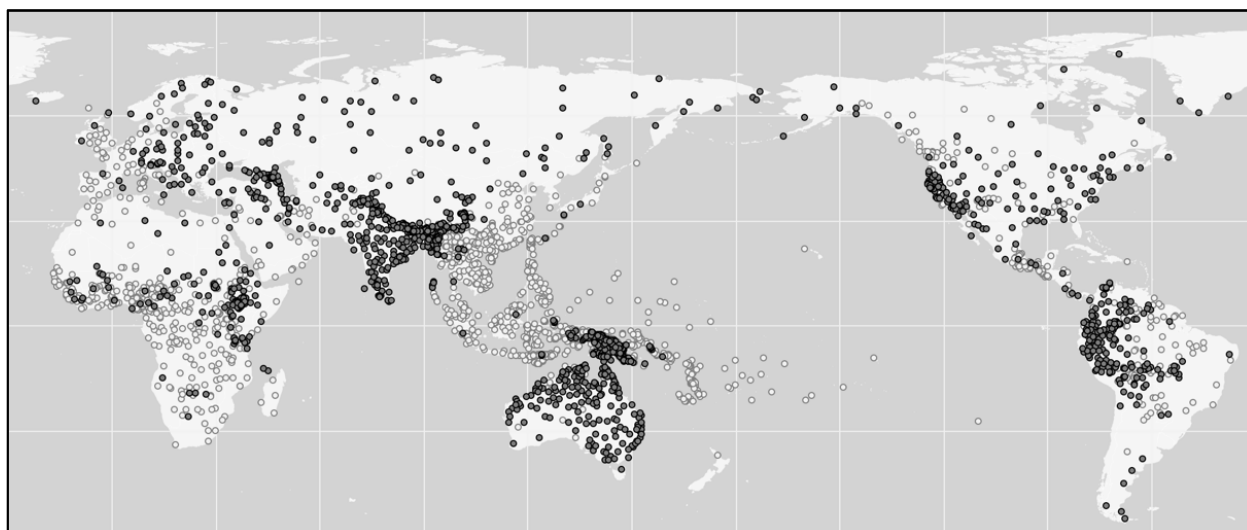
Map A9.1.8: Postpositions.



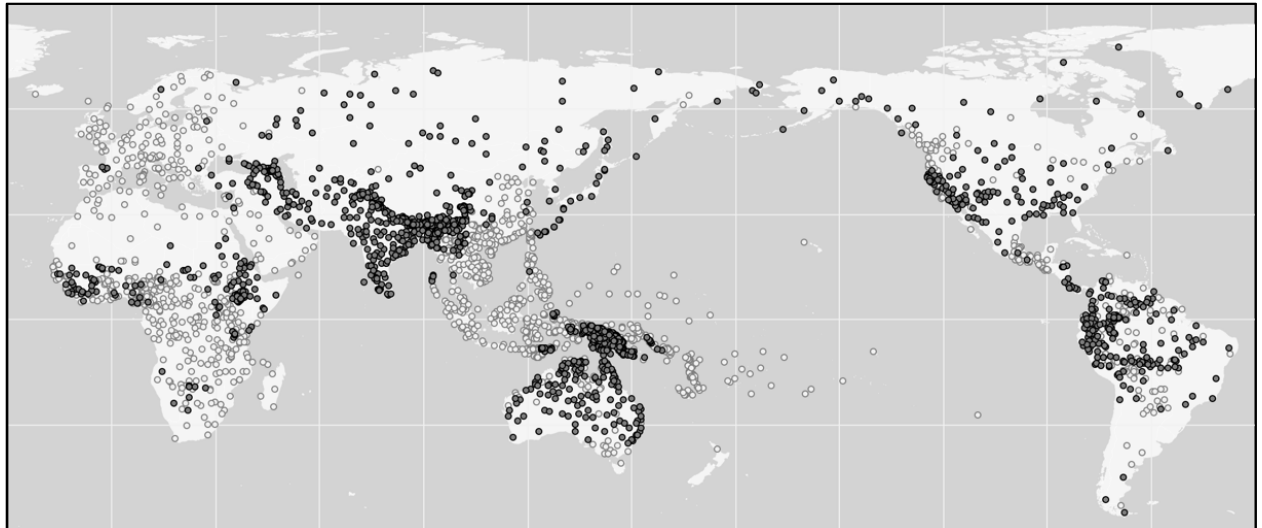
Map A9.1.9: Oblique V order.



Map A9.1.10: Number of cases.



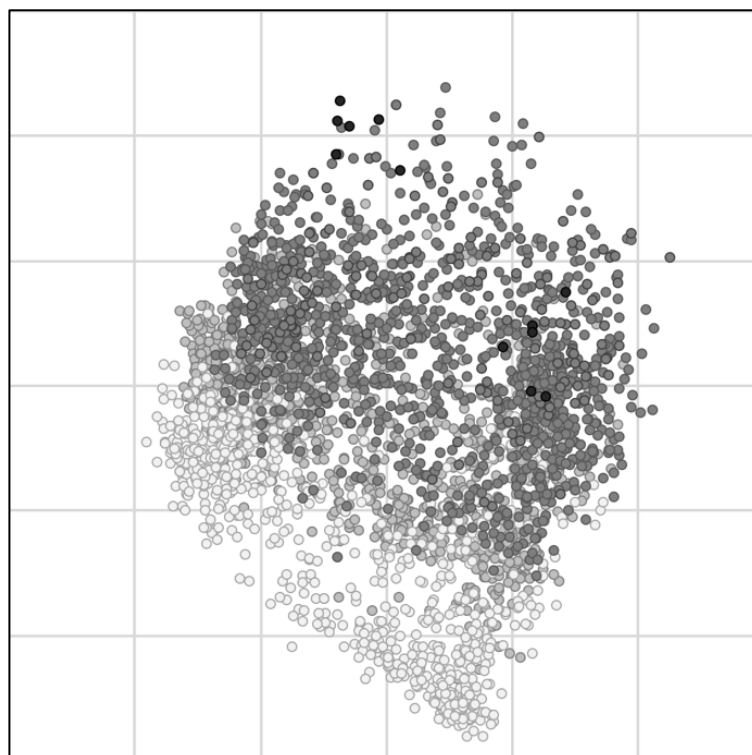
Map A9.1.11: Postnominal case.



Map A9.1.12: SOV order.

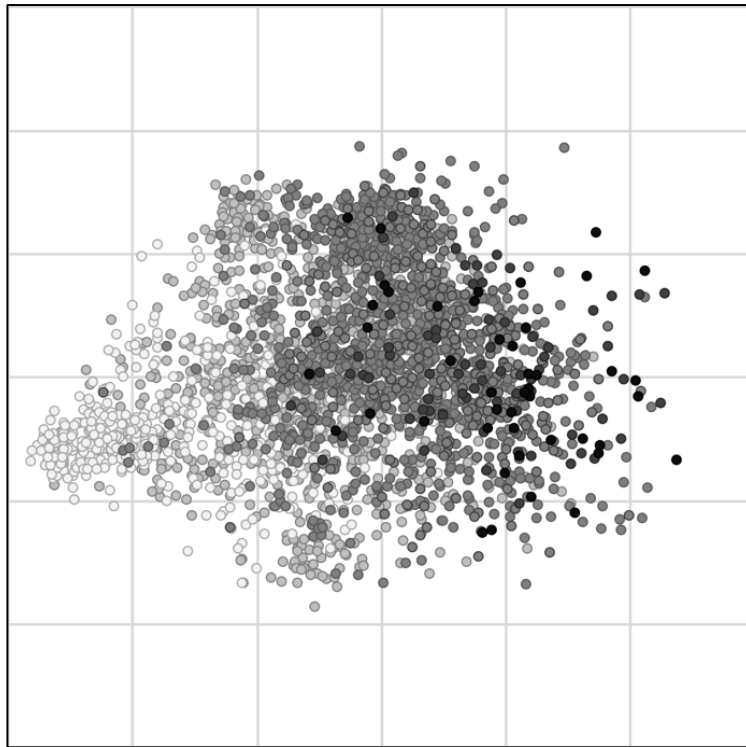
### ***A9.2 Features associated with Dimension 2***

Dimension plots of features with strong associations with Dimension 2.  
Pairs of dimensions listed for each chart.

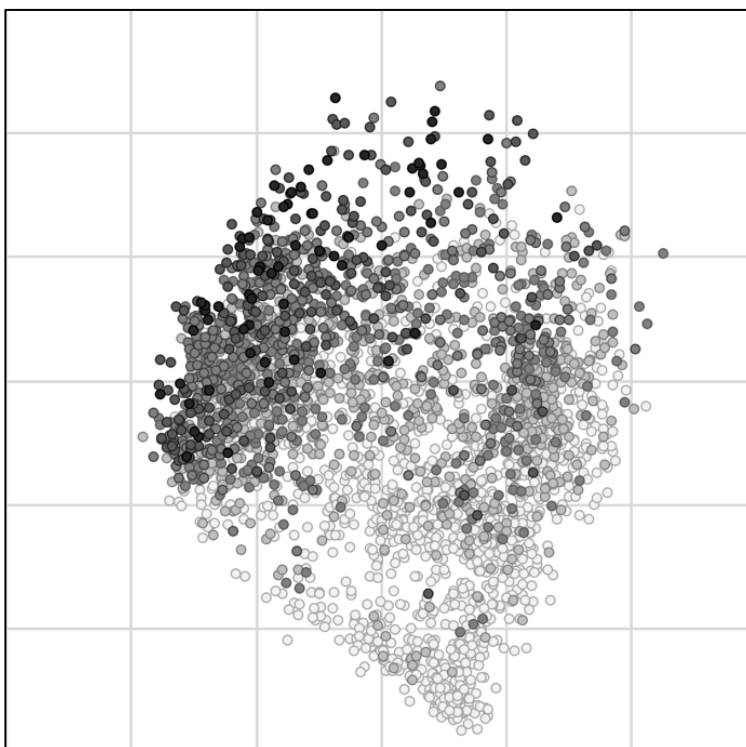


**Figure A9.2.1:** Total verbal agreement positions in Dimensions 2 and 3 ( $r^2 = 0.37$ ).





**Figure A9.2.2:** Total verbal inflectional synthesis in Dimensions 2 and 3 ( $r^2 = 0.34$ ).



**Figure A9.2.3:** Total Modality affixes in Dimensions 1 and 2 ( $r^2 = 0.30$ ).

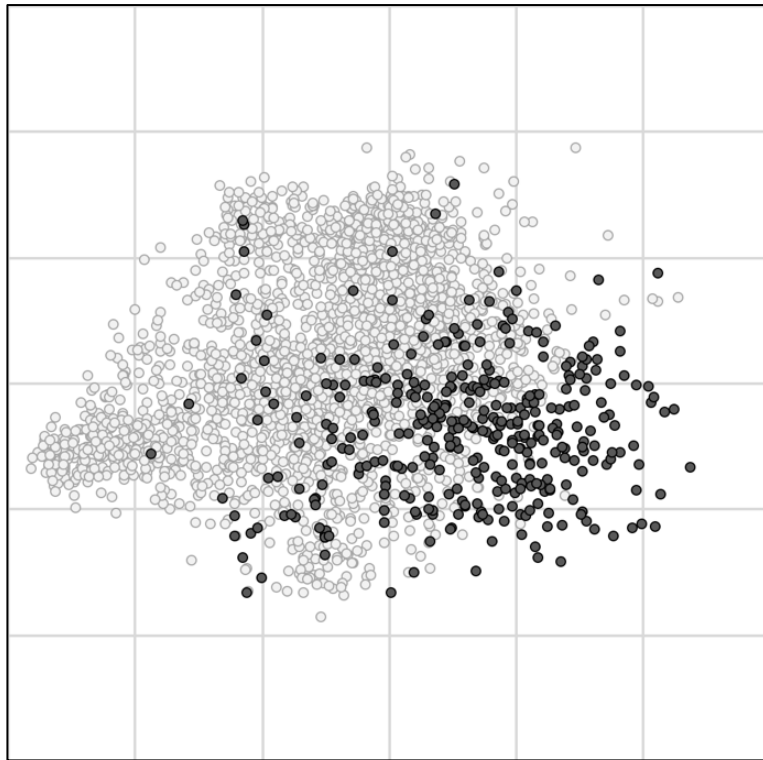


Figure A9.2.4: Incorporation in Dimensions 2 and 3 ( $r^2 = 0.22$ ).

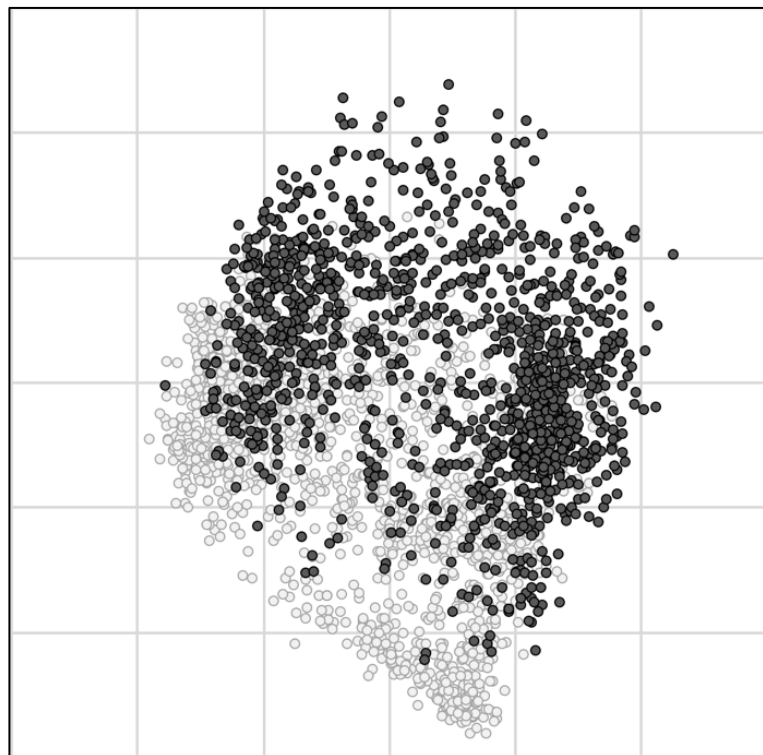
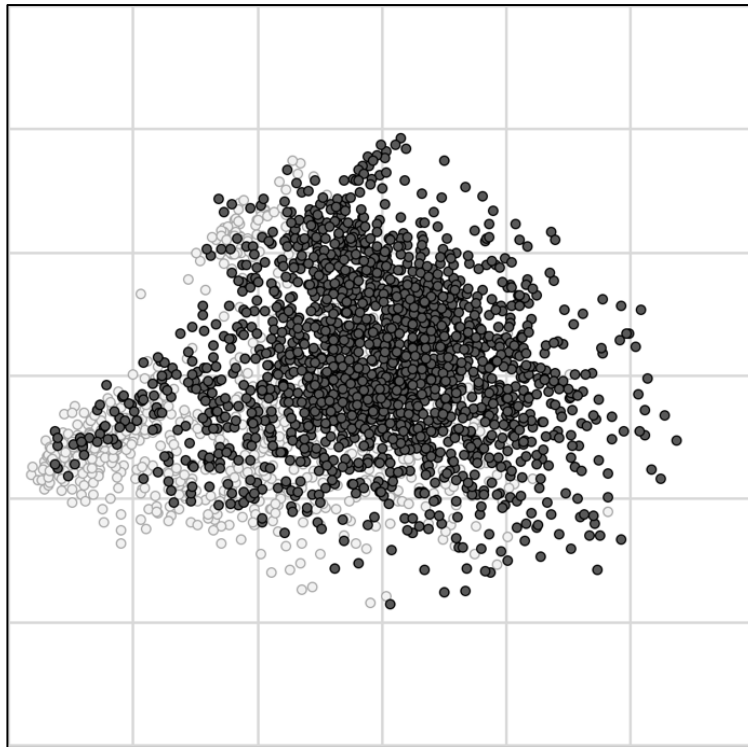
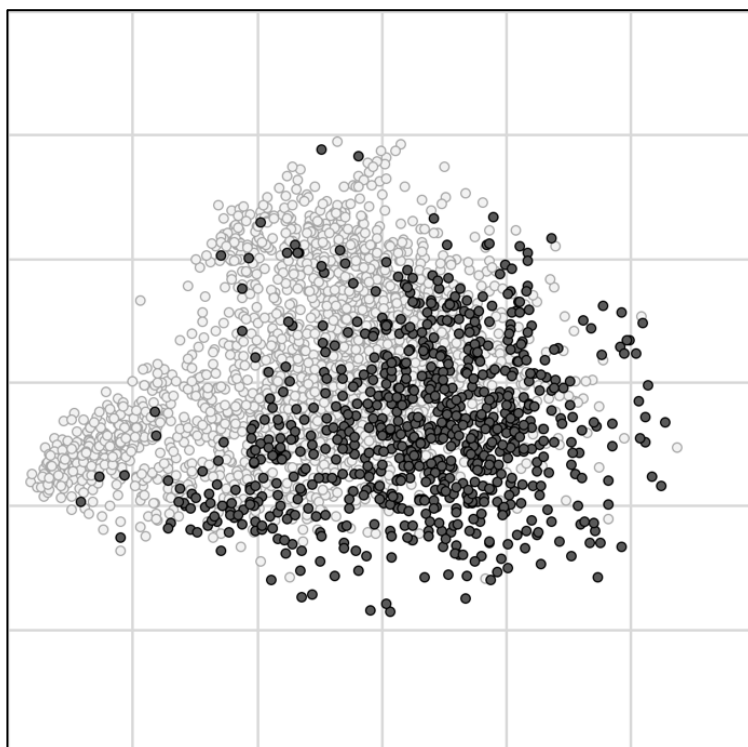


Figure A9.2.5: Applicatives in Dimensions 1 and 2 ( $r^2 = 0.20$ ).



**Figure A9.2.6:** Morphological causatives in Dimensions 2 and 4 ( $r^2 = 0.16$ ).



**Figure A9.2.7:** Possessive prefixes on nouns in Dimensions 2 and 4 ( $r^2 = 0.15$ ).

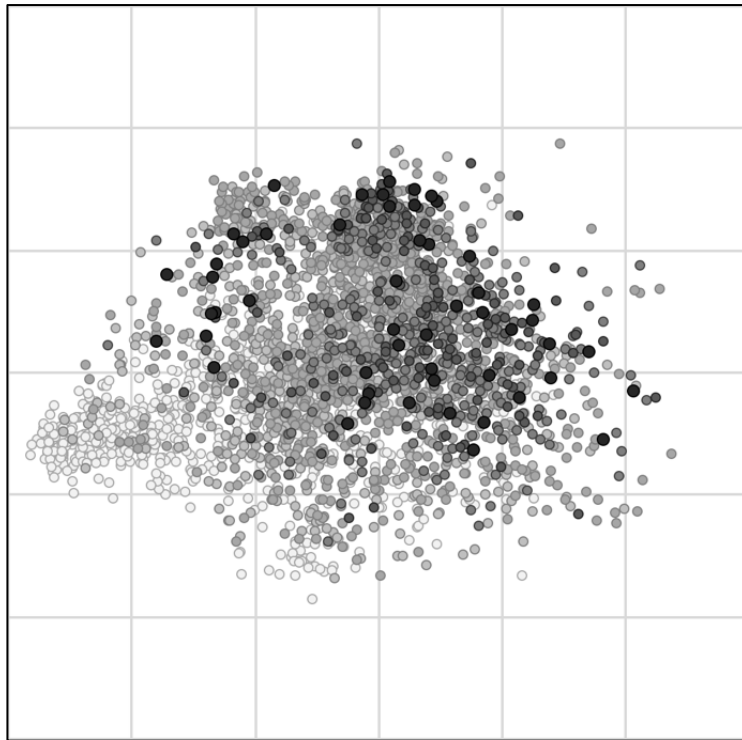


Figure A9.2.8: Total tense distinctions in Dimensions 2 and 3 ( $r^2 = 0.11$ ).

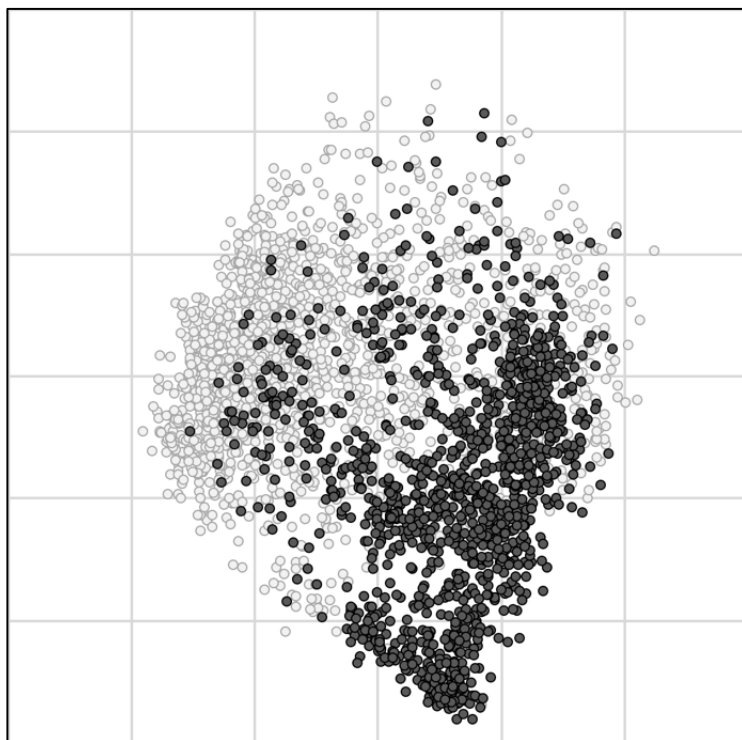


Figure A9.2.9: SVO order in Dimensions 1 and 2 ( $r^2 = 0.14$ ).

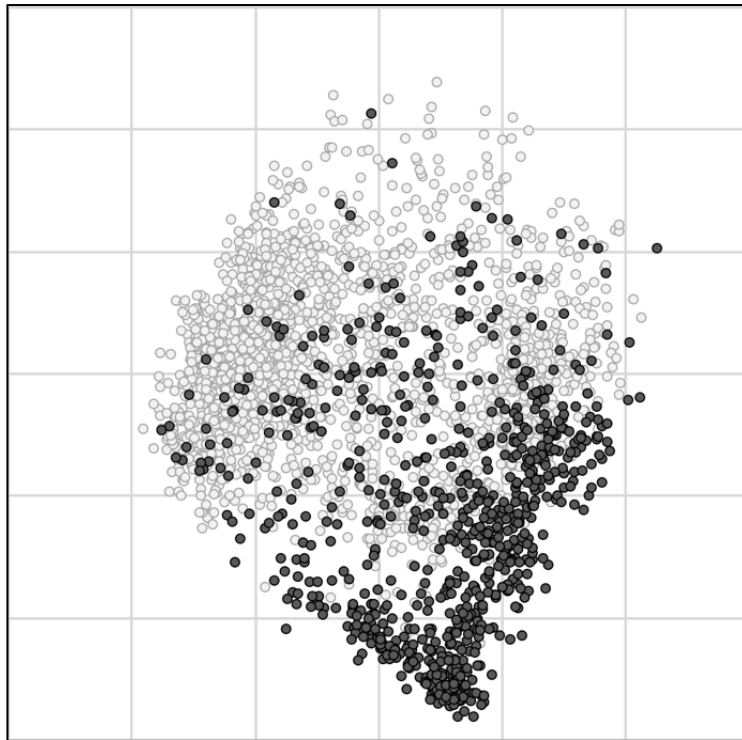


Figure A9.2.10: Symmetrical clauses: purpose in Dimensions 1 and 2 ( $r^2 = 0.17$ ).

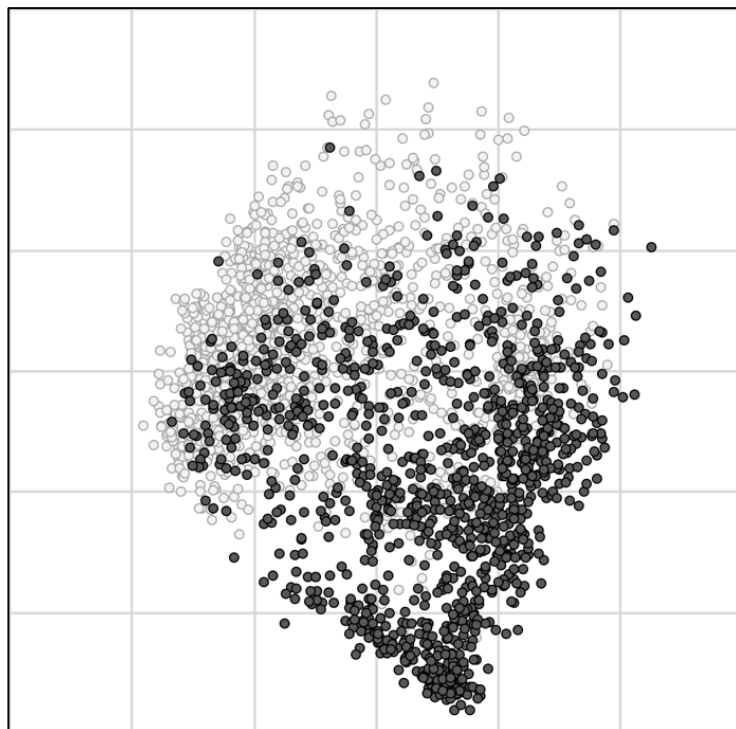


Figure A9.2.11: Symmetrical clauses: temporal in Dimensions 1 and 2 ( $r^2 = 0.18$ ).

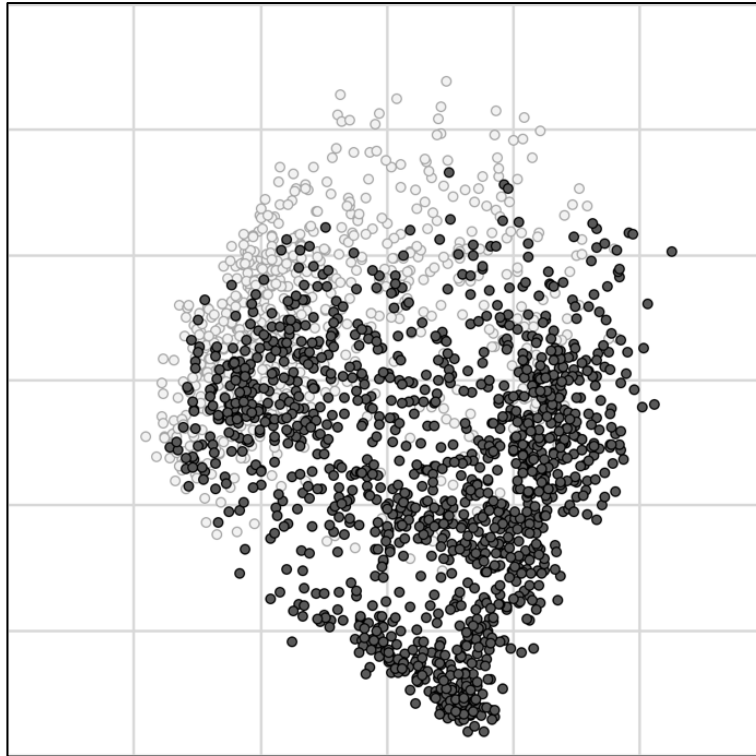
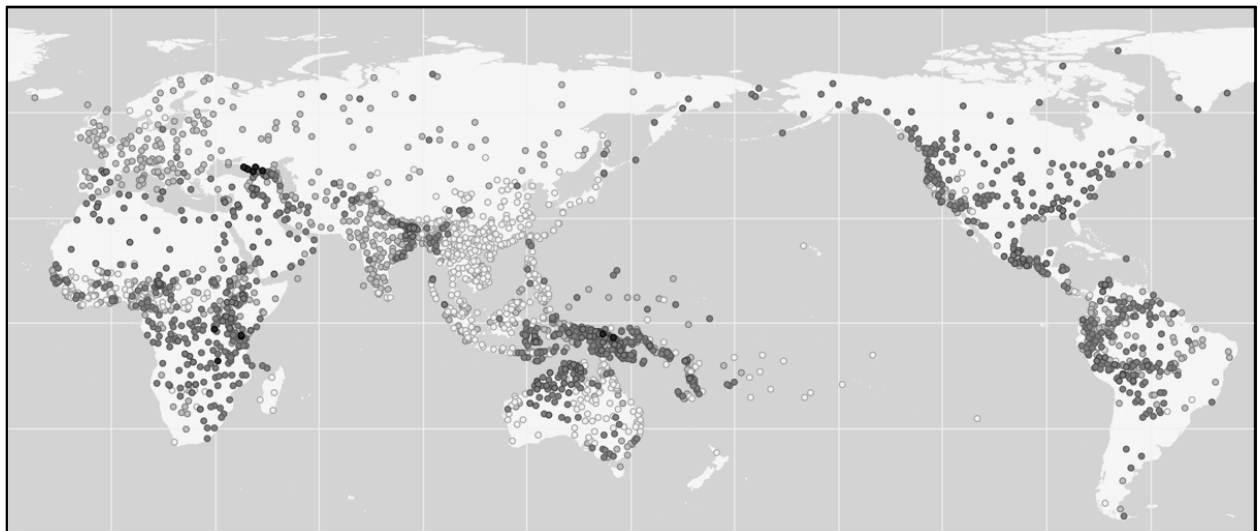
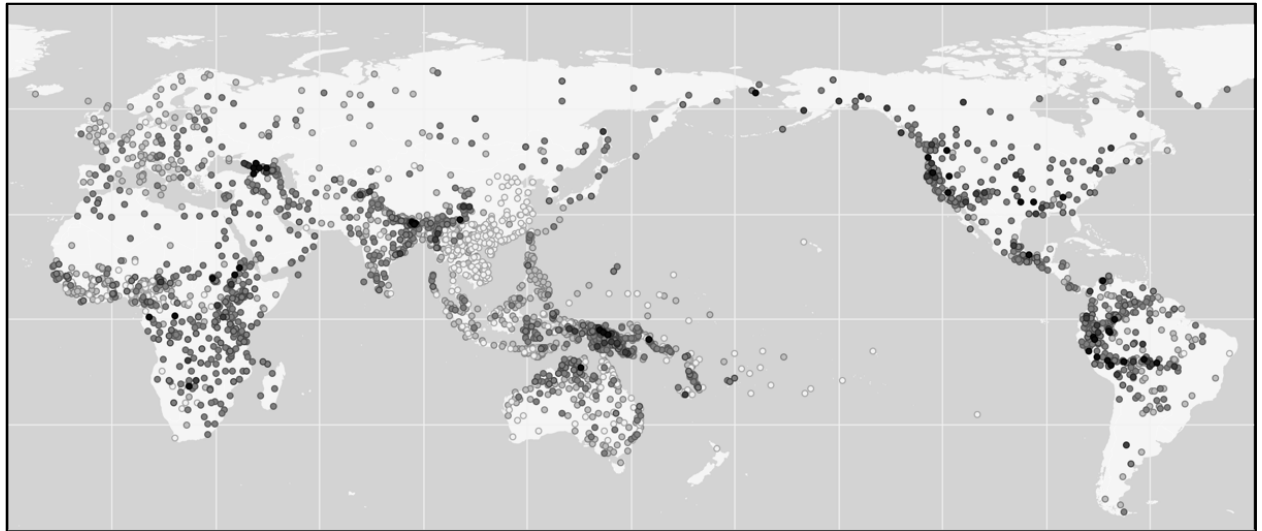


Figure A9.2.12: Symmetrical clauses: reason in Dimensions 1 and 2 ( $r^2 = 0.21$ ).

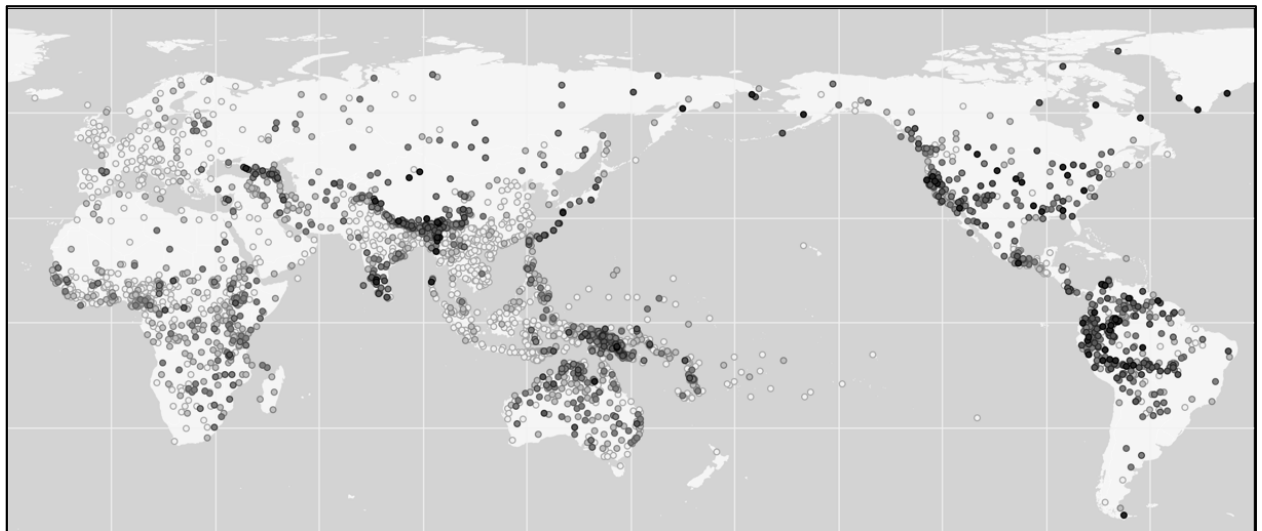
Maps of features with strong associations with Dimension 2:



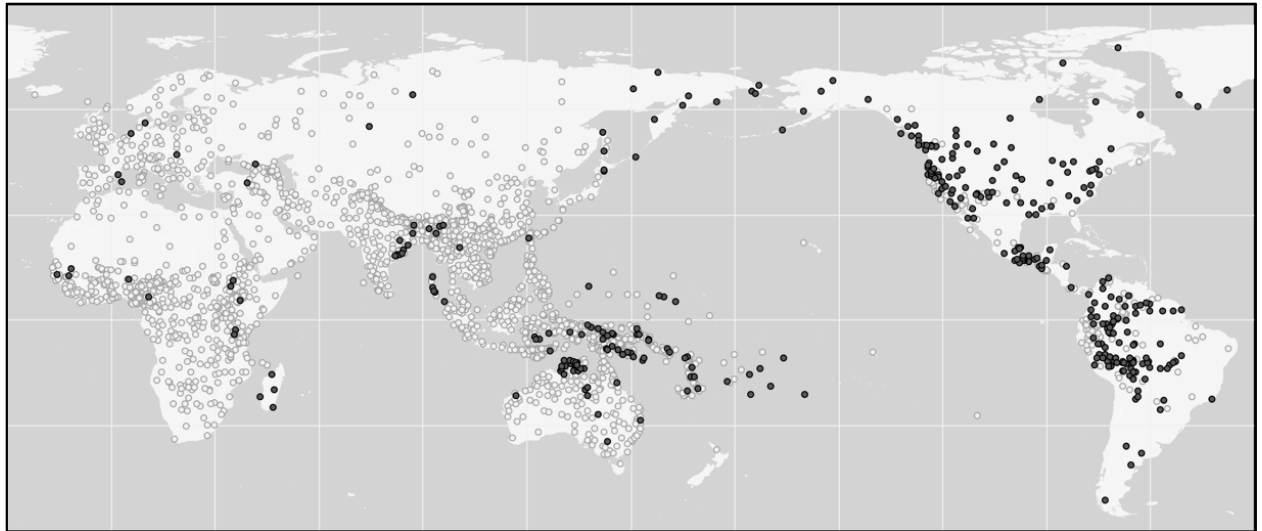
Map A9.2.1: Total verbal agreement positions.



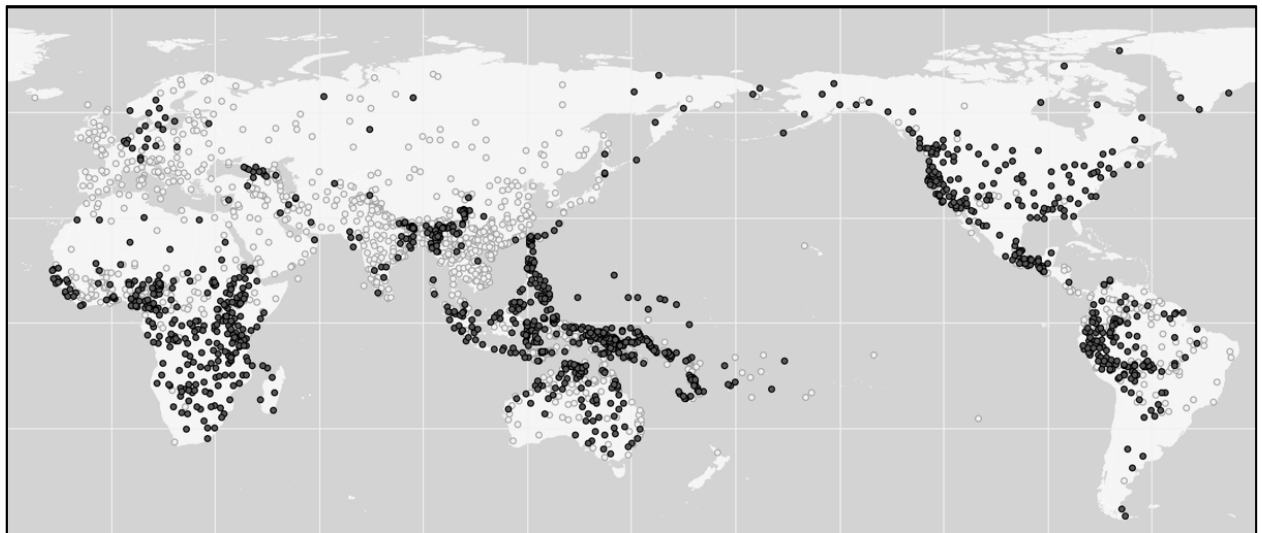
**Map A9.2.2:** Total verbal inflectional synthesis.



**Map A9.2.3:** Total Modality affixes.

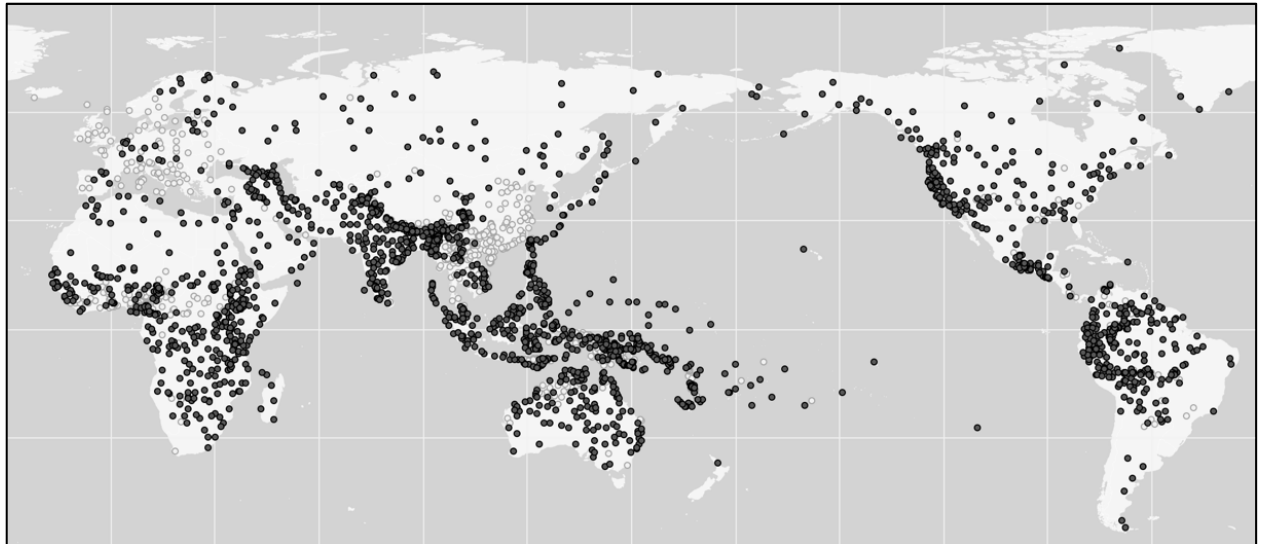


**Map A9.2.4: Incorporation.**

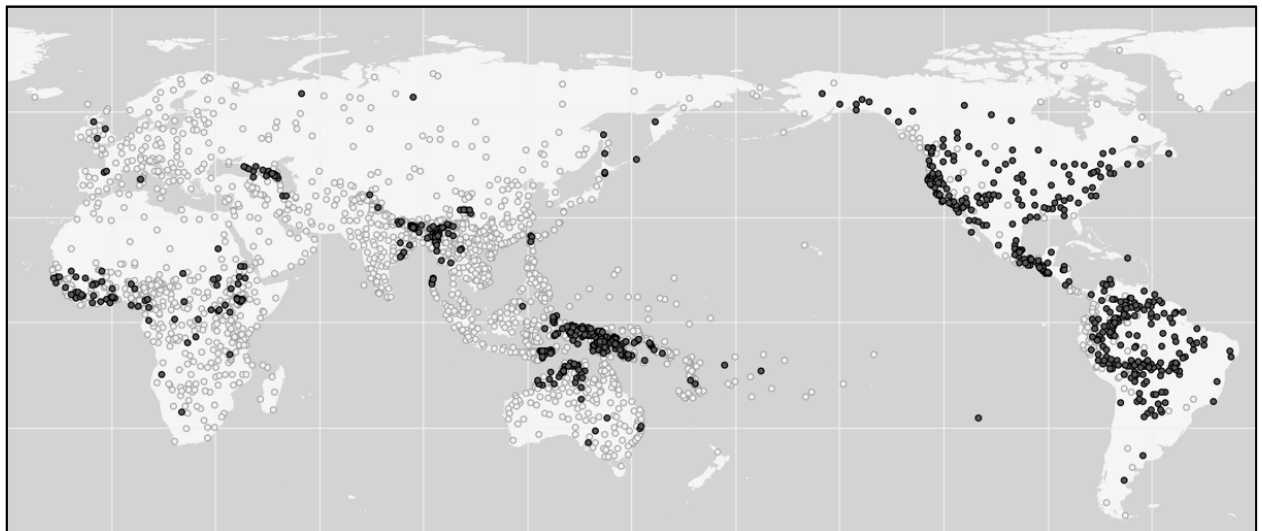


**Map A9.2.5: Applicatives.**

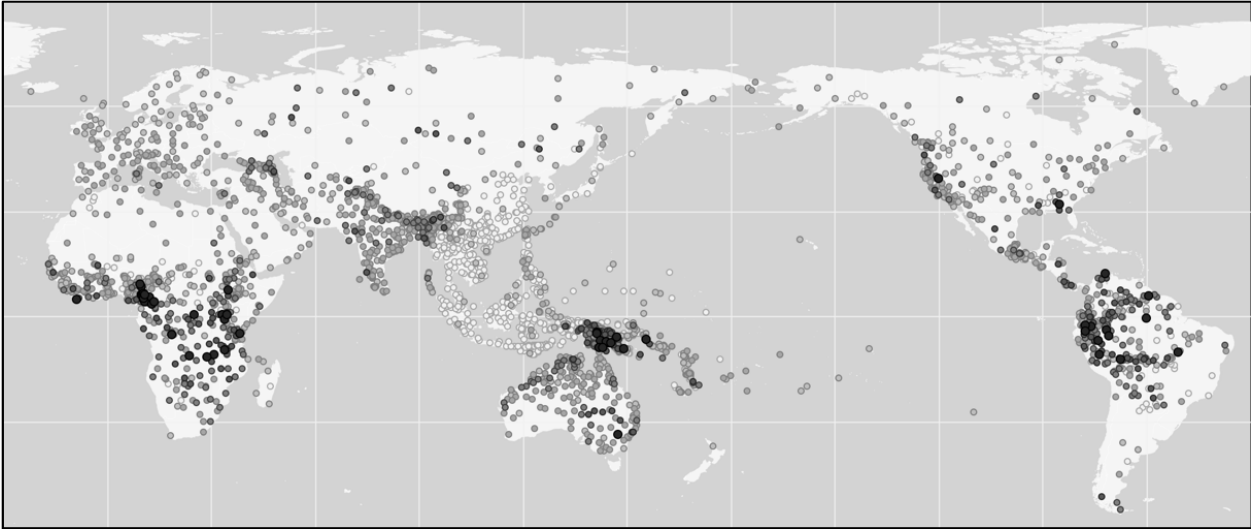




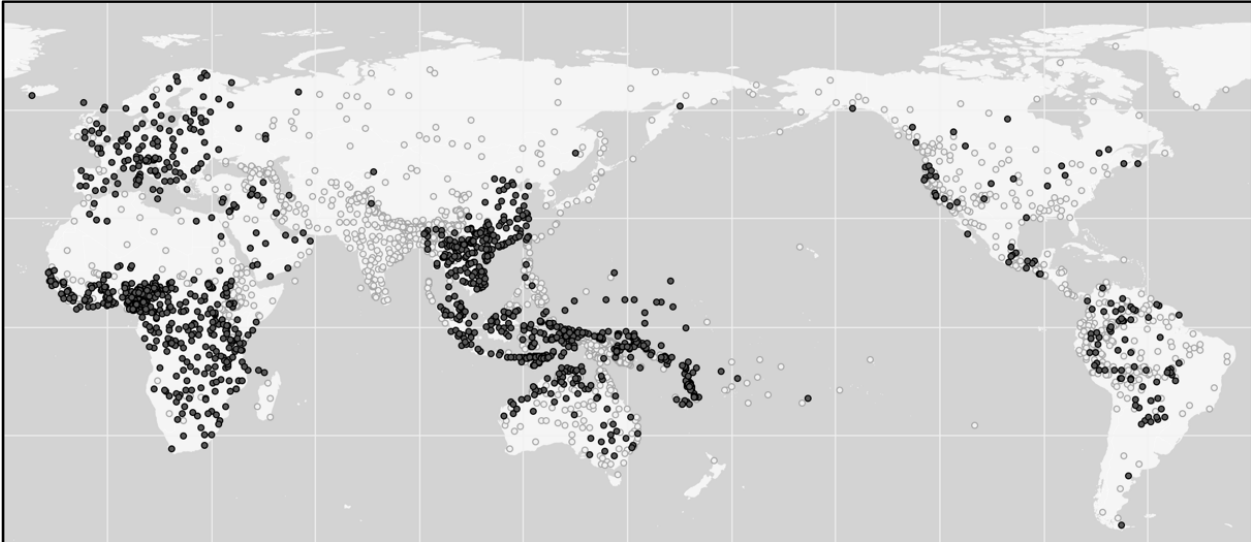
**Map A9.2.6:** Morphological causatives.



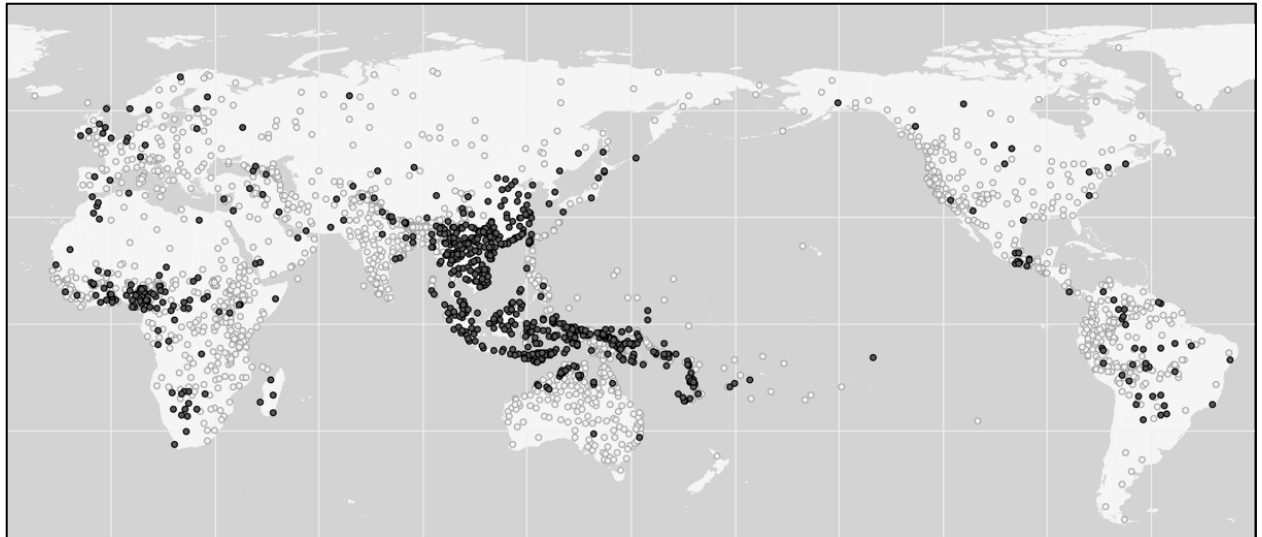
**Map A9.2.7:** Possessive prefixes on nouns.



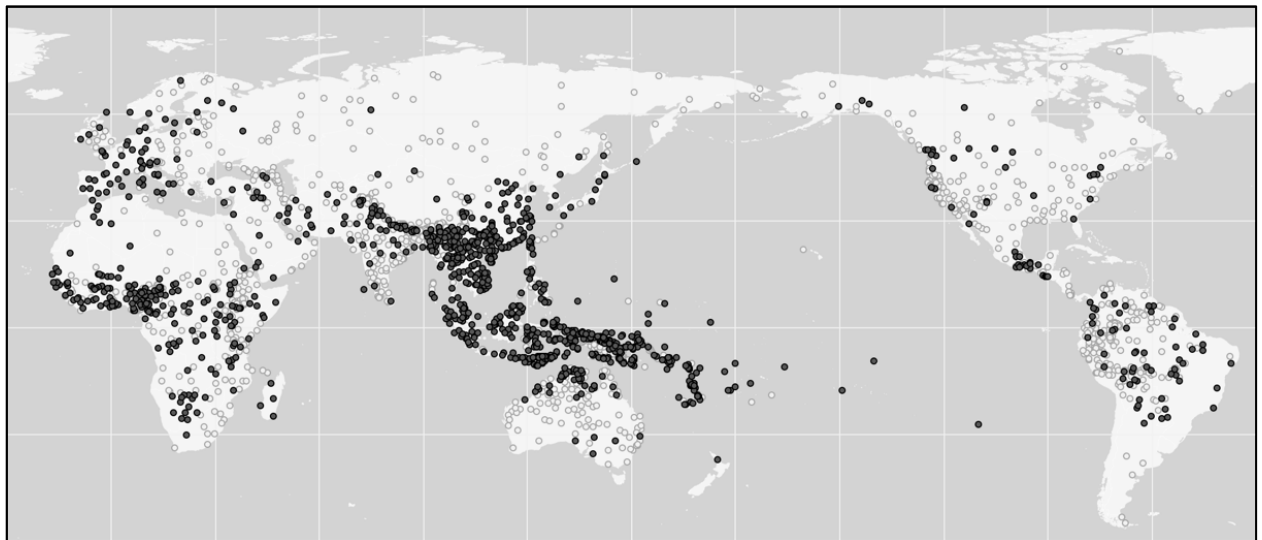
Map A9.2.8: Total tense distinctions.



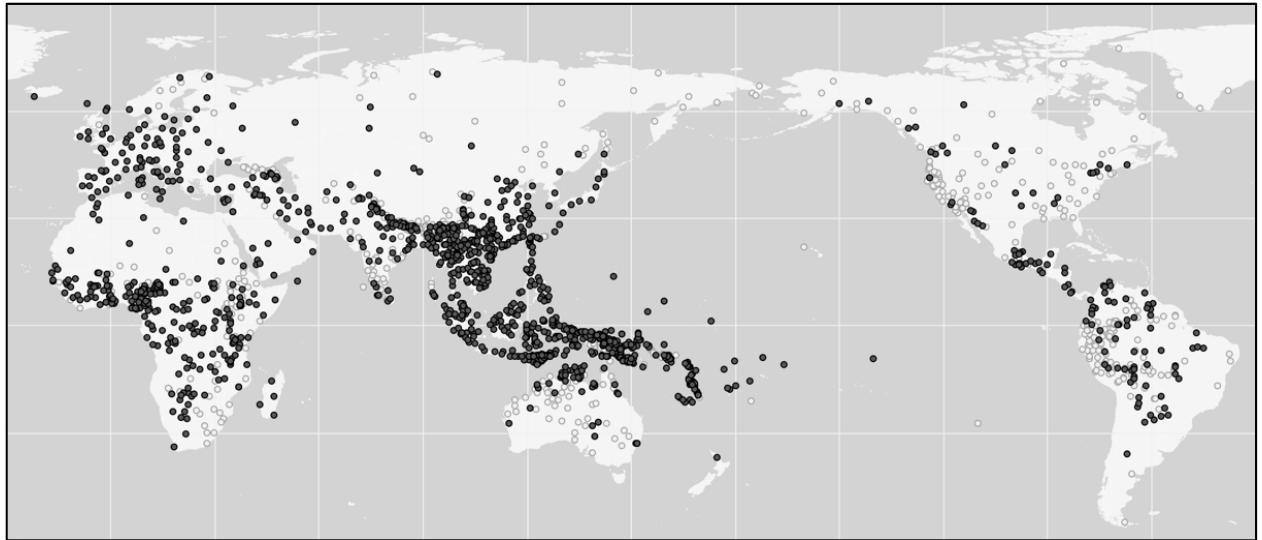
Map A9.2.9: SVO order.



**Map A9.2.10:** Symmetrical clauses: Purpose.



**Map A9.2.11:** Symmetrical clauses: Temporal.



Map A9.2.12: Symmetrical clauses: Reason.

### ***A9.3 Features associated with Dimension 3***

Dimension plots of features with strong associations with Dimension 3.  
Pairs of dimensions listed for each chart.

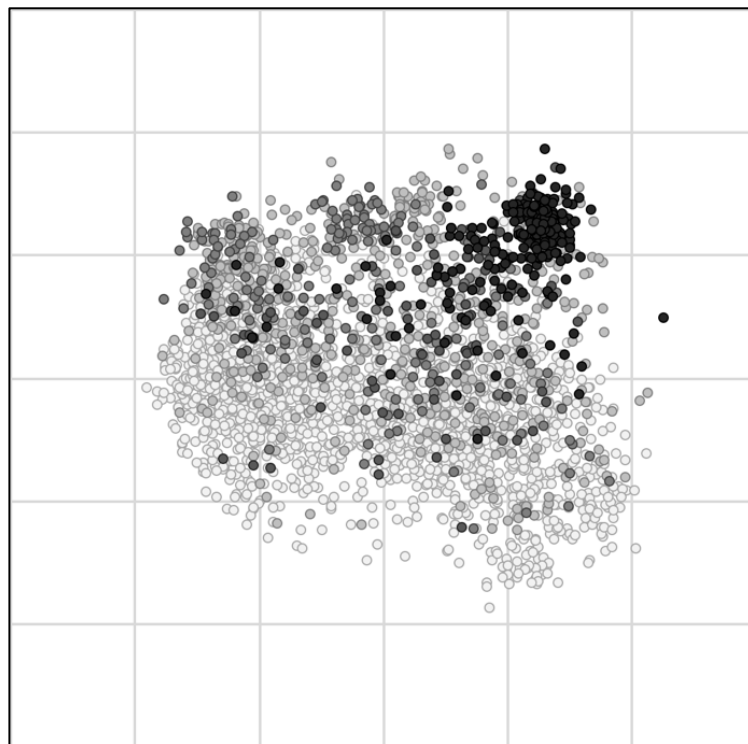
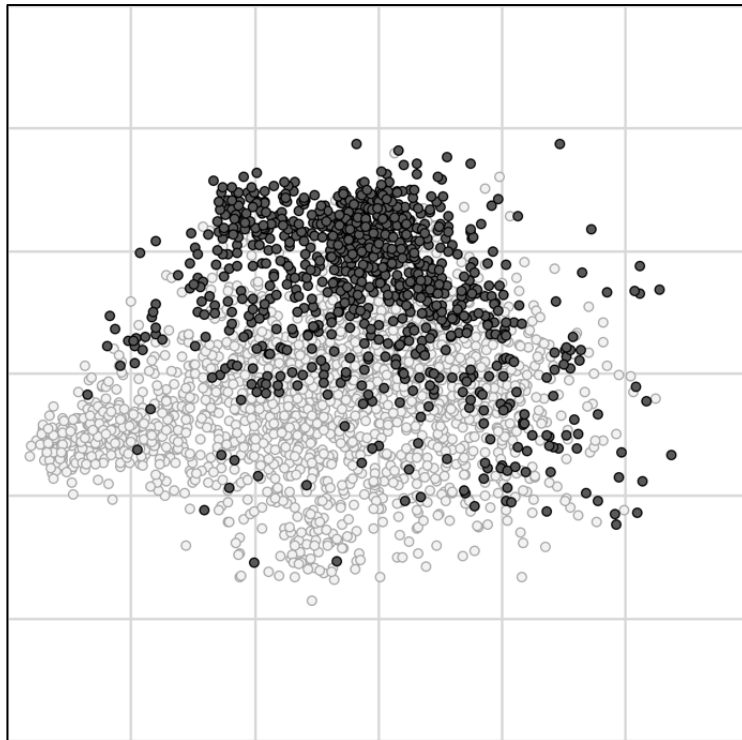
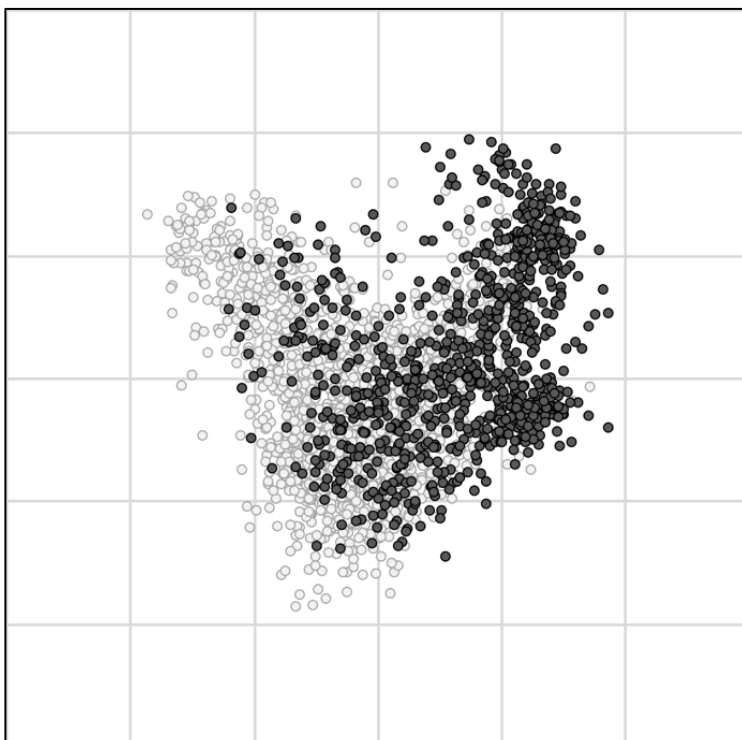


Figure A9.3.1: Genders in Dimensions 1 and 3 ( $r^2 = 0.28$ ).



**Figure A9.3.2:** Obligatory plural marking on nouns in Dimensions 2 and 3 ( $r^2 = 0.27$ ).



**Figure A9.3.3:** 3SG pronominal gender in Dimensions 3 and 4 ( $r^2 = 0.25$ ).

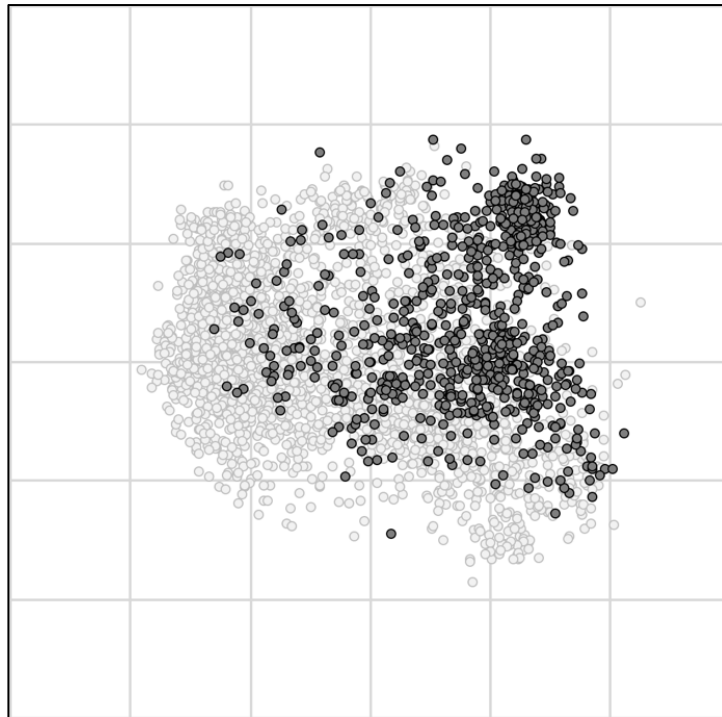


Figure A9.3.4: Accusative verb alignment in Dimensions 1 and 3 ( $r^2 = 0.22$ ).

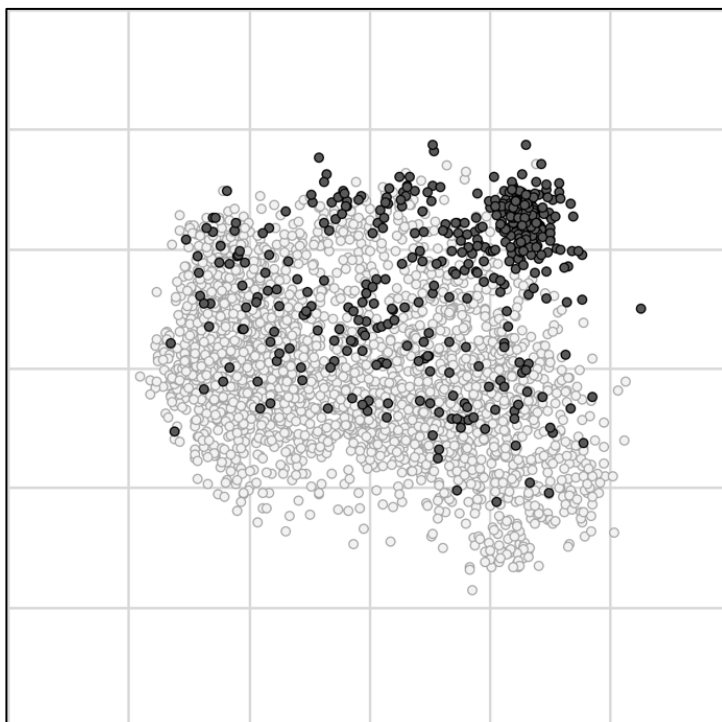
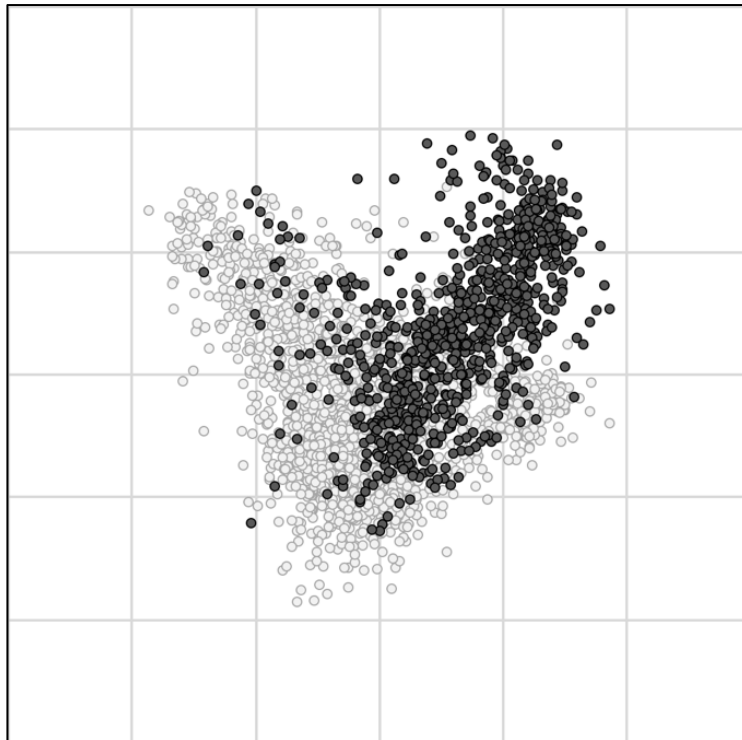
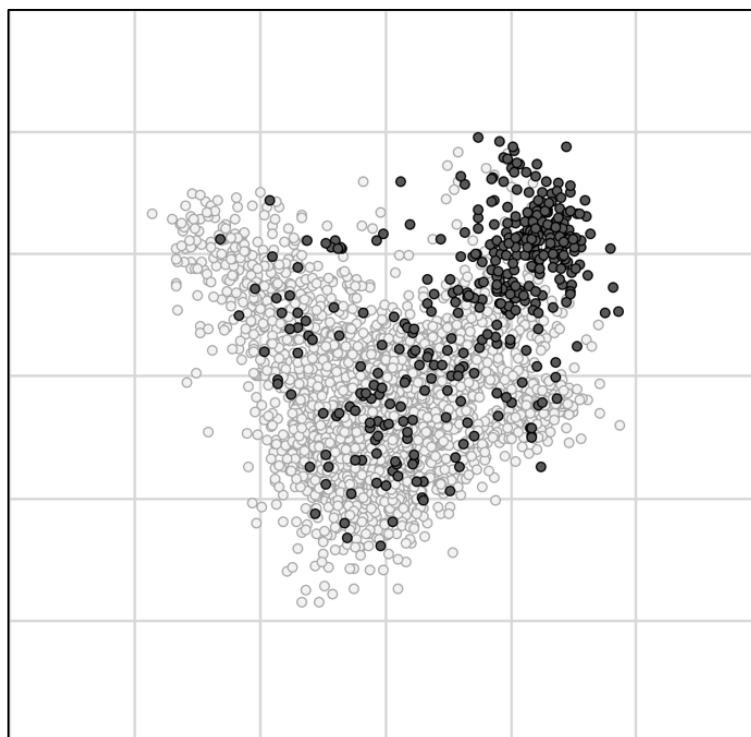


Figure A9.3.5: 3PL pronominal gender in Dimensions 1 and 3 ( $r^2 = 0.17$ ).



**Figure A9.3.6:** Suffixal subject agreement on verbs in Dimensions 3 and 4 ( $r^2 = 0.13$ ).



**Figure A9.3.7:** Relative pronouns in Dimensions 3 and 4 ( $r^2 = 0.13$ ).

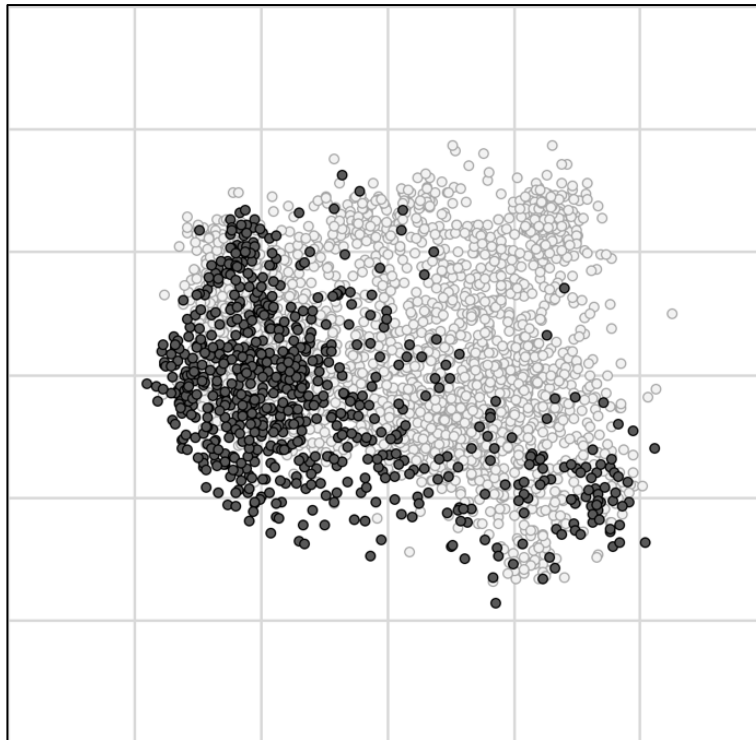


Figure A9.3.8: Ergativity in Dimensions 1 and 3 ( $r^2 = 0.13$ ).

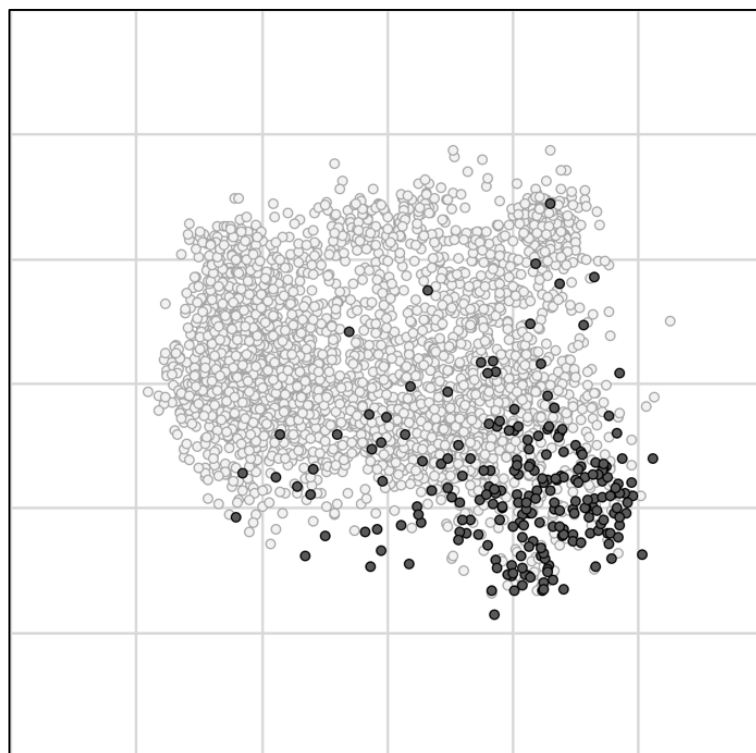
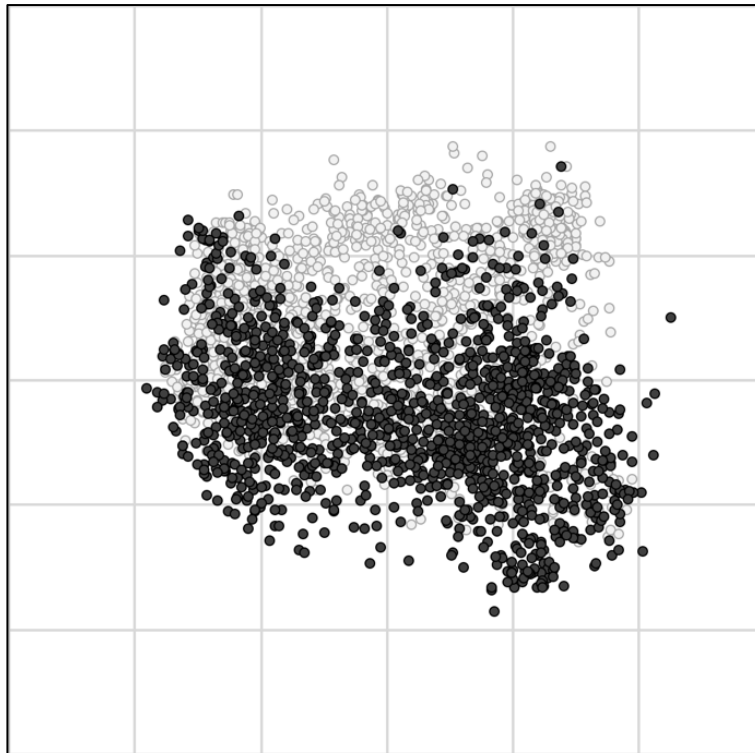
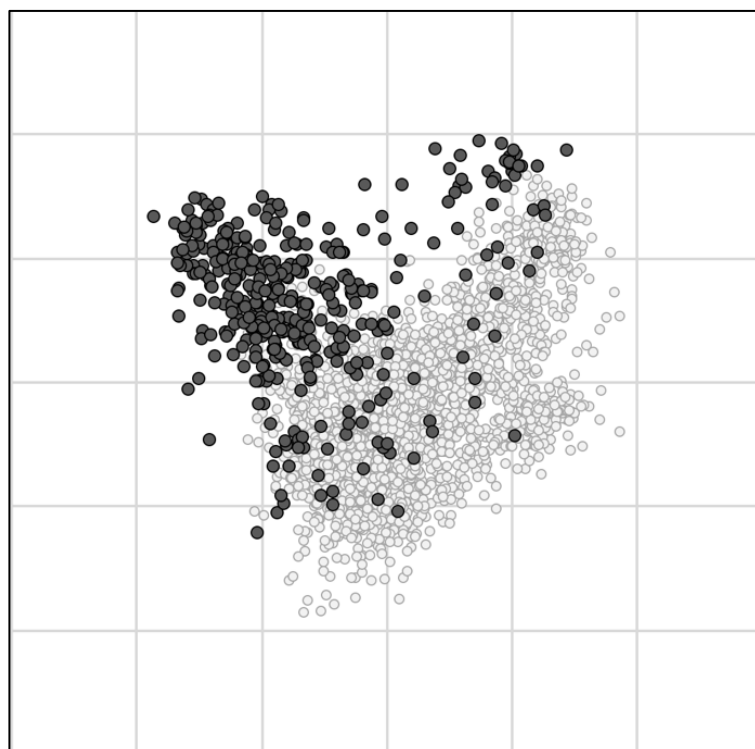


Figure A9.3.9: VOS order in Dimensions 1 and 3 ( $r^2 = 0.15$ ).



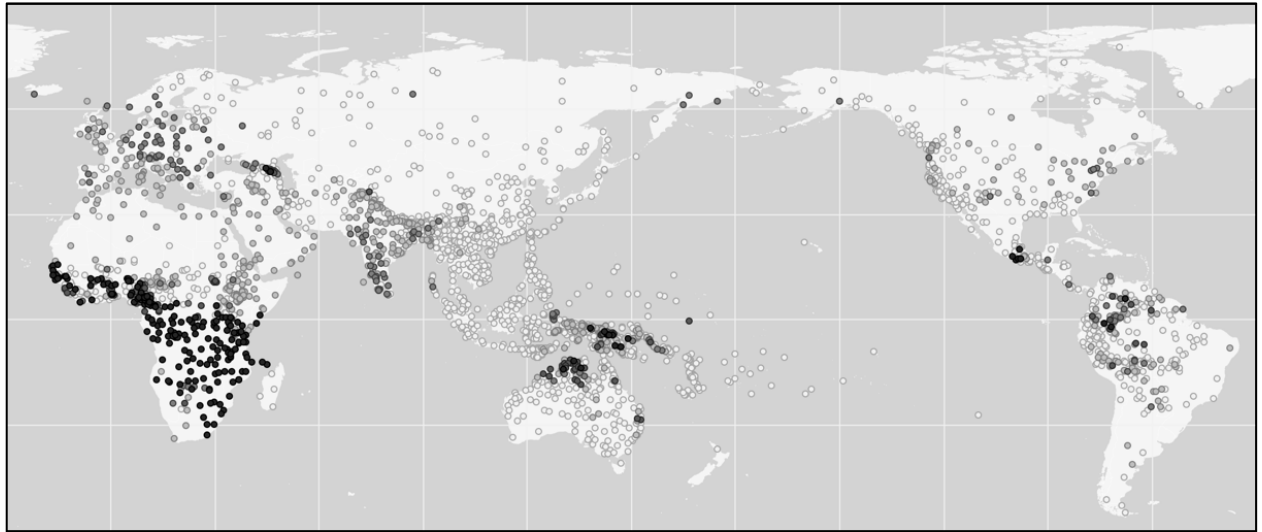


**Figure A9.3.10:** Inclusive/Exclusive contrasts in Dimensions 1 and 3 ( $r^2 = 0.18$ ).

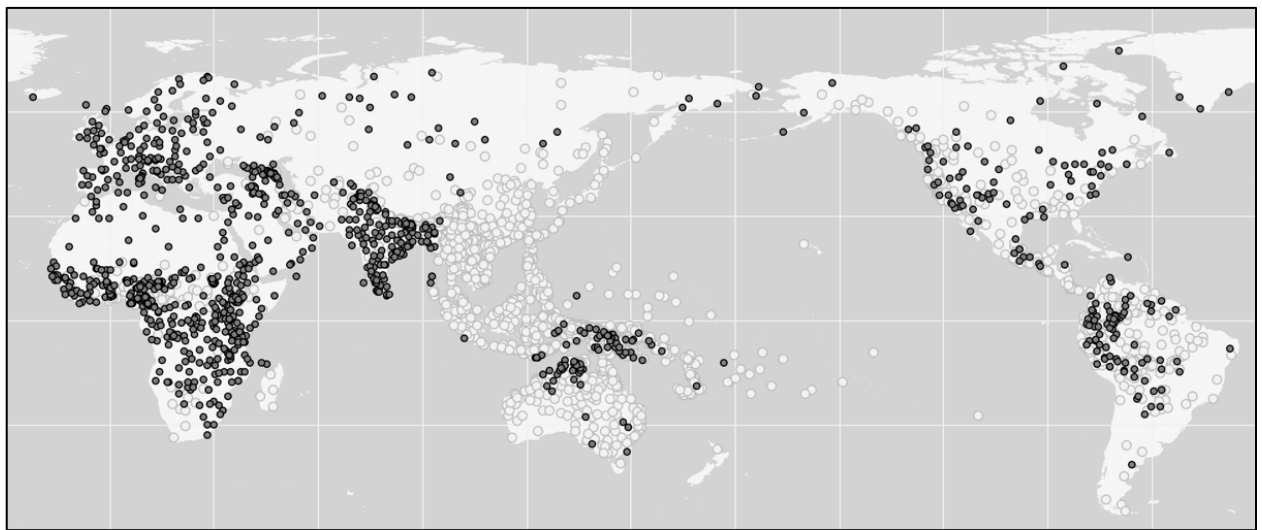


**Figure A9.3.11:** Clause-initial negation in Dimensions 3 and 4 ( $r^2 = 0.18$ ).

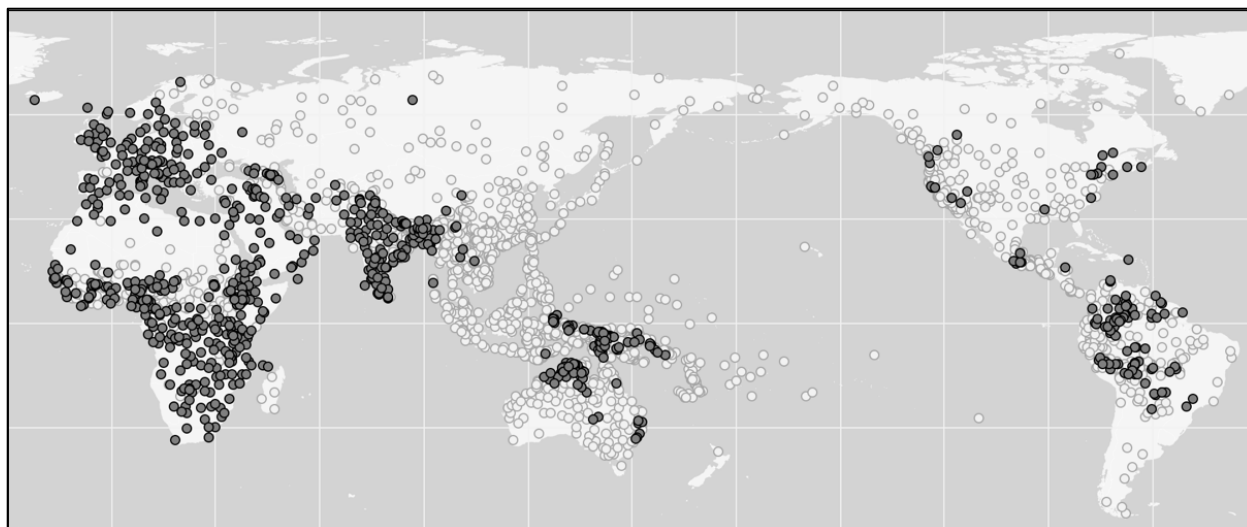
Maps of features with strong associations with Dimension 3:



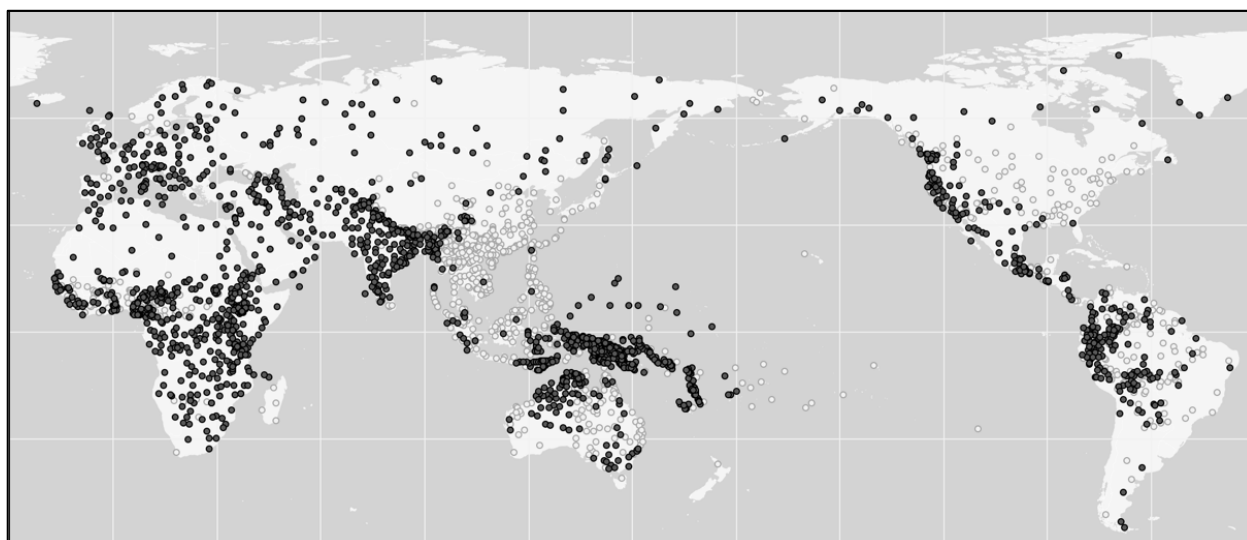
**Map A9.3.1: Genders.**



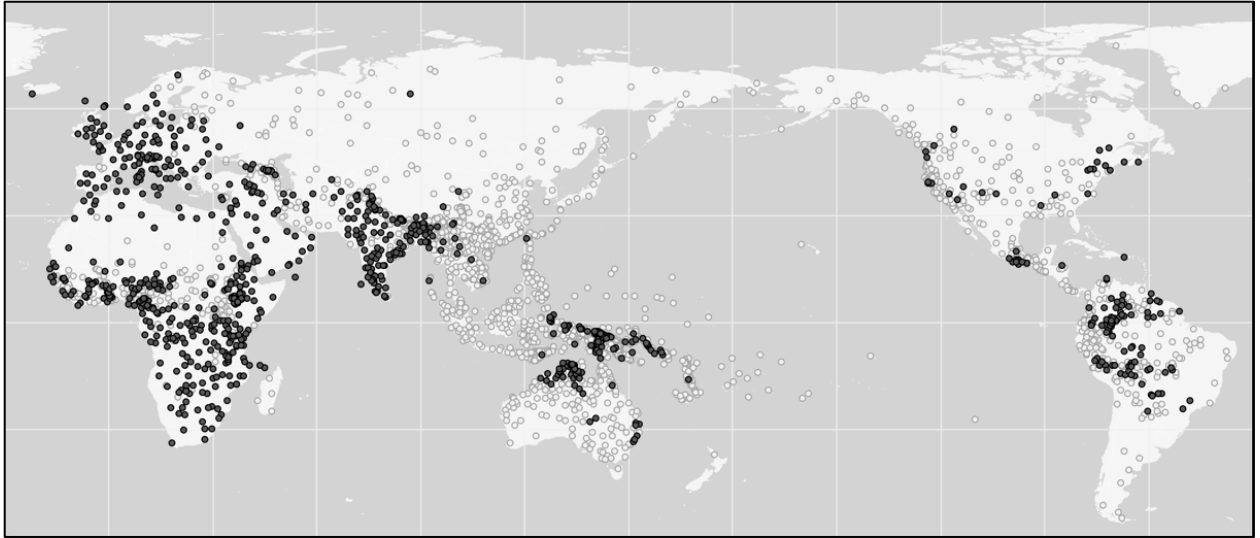
**Map A9.3.2: Obligatory plural marking on nouns.**



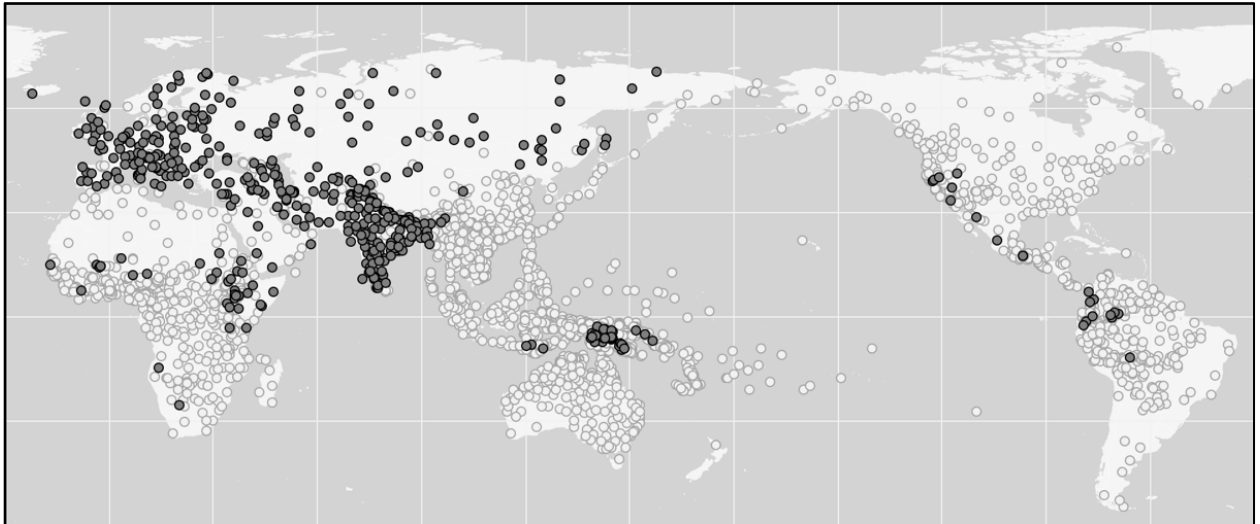
Map A9.3.3: 3SG pronominal gender.



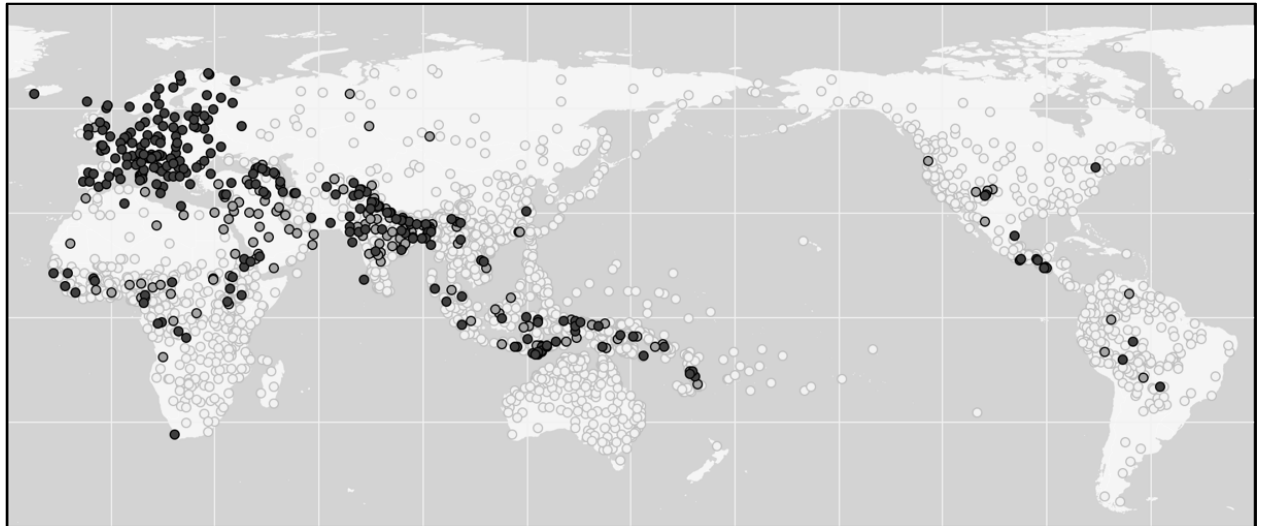
Map A9.3.4: Verb alignment: accusative.



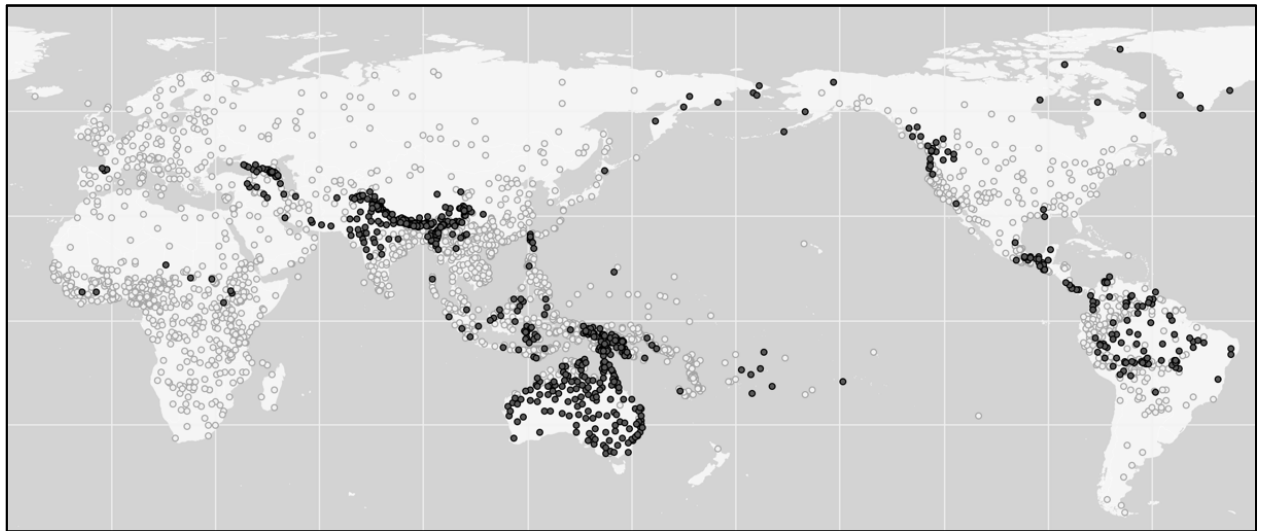
Map A9.3.5: 3PL pronominal gender.



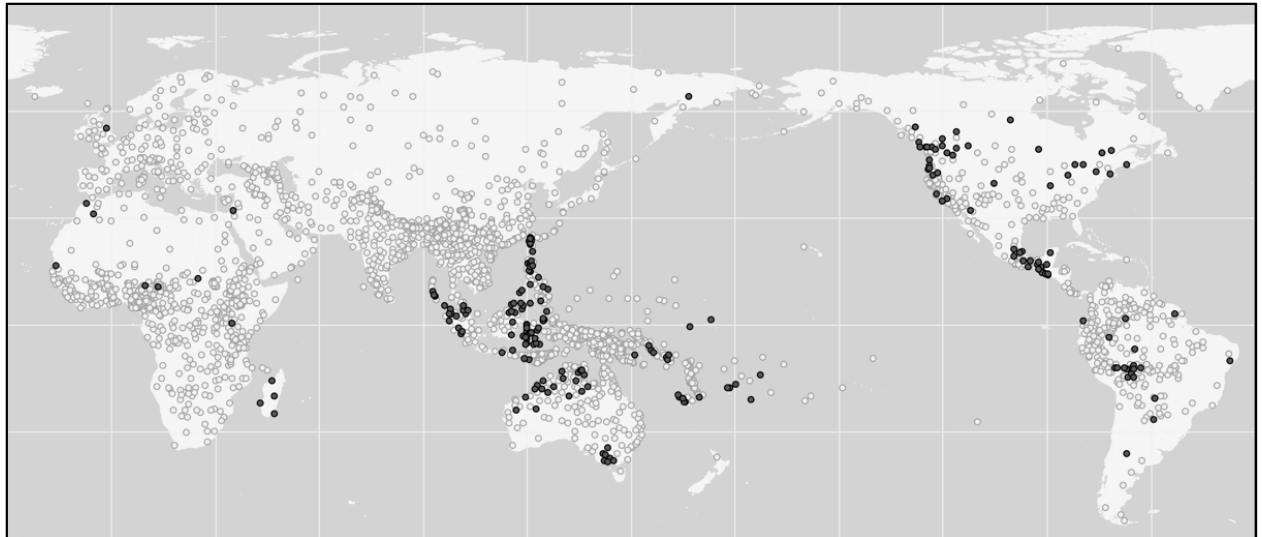
Map A9.3.6: Suffixal subject agreement on verbs.



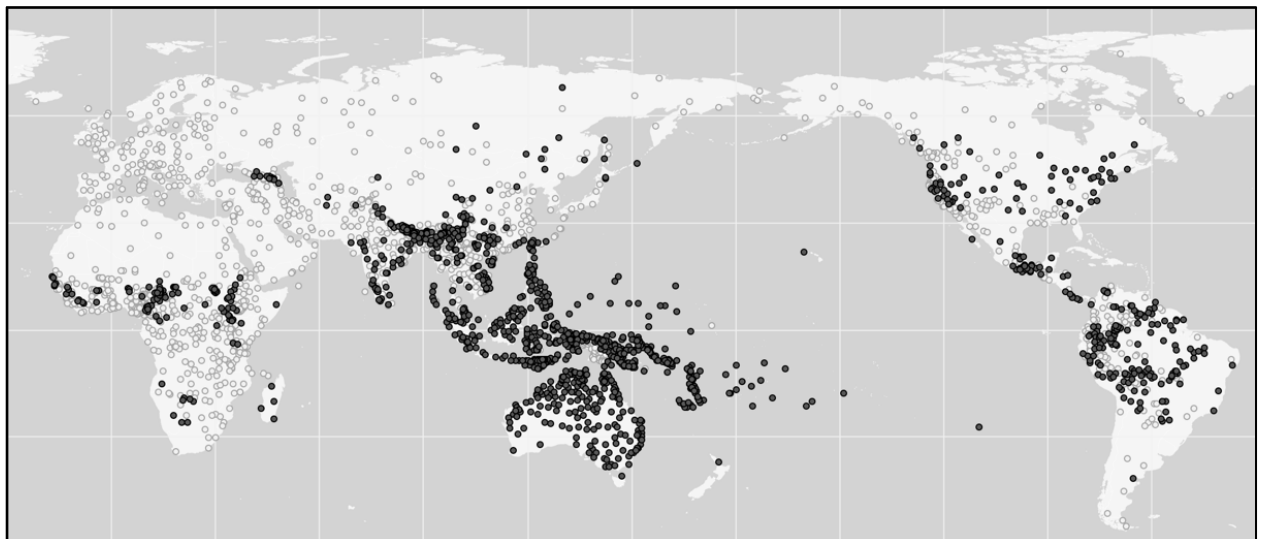
**Map A9.3.7:** Relative pronouns.



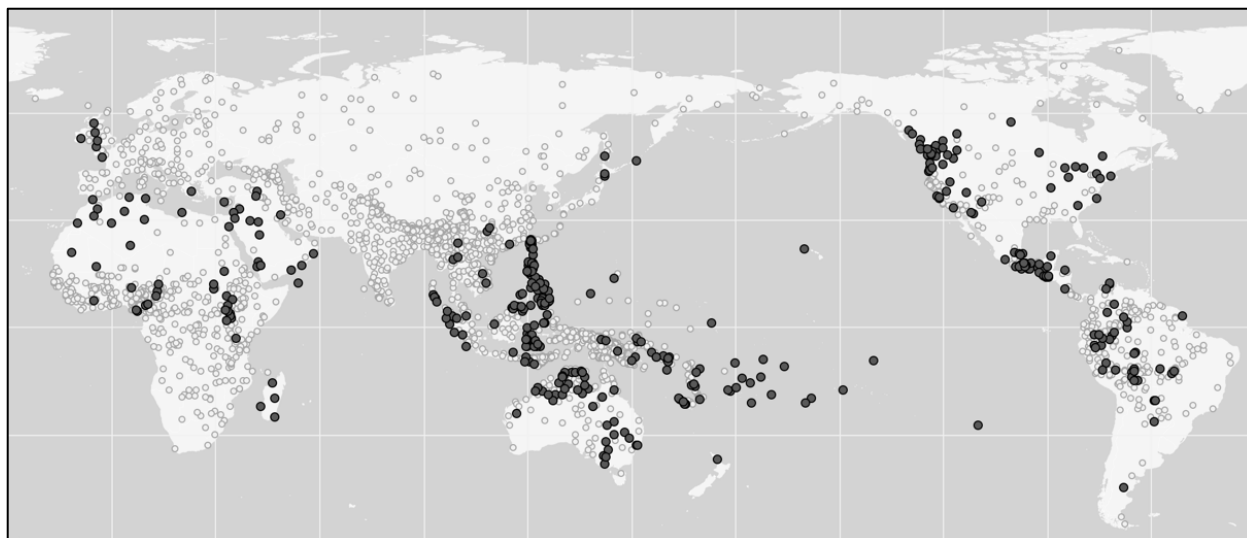
**Map A9.3.8:** Ergativity.



Map A9.3.9: VOS order.



Map A9.3.10: Inclusive/Exclusive contrasts.



Map A9.3.11: Clause-initial negation.

#### ***A9.4 Features associated with Dimension 4***

Dimension plots of features with strong associations with Dimension 4.  
Pairs of dimensions listed for each chart.

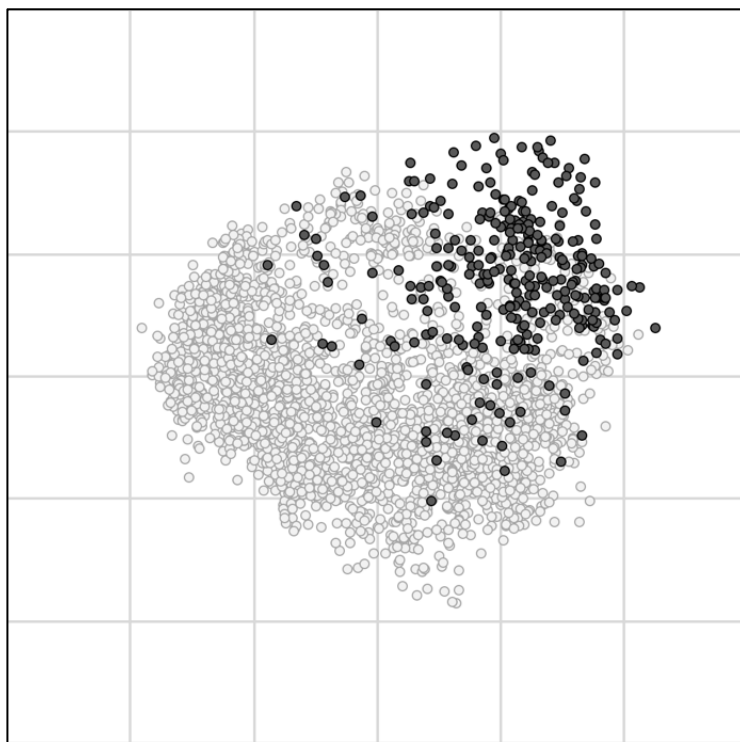
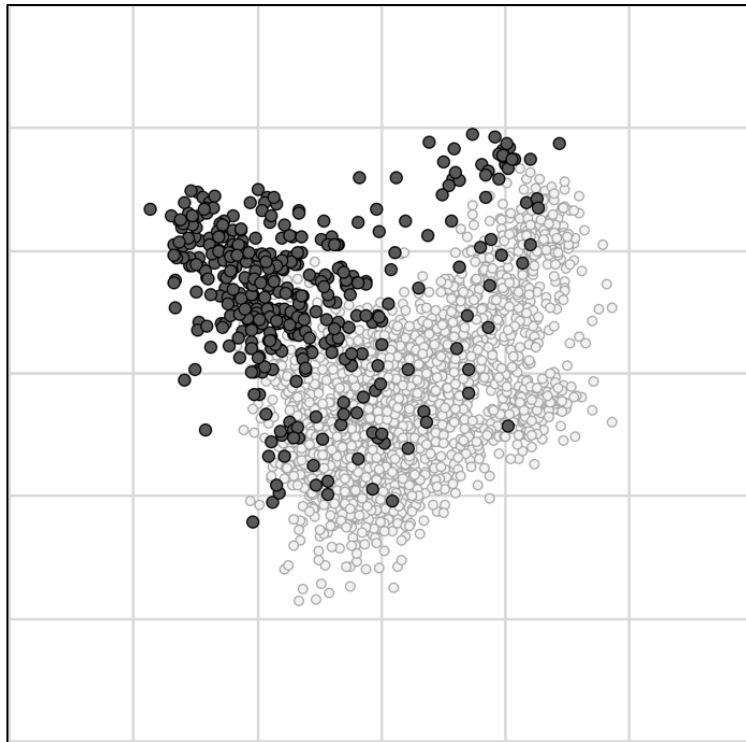
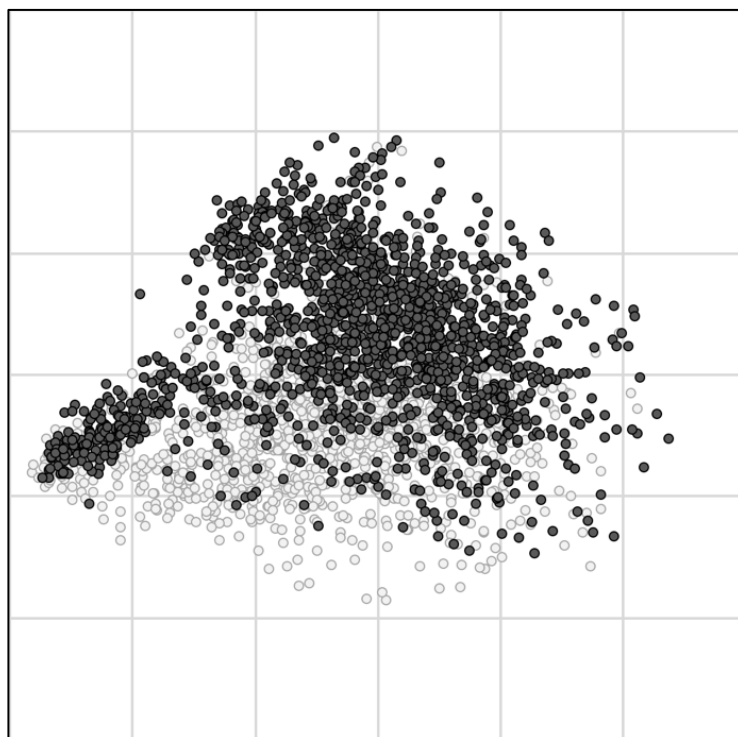


Figure A9.4.1: VSO order in Dimensions 1 and 4 ( $r^2 = 0.22$ ).



**Figure A9.4.2:** Clause-initial negation in Dimensions 3 and 4 ( $r^2 = 0.20$ )



**Figure A9.4.3:** Numeral N order in Dimensions 2 and 4 ( $r^2 = 0.19$ ).



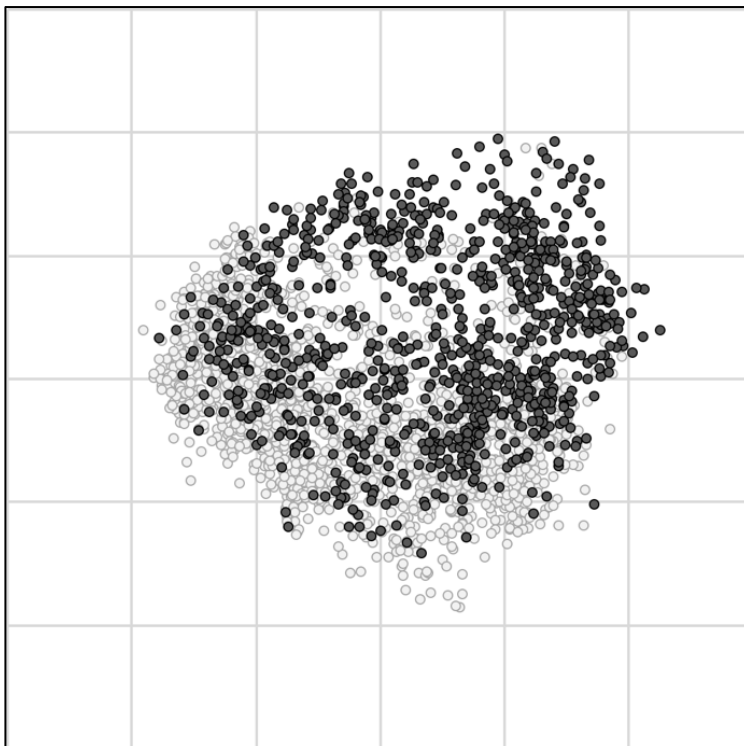


Figure A9.4.4: Clause-initial Wh in Dimensions 1 and 4 ( $r^2 = 0.11$ ).

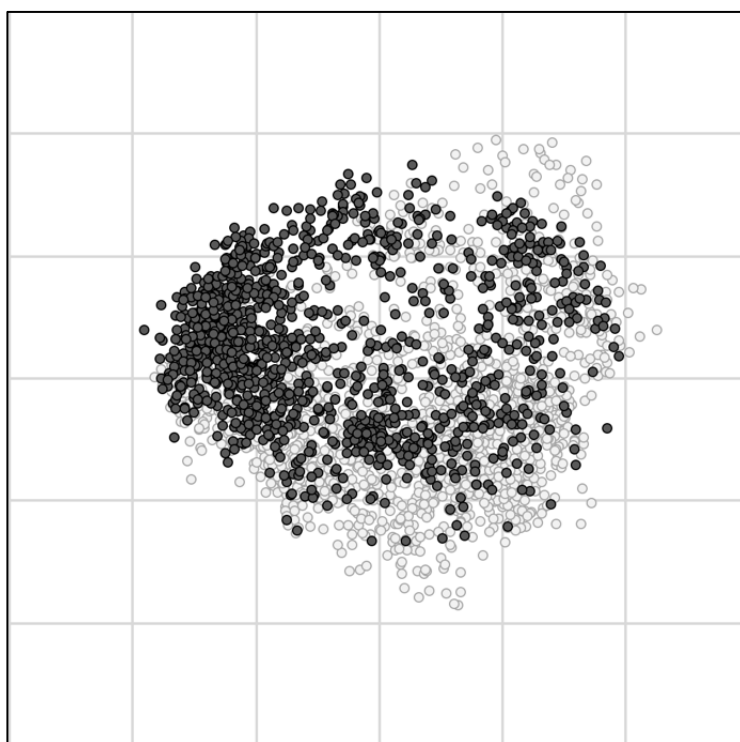


Figure A9.4.5: Adjective N order in Dimensions 1 and 4 ( $r^2 = 0.11$ ).

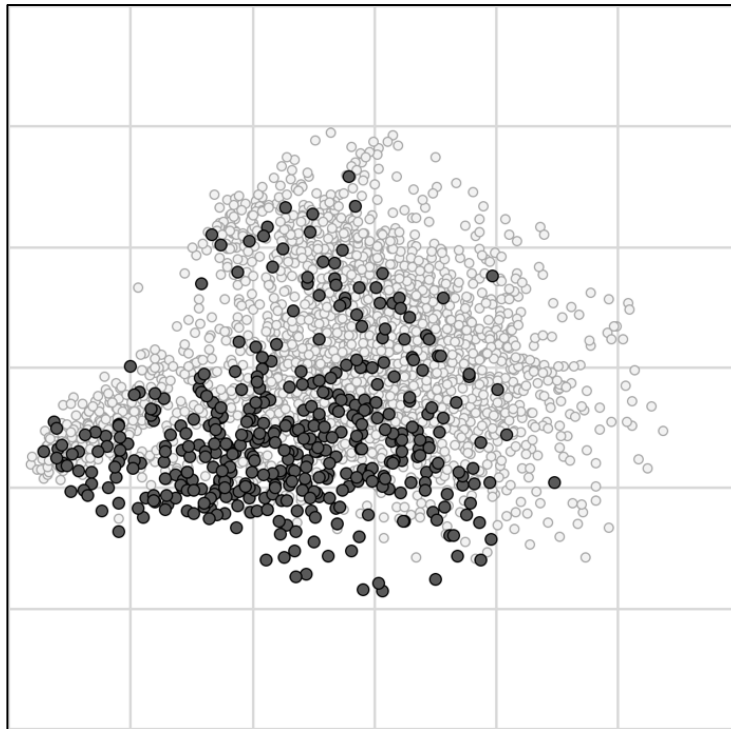


Figure A9.4.6: Clause-final negation in Dimensions 2 and 4 ( $r^2 = 0.10$ ).

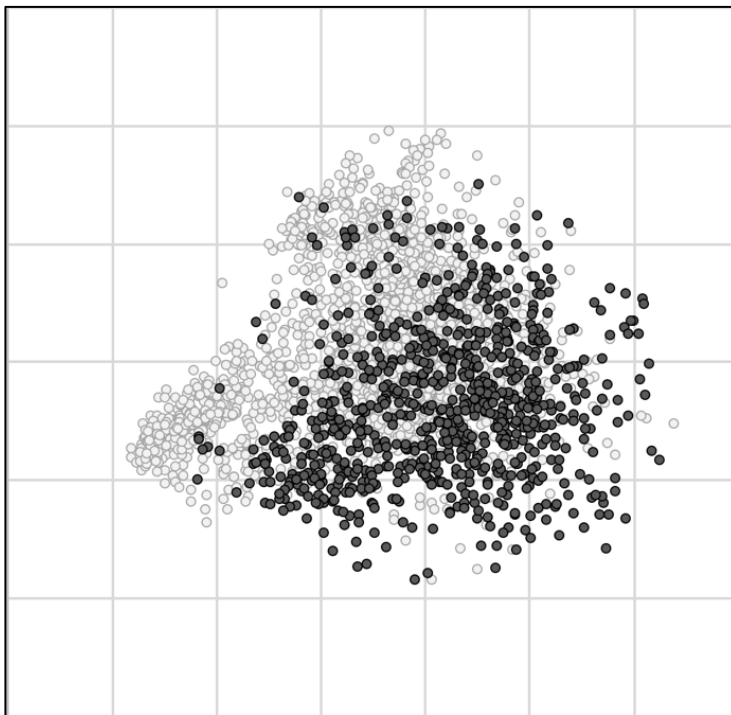
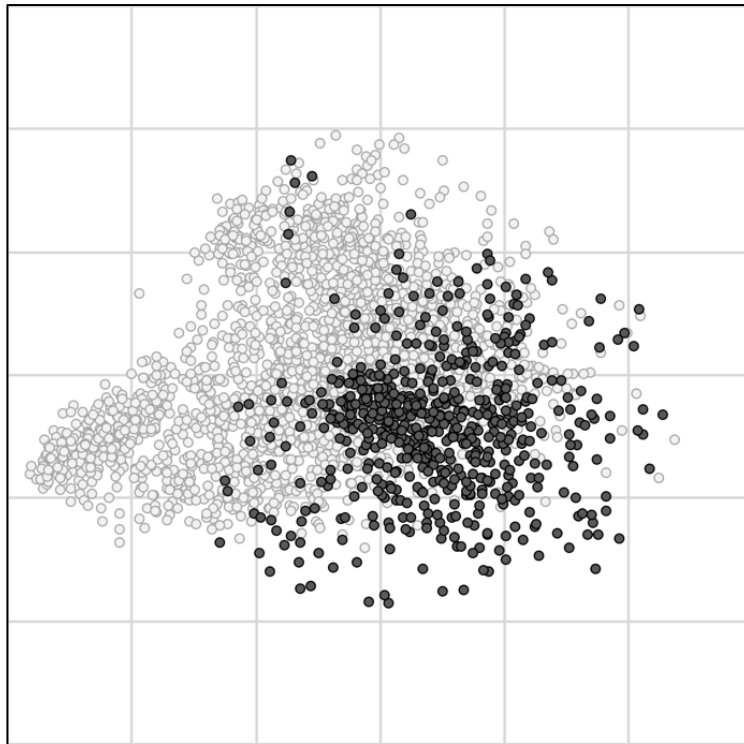
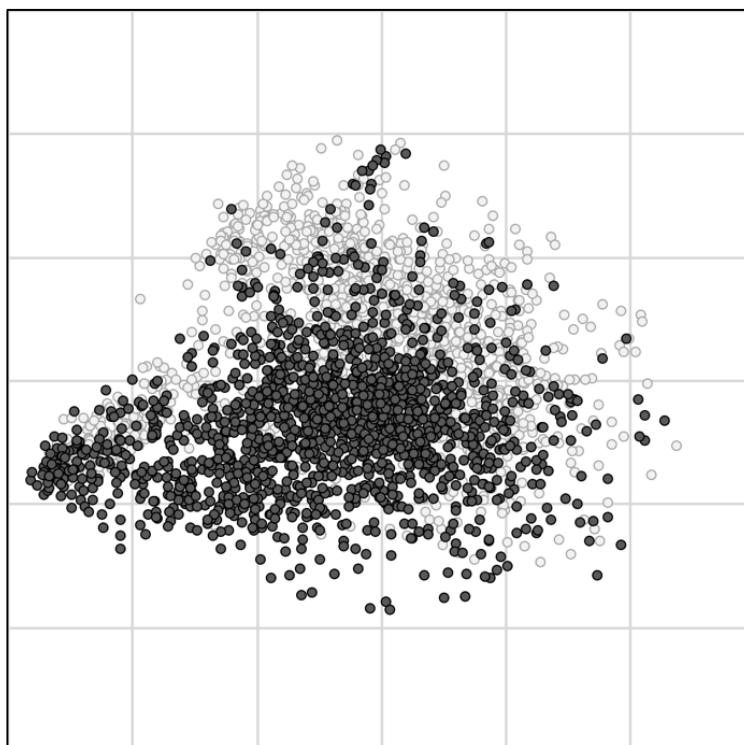


Figure A9.4.7: Inalienable possession in Dimensions 2 and 4 ( $r^2 = 0.10$ ).



**Figure A9.4.8:** Object agreement prefixes in Dimensions 2 and 4 ( $r^2 = 0.14$ ).



**Figure A9.4.9:** N Numeral order in Dimensions 2 and 4 ( $r^2 = 0.14$ ).

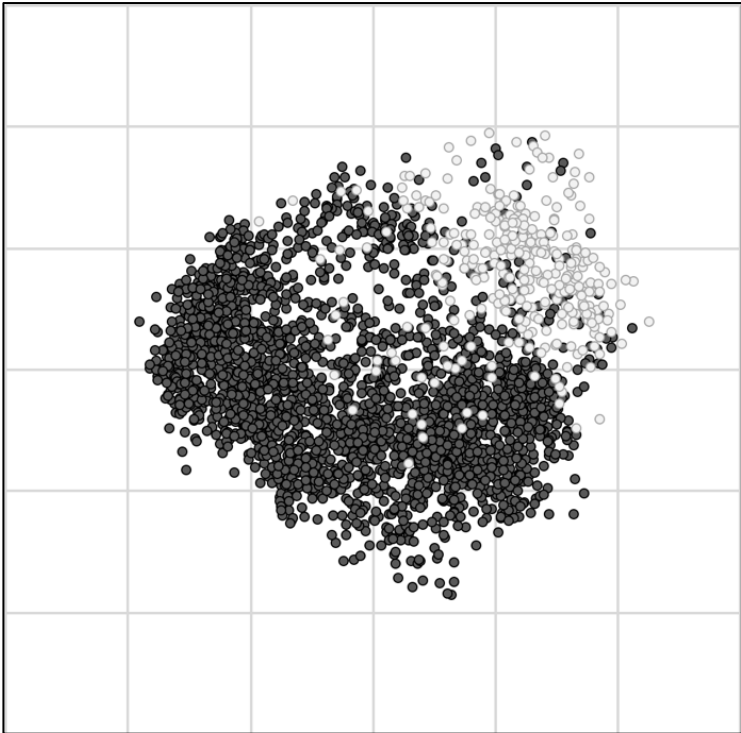
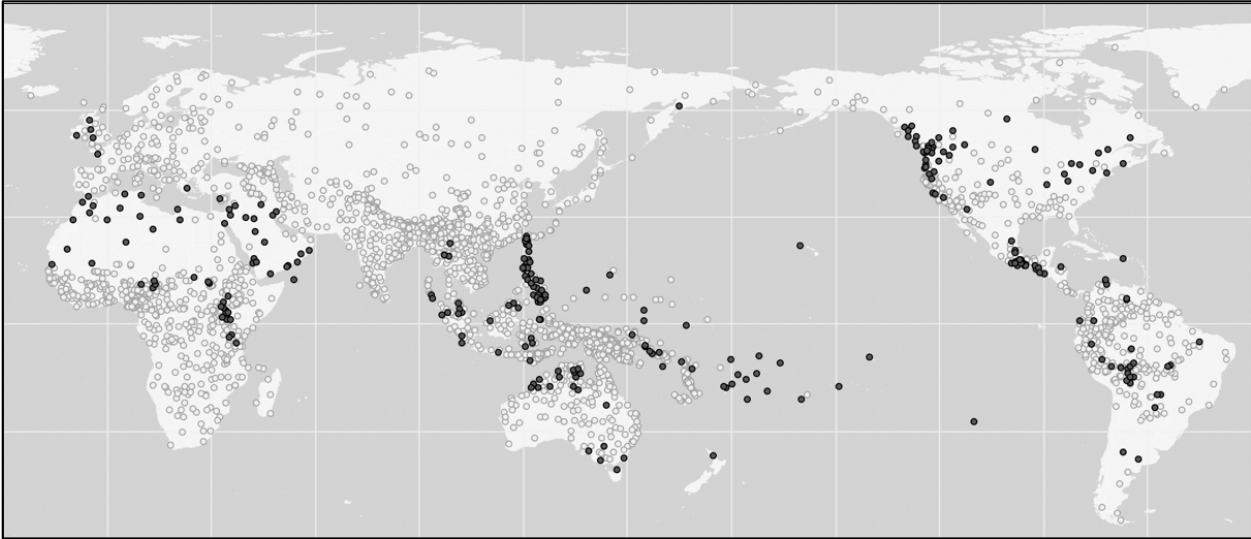
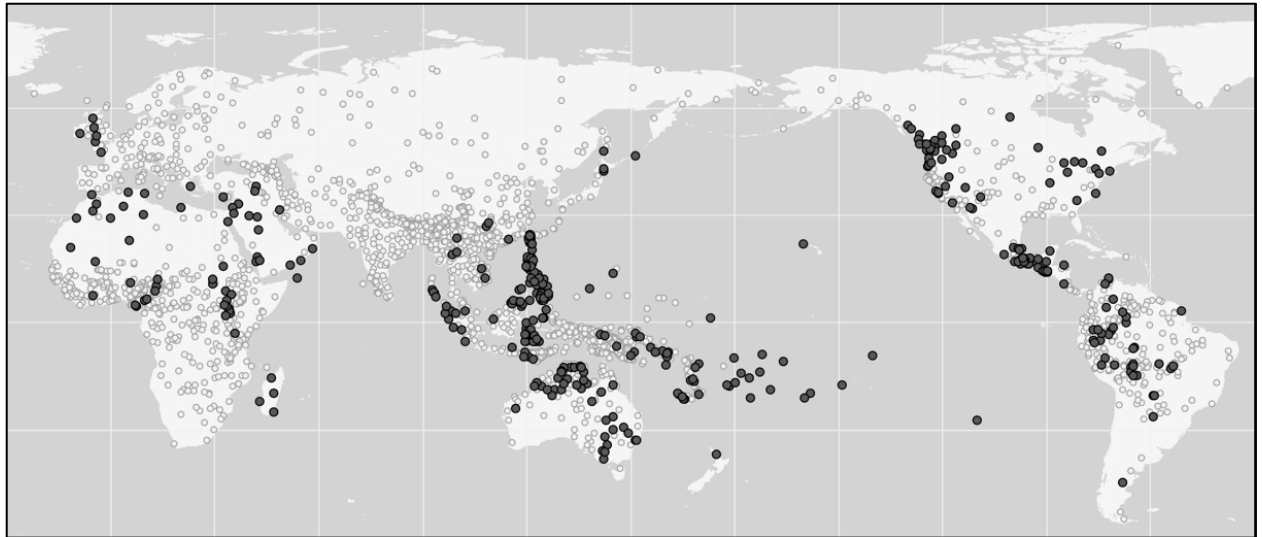


Figure A9.4.10: SV order in Dimensions 1 and 4 ( $r^2 = 0.21$ ).

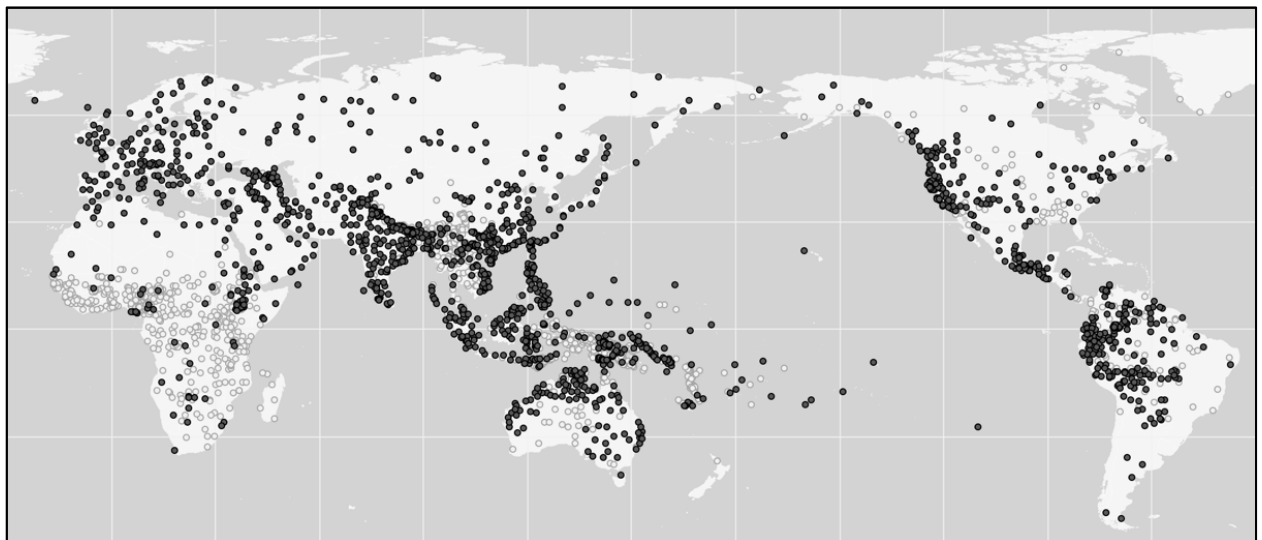
Maps of features with strong associations with Dimension 4.



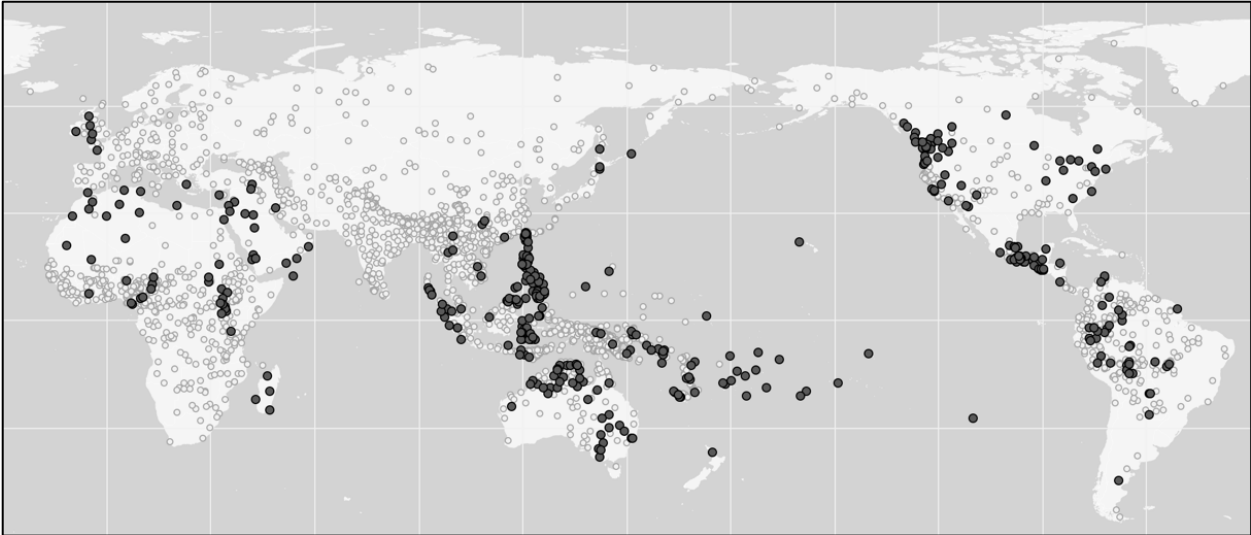
Map A9.4.1: VSO order.



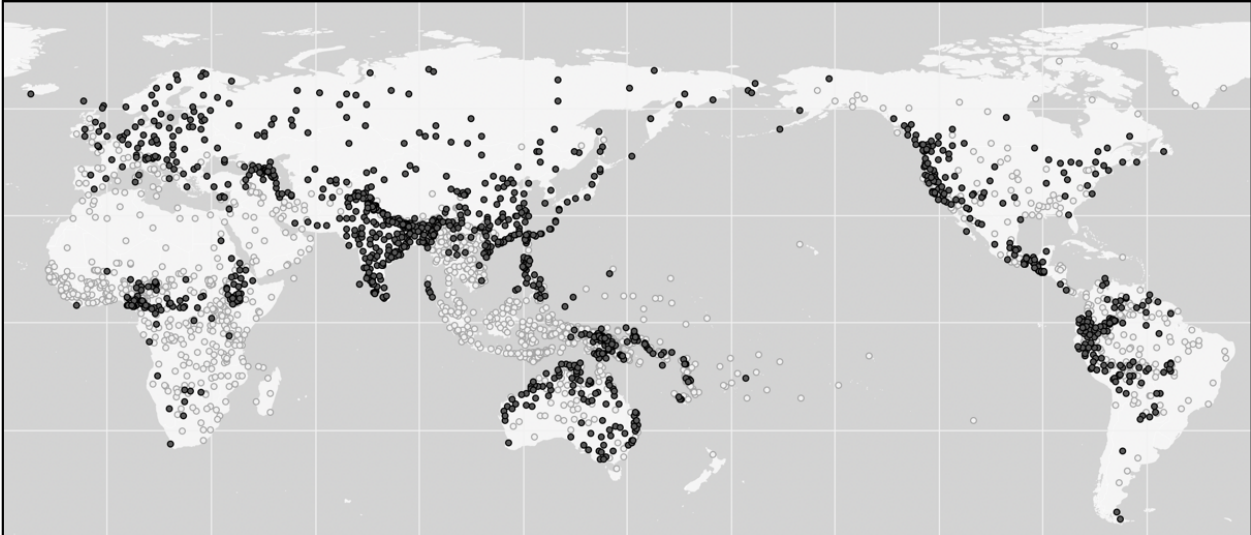
Map A9.4.2: Clause-initial negation.



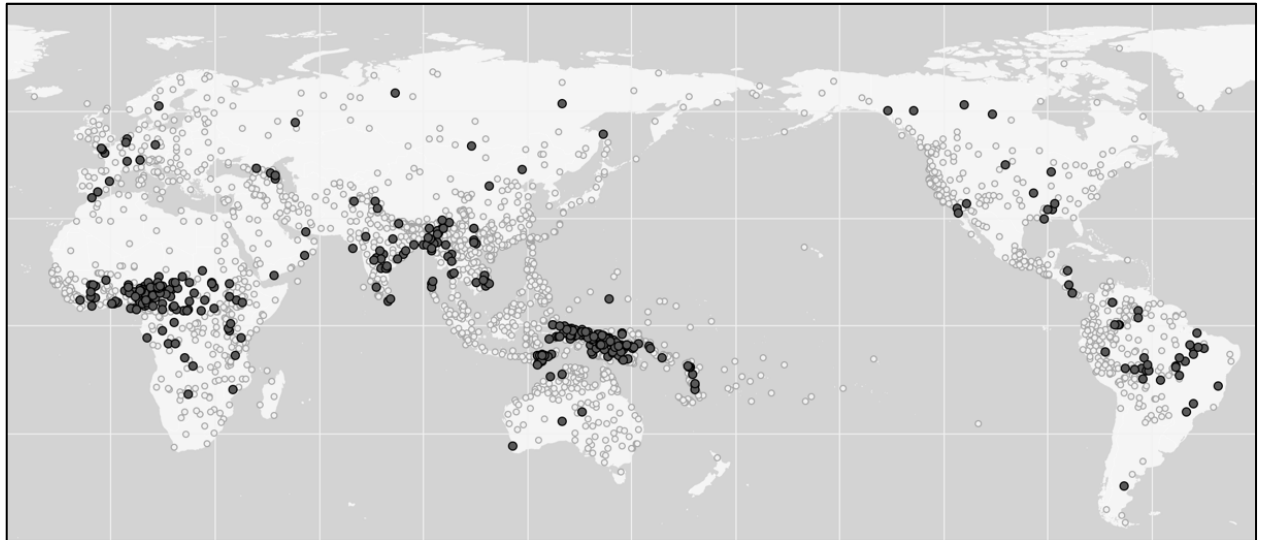
Map A9.4.3: Numeral N order.



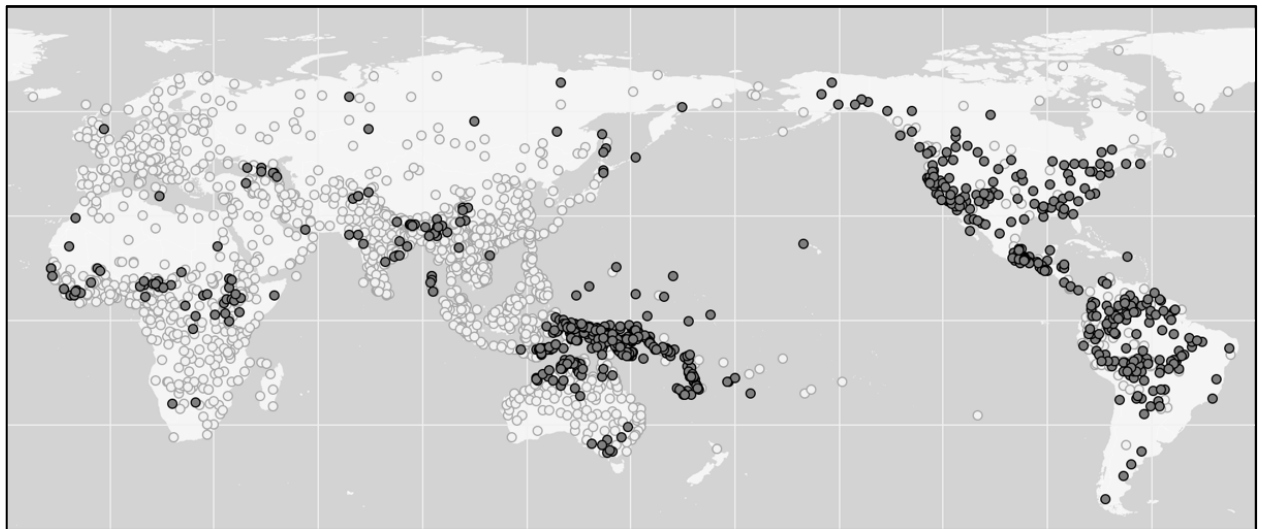
Map A9.4.4: Clause-initial Wh.



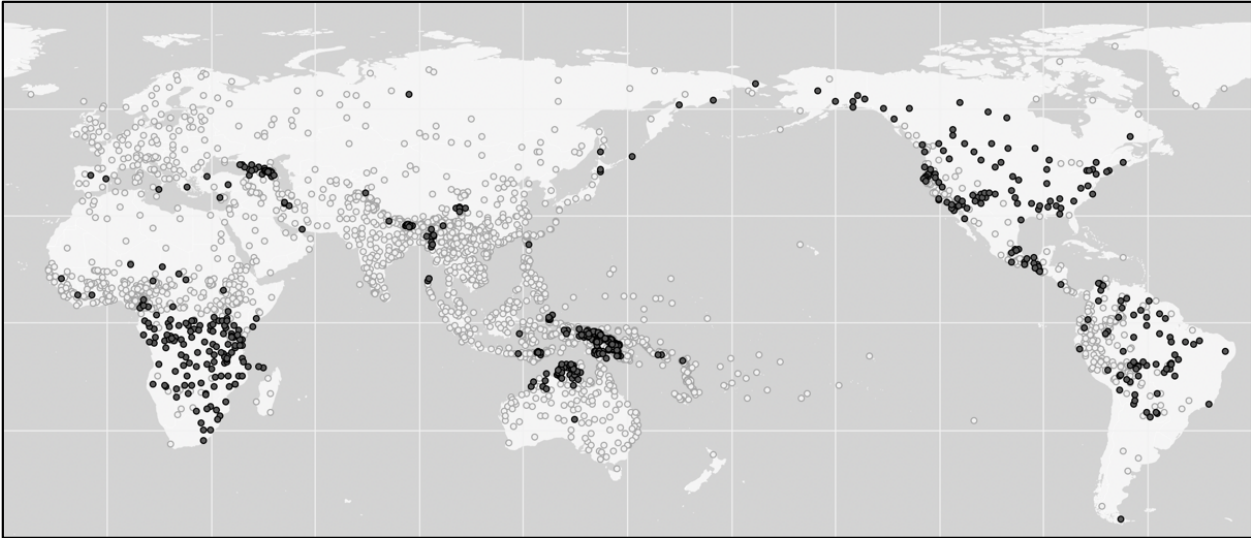
Map A9.4.5: Adjective N order.



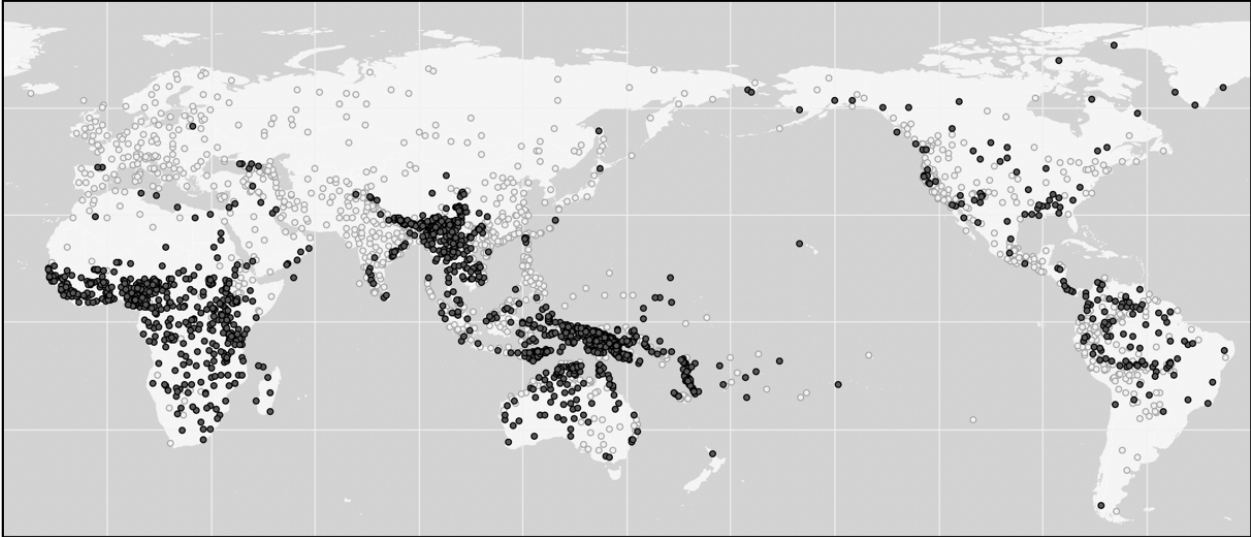
**Map A9.4.6:** Clause-final negation.



**Map A9.4.7:** Inalienable possession.

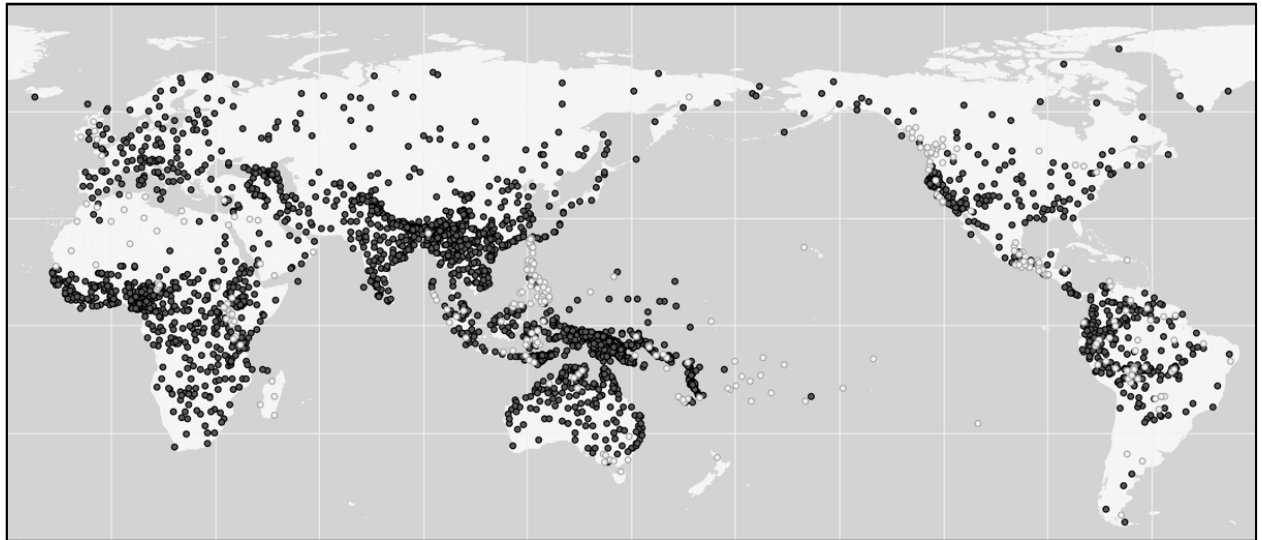


Map A9.4.8: Object agreement prefix.



Map A9.4.9: N Numeral order.





Map A9.4.10: SV order.

## Appendix 10. Source code

The source code used can be found at <https://osf.io/u9qbe/>.

# Towards a typology of continuative expressions

ANASTASIA PANOVA

STOCKHOLM UNIVERSITY

Submitted: 29/01/2023 Revised version: 22/07/2023

Accepted: 03/08/2023 Published: 27/12/2023



Articles are published under a Creative Commons Attribution 4.0 International License (The authors remain the copyright holders and grant third parties the right to use, reproduce, and share the article).

## Abstract

This paper investigates the cross-linguistic diversity of continuative ('still') expressions. Based on a genealogically stratified sample of 120 languages, the continuative expressions are systematically analyzed according to the four following parameters: morphosyntactic type, emphatic vs. non-emphatic status, other (non-continuative) uses and semantic effects when combined with negation. The study shows that the most widespread type of continuative expressions is represented by monosemous emphatic continuative adverbials which in combination with negation acquire a 'not yet' meaning. In many languages, however, we also find continuative expressions which have followed evolutionary pathways towards morphologization, non-emphatic uses, rich polysemy networks, and less trivial types of interaction with negation. The paper discusses possible areal, genealogical and structural factors which might contribute to the "maturation" of continuative expressions in the world's languages.

**Keywords:** continuative; phasal polarity; typology; polysemy; maturation.

## 1. Introduction

This article is a study of linguistic encoding of the semantics of continuation: a domain that has not yet been a topic of a dedicated typological investigation. The term *continuative* covers expressions used in reference to the situations which started to exist before the reference time and exist at reference time (for a more detailed

definition and discussion, see Section 2.1). Examples of such expressions, hereinafter referred to as *continuative expressions*, are given in (1)-(4). It can be seen that continuative expressions vary greatly with respect to their morphosyntactic status.

- (1) Spanish (spa; Indo-European, Italic; van der Auwera 1998: 30)<sup>1</sup>

*Juan duerme todavía.*

Juan sleeps still

‘Juan is still asleep.’

- (2) Balanta-Ganja (bjt; Atlantic-Congo, North-Central Atlantic; Creissels and Biaye 2016: 201)

*bá-n-tígtà-nà yâaθ.*

INCL-INACP-AUX<sub>CONT</sub>-INCL work

‘We keep working.’

- (3) Yine (pib; Arawakan, Southern Maipuran; Hanson 2010: 245; glosses adapted)

*r-halna-wa*

3-fly-IPFV

‘He is still flying.’ / ‘He continues flying.’

- (4) Nanga (nzz; Dogon, Nangan Dogon; Heath 2016: 226)

[*nújèy<sup>n</sup> yɲà*] [ò:<sup>L</sup> gó] bù-∅

[now INST] [field<sup>L</sup> LOC] be-3SG.SBJ

‘He/She is still in the fields.’

In the late 1990s, several studies dealt with continuative expressions from a cross-linguistic perspective within a broader semantic domain — phasality or phasal polarity (van Baar 1997; van der Auwera 1998; Plungian 1999) — which, apart from the continuative (~ ‘still’), also includes ‘already’, ‘no longer’ and ‘not yet’. More recent studies have focused on specific phasal meanings. For example, Veselinova (2015) addresses ‘not yet’ expressions in the languages of the world. Dahl and Wälchli (2016) investigate the iimitive (‘already’) meaning. The continuative meaning,

<sup>1</sup> Throughout the article the language names are provided according to Glottolog 4.8 (Hammarström et al. 2021). Transcription and glosses in the linguistic examples are provided as in the sources unless otherwise stated.

however, has never been a topic of a dedicated large-sample typological investigation. The present study is intended to fill this gap.

The research questions addressed in this paper deal with the problem of structural diversity and linguistic preferences. The specific goal of the study is to describe the cross-linguistic variation in the properties of continuative expressions, and what properties of continuative expressions are more or less typical of the world's languages. With respect to these questions, the continuative semantics is particularly interesting because it can be expressed by both lexical items and grammatical markers. Thus, the challenge of this study is to conduct a consistent cross-linguistic analysis of the highly diverse class of linguistic expressions combining methods of both lexical and grammatical typology.

The structure of the paper is as follows. Section 2 discusses the relevant theoretical concepts and the methodology used in this study. Section 3 is a detailed description of the analysis of the continuative expressions along four parameters: morphosyntactic type, emphatic vs. non-emphatic status, uses outside the continuative domain and semantic effects when combined with negation. Section 4 provides a comprehensive account of the typology of continuative expressions from a diachronic perspective. Finally, Section 5 discusses the main findings of the study.

## **2. Theoretical and methodological preliminaries**

### ***2.1. The continuative meaning: a definition***

According to van Baar (1997), van der Auwera (1998), Plungian (1999) among others, the continuative meaning belongs to the phasal domain (also known as phasality or phasal polarity). The phasal domain consists of the four values: 'already', 'still', 'no longer' and 'not yet'. As shown in Table 1, phasal markers denote "existence or non-existence of a situation at several moments, as compared to some other moments" (Plungian 1999: 315). For example, 'already' indicates that the situation existed (+) at the reference time and that it did not exist (-) at some moment preceding the reference time.

$t_i$ (preceding moment)	$t_0$ (reference time)	meaning	van der Auwera 1998	Plungian 1999
-	+	'already'	inchoative	inchoative
+	+	'still'	continuative	continuative
-	-	'not yet'	continuative negative	cunctative
+	-	'no longer'	discontinuative	terminative

Table 1: Phasal values.

The phasal domain remains a rather vague semantic area for (at least) three reasons which will now be addressed in more detail.

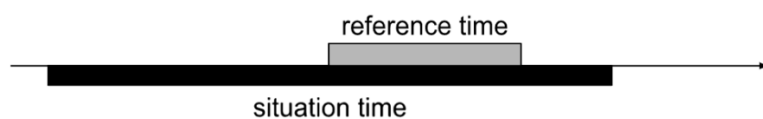
### 2.1.1. Phasal polarity and aspect

First, it is not clear whether phasal polarity is a part of the aspectual domain or is a functional domain in its own right (Plungian 1999: 313-315). Following Klein (1994), in this study I define aspect as a relation between *reference time* and *situation time*. For example, in (5) reference time is the moment of coming into the room, whereas situation time is the whole period during which John was sleeping. Reference time is fully included into situation time (5'), and this type of the relation between reference time and situation time represents the *imperfective* aspect.

(5) [Context: I came into the room and saw...]

*John was sleeping.*

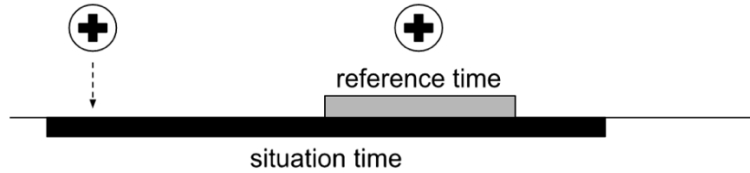
(5')



When the adverb *still* is added to the sentence, it brings the following information: the situation existed at reference time, and it existed at some moment preceding reference time. As shown in (6'), the semantic contribution of *still* does not interfere with aspect, the scheme of the imperfective remains the same.

(6) [Context: I came into the room and saw...]  
*John was **still** sleeping.*

(6')



Thus, according to this view of aspectual domain, phasal polarity and aspect are two distinct categories which complement each other.

Note, however, that not all combinations of aspectual and phasal values are available, cf. Figure 1.

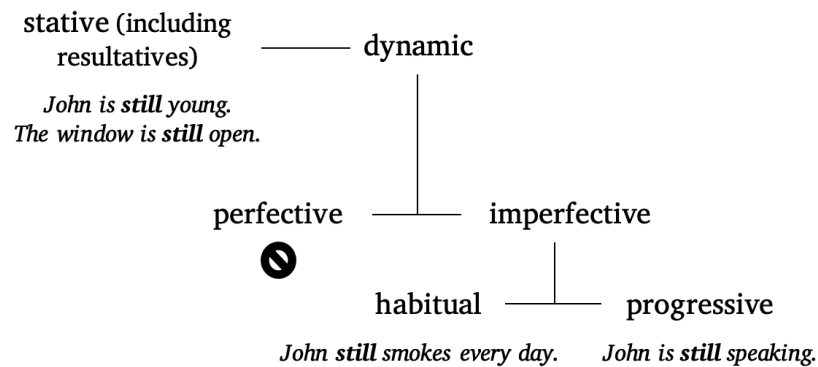


Figure 1: Compatibility of *still* with some actional and aspectual categories.

Importantly for this study, the continuative meaning is not compatible with the perfective aspect,<sup>2</sup> cf. (7)-(8).

(7) \**John still arrived.* (telic)

(8) \**John still slept.* (atelic) (the perfective interpretation is impossible; the reader is asked to ignore the habitual reading)

<sup>2</sup> Perfective implies that the boundaries of situation time are included in reference time (Klein 1994), and this condition is incompatible with the continuative meaning (it requires the situation to be true at some moment before reference time).

Examples (7)-(8) demonstrate that, contrary to the analysis in Michaelis (1993: 198), the reason for such restrictions on *still* is indeed perfectivity and not telicity: *sleep* is an atelic verb but its combination with *still* in the perfective context is forbidden.

### 2.1.2. Phasal verbs

The second issue crucial to understand the structure of the phasal domain is the relationship between phasal expressions, e.g., *still*, and the so-called phasal verbs, e.g., *continue*. There may be a substantial difference between these two types of expressions (cf. discussion of *already* and *start* in Gorbunova 2014), or their semantics may overlap.

The English verb *continue* is ambiguous. When used imperfectively as in (9a), it seems to be semantically identical to *still*. However, *continue* may also be used perfectly, being interpreted as ‘continue after a break’ (9b).<sup>3</sup>

- (9) a. *The sudsy water **continues** working while it is slippery and you can still make bubbles by agitating it.* [BNC]  
 b. *After dinner we **continued** to fiddle around with tackle and were joined by Mr. Ferguson and his son, Paul, who were also booked in for the same week.* [BNC]

In this study only the imperfective uses of the phasal verbs like *continue* (9a), which do not presuppose any interruptions, are treated as examples of the continuative meaning. The meaning ‘continue after a break’ (9b), in turn, is considered a distinct semantic value and is not discussed in the paper.<sup>4</sup>

### 2.1.3. Counter-expectations

It has been suggested that the phasal polarity semantics involves a component of counter-expectations (Plungian 1999: 318). As was shown by van der Auwera (1998, 2021), that is not exactly true: the expectation of the contrary is not an obligatory semantic component of phasal markers, although it may sometimes be present.

<sup>3</sup> Xiao and McEnery (2004: 233) describe the same ambiguity with respect to the continuative *-xiaoqu* in Mandarin Chinese (Sino-Tibetan, Sinitic).

<sup>4</sup> In Stoyanova (2013: 128-129) the meaning ‘continue after a break’ is discussed as part of the repetitive semantic domain.

Van der Auwera distinguishes two possible scenarios where continuative markers may be used: a neutral scenario and a counterfactual (= counter-expected) scenario. Examples (10)-(11), taken from the overview of van der Auwera's discussion by van Baar (1997: 31-32),<sup>5</sup> illustrate neutral and counterfactual scenarios respectively.

(10) [Peter is going to fly from London to Amsterdam at 4 p.m. John and Peter meet at the airport at 3 p.m. At 3 p.m. it is possible for John to say:]

*(Yes, I know.) Peter is **still** in London.*

(11) [Peter is going to fly from London to Amsterdam at 4 p.m. John and Peter meet at the airport at 3 p.m. Then Peter makes an ad hoc decision to leave for Amsterdam on a later plane, which departs at 7 p.m. Suppose that their appointment was arranged in order to discuss some urgent matter which had to be transferred to Amsterdam as soon as possible. If John finds out at 6 p.m. that Peter will take a later plane, it is possible for John to say:]

*(Damn!) Peter is **still** in London.*

In English, as shown in (10)-(11), the adverb *still* can be used in both scenarios, while in some other languages continuative expressions may be available only in one of them. Thus, according to van der Auwera, the ability of phasal markers to be used in neutral and counterfactual scenarios is a parameter of typological variation. This approach is adopted in the present paper.

#### 2.1.4. *The definition*

Based on the discussion above, the definition of the continuative meaning can be formulated as follows:

(12) Continuative is a phasal value which indicates that

- (a) the situation X exists at reference time,
- (b) the situation X existed at the moment  $t_i$  preceding reference time,
- (c) the situation X has not been interrupted between  $t_i$  and reference time.

---

<sup>5</sup> The year of publication of van Baar's dissertation (1997) may be misleading. Although van der Auwera's article was published in 1998, van Baar discusses it in great detail.



To be absolutely clear on the expectations issue (see 2.1.3), it may be added that the state of affairs in (a)-(c) may (but does not have to) be compared with someone's expectations.

## 2.2. Methods and data

### 2.2.1. Methodology

The aim of this study is a large-scale typological investigation of a particular lexical-grammatical domain based on a stratified language sample. This approach allows to capture the world-wide variation of the parameters deemed relevant for the domain and to determine the observed relative frequency and areal distribution of their different values. I focus on four parameters which are relevant to the typological profile of a continuative expression and information on which can be found in the sources (see a more detailed description of each of the parameters in Section 3):

- (13) (a) morphosyntactic type,  
 (b) emphatic vs. non-emphatic status,  
 (c) uses outside the continuative domain,  
 (d) meaning in combination with negation.

Information concerning these parameters for each individual language was obtained from grammatical descriptions and dictionaries. When necessary, information provided by experts on and speakers of particular languages was also used.

When it was possible, I searched for translational equivalents of the lexical items presented in Table 2 (cf. the use of the same method in e.g., Khanina 2008).

language of the source	continuative expressions
English	<i>still, continue, keep (on), stay, remain</i>
French	<i>encore, continuer, rester</i>
Spanish	<i>todavía, aún, seguir, continuar</i>
Portuguese	<i>ainda, continuar</i>
Russian	<i>(vsě) eščë [(vsě) eščë], prodolžat' [продолжать], пока [пока]</i>

**Table 2:** Translational equivalents used when searching for continuative expressions.

When it was not possible to search for translational equivalents, I looked through sections dedicated to aspect, derivational morphology, auxiliary verbs, adverbials and particles. If some expression (in at least one of its meanings) fitted the definition of the continuative given in Section 2.1.4, it was included in the database.<sup>6</sup>

Some reference grammars contained a special section about continuative expressions where at least some of the parameters (a)-(d) were discussed. If there was no description of continuative expressions or if it was not detailed enough, the relevant information could often (but not always) be retrieved from examples found in the sources.

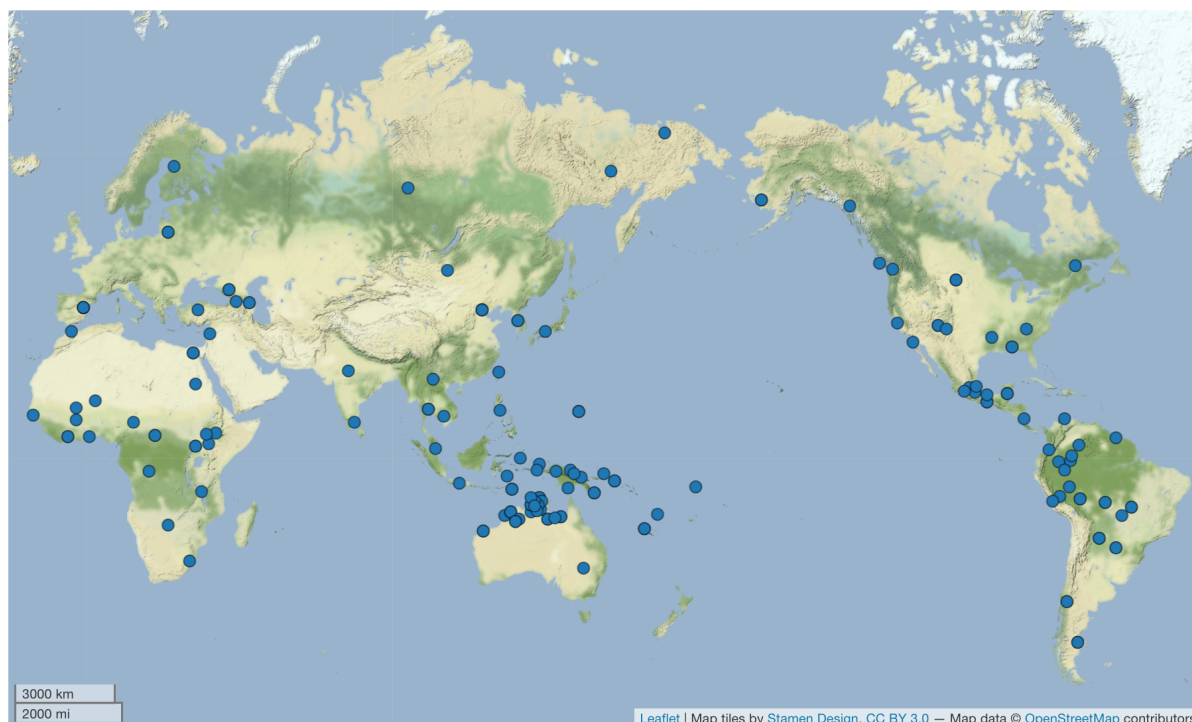
### 2.2.2. *Sampling*

To construct a genealogically stratified sample, I included one language per family by default and added more languages from the families which are the most diverse (Indo-European, Austroasiatic, Atlantic-Congo, Afro-Asiatic, Otomanguean, Athabaskan-Eyak-Tlingit, Arawakan, Pama-Nyungan). I used genealogical classifications in Glottolog 4.8 (Hammarström et al. 2021). To make the sample geographically balanced, I made sure that all macro-areas were represented by an equal number of languages. In dividing the world into macro-areas, I followed Hammarström & Donohue (2014) who distinguish Eurasia, Africa, North America, South America, Australia and Papunesia. The overall number of languages in the sample is determined by the quality of the available language descriptions. In particular, after applying the genealogical “filter” to the languages of Australia, the number of languages having descriptions which mention continuative expressions is hardly above 20, and a similar situation is observed with the languages of South America. As a result, I decided to take 20 languages per macro-area, and, in case of more than 20 good candidates, I included those which are geographically more distant from each other. The geographical distribution of all 120 languages included in the sample is shown in Figure 2.<sup>7</sup>

---

<sup>6</sup> The database is available online at <https://anapanifica.github.io/continuative>, the raw dataset is available at <https://doi.org/10.5281/zenodo.8352034>.

<sup>7</sup> All maps were created with the R package “lingtypology” (Moroz 2017).



**Figure 2:** Languages included in the sample.

## 2. Analysis

### 2.1. *Morphosyntactic type*

#### 2.1.1. *Defining the types*

The continuative meaning was defined in Section 2.1.4 in such a way that does not allow a continuative marker to constitute an independent predication itself: it must modify a predicate. Thus, the range of possible types of continuative expressions is restricted to those types of morphosyntactic elements that can function as predicate modifiers. The continuative markers identified in the sources can be classified into the following types of predicate modifiers: affixes, auxiliaries and adverbs/particles. Importantly, when using the terms “auxiliary” and “adverbs/particles” I mean exclusively morphosyntactic properties of continuative expressions and not their place on the lexical-grammatical scale (for example, particles may be both lexical and grammatical markers, but this difference is ignored in the annotation).

Working criteria used for assigning the morphosyntactic types are summarized in Table 3.

morphologically bound	morphologically free (including clitics)	
affixes	marking typical for verbs in the given language (e.g., agreement)	no verb-like marking
	auxiliaries	adverbs/particles

**Table 3:** Morphosyntactic types of continuative expressions.

If the continuative meaning is expressed by an element interpreted in the source as an affix on the predicate, this element is classified as affix, cf. (14). Several cases where the author’s decision about the morphological status of an element seems debatable are discussed in Section 3.1.2.

(14) Central Alaskan Yupik (esu; Eskimo-Aleut, Yupik; Miyaoka 2012: 1232, glosses added)

*tai-gur-tuq*

come-CONT-IND.3SG

‘He is still coming, keeps coming.’

Second, morphologically free continuative markers which can be identified as verbs in the given language, e.g., agree with the subject of the clause and/or have TAM markers, etc., and which combine with another (lexical) predicate, are labelled auxiliaries, cf. (2) repeated here as (15).

(15) Balanta-Ganja (Atlantic-Congo, North-Central Atlantic; Creissels and Biaye 2016: 201)

*Bá-n-tígtà-nà*

*yâaθ.*

INCL-INACP-AUX<sub>CONT</sub>-INCL

work

‘We keep working.’

Most of the rest of the continuative expressions fit into the category of adverbs/particles. To draw a boundary between adverbs and particles is hardly possible because in the literature on specific languages there are no common methodological grounds for using the terms. In addition, this type includes adpositional phrases and combinations of adverbs/particles with intensifiers. Examples of continuative adverbs/particles are given in (16)-(17).

(16) Montagnais (moe; Algic, Algonquian-Blackfoot; Oxford 2007: 209)

*Tâpue* **eshku** *mishta-minuâteu*.  
truly **still** really-love.3 > 3'  
'He truly still loves her.'

(17) Hup (jup; Naduhup, Eastern Naduhup; Epps 2008: 584)

*dóʔ = d'əh* *b'óy-óy* **té**  
child = PL study-DYNM YET  
'The children are still studying/at school.'

Finally, there are six continuative expressions whose morphosyntactic status cannot be defined based on the data provided in the sources. I mark such cases with the label “not clear”.

### 3.1.2. Areal distribution

Figure 3 shows the distribution of morphosyntactic types of continuative expressions across macro-areas.

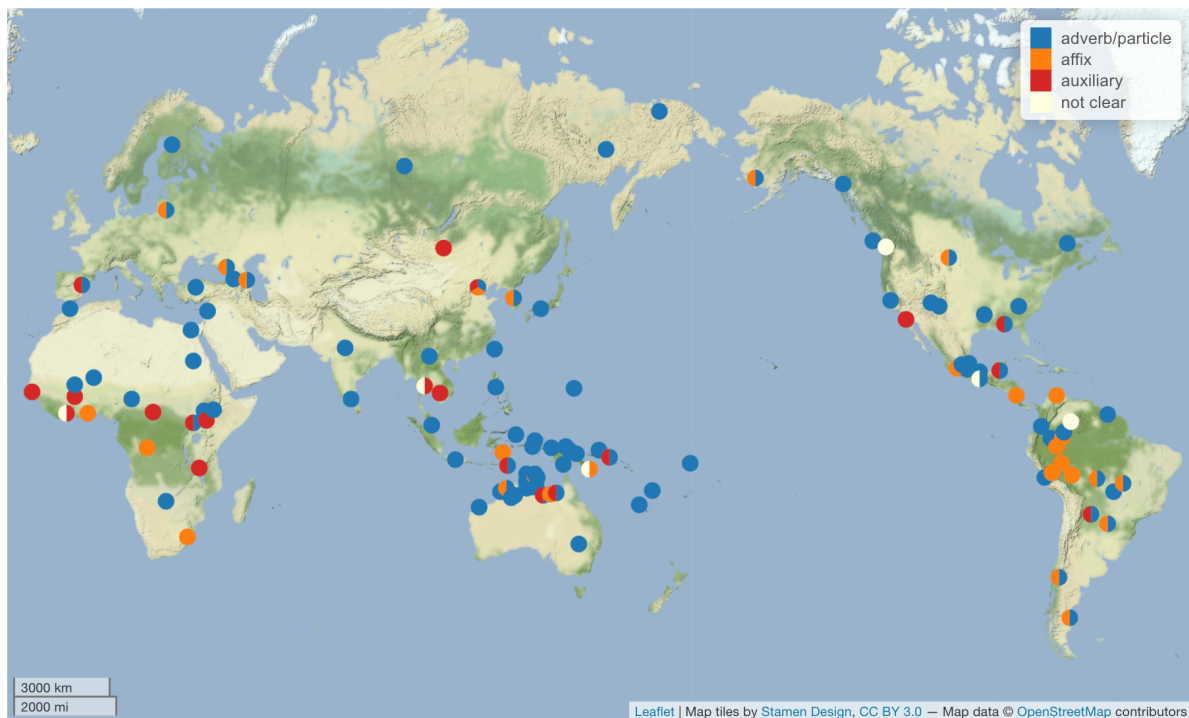


Figure 3: Morphosyntactic types of continuative expressions.

It can be seen that adverbs/particles constitute the most widespread morphosyntactic type of continuative expressions: they occur in many languages in the sample and are present in all macro-areas. Moreover, one may suspect that the adverbial strategy is actually possible in many more or even all languages in the sample. Since most of my sources are grammatical descriptions, and adverbial phrases tend to be lexical rather than grammatical items, some adverbs/particles might have been overlooked. This assumption is plausible from yet another perspective: although many languages have continuative adverbs/particles whose diachronic sources are no longer transparent, supposedly all languages can derive a continuative as a periphrastic expression, e.g., as a combination ‘until’/‘to’/‘and’ + ‘now’, cf. (18).<sup>8</sup>

(18) Mongolian (mon; Mongolic-Khitani, Mongolic; Pjurbeev 2001; glosses and transcription added)

*odoo boltol / odoo xiirtel*  
now until / now to  
‘Up to now; to this day; still’

Having established that the ‘adverbs/particles’-type is the default morphosyntactic type of continuative expressions, let us now turn to the affix-type and auxiliary-type in each of the macro-areas.

Among 20 Eurasian languages included in the sample there are nine languages which feature either a continuative auxiliary or a continuative affix. The boundary between these two types is not always clear, especially in languages which lack inflectional morphology. For example, the Mandarin Chinese continuative (-)*xiaqu* is usually interpreted as one of the suffixes deriving verbal compounds (Li & Thompson 1989: 61-62; Ross & Ma 2014: 120), cf. (19a). However, the same element can occur in its lexical meaning as a morphologically independent predicate (19b). Thus, the affixal status of (-)*xiaqu* is not self-evident (at least, from a purely morphological perspective).

---

<sup>8</sup> Although ‘now’-based periphrastic continuatives may be restricted to present tense, they are nevertheless considered continuative expressions in the framework of the study (cf. van der Auwera 2021: 26 on this issue for phasal expressions in general).

## (19) Mandarin Chinese (cmn; Sino-Tibetan, Sinitic)

a. *kàn-xià-qù*

read-descend-go

'Keep reading.' (Li &amp; Thompson 1989: 61)

b. *wǒ xiān xiàqù dǎng-dǎng zài shuō*

I first get\_off enquire-enquire then say

'Let me get off (the car) and ask about it first before taking action.' (Xiao &amp; McEnery 2004: 227)

Similar issues can be discussed with regard to almost all continuative auxiliaries/affixes in South East Asian languages included in the sample (Central Khmer, Thai, Korean and Halh Mongolian).<sup>9</sup>

In the European part of the macro-area there are four languages in the sample which possess continuative affixes or auxiliaries: Spanish (Indo-European, Italic), Lithuanian (lit; Indo-European, Balto-Slavic), Abaza (abq; Abkhaz-Adyge, Abkhaz-Abaza) and Lezgian (lez; Nakh-Daghestanian, Daghestanian). The degree of morphologization of the continuative markers in these languages clearly varies. The auxiliary verbs *continuar* 'continue' and *seguir* 'follow, continue' in Spanish do not show any effects of morphologization. Moreover, even their auxiliary status seems to be relatively new: according to van der Auwera (1998: 30), "Spanish *continuar*, when followed by the gerundio, could be considered to be an auxiliary or semi-auxiliary", while "English *continue* may still be a lexical verb". In contrast, the Lezgian continuative auxiliary *ama* 'stays' is univerbated with the preposed dependent verb inflected for aspect (20), see Haspelmath (1993: 145) and Maisak & Verhees (ms); but in other Lezgian languages, e.g., in Agul (21), the auxiliary 'stay' still functions as a morphologically independent verb (Maisak & Verhees ms).

## (20) Lezgian (Nakh-Daghestanian, Daghestanian; Haspelmath 1993: 210; glosses adapted)

*Jusuf.a k'walax-zama.*

Jusuf.ERG work-IPFV.CONT

'Jusuf is still working.'

<sup>9</sup> Central Khmer (khm; Austroasiatic, Khmeric), Thai (tha; Tai-Kadai, Kam-Tai), Korean (kor; Koreanic); Halh Mongolian (khk; Mongolic-Khitian, Eastern Mongolic).

(21) Agul (agx; Nakh-Daghestanian, Daghestanian; Merdanova 2004: 115, cited by Maisak & Verhees ms)

<i>dad.a</i>	<i>k:azit</i>	<i>ruχ.a-j</i>	<b><i>ame-a</i></b>
father.ERG	newspaper	read.IPFV-CVB	<b>stay-PRS</b>

‘Father is still reading the newspaper.’

The diachronic source of the Abaza continuative *-rkʷ(a)* (22) is not absolutely clear, although there is some partial support for its verbal origins (Genko 1955: 140). In the modern language, its affixal status is fairly certain.

(22) Abaza (Abkhaz-Adyge, Abkhaz-Abaza; Klyagina & Panova 2019: 7; glosses adapted)

<i>a-kʷa</i>	<i>ʕa-kʷa-rkʷ-əw-n</i>
DEF-rain	CSL-rain-CONT-IPFV-PST

‘[When I was going home] it was still raining.’

The Lithuanian prefix *tebe-* is a rare example of the continuative affix originating from an adverb. According to Ostrowski (2011, 2016: 176-177), the verbal prefix *be-*, reinforced by the deictic element *te-*, goes back to the adverb *be* ‘yet, still’. As can be seen from texts, the prefix superseded the corresponding adverb in the first half of 19<sup>th</sup> century.

Auxiliaries and affixes are very frequent types of continuative expressions in the languages of Africa. With the exception of the Afro-Asiatic languages, which generally prefer adverbial phrases, all the seven Atlantic-Congo languages included in the sample as well as three out of eight languages belonging to the smaller African families (Gban, Mande; Ma’di, Central Sudanic; Turkana, Nilotic)<sup>10</sup> feature either a continuative auxiliary or a continuative affix. Moreover, in most of these languages, and especially in Bantu, continuative markers play a prominent role in TAM systems, so they are discussed both in grammatical descriptions and in specific papers, see Nurse (2008: 145-148), Maho (2008), Löfgren (2019), among others. Two main morphosyntactic types of continuative expressions in Bantu are exemplified in (23)-(24).<sup>11</sup> In Zulu the continuative marker *sa-* appears directly in the verb; in Nyakyusa-

<sup>10</sup> Gban (ggu; Mande, Eastern Mande), Ma’di (mhi; Central Sudanic, Moru-Madi), Turkana (tuv; Nilotic, Eastern Nilotic).

<sup>11</sup> The third possible option can be seen in Swahili (swh; Atlantic-Congo, Volta-Congo) where the borrowed adverb *bado* ‘still’ almost replaced an older continuative auxiliary *-ngali* (Nurse 2008: 145; Zahran & Bloom Ström 2022).



Ngonde the continuative marker *tu-kaalɪ* (glossed as SP1PL-PERS(persistentive)) is a separate word consisting of the subject prefix and the root *-kaalɪ*.<sup>12</sup>

(23) Zulu (zul; Atlantic-Congo, Volta-Congo, Bantoid; Ziervogel et al. 1967: 91; glosses added)

- a. *u-sa-phek.a*  
 3SG.G2.SBJ-CONT-cook  
 ‘She still cooks/ She is still cooking.’
- b. *ba-sa-wadla*                      *ama-swidi*  
 3PL.G2.SBJ-CONT-eat      PL.G4-sweet  
 ‘They are still eating the sweets.’

(24) Nyakyusa-Ngonde (nyy; Atlantic-Congo, Volta-Congo; Persohn 2021: 133; glosses adapted)

- tu-kaalɪ*                      *tu-ku-bop-a.*  
 SBJ1PL-PERS      SP1PL-PRS-run-FV  
 ‘We are still running / still run.’

Historically, continuative markers attested in Bantu (often called “persistives” in the literature) go back to the Proto-Bantu marker *\*-kí(-)* (Meeussen 1967: 109; Nurse 2008: 145-148). According to (Maho 2008: 296), it originally had an imperfective and/or progressive meaning and functioned as an auxiliary in the construction structurally similar to the one in (24).

Outside Bantu, several African languages, e.g., Balanta-Ganja (Atlantic-Congo, North-Central Atlantic) and Ewe (ewe; Atlantic-Congo, Volta-Congo, Kwa Volta-Congo), show dedicated continuative auxiliaries similar to (24), while others — at least, Sango (sag; Atlantic-Congo, Volta-Congo, North Volta-Congo), Ma’di (Central Sudanic, Moru-Madi) and Gban (Mande, Eastern Mande) — demonstrate a different strategy, using the verb ‘stay, remain’. Example (25) illustrates the case of Sango: in (25a) the verb *ngba* is a lexical verb, in (25b) it functions as an auxiliary, taking the nominalized verb form as a complement.

<sup>12</sup> The root *-kaalɪ* is very likely to contain the copula *ɪ* (Persohn 2021: 133), cf. Nurse (2008: 147) on this pattern in other Bantu languages.

(25) Sango (Atlantic-Congo, Volta-Congo, North Volta-Congo)

- a. *mbi yí ála ngbá na ndo só pepe*  
 1SG want 3PL **stay** PREP ? DEM NEG

‘I don’t want them to stay here.’ (Samarin 1970: 127; glosses added)

- b. *mbi ngba ti hu-ngo pino*  
 1SG **remain** SBJ see-NMLZ suffering

‘I am still suffering.’ (Nassenstein & Pasch 2021: 114; glosses adapted)

Finally, one continuative marker in Africa in the database is classified as ‘not clear’. This is a predicative marker *lé* in Gban (Mande). Predicative markers are portmanteau morphemes expressing TAM and polarity and occurring in the post-subject position. An example illustrating the predicative marker *lé* in the continuative meaning is given in (26).

(26) Gban (Mande, Eastern Mande; Fedotov 2015: 4; glosses adapted)

- zǐǐǐó ð lé blè*  
 then 3SG[...] **CONT** IPFV\walk

‘[They walked all morning] and are still walking.’

Overall, the number of continuative auxiliaries/affixes in Africa gradually increases from north to south: continuative auxiliaries/affixes are not widespread in North Africa but in Central and especially South Africa (with the notable exception of the Khoisan language Ts’ixa<sup>13</sup> genealogically distant from most languages spoken in this area) they represent the dominant strategy of expressing the continuative meaning.

In North America non-adverbial continuatives are rare and scattered throughout the macro-area. Central Alaskan Yupik (esu; Eskimo-Aleut, Eskimo), Purepecha (tsz; Tarascan) and Lakota (lkt; Siouan, Core Siouan) are the only North American languages in the sample with continuative affixes (however, the Lakota continuative suffix *-akhe* is non-productive (Ullrich 2018: 190)), and two more languages have continuative auxiliaries (27)-(28). In addition, Yucatec Maya (yua; Mayan, Core Mayan) employs the auxiliary *sègir* ‘continue’, borrowed from Spanish.

<sup>13</sup> Ts’ixa (Khoe-Kwadi, Khoe).

(27) Tipai (dih; Cochimi-Yuman, Yuman; Miller 2001: 293; glosses adapted)

*nyaach saaw xkiway*  
 I + SBJ eat **still.do**  
 ‘I’m still eating.’

(28) Creek (mus; Muskogean; Martin 2011: 306)

*a:fack-itá hámk-it ahô:sk-i: mónk-ati:-s*  
 happy-INF one-T left.over.FGR-DUR **still-PAST5-IND**  
 ‘One game still remained.’

South America turns out to be the macro-area with the largest number of morphologically bound continuative expressions. Continuative affixes are attested in 12 of the 20 South American languages included in the sample, and only one language — Nivaclé (cag; Matacoan, Mataguayo I) — demonstrates a continuative auxiliary (based on the verb ‘stay’ (Fabre 2016: 360)). South America is often subdivided into two linguistic areas with different typological profiles: Amazonia and the Andes (Dixon & Aikhenvald 1999: 8-10). However, continuative affixes found in Amazonian and Andean languages are structurally very similar, see examples in (29)-(30). Usually these are optional suffixes occupying a specific slot in the verbal template along with a number of other suffixes having rather “lexical” meanings, e.g., ‘again’, ‘for a long time’, ‘regretfully’, etc. Diachronic sources of South American continuative suffixes are not discussed in the literature, which might suggest that they are not synchronically transparent.

(29) Tanimuca-Retuarã (tnc; Tucanoan, Eastern Tucanoan; Eraso 2015: 263)

[Amazonia]  
*ʃi-bá’írábé-~júhú-ruʃú*  
 1S-work-NO.COMPL-FUT  
 ‘I will still be working.’

(30) Mapudungun (arn; Araucanian; Smeets 2008: 172, glosses simplified) [the Andes]

*müle-ka-y ta-mi chaw?*  
 be-CONT-IND-3 the-POSS.2SG father  
 ‘Is your father still there?’

In Australia continuative auxiliaries and continuative affixes are rare. The database includes two continuative expressions encoded as auxiliaries: *wirdija* in Kayardild (gyd; Tangkic, Southern Tangkic) and *mirra* in Wambayan (wmb; Mirndi, Ngurlun). Both verbs have a wide range of meanings: ‘stay, reside’ (in locative clauses), ‘be’, ‘become’, and only in ascriptive clauses<sup>14</sup> it is used in the continuative meaning (– ‘be still’) (Evans 1995: 321; Nordlinger 1998: 178). Two examples of continuative affixes are also not straightforward. The continuative *-wa* in Garrwa (wrk; Garrwan) is described as a suffix which is exclusive to verbs (Mushin 2012: 199). However, there are examples where it also attaches to the temporal adverbial *wabula* ‘olden times’ (Mushin 2012: 321), which makes it more similar to a clitic. The continuative *djal-* in Bininj Kun-Wok (gup; Gunwinyguan) appears in the verb as a prefix (31) but “when it restricts nouns, it is a separate word rather than a prefix” (Evans 2003: 516).

(31) Bininj Kun-Wok (gup; Gunwinyguan; Evans 2003: 518)

*A-marne-djal-djare*

1/3-BEN-**just**-want.NPST

‘I still love him/her’ (first interpretation offered) / ‘Only I love him/her’ / ‘I love only him/her.’

Papunesia is a very diverse macro-area, and it does not always make sense to discuss it as a whole. However, as for continuatives, the general tendency to lack continuative auxiliaries and affixes seems to hold for all major areal and genealogical linguistic units distinguished in the macro-area. The detailed data on continuative expressions in Malayo-Polynesian languages of South East Asia (MPSEA) are provided in (Veselinova et al. to appear). The distribution of morphosyntactic types of continuative expressions in their sample is presented in Table 4.

As Veselinova et al. (to appear) note, “the preference is clearly for morphologically free expressions; bound ones exist but are relatively few”. Leaving aside differences in encoding, my data show the same pattern: seven out of eight Austronesian languages included in my sample have continuative expressions classified as adverbial phrases, and only *Tukang Besi North* (khc; Austronesian, Malayo-Polynesian) — which happened to be included both in my study and in the study of Veselinova et al. (to appear) — has the continuative suffix *-ho* (Donohue 1999: 173-174).

---

<sup>14</sup> Ascriptive clauses are clauses where the predicate “attributes a certain property to the subject” (Nordlinger 1998: 173).

type	number of languages
free markers	
adverb	15 languages
aspect marker	11 languages
free gram	9 languages
single morpheme	5 languages
auxiliary	4 languages
particle	3 languages
aspectual adverb	1 language
periphrastic construction/adverbial	1 language
several markers	
1. AUX-like 2. aspect marker	1 language
1. adverb 2. combination of adverb and clitic 3. clitic	1 language
1. AUX-like 2. aspect marker 2. aspect clitic	1 language
1. free aspect marker 2. proclitic aspect marker	1 language
enclitics	7 languages
suffix	1 language

**Table 4:** Types of STILL expressions in 61 MPSEA languages (Veselinova et al. to appear).<sup>15</sup>

Non-Austronesian languages included in the sample show a similar picture. There are only three non-adverbial continuatives and, in addition, none of them has continuative as a core meaning. The only putative example of the morphologically bound continuative marker is the suffix *-an* in Daga (dgz; Dagan, Central Dagan) described in (Murane 1974: 62) as the marker of the Prolonged Action Tense, but it is mainly used in the sense ‘do smth until’. Two continuative auxiliaries found in the languages Bunak (bfn; Timor-Alor-Pantar) and Siwai (siw; South Bougainville, Buinic) show the continuative meaning only in examples with stative predicates (32)-(33), while in other cases they have meanings in the domain of pluractionality (cf. Section 3.3 on polysemy of continuative markers).

(32) Bunak (Timor-Alor-Pantar; Schapper 2022: 469)

<i>Baqi</i>	<i>u</i>	<i>niq</i>	<i>oa,</i>	<i>baqi</i>	<i>heser</i>	<i>liol.</i>
NPRX.AN	live	NEG	PFV	NPRX.AN	dead	<b>continue</b>

‘He didn’t live any more, he kept on being dead.’

<sup>15</sup> The data come from the map layer “MPSEA\_STILL: bound or free expressions”, <https://arcg.is/OjnvHm>.

(33) Siwai (South Bougainville, Buinic; Onishi 1994: 500; glosses adapted)

<i>ong</i>	<i>noo</i>	<i>toko = tokoh-ah</i>	<i>tu-ro-ng?</i>
DEM.M	possibly	REDUP = be.hot-PART	COP.3SG-PERF-M

‘Is that possibly still hot?’

### 3.1.3. Discussion

The preferences for specific morphosyntactic types of continuative expressions can be to a certain degree explained by appealing to the typological profiles of the languages. Two linguistic characteristics which seem to be especially relevant for developing continuative auxiliaries and continuative affixes are multiple verb (i.e., auxiliary or serial verb) constructions and polysynthetic morphology.

Predictably, continuative auxiliaries tend to be found in languages with well-developed systems of multiple verb constructions. Good examples are Mandarin Chinese and Thai in Southeast Asia and the Papuan language Bunak, which are known for directional serial verb constructions and also use motion verbs as continuative markers (Li & Thompson 1989: 61; Iwasaki and Ingkaphirom 2005: 157-158; Schapper 2022: 468). Continuative auxiliaries that originated from the position verbs ‘be at’, ‘stay’, ‘sit’ or from copulas usually present a part of a diverse system of auxiliary verb constructions, as in Creek (Muskogean; Martin 2011: 306-307) or Ma’di (Central Sudanic; Blackings & Fabb 2003: 246-250). It must be specially noted that the Atlantic-Congo type of continuative auxiliaries can also be associated with the extensive use of multiple verb constructions in many of these languages: when an “old” continuative affix is fused with copula and then combined with a lexical verb, a new multiple verb construction emerges, cf. (24) above and the still transparent continuative construction in Ewe (34).

(34) Ewe (Atlantic-Congo, Volta-Congo, Kwa Volta-Congo; Ameka 2018, cited by Kramer 2021: 7; *ko* ‘only’ is an intensifier)

<i>é-ga-le</i>	<i>aha</i>	<i>no-m</i>	<i>ko</i>
3SG-REP-be.at:PRS	alcohol	drink-PROG	only

‘He is still drinking alcohol.’

Polysynthesis and, specifically, elaborate derivational morphology is a characteristic of many languages featuring continuative affixes. In particular, morphologically

bound continuatives are found in such polysynthetic languages as Abaza (Abkhaz-Adyge), Central Alaskan Yupik (Eskimo-Aleut) and many South American languages. One may hypothesize that the widespread presence of continuative affixes in South America can be a part of the more general tendency concerning verbal derivational morphology as a whole. For example, as Müller (2014) reports, South America is the macro-area where the number of languages having special desiderative affixes is considerably higher than in the rest of the world.

Of course, polysynthesis and multiple verb constructions are not sufficient criteria for the development of a continuative affix/auxiliary: there are languages that show these features but still exclusively use continuative adverbials, e.g., polysynthetic Navajo (nav; Athabaskan-Eyak-Tlingit, Athabaskan-Eyak) or serializing Nêlêmwa-Nixumwak (nee; Austronesian, Malayo-Polynesian, Oceanic). However, in general it seems that a continuative affix or auxiliary can be easily developed only in a language with the relevant morphosyntactic profile.

### 3.2. *Emphatic vs. non-emphatic status*

#### 3.2.1. *Non-emphatic continuatives*

In the literature on phasal polarity it has been noted that some phasal expressions tend to regularly occur in contexts where their presence seems to be redundant. For example, Dahl and Wälchli (2016) show that iamitives (a type of ‘already’-markers) frequently occur with “natural development predicates, that is, predicates that become true sooner or later under normal circumstances” (Dahl & Wälchli 2016: 326). In (35) the change of state is a part of the semantics of the predicate ‘rot’, and, nevertheless, Indonesian *sudah* ‘already’ appears in such contexts almost obligatorily. Importantly, it does not seem to add any emphasis to the statement.

(35) Indonesian (ind; Austronesian, Malayo-Polynesian; Dahl & Wälchli 2016: 328)

<i>Kamu</i>	<i>tidak</i>	<i>bisa</i>	<i>memakan-nya.</i>	<i>Itu</i>	<b><i>sudah</i></b>	<i>busuk.</i>
You	not	can	eat-it	that	IAM	rotten

‘You cannot eat it. It is rotten.’

A tendency to accompany natural development predicates has also been mentioned for some nondum (‘not yet’) markers by Veselinova (2015: 20).

Likewise, a tendency to regularly occur in contexts already implying the continuative semantics without adding any emphasis seems to be a characteristic of some continuative expressions. Continuatives of this type are most often found in the contexts ‘still alive’ (36), ‘still young’ (37) and ‘still morning’ (38). Less frequent contexts are ‘still night/dawn’, ‘still a virgin/unmarried’, ‘still in a belly/at the breast’, etc. All mentioned predicates already contain the semantic components typical for the continuative meaning: the situation existed at some moment before reference time, it was not interrupted, and it is expected to change in the future.

(36) Bunak (Timor-Alor-Pantar; Schapper 2022: 508)

*Lui Bert u taq.*  
Louis Berthe live CONT  
‘Louis Berthe was still alive.’

(37) Mapudungun (Araucanian; Smeets 2008: 313, glosses adapted)

*ĩnché rumé lliika-nten-nge-wma pichi-ka-lu*  
I very get.afraid-NMLZ-VERB-SCVN small-CONT-SVN  
‘I really used to be someone who easily got scared when I was young.’

(38) Yeri (yev; Nuclear Torricelli, Wapei-Palei; Wilson 2017: 196)

*awo ko maleikia-pi kua*  
yes still morning-ADD still  
‘Yes, it is still morning.’

In the framework of this study, the continuative expressions showing a high degree of obligatoriness in contexts like (36)-(38) will be called *non-emphatic*. In contrast, the continuative expressions which are usually omitted in such contexts (and, when present, have an obvious emphatic function) will be called *emphatic*.

One of the further development paths of non-emphatic continuatives is the gradual loss of productivity: they become available with a restricted set of stative predicates and appear only in subordinate while-clauses or secondary predications (in the depictive function). An example of the non-productive continuative comes from Lakota: the continuative suffix *-akhe* is found only in five constructions with the meanings ‘while still fresh’, ‘with clothes still on’, ‘while still healthy’, ‘while it still down’, ‘while still alive’ (39) (Ullrich 2018: 190).



(39) Lakota (Siouan, Core Siouan; Ullrich 2018: 279)

<i>ní-akhe</i>	<i>thiyáta</i>	<i>ya-khí-pi</i>	<i>ktA</i>
alive-DER.CONT	home	2A-arrive.back.there-PL	FUT.IRR

‘You will get back home alive.’

Judging by examples from grammatical descriptions, the continuatives *-aanjanu* in Worrorra (40) and *té* in Hup (41) also tend to be frozen with certain stative predicates (although there are no restrictions on their combinations with dynamic predicates).

(40) Worrorra (wro; Worrوران, Western Worrوران; Clendon 2014: 269)

- a. *wangalang-aanjanu*  
child-ESS  
‘While (someone) was a child.’
- b. *lewarra-aanjanu*  
daylight-ESS  
‘While there’s still daylight.’

(41) Hup (Naduhup, Eastern Naduhup; Epps 2008: 585)

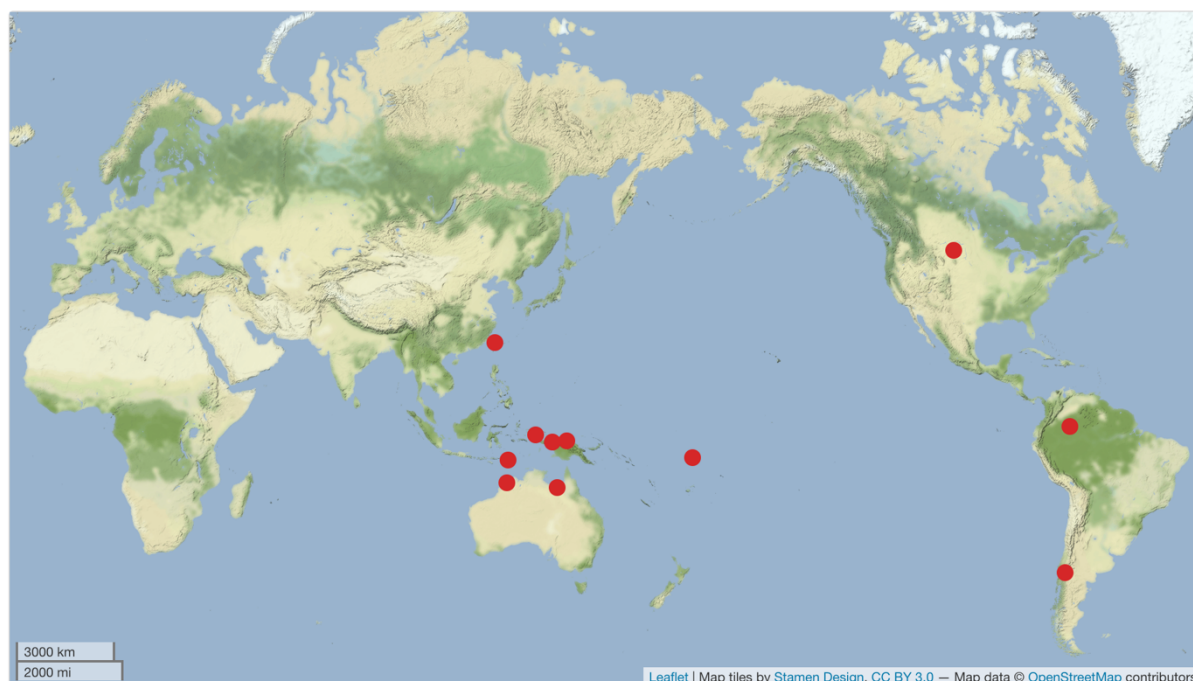
- a. *tih = pæccéw = d’əh*                      *té*  
3SG = adolescent.boy = PL      YET  
‘Still young (boys).’
- b. *wág té*  
day YET  
‘Still day/light.’

### 3.2.2. How to find non-emphatic continuatives

For the purposes of this study, we need a method which allows us to determine an emphatic vs. non-emphatic status of all continuative expressions in the sample. The approach proposed in this section is based on examples presented in grammatical descriptions. Since non-emphatic continuative expressions tend to frequently occur in the contexts like ‘be alive’ and ‘be young’, a substantial number of examples of the specific continuative expression in such contexts in the grammatical description may

signal the non-emphatic status of that continuative expression. Despite an obvious bibliographical bias, this method has an important advantage: it is applicable to all languages the descriptions of which contain at least several sentential examples with continuative expressions, i.e., almost all languages of the sample.

Figure 4 shows the distribution of continuative expressions which occur in at least three different contexts, already implying the continuative semantics, throughout the grammatical description. For example, the marker *koi* in Tuvalu (tvl; Austronesian, Malayo-Polynesian) in the reference grammar (Besnier 2000) occurs in the contexts ‘still young’ (Besnier 2000: 70), ‘still predawn’ (Besnier 2000: 82) and ‘still alive’ (Besnier 2000: 120), thus it is included to Figure 4.



**Figure 4:** Non-emphatic continuatives.

Even taking into account the vagueness of this method, the areal patterns observed in Figure 4 do not seem accidental. Most continuative expressions typically used in natural development contexts are found in Papunesia and Australia, three more examples come from the Americas. Interestingly, Papunesia is an area where iamitive (‘already’) markers are also often found, cf. “Philippine” and “Indonesian” types of iamitives identified in (Dahl & Wälchli 2016).

### 3.3. Polysemy of continuative expressions

#### 3.3.1. Iterative ('continuously, repeatedly, always'), repetitive ('again'), additive ('more, also')

Most often the continuative expressions have additional meanings related to pluractionality, i.e., repetition of the same or, at least, comparable situation one or more times. The semantic link between the continuative meaning and pluractionality is easily explained: both meanings imply the existence of some situation at several temporal points, but the continuative requires it to be precisely the same uninterrupted situation, while pluractionality allows it to be different situations (cf. McGregor 1990: 470). Languages with continuative markers which are also used to express pluractional meanings are shown in Figure 5.<sup>16</sup>

The continuative expressions which also have the *iterative* meaning ('constantly, repeatedly, always') are found in eight languages in the sample.<sup>17</sup> For example, in Kesawai (Nuclear Trans New Guinea) the enclitic =*apaie*, usually occurring in the serial construction with the verb *te* 'do', conveys the meaning close to English *still*, as in (42a). However, in the habitual/iterative contexts like (42b) the continuative meaning of =*apaie* is not seen anymore, instead it is paraphrased as 'continuously'.

(42) Kesawai (Nuclear Trans New Guinea, Madang; Priestley 2008: 382-383; glosses adapted)

- a. *Hekeni ketin = apaie te-r-i.*  
 firewood cut:split = **continuously** do-PRS-1S  
 'I'm still cutting firewood.'
- b. *Mo esame pi mipii somoru paru = apai tu-pu-r-a.*  
 this dog time many night bark = **continuously** do-HAB-PRS-3SG  
 'Often (many times) this dog barks at night continuously.'

<sup>16</sup> If a language has several continuative expressions, their non-continuative meanings are shown together (this applies to all maps in Section 3.3).

<sup>17</sup> Gooniyandi (gni; Bunaban), Central Alaskan Yupik (Eskimo-Aleut, Eskimo), Bininj Kun-Wok (gup; Gunwinyguan), Mawng (mph; Iwaidjan Proper), Mullukmulluk (mpb; Northern Daly), Kesawai (xes; Nuclear Trans New Guinea, Madang), Ngiyambaa (wyb; Pama-Nyungan, Southeastern Pama-Nyungan), Bunak (Timor-Alor-Pantar).

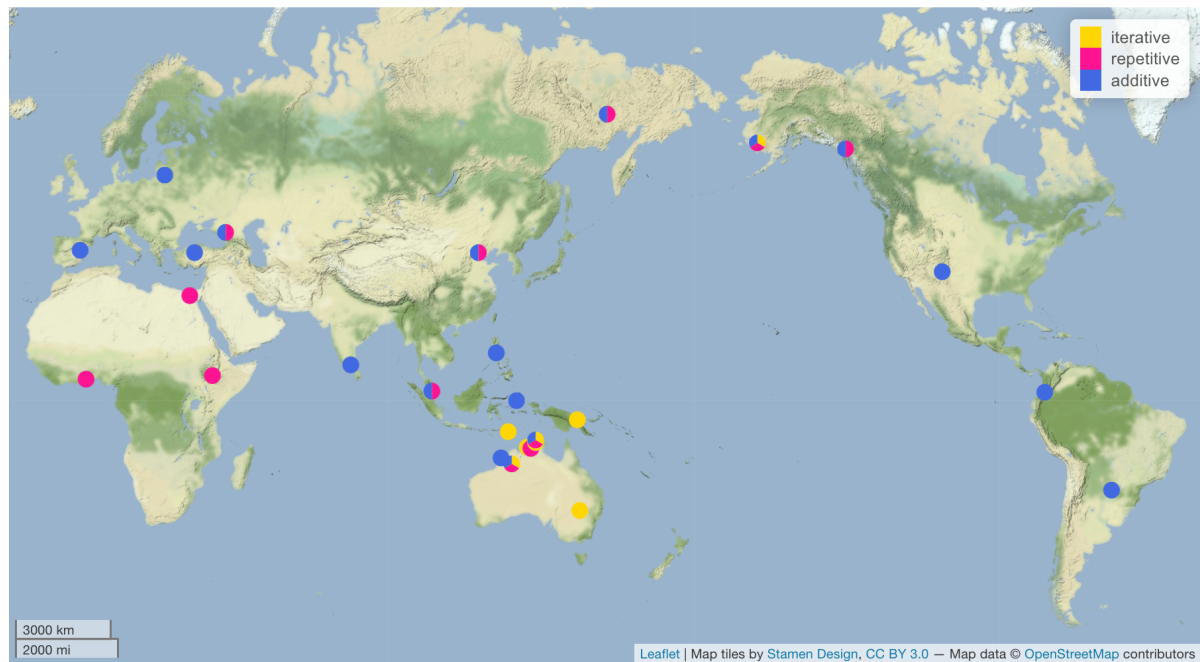


Figure 5: Additive, repetitive and iterative meanings of the continuative expressions.

Remarkably, the English *still* represents one more example of the close relationship between continuative and iterative meanings. With a reference to Kemmer (1990), Michaelis (1993: 205) notes that “temporal *still* at one time served as a durational adverb akin to *constantly* or *continually*”.

The continuative-iterative ambiguity of some markers often poses a problem for determining what can be considered a continuative expression. For instance, in Paakantyi (Pama-Nyungan, not in the sample) there is the suffix *-ɲana* which “mainly describes a prolonged process” (Hercus 1982: 195) but in one of the examples provided (43) it seems to be used in the continuative meaning. In principle, the example (43) can be interpreted either as ‘the dog started to sniff at the rat’s nest some time ago and it is still sniffing’ (continuative) or as ‘the dog is being sniffing at the rat’s nest continuously’ (iterative). Since there is no evidence in favor of the continuative interpretation, this marker was not included in the database of the continuative markers.

(43) Paakantyi (drl; Pama-Nyungan, Yarli-Baagandj; Hercus 1982: 195)

<i>gaḷi</i>	<i>bāra-la-ɲana</i>	<i>bulgu-na</i>	<i>yabara</i>
dog	smell-TOP-ASP	rat-GEN	camp

‘The dog keeps sniffing at the rat’s nest.’

The *repetitive* meaning ('again') is expressed by the same item as the continuative meaning in 10 languages of the sample.<sup>18</sup> Example (44) shows a copula verb with the repetitive/continuative prefix *gà-* in Ewe (Kwa Volta-Congo). Another well-known example of this polysemy is the French adverbial *encore* 'again, still'.

(44) Ewe (Atlantic-Congo, Volta-Congo, Kwa Volta-Congo; Rongier 2004: 75)

*Egàle zɔzɔm.*

'He walks again.' / 'He keeps walking.'

Finally, 17 languages of the sample possess markers which have both the continuative meaning and the *additive* meaning ('also, more').<sup>19</sup> An example of such a marker is given in (45): the particle *ql* in Towa (Kiowa-Tanoan) can mean both 'still' and 'also'.

(45) Towa (Kiowa-Tanoan; Yumitani 1998: 247; glosses adapted)

a. *ql* *ì-wéñí*

**still** INTR[1DU]-strong/STAT

'We are still strong/healthy.'

b. *vî?wè* *tyê'tiba-š* *ql* *səpə-pé'*

both box-INV **also** TR[1DU:3INV]-make/PFV

'We both also made a box.'

The polysemy 'still'/'more' is illustrated with the suffix *-ve* in Paraguayan Guaraní (Tupian) (46). Another good example is the Turkish continuative *daha* which,

<sup>18</sup> Abaza (Abkhaz-Adyge, Abkhaz-Abaza), Kambaata (ktb; Afro-Asiatic, Cushitic), Coptic (cop; Afro-Asiatic, Egyptian), Tlingit (tli; Athabaskan-Eyak-Tlingit), Ewe (Atlantic-Congo, Volta-Congo, Kwa Volta-Congo), Semelai (sza; Austroasiatic, Aslian), Gooniyandi (Bunaban), Wageman (waq; Isolate, Australia), Mawng (Iwaidjan Proper), Southern Yukaghir (yux; Yukaghir, Southern Yukaghir).

<sup>19</sup> Abaza (Abkhaz-Adyge, Abkhaz-Abaza), Lithuanian (Indo-European, Balto-Slavic), Tlingit (tli; Athabaskan-Eyak-Tlingit), Semelai (sza; Austroasiatic, Aslian), Tagalog (tgl; Austronesian, Malayo-Polynesian), Totoro (ttk; Barbacoan, Coconucan), Gooniyandi (Bunaban), Tamil (tam; Dravidian, South Dravidian), Spanish (Indo-European, Italic), Mawng (Iwaidjan Proper), Towa (tow; Kiowa-Tanoan), Tidore (tvo; North Halmahera, Northern North Halmahera), Bardi (bcj; Nyulnyulan, Western Nyulnyulan), Mandarin Chinese (Sino-Tibetan, Sinitic), Paraguayan Guaraní (gug; Tupian, Maweti-Guaran), Turkish (tur; Turkic, Common Turkic), Southern Yukaghir (yux; Yukaghir, Southern Yukaghir).

similarly to *-ve* in Paraguayan Guaraní (46b), is extensively used as a marker of comparative constructions (Göksel & Kerslake 2005: 176-177).

(46) Paraguayan Guaraní (Tupian, Maweti-Guaran)

a. *o-ho-se-ve*

3ACT-go-DES-CMPR

‘He still wants to go.’ / ‘He wants to go on.’ (Gerasimov 2020: 2; glosses adapted)

b. *che a-mba'apo-ve ndehgui*

I 1SG.ACT-work-more from.you

‘I work more than you.’ (Estigarribia 2020: 249)

From a historical perspective, the meaning ‘also, more’ tends to be older than the continuative meaning. In his study of European languages, van der Auwera (1998: 75-76) lists a number of continuative adverbials which originate from expressions denoting addition or comparison, whereas cases of the semantic development in the other direction, to my knowledge, are not documented.

3.3.2. *Temporal (non-)simultaneity* (‘while’, ‘before’, ‘a while ago’, etc.)

The continuative expressions often serve as markers of temporal simultaneity or non-simultaneity: they may express such meanings as ‘while (still)’, ‘before’, ‘a while ago’, ‘just’, ‘recently’, cf. Figure 6.

Note that in these functions continuative expressions might have syntactic properties fundamentally different from just being a predicate modifier: they may behave not only as standard adverbial expressions but also take a nominal or a whole clause as an argument, cf. English *before* vs. *before Christmas* vs. *before Christmas comes*.

The most common meaning from this semantic group is ‘while (still)’: it is attested in seven languages of the sample.<sup>20</sup> Syntactically the ‘while’-clause can be more or less independent from the other clause: for example, in (47b) the ‘while’-predicate can be analyzed as a depictive, whereas in (48b) the biclausal analysis is preferable.

---

<sup>20</sup> Tagalog (Austronesian, Malayo-Polynesian), Tuvalu (tvl; Austronesian, Malayo-Polynesian), Creek (Muskogean), Ngarinman (nbj; Pama-Nyungan, Desert Nyungic), Lakota (Siouan, Core Siouan), Tadaksahak (dsq; Songhay, Northwest Songhay), Siwai (South Bougainville, Buinic).

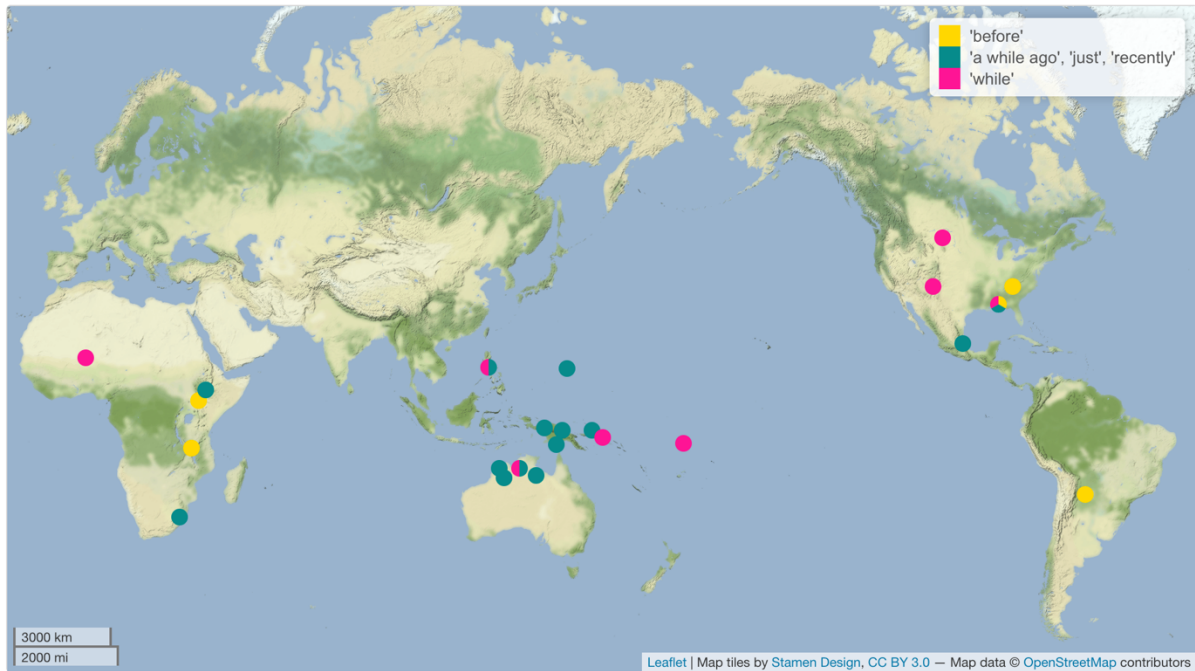


Figure 6: Continuative expressions functioning as markers of temporal (non-)simultaneity.

(47) Ngarinman (Pama-Nyungan, Desert Nyungic; Meakins & Nordlinger 2014: 387; glosses adapted)

- a. *No nyawa = ma = rna janga = rni.*  
 no this = TOP = 1MIN.SBJ sick = ONLY<sup>21</sup>  
 ‘No I’m still sick.’
- b. *Lab ma-na magin-jirri = rni.*  
 pick.up get-PST sleep-ALL = ONLY  
 ‘He takes him while he’s still asleep.’

(48) Tuvalu (Austronesian, Malayo-Polynesian; Besnier 2000: 90, 488; glosses adapted)

- a. *Koi fai vaa ssuaa maaloo ki ssuaa maaloo.*  
 still have poor.relationship a + other government to a + other government  
 ‘The relationship between these countries is still bad.’

<sup>21</sup> The continuative is glossed ONLY because it also has a restrictive meaning, see Section 3.3.3.

- b. [Kee naa vau koe] [koi nofo au i konei].  
 SBJC please come you still stay I at here  
 ‘Please come while I’m still here.’

Examples with the continuative expressions in the ‘before’ function are attested in five languages: Turkana (Nilotic, Eastern Nilotic), Nyakyusa-Ngonde (Atlantic-Congo, Volta-Congo, Bantoid), Cherokee (chr; Iroquoian), Creek (Muskogean) and Nivacle (Matacoan, Mataguayo I). The syntactic structure of these examples is also highly variable. In (49) the continuative marker in Turkana modifies a nominal (‘tomorrow’) as a preposition; the ‘before’-clause in Nyakyusa-Ngonde presented in (50) is formed by the element *bo* ‘as’, the continuative auxiliary and the verb in the infinitive form.

- (49) Turkana (Nilotic, Eastern Nilotic; Dimmendaal 1983: 360)

*tò-boŋ-ù*            *è-ròkò*    *mòyì*  
 IMP-return-VEN    **still**        tomorrow  
 ‘Return before tomorrow.’

- (50) Nyakyusa-Ngonde (nyy; Atlantic-Congo, Volta-Congo, Bantoid; Persohn 2017: 187)

*m-ba-kooliile*            *ukuti*        *m-ba-lagil-e*                    *a-ma-syu*  
 1SG-2PL-call.PFV        COMP        1SG-2PL-dictate-SUBJ        AUG-6-word  
*bo*    ***n-gaal***        *u-ku-fw-a*  
 as    **1SG-PERS**        AUG-15(INF)-die-FV  
 ‘I’ve called you (pl.) to give you instructions before I die.’

At first glance, the meaning ‘before’ seems more natural for nondum (‘not yet’) markers, cf. *I will return while it is not Sunday yet* > *I will return before Sunday*, and indeed, e.g., in Indonesian the word for ‘before’ *sebelum* is based on *belum* ‘not yet’ (Sneddon et al. 2010: 199). That the continuative expressions often have the ‘before’ meaning can be explained in two ways. First, the semantic shift can happen according to the model ‘still to do > not yet done’ discussed in Section 3.3.3. This seems to be the case in (50) where the embedded clause is formed with the infinitive, i.e., literally it means ‘while I am still to die’. Second, according to Jin & Koenig (2020), ‘before’-clauses represent one of the contexts where the phenomenon of expletive negation frequently occurs. In other words, ‘before’-clauses involving negation and non-



involving negation often denote the same situation, and this factor could also contribute to the shift from ‘still’ to ‘not yet’ in the temporal clauses.

A whole group of meanings related to localization of the situation in the past (‘a while ago’, ‘just a moment ago’, ‘just’, ‘(immediately) after’) usually occurs in the perfective contexts, cf. (51). The mechanism of this semantic shift is not clear.

(51) Huehuetla Tepehua (tee; Totonacan, Tepehua; Kung 2007: 468; glosses adapted)

- a. *xa-k-maq-sqoli-y + ka7*  
 PST-1SUBJ-CAUS-whistle-IPFV + JST  
 ‘I still played [music].’
- b. *waa min-li + ka7*  
 FOC come-PFV + JST  
 ‘He just arrived.’

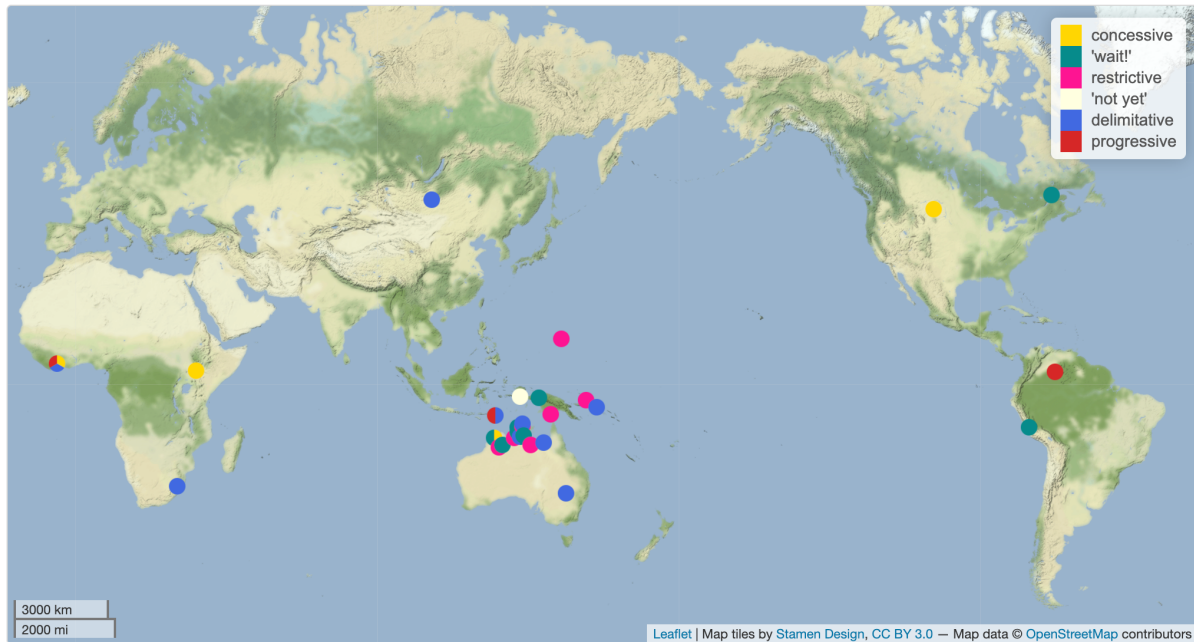
### 3.3.3. Other meanings (concessive, restrictive, delimitative, progressive, etc.)

Several more non-continuative meanings of continuative expressions attested in the data can be seen in Figure 7.

The *concessive* meaning (‘despite, nevertheless’) of the continuative expressions can be illustrated by the English continuative *still*: it is the meaning of *still* in the perfective contexts where the continuative interpretation is not available, cf. (52).

(52) *We told Bill not to come, but he still showed up.* (Michaelis 1993: 193)

The concessive meaning is tightly connected to the “counter-expectation” semantic component which is sometimes discussed as a part of the semantics of the phasal domain in general (Plungian 1999, see Section 2.1.3). Indeed, in the languages in the sample the concessive meaning is often combined with the continuative meaning and results in the sense ‘something is still happening, although it is expected to be over’, cf. (53). Bower (2012: 650) describes the meaning of the Bardi (Nyulnyulan) continuative *gardi* in (53) as follows: “it refers to actions or states which persist, despite the action of the previous clause”.



**Figure 7:** Continuatives which can also express concessive, restrictive, delimitative, progressive, ‘not yet’ and ‘wait!’ meanings.

(53) Bardi (Nyulnyulan, Western Nyulnyulan; Bower 2012: 650; glosses adapted)

<i>Ginyinggon</i>	<i>i-ng-arr-bala-nyji-n,</i>	<i>arra</i>		
then	3-PST-AUG-fight-REFL-CNTS	NEG		
<i>oo-la-rr-m-ala-nyji-n,</i>	<b>gardi</b>	<i>ragal</i>	<i>irr-garda.</i>	
3-IRR-AUG-REFL-fight-REFL-CNTS	<b>still</b>	uninjured	3AUG-body	

‘Then they fought, but it wasn’t a serious fight; their bodies were uninjured.’

Moreover, sometimes languages develop two continuative markers, one of which has a “plain” continuative semantics, while the other one has an obligatory counter-expectation semantic component, see *gaa* vs. *gat* in Nêlêmwa-Nixumwak (Austronesian; Bril 2016: 93) and *jon* vs. *hum* in Mankanya (knf; Atlantic-Congo, North-Central Atlantic, not in the sample; Gaved 2020: 180-184). Overall, continuative expressions which are described as having the concessive/counter-expectation meaning (at least in some of the contexts) are found in four languages: Gban (Mande), Bardi (Nyulnyulan), Nêlêmwa-Nixumwak (Austronesian) and Turkana (Nilotic).

Continuative expressions in three Papuanesian (Chamorro, Austronesian; Taulil, Taulil-Butam, tuh; Anta-Komnzo-Wára-Wérè-Kémä, Yam) and four Australian (Gooniyandi, Bunaban; Wambayan, Mirndi; Bininj Kun-Wok, Gunwinyuan;

Ngarinman, Pama-Nyungan) languages also function as *restrictive* ('only, just') markers. Example (54) shows the continuative and restrictive uses of the particle *ha'* in Chamorro.

(54) Chamorro (cha; Austronesian, Malayo-Polynesian; Chung 2020: 344, 514)

- a. *Mungnga hit manburuka mientras ki*  
 don't we.INCL AGR.INF.make.noise while PRT  
*mamaigu' ha' i neni.*  
 AGR.sleep.PROG EMP the baby  
 'Let's (incl.) not make noise while the baby is still sleeping.'
- b. *Para hami ha' esti na inetnun.*  
 for us.EXCL EMP this L group  
 'This gathering is only for us (excl).'

The relations between the continuative and restrictive meanings are discussed by van Baar (1997: 110-113). In particular, van Baar (1997: 111) analyzes the case of the particle *(-pa/-wa)-rni* in Gurindji (gue; Pama-Nyungan, Desert Nyungic) and concludes that its continuative and restrictive meanings are connected through several intermediate meanings which this particle also demonstrates. Thus, the full semantic scale can be formulated as follows: 'only' — 'right, exactly, really' — 'all the time' — 'still'. Likewise, Evans (1995: 248-249) suggests a diachronic path from continuative to restrictive for the affix *djal-* in Bininj Kun-Wok (Gunwinyguan): 'keep doing A until B', 'still be doing A at reference time, keep on doing A' > 'only do A and no more', 'only do A and not something else one might expect' > 'only'. In addition, both van Baar (1997) and Evans (1995) point out that in some contexts the 'all' meaning, close to the continuative, turns out to be synonymous to 'only', cf. *all that happened was...*. Apparently, such contexts could also facilitate the continuative-restrictive semantic shift.

It is worth noting two non-continuative meanings of continuative expressions which belong to the aspectual domain: delimitative ('for a while, for some period of time') and progressive. Both meanings are close to the continuative in terms of types

of situations they can modify.<sup>22</sup> The *delimitative* meaning, illustrated in (54), is attested in Halh Mongolian (Mongolic-Khitan), Zulu (Atlantic-Congo, Volta-Congo, Bantoid), Gban (Mande), Wardaman (wrr; Yangmanic) and Siwai (South Bougainville).

(54) Halh Mongolian (Mongolic-Khitan, Mongolic; Kullmann 1996: 138; transcription and glosses added)

*ta ger-eesee zahia av-saar l bay-na uu*  
 2SG home-ABL letter receive-CVB PRT COP-NPST Q

‘Are you still receiving letters from home?’ / ‘Have you been receiving letters from home lately?’

The *progressive* meaning is mentioned with respect to the continuative *lé* in Gban (Mande; Fedotov 2015). The second probable example is the marker *-ju* in Puinave (isolate): in (55a) it is used in the continuative meaning, in (55b) it can be interpreted as progressive.

(55) Puinave (pui; Isolate, South America; Higuaita 2008: 262; glosses adapted)

a. *mam-da ka-peu-é-ju* ~ *mam-da ka-ju-peu-é*  
 PR2SG-ASR 3PL-load-AGT-IPFV  
 ‘You are still loading them.’

b. *ja-bêp-di-da-ju ó’o*  
 3SG-work-PST-ASR-IPFV PRNE  
 ‘He was working [when the speaker stopped seeing him].’

The continuative-progressive polysemy also played a significant role in the history of the continuative marker *(te)be-* in Lithuanian. As shown in (Arkadiev 2011, 2019;

---

<sup>22</sup> One more meaning close to the continuative in this respect is the non-phasal meaning ‘keep on doing’ called “continuative” by Bybee et al. (1994). It “involves a continued input of energy and implies that the situation is continued longer than normal” (Bybee et al. 1994: 170), but, as far as I can tell, it does not presuppose the moment *t<sub>i</sub>* preceding reference time when the situation X was also true. This meaning can be illustrated, for example, by the suffix *-poki/ -pokya* in Ese Ejja (ese; Pano-Tacanan, Tacanan; Vuillermet 2012: 478-480) or by the construction *V vienā V-šanā* (lit. V in one V-ing) in Latvian (lav; Indo-European, Balto-Slavic; not in the sample; Nau 2019). My database does not contain clear examples of the “continuative” meaning in terms of (Bybee et al. 1994) colexified with the (phasal) continuative, so it is not considered further.

Holvoet & Kavaliūnaitė 2021), the prefix *be-* has a wide range of construction-specific meanings, including continuative (when used in this meaning, *be-* is reinforced by *te-*, see Section 3.1.2), progressive, avertive (“something was going to happen but did not”) and mirative. Apparently, the avertive construction with *be-* historically developed from the progressive construction with *be-* in the past tense due to conventionalization of interruption implicature (Arkadiev 2011: 49-50; Arkadiev 2019: 103-104). Example (56) shows the progressive-avertive use of *be-* in the context with an interrupted process in Old Lithuanian.

(56) Old Lithuanian (Indo-European, Balto-Slavic; Arkadiev 2011: 49; glosses adapted)

*Tawa tarn-as buw-a be-gan-ans*  
 your servant-NOM.SG AUX-PST CONT-pasture-PRS.PA.NOM.SG.M  
*aw-is sawa Tiew-o, ir ateij-a Lėw-as.*  
 sheep-ACC.PL his father-GEN and come-PST lion-NOM.SG  
 ‘Your servant has been keeping his father’s sheep, and a lion came...’

In six Australian languages and one Papunesian language (Western Dani, Nuclear Trans New Guinea) isolated continuative markers are used as exclamations ‘wait!’ or ‘hang on!’, cf. (57).<sup>23</sup> Supposedly, this meaning is a further development of the concessive meaning: one can introduce her speech in this way if what is going to be said contradicts what has been said before by another person.

(57) Limilngan (Limilngan-Wulna; Harvey 2001: 141, 142; glosses adapted)

- a. *Ø-ayum-iji i-yi-jukgula-rri ulik i-y-im-ambijiwi-rri*  
 IV-go\_back-here 3 < 3AUG-shoot-PL still 3 < 3AUG-IPFV-hit-PL  
 ‘(The planes) had come back. They were shooting. They were still fighting.’
- b. *Captain Gray-in il-ami-ny, ulik,*  
 Captain Gray II-say-PP wait  
 ‘Captain Gray said: Wait!’

Interestingly, in certain contexts continuative markers may express the opposite phasal meaning ‘not yet’. To start with, the English *still* becomes semantically close

<sup>23</sup> Worrorra (Worroran, Western Worroran), Mangarrayi (mpc; Mangarrayi-Maran), Kitja (gia; Jarrakan), Limilngan (lmc; Limilngan-Wulna), Wardaman (Yangmanic), Wageman (Isolate, Australia).

to *not yet* when combined with telic predicates, cf. *I am still crossing the street ~ I haven't crossed the street yet*. There are several more cases attested in the literature: in Kalamang (West Bomberai) the continuative expression is used in the sense 'not yet' when it occurs as a one-word fragment answer to a negative question (58), see also Fanego (2021: 342) on the same pattern in Tachelhit (shi; Afro-Asiatic, Berber, not in the sample).

(58) Kalamang (kgv; West Bomberai; Visser 2020: 388)

a. A: *ka tok sekola*  
2SG still go.to.school  
'Do you still go to school?'

B: *tok*  
*still*  
'Yes [I still go to school].'

b. A: *ka tok sekola = nin*  
2SG yet go.to.school = NEG  
'Don't you go to school yet?'

B: *tok*  
*not.yet*  
'Not yet.'

According to Nurse (2008: 148), some of the continuative markers in Bantu, following the model "We are still to buy > We haven't bought yet", changed their meaning to 'not yet'. Another interesting case is discussed in Veselinova et al. (to appear): the continuative *morã* in Lamaholot (slp; Austronesian, Malayo-Polynesian, not in the sample), when used with atelic predicates, has the continuative meaning, while when used with telic predicates, it has the meaning 'not yet'.

This is, of course, not an exhaustive list of possible meanings of continuative expressions. Due to space constraints, I do not discuss in detail the relatively rare meanings 'first', 'later', 'always', 'throughout', 'together', 'even', 'same', 'forever', 'barely' and several others.

### 3.3.4. Areal patterns

As for polysemy of continuatives in a geographical perspective, two macro-areas clearly stand out as exceptional — Australia and Papunesia. First, according to my data, two non-continuative meanings — restrictive (‘only’) and ‘wait!’ — occur exclusively in Australia and Papunesia (see the previous section). Second, many continuative expressions found in these two macro-areas are enormously multi-functional. The extreme case is the marker *-nyali* in Gooniyandi (Bunaban), which, according to McGregor (1990: 469), has 14 meanings. Other continuative markers normally have at least three-four meanings including continuative. Apparently, the existence of such a polyfunctional marker in the majority of Australian and Papunesian languages can be considered a phenomenon of areal nature.

### 3.4. Semantic effects when combined with negation

The so-called Duality Hypothesis (Löbner 1989) predicts that continuative markers in negative contexts can mean either ‘not yet’ or ‘no longer’. Two meanings are possible because of the different scope of semantic operators: ‘still (not)’ = ‘not yet’, ‘not (still)’ = ‘no longer’. Both strategies are found in the languages of the sample (59)-(60).

(59) Kalamang (West Bomberai; Visser 2020: 391)

a. *ma tok nawanggar*  
 3SG still wait  
 ‘He still waits.’

b. *Nyong esun tok bo-t = nin*  
 N. father.3POSS yet go-T = NEG  
 ‘Nyong’s father doesn’t go yet.’

(60) Lezgian (Nakh-Daghestanian, Daghestanian; Haspelmath 1993: 210; glosses adapted)

*Jusuf.a k’walax-zama-č*  
 Jusuf.ERG work-IPFV.CONT-NEG  
 ‘Jusuf is no longer working.’

In some sources continuative expressions accompanied by the markers of negation are translated into English as ‘still not’, cf. an example from Cherokee (61). As shown by van der Auwera (1993: 625; 2021: 32), the meaning ‘still not’ is not identical to ‘not yet’: ‘still not’ is more emphatic because, in contrast to ‘not yet’, it obligatorily presupposes speaker’s expectation of the contrary (see Section 2.1.3). However, for the purposes of this study, in the encoding of the data I unite the meanings ‘not yet’ and ‘still not’ into one value ‘not yet (still not)’.

(61) Cherokee (Iroquoian; Montgomery & Anderson 2008: 185)

*tla + si*            *yi-uunii-anvhta*  
 NEG + **still**      IRR-3B.PL-know:PRC  
 ‘They still don’t know.’

In addition to ‘not yet (still not)’ and ‘no longer’, three more values of this parameter are distinguished. The value ‘no longer, not yet (still not)’ covers cases where a language has two separate constructions for expressing the meanings ‘no longer’ and ‘not yet (still not)’ based on the same continuative marker, as in Turkana (62).

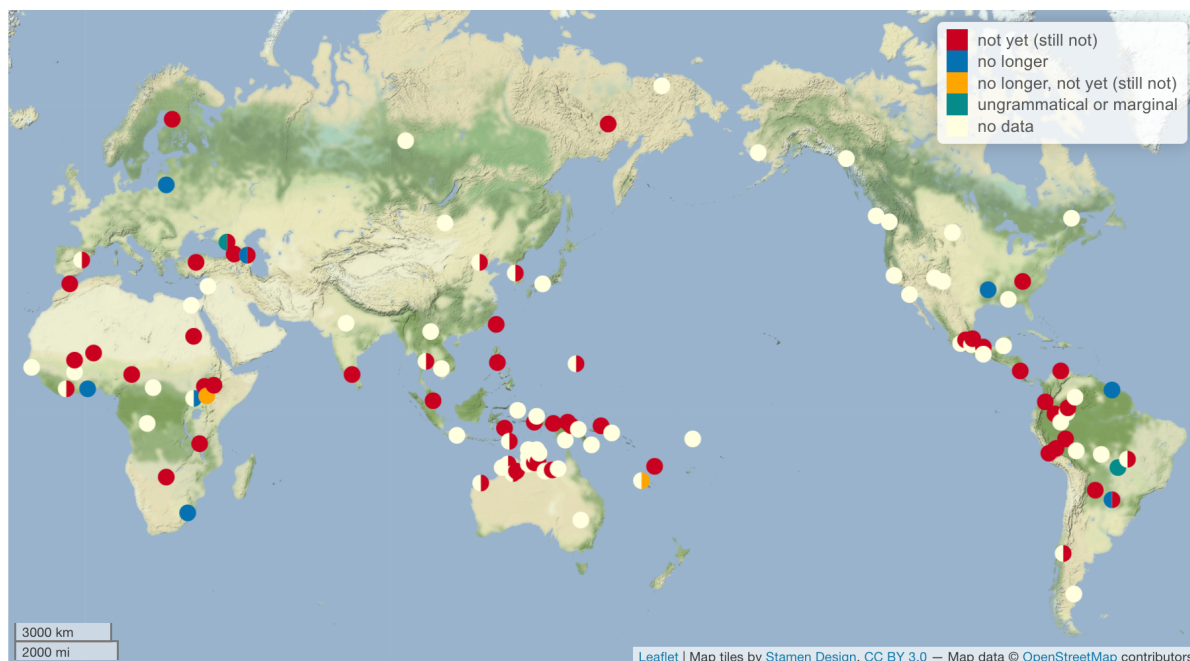
(62) Turkana (Nilotic, Eastern Nilotic; Dimmendaal 1983: 458-459, cited by Kramer 2017: 6; glosses adapted)

- a. *è-ròkò*    *apèse*    *ɲ-è-nap-à*            *ewòrù*    *keɲ*    *lòkiùset*  
**3-still**    girl       not-3-wear-v    cloth    her    wedding  
 ‘The girl does not wear her wedding dress yet.’
- b. *ɲ-è-roko*    *apèse*        *è-nàp-it*            *ewòrù*    *keɲ*    *lòkiùset*  
**not-3-still**    girl            3-wear-ASP    cloth    her    wedding  
 ‘The girl no longer wears her wedding dress.’

The value ‘ungrammatical / marginal’ denotes the situation when the continuative marker is not compatible with negation or its use in the negative contexts is estimated as marginal.

The information on the semantics of continuatives combined with negation is available for 71 out of 159 continuative expressions in the database, and for the other 88 continuative expressions this field was left blank. The geographical distribution of types of continuatives according to this parameter is shown in Figure 8.





**Figure 8:** Semantics of continuative expressions in the negative contexts.

Figure 8 shows that continuative expressions in the negative contexts most frequently have the meaning ‘not yet (still not)’. This pattern appears in the majority of languages in all macro-areas. That derivation of ‘not yet’ expressions from continuative markers is a very widespread strategy in the languages of the world has been noted in the previous literature. In particular, van Baar (1997: 179) mentions that in his sample more than 50% of ‘still’ markers are used to form the meaning ‘not yet’. Diachronically the development of ‘not yet’ markers based on ‘still’ markers has been traced in such languages as Bantu (Veselinova & Devos 2021: 474-477), Austronesian (Veselinova et al. to appear) and even English. In English the adverbial *yet* is an old continuative marker which nowadays is mostly used in the expression *not yet*, whereas in the continuative function it has been almost fully replaced by the new continuative *still* (König & Traugott 1982, van der Auwera 1998: 53). Apparently, similar processes happen in the languages in the sample, e.g., in Wayuu (Arawakan) the use of the continuative suffix *-yülia* in positive contexts is restricted to specific locative phrases, while when combined with negation it can function as a ‘not yet’ marker in all types of clauses (Mansen & Mansen 1984: 535-539).

The meaning ‘no longer’ is generally less preferable and can be considered relatively frequent only in Africa where it is attested five times (cf. also Löfgren (2019: 29) who shows that in Bantu ‘no longer’ is a more frequent option). The diachronic relations between continuatives and ‘no longer’ markers appear to be less

straightforward than between continuatives and ‘not yet’ markers. Specifically, van Baar (1997) calls into question the existence of the diachronic path from continuatives to ‘no longer’ markers. He suggests that “whenever there is a coverage of the STILL/NO LONGER by means of one and the same expression, this is either the result of the development of NO LONGER into STILL or the result of independent development of two different PhP-uses of the same expression” (van Baar 1997: 195). In my sample there are no cases where there would be enough evidence to determine the (in)dependence of the diachronic development of phasal markers and its possible direction, but see van Baar’s (1997: 191-192) suggestions on the evolution of the continuative *ga* in Ewe (Atlantic-Congo, Volta-Congo, Kwa Volta-Congo) from repetitive to ‘no longer’ and then to continuative marker.

Finally, two languages are marked on the map as having continuative expressions which are normally not combined with negation. In fact, the number of such languages must be larger. I suppose that quite a few reference grammars which lack the description of continuatives in the negative contexts do not include it because continuatives are not (widely) used in the negative contexts. One of the reasons for this incompatibility is the existence of the distinct markers expressing the ‘not yet’ (and ‘no longer’) meanings, so the combination of continuative with negation turns out to be redundant. For instance, the Papuan language Moskona (mtj; East Bird’s Head, Meax; Gravelle 2010: 151) has the continuative adverbial *ros* and the separate adverbial *néesa* ‘not yet’. Even though Gravelle (2010) does not say explicitly what happens to *ros* when it is used in the negative context, one may suggest that it is not used as a ‘not yet’ expression because this function is fulfilled by *néesa*.

#### **4. Discussion: maturation of the continuative expressions**

In this section I will pursue an integrative approach of the interplay of parameters of continuative expressions using Dahl’s (2004) notion of *maturity* process. Dahl defines “mature” linguistic phenomena as “those that presuppose a non-trivial prehistory: that is, they can only exist in a language which has passed through specific earlier stages” (Dahl 2004: 2) and adduces such examples of mature phenomena as inflectional morphology, incorporation, and agreement (Dahl 2004: 111-115). Further, Bisang (2015) suggests to distinguish between two types of maturity: morphosyntax-based maturity (overt complexity, on which Dahl focuses) and pragmatics-based maturity (hidden complexity, in Bisang’s terms). Pragmatics-based

maturity is driven by economy and can be illustrated by such phenomena as radical pro-drop and optional (in)definiteness marking in East and Mainland Southeast Asian languages (Bisang 2015: 180-181). In this section, both morphosyntax-based and pragmatics-based maturity are considered.

The least mature type of continuative expressions is synchronically compositional adverbial expressions based on the word ‘now’, e.g., *de yanna* ‘still, until now’ in Isthmus Zapotec (zai; Otomanguean, Eastern Otomanguean; Pickett 2007: 97). As suggested in Section 3.1.2, adverbial expressions of this type, if not yet conventionalized as a lexical item, can be coined at any moment in any language which has the word ‘now’, thus the time needed for their development is minimal. Just created, they do not show any signs of morphologization, are emphatic, do not have non-continuative functions and take the negated predicate in their scope.

The group of continuative expressions showing an initial stage of maturity are verbs with the meanings ‘stay’, ‘remain’, ‘continue’: they can function as continuative markers from the moment they become able to take a predicative complement. Apparently, English verbs like *stay*, *remain*, *continue* and *keep* represent examples of this group of continuative expressions.

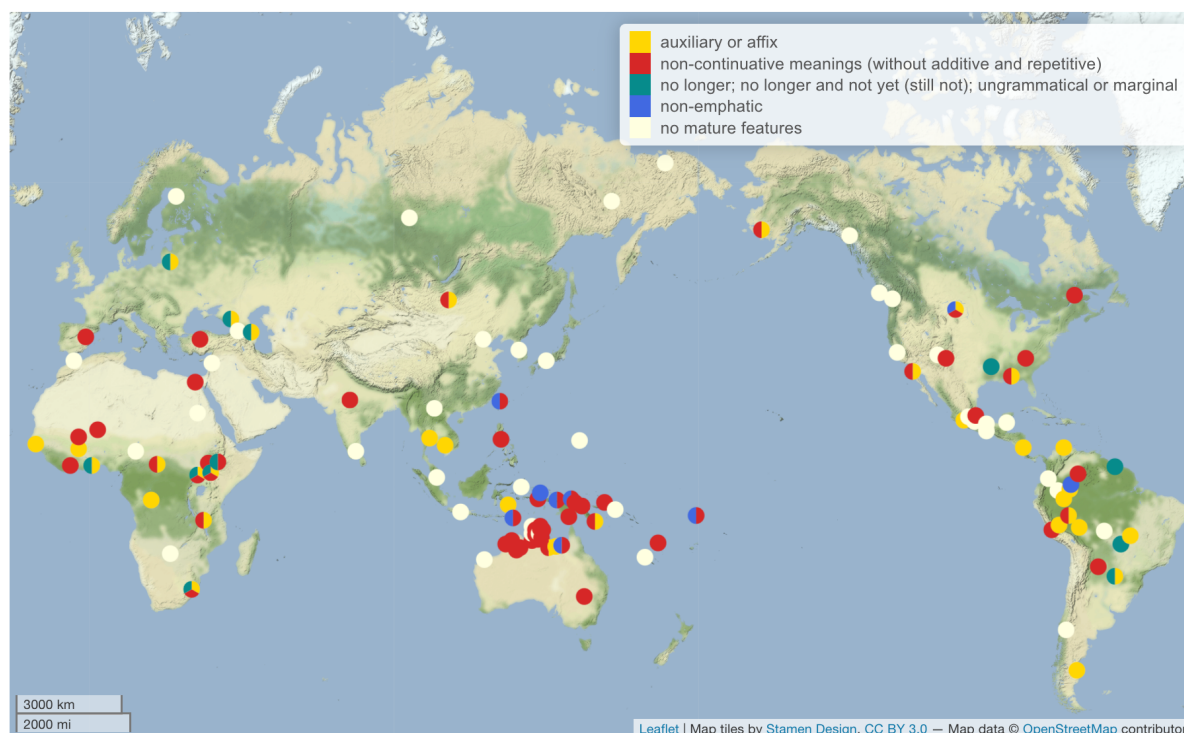
There are many further maturation pathways by which continuative expressions can (although not necessarily have to) gradually change and acquire new features. As for their form, they often undergo univerbation with adjacent elements, cf. the adverb *innum* ‘still’ in Tamil (Dravidian), which historically contains the root *in* ‘the present moment’ and the particle *um* (Dubjanskij 2013: 108). As for their semantics and functions, such expressions may become obligatory in continuative contexts and turn into fully-fledged grammatical markers, see, e.g., Gorbunova (2014), who argues for the grammatical status of the particles *na* ‘still’ and *la* ‘already’ in Atayal (tay; Austronesian, Atayalic). One more example of maturation is the morphological integration of the continuative expression into the predicate, which may result in that the predicate together with the continuative fall under the scope of negation and yield the ‘no longer’ interpretation. Finally, the existence of a high number of non-continuative uses, particularly typical for continuative expressions in the languages of Australia and Papunesia (Section 3.3.5), is also the result of the evolution which must have taken considerable time to occur. Note, however, that the starting point of such developments need not necessarily be a marker with the continuative meaning, its original function underlying a polysemy network (including the continuative as one of its meanings) may be different.

Table 5 represents an attempt to formalize the distinction between less mature and more mature continuatives based on the values of each of the discussed parameters. Note that since the semantic extension of the continuative expressions to additive and repetitive domains can be considered rather trivial, it is associated with less mature features.

parameters	less mature values	more mature values
morphosyntactic type	adverbial phrases	auxiliaries; affixes
emphatic vs. non-emphatic status	emphatic	non-emphatic
non-continuative meanings of continuative expressions	not attested; only additive and/or repetitive	other non-continuative meanings
meaning when combined with negation	not yet (still not)	no longer; no longer and not yet (still not); ungrammatical or marginal

**Table 5:** Less mature and more mature features of continuative expressions.

Figure 9 shows what mature features (if any) are attested for the continuative expressions in the sample. For better visibility only one continuative expression per language (the one that shows more mature features) is indicated on the map.



**Figure 9:** Mature features of continuative expressions.

It is possible to see several areal patterns that have already been discussed in the previous sections: for example, non-emphatic continuative markers tend to occur in Papunesia and Australia, the ‘no longer’ interpretation is most typical for continuative expressions in Africa and the Americas. In addition, Figure 9 shows that the morphosyntactic type “auxiliary or affix” has a weak tendency to combine with the ‘no longer’ interpretation. This link can be explained: on the morphosyntactic grounds, we expect that adverbial continuatives take the negated predicate in their scope and, as a result, convey the meaning ‘still (not)’ > ‘not yet’. Continuative affixes and auxiliaries, in turn, are more likely to themselves fall under the scope of the negative marker and hence express the meaning ‘not (still)’ > ‘no longer’.

As for the continuatives not showing any mature features, it should be kept in mind that their real number is likely to be much higher. The first reason of that inaccuracy is due to their frequent co-existence with more mature continuatives; thus, they are not visible in Figure 9. The second reason is that synchronically compositional continuatives of the type ‘and’ + ‘now’ may be not specifically mentioned in the sources and thus are not included in the database.

## **5. Concluding remarks**

In this paper, I presented a typological study of continuative expressions based on a balanced sample of 120 languages. The continuative expressions were analyzed according to specific parameters. It has been shown that, in terms of morphosyntactic properties, in the vast majority of languages it is possible to encode continuative semantics by an adverb or particle. However, many languages also develop continuative auxiliaries and/or continuative affixes; this is more typical of languages which already feature an elaborate system of auxiliaries and/or affixes. Continuative expressions may be emphatic and non-emphatic. Examples found in the sources indicate that some continuative expressions obligatorily accompany predicates already implying the continuative semantics, and this feature speaks in favor of their non-emphatic status. In addition to the continuative meaning, continuative expressions occur in various non-continuative functions. The most frequent meanings of continuative expressions outside of continuation are related to pluractionality (‘again’, ‘also’, ‘continuously’); other possible meanings are temporal (non-)simultaneity (e.g., ‘while’ and ‘before’), ‘not yet’, restrictive (‘only’), concessive (‘despite’), the meaning ‘wait!’, etc. The rich polysemy of continuative markers is especially common in languages of Australia and Papunesia. When combined with

negation, continuative expressions most frequently have the meaning ‘not yet’ (or the semantically very close meaning ‘still not’), much less frequently — the meaning ‘no longer’.

From a more integrative perspective, continuative expressions vary with respect to the parameter of maturity (Dahl 2004), i.e., the degree of non-triviality of their historical development. While there are many continuative expressions representing the least mature types of continuatives, such as adverbials derived from the word ‘now, the present moment’, we also find continuative expressions which follow one or several maturation pathways towards morphologization, non-emphatic uses, polysemy, less trivial interaction with negation. This study has shown that in all the parameters discussed above the more “mature” values are distributed unevenly across the languages of the sample, and areal, genealogical and structural factors affect the probability of the maturation of continuative expressions. It can be further hypothesized that the non-default, mature properties of continuative expressions, such as ‘being an affix’ or ‘have the additional ‘before’ meaning’, work similarly to more fundamental features of linguistic systems, such as, for example, the presence of ejective consonants, ergativity or VOS word order. Features of this type, first, need specific sociolinguistic conditions to develop: it is generally agreed upon that the probability of development of mature linguistic phenomena is higher in closed or “esoteric” communities, i.e., characterized by small size, dense social networks and low contact (Trudgill 2011). Second, mature phenomena often turn out to be diachronically unstable, i.e., they have low propensity to be inherited and/or borrowed (Nichols 2003) and are therefore prone to be lost. A more detailed account of the social and historical factors influencing the development of continuative expressions requires further studies focusing on the continuative expressions in specific linguistic areas or specific language families. Taking into account the cross-linguistic diversity of continuative expressions described in this study, it will be possible to estimate to what extent continuative expressions vary in geographically and genealogically close languages with respect to different social and historical circumstances.

### **Acknowledgements**

I am grateful to Ljuba Veselinova, Michael Daniel, Maria Koptjevskaja Tamm, Bernhard Wälchli, Peter Arkadiev and two anonymous reviewers for their valuable comments on earlier versions of this paper. All errors are mine.

## Abbreviations

1, 2, 3 = 1 <sup>st</sup> , 2 <sup>nd</sup> , 3 <sup>rd</sup> person	AUX = auxiliaire	SUB = subordination
FGR = falling tone grade	IND = indicative	DEM = demonstrative
PAST5 = remote past	PRS = present	MIN = minimal
A = actor	B = set B Prefix	SUBJ = subjunctive
FOC = focus	INF = infinitive	DER = morphological derivation
PERF = perfect	PRT = particle	NEG = negation
ABL = ablative	BEN = benefactive	SVN = subjective verbal noun
FUT = future	INST = instrumental	DES = desiderative
PERS = persistive	PST = past	NMLZ = nominalization
ACC = accusative	CAUS = causative	T = thematic clitic
FV = final vowel	INTR = intransitive	DU = dual
PFV = perfective	Q = interrogative	NO.COMPL = incomplete/not fulfilled
ACT = active	CMPR = comparative	TOP = topic
G2, G4 = gender	INV = inverse number	DUR = durative
PL = plural	REFL = reflexive	NOM = nominative
ADD = additive	CNTS = continuous	TR = transitive
GEN = genitive case	IPFV = imperfective	DYNM = dynamic
POSS = possession	REP = repetitive	NPRX.AN = animate non- proximal demonstrative
AGR = agreement	COMP = complementizer	V = verb
HAB = habitual	IRR = irrealis	EMP = emphatic particle
PP = past realis perfective	SBJ = subject	NPST = non-past
AGT = agentive	CONT = continuative	VEN = ventive extension
IAM = iamitive	JST = 'just'	ERG = ergative
PR = pronoun	SBJC = subjunctive complementizer	ONLY = restrictive
ALL = allative	COP = copula	VERB = verbalizer
II, IV, 6, 15 = noun classes	L = linker	ESS = essive
PRC = present continuous	SCVN = completive subjective verbal noun	PA = active participle
ASP = aspect marker	CSL = cislocative	YET = persistive
IMP = imperative mood	<sup>l</sup> = low (tone)	EXCL = exclusive
PREP = preposition	SG = singular	PART = participle
ASR = assertive	CVB = converb	
INACP = incomplete pronoun	LOC = locative	
PRNE = non-specific pronoun	STAT = stative	
AUG = augment	DEF = definite article	
INCL = inclusive	M = masculine	
PROG = progressive		

## References

- Ameka, Felix K. 2018. Phasal polarity in Ewe: Diversity of constructions and dialect differences. Paper presented at the International Conference on The expression of Phasal Polarity in sub-Saharan African languages. University of Hamburg, 3–4 February 2018.
- Arkadiev, Peter. 2011. On the aspectual uses of the prefix *be-* in Lithuanian. *Baltic Linguistics* 2. 37–78. <https://doi.org/10.32798/bl.426>
- Arkadiev, Peter. 2019. The Lithuanian “*buvo* + be-present active participle” construction revisited. *Baltic Linguistics* 10. 65-108. <https://doi.org/10.32798/bl.361>
- Besnier, Niko. 2000. *Tuvaluan: A Polynesian Language of the Central Pacific*. London & New York: Routledge.
- Bisang, Walter. 2015. Hidden complexity – The neglected side of complexity and its implications. *Linguistics Vanguard* 1 (1). 177-187. <https://doi.org/10.1515/lingvan-2014-1014>
- Blackings, Mairi & Nigel Fabb. 2003. *A Grammar of Ma'di*. Berlin, New York: De Gruyter Mouton. <https://doi.org/10.1515/9783110894967>
- Bowern, Claire Louise. 2012. *A Grammar of Bardi*. Berlin, Boston: De Gruyter Mouton. <https://doi.org/10.1515/9783110278187>
- British National Corpus (BNC)  
[https://app.sketchengine.eu/#dashboard?corpname=preloaded%2Fbnc2\\_tt21](https://app.sketchengine.eu/#dashboard?corpname=preloaded%2Fbnc2_tt21)
- Bril, Isabelle. 2016. Tense, Aspect and Mood in Nêlêmwa (New Caledonia): Encoding Events, Processes and States. In Zlatka Guentchéva (ed.), *Aspectuality and temporality: Descriptive and theoretical issues*, 63–106. Amsterdam: John Benjamins. <https://doi.org/10.1075/slcs.172.03bri>
- Bybee, Joan & Revere Perkins & William Pagliuca. 1994. *The evolution of grammar: Tense, aspect, and modality in the languages of the world*. Chicago & London: The University of Chicago Press.
- Chung, Sandra. 2020. *Chamorro Grammar*. Santa Cruz: University of California. <https://doi.org/10.48330/E2159R>
- Clendon, Mark. 2014. *Worrorra: A language of the north-west Kimberley coast*. Adelaide: University of Adelaide.
- Creissels, Denis & Séckou Biaye. 2016. *Le balant ganja: phonologie, morphosyntaxe, liste lexicale, textes*. Dakar: IFAN CH.A.DIOP.



- Dahl, Östen. 2004. *The Growth and Maintenance of Linguistic Complexity*. Amsterdam & Philadelphia: John Benjamins. <https://doi.org/10.1075/slcs.71>
- Dahl, Östen & Bernhard Wälchli. 2016. Perfects and iamitives: two gram types in one grammatical space. *Letras De Hoje* 51(3). 325–348. <https://doi.org/10.15448/1984-7726.2016.3.25454>
- Dimmendaal, Gerrit Jan. 1983. *The Turkana Language*. Dordrecht: Foris Publications.
- Dixon, Robert M. W. & Aleksandra Y. Aikhenvald 1999. Introduction. In Robert M. W. Dixon & Aleksandra Y. Aikhenvald (eds.), *The Amazonian languages*, 1–21. Cambridge: Cambridge University Press.
- Donohue, Mark. 1999. *A Grammar of Tukang Besi*. Berlin, New York: Mouton de Gruyter. <https://doi.org/10.1515/9783110805543>
- Dubjanskij, Aleksandr M. 2013. Tamil'skij jazyk [Tamil language]. In Nikita V. Gurov & Aleksandr M. Dubjanskij & Andrej A. Kibrik & Elena B. Markus (eds.), *Jazyki mira. Dravidijskie jazyki* [Languages of the World. Dravidian languages], 47–150. Moscow: Academia.
- Epps, Patience. 2008. *A Grammar of Hup*. Berlin, New York: Mouton de Gruyter. <https://doi.org/10.1515/9783110199079>
- Eraso, Natalia. 2015. *Gramática Tanimuka, una lengua de la Amazonía Colombiana*. Lyon: Université Lumière Lyon 2. (Doctoral dissertation.)
- Estigarribia, Bruno. 2020. *A Grammar of Paraguayan Guarani*. London: UCL Press.
- Evans, Nicholas. 1995. *A Grammar of Kayardild: With Historical-Comparative Notes on Tangkic*. Berlin, New York: De Gruyter Mouton. <https://doi.org/10.1515/9783110873733>
- Evans, Nicholas. 2003. *Bininj Gun-Wok: A Pan-Dialectal Grammar of Mayali, Kunwinjku and Kune*. Canberra: Research School of Pacific; Asian Studies, Australian National University.
- Fabre, Alain. 2016. *Gramática de la lengua Nivacle (familia Mataguayo, Chaco Paraguayo)*. Kangasala, Finlandia. (Available online at [https://etnolingustica.wdfiles.com/local--files/biblio%3Afabre-2016/Fabre\\_2016\\_Gramatica\\_Nivacle.pdf](https://etnolingustica.wdfiles.com/local--files/biblio%3Afabre-2016/Fabre_2016_Gramatica_Nivacle.pdf) Accessed on 01-01-2023)
- Fanego, Axel. 2021. Phasal Polarity in Amazigh varieties. In Raija Kramer (ed.), *The Expression of Phasal Polarity in African Languages*, 335–364. Berlin, Boston: De Gruyter Mouton. <https://doi.org/10.1515/9783110646290-015>
- Fedotov, Maksim. 2015. To be continued... : priključenija kontinuativnogo pokazatelja v jazyke gban [To be continued...: adventures of the continuative

- marker in Gban]. Paper presented at Twelfth Conference on Typology and Grammar for Young Scholars. Saint Petersburg, 19–21 November 2015.
- Gaved, Timothy. 2020. *A grammar of Mankanya: An Atlantic language of Guinea-Bissau, Senegal and the Gambia*. Amsterdam: LOT.
- Genko, Anatolij N. 1955. *Abazinskij jazyk. Grammatičeskij očerk narečija tapanta* [The Abaza language. A grammatical sketch of the Tapanta dialect]. Moscow: Izd. Akademii nauk SSSR.
- Gerasimov, Dmitrij. 2020. Mini-course on Paraguayan Guarani at HSE. Handout.
- Gorbunova, Irina M. 2014. Kategorija fazovoj poljarnosti v atajal'skom jazyke [A category of phasal polarity in Atayal]. *Voprosy Jazykoznanija* 3. 34–54.
- Göksel, A. & Celia Kerslake. 2004. *Turkish: A comprehensive grammar*. London, New York: Routledge.
- Gravelle, Gloria J. 2010. *A Grammar of Moskona: An East Bird's Head Language of West Papua, Indonesia*. Vrije Universiteit Amsterdam. (Doctoral dissertation.)
- Hammarström, Harald & Mark Donohue. 2014. Some Principles on the Use of Macro-Areas in Typological Comparison. *Language Dynamics and Change* 4(1). 167-187. <https://doi.org/10.1163/22105832-00401001>
- Hammarström, Harald & Robert Forkel & Martin Haspelmath & Sebastian Bank. 2021. *Glottolog* 4.8. Leipzig: Max Planck Institute for Evolutionary Anthropology. <https://glottolog.org>
- Hanson, Rebecca. 2010. *A grammar of Yine (Piro)*. Bundoora, Victoria: La Trobe University. (Doctoral dissertation.)
- Harvey, Mark. 2001. *A Grammar of Limilngan: A Language of the Mary River Region Northern Territory Australia*. Canberra: Research School of Pacific; Asian Studies, Australian National University.
- Haspelmath, Martin. 1993. *A Grammar of Lezgian*. Berlin, New York: Mouton de Gruyter. <https://doi.org/10.1515/9783110884210>
- Heath, Jeffrey. 2016. A grammar of Nanga (Dogon language family, Mali). Unpublished MS.
- Hercus, L. A. 1982. *The Bāgandji Language*. Canberra: Research School of Pacific; Asian Studies, Australian National University.
- Higuita, Jesús Mario Girón. 2008. *Una Gramatica Del Wansöjöt (Puinave)*. Utrecht: LOT.
- Holvoet, Axel & Gina Kavaliūnaitė. 2021. The Lithuanian mirative present and its history. *Baltic Linguistics* 12. 413–439.

- Iwasaki, Shoichi & Preeya Ingkaphirom. 2005. *A Reference Grammar of Thai*. Cambridge: Cambridge University Press.
- Jin, Yanwei & Jean-Pierre Koenig. 2020. A cross-linguistic study of expletive negation. *Linguistic Typology* 25 (1). 39–78. <https://doi.org/10.1515/lingty-2020-2053>
- Kemmer, Suzanne. 1990. *Still*. Paper presented at the Fourth Annual UC Berkeley-UC San Diego Cognitive Linguistics Workshop.
- Khanina, Olesya. 2008. How universal is wanting? *Studies in Language* 32(4). 818–865. <https://doi.org/10.1075/sl.32.4.03kha>
- Klein, Wolfgang. 1994. *Time in Language*. London: Routledge.
- Klyagina, Evgeniya & Anastasia Panova. 2019. Phasal Polarity in Abaza. *HSE Working Papers in Linguistics* 89. 1-24.
- König, Ekkehard & Elizabeth Closs Traugott. 1982. Divergence and apparent convergence in the development of *yet* and *still*. In *Annual Meeting of the Berkeley Linguistics Society*, 170–179. <https://doi.org/10.3765/bls.v8i0.2029>
- Kramer, Raija. 2017. Position paper on Phasal Polarity expressions. Hamburg: University of Hamburg. Unpublished MS. <https://www.aai.unihamburg.de/afrika/php2018/medien/position-paper-on-php.pdf>
- Kramer, Raija. 2021. Introduction: The expression of phasal polarity in African languages. In Kramer, Raija (ed.), *The Expression of Phasal Polarity in African Languages*, 3–24. Berlin, Boston: De Gruyter Mouton. <https://doi.org/10.1515/9783110646290-002>
- Kullmann, Rita & Dandii-Yadam Tserenpil. 1996. *Mongolian Grammar*. Hong Kong: Jensco.
- Kung, Susan Smythe. 2007. *A Descriptive Grammar of Huehuetla Tepehua*. The University of Texas at Austin. (Doctoral dissertation.)
- Li, Charles N. & Sandra A. Thompson 1989. *Mandarin Chinese: A Functional Reference Grammar*. Berkeley, Los Angeles, London: University of California Press.
- Löbner, Sebastian. 1989. German *Schon - Erst - Noch*: An integrated analysis. *Linguistics and Philosophy* 12. 167–212. <https://doi.org/10.1007/BF00627659>
- Löfgren, Althea. 2019. Phasal Polarity in Bantu Languages: A typological study. Stockholm: Stockholm University. (MA thesis.)
- Maisak, Timur & Samira Verhees. The Still Not Present in Andi: discerning the grammaticalization source. Unpublished MS.

- Maho, Jouni Filip. 2008. Comparative TAM morphology in Niger-Congo: The case of persistive, and some other markers in Bantu. In Folke Josephson & Ingmar Söhrman (eds.), *Interdependence of Diachronic and Synchronic Analyses*, 283–298. Amsterdam: John Benjamins. <https://doi.org/10.1075/slcs.103.14mah>
- Mansen, Karis B. & Richard A. Mansen 1984. *Aprendamos Guajiro: Gramática Pedagógica de Guajiro*. Bogotá: Editorial Townsend.
- Martin, Jack B. 2011. *A Grammar of Creek (Muskogee)*. Lincoln, London: University of Nebraska Press.
- McGregor, William B. 1990. *A Functional Grammar of Gooniyandi*. Amsterdam: John Benjamins. <https://doi.org/10.1075/slcs.22>
- Meakins, Felicity & Rachel Nordlinger. 2014. *A Grammar of Bilinearra: An Australian Aboriginal Language of the Northern Territory*. Berlin, Boston: De Gruyter Mouton. <https://doi.org/10.1515/9781614512745>
- Meeussen, Achille E. 1967. Bantu grammatical reconstructions. *Africana linguistica* 3. 79–121.
- Merdanova, Solmaz R. 2004. *Morfologija i grammatičeskaja semantika agul'skogo jazyka (na materiale xpjukskogo govora)* [Morphology and grammatical semantics of Agul (data from the Huppuq dialect)]. Moscow: Sovetskij pisatel'.
- Michaelis, Laura A. 1993. 'Continuity' within Three Scalar Models: The Polysemy of Adverbial Still. *Journal of Semantics* 10(3). 193–237. <https://doi.org/10.1093/jos/10.3.193>
- Miller, Amy. 2001. *A Grammar of Jamul Tiipay*. Berlin, New York: De Gruyter Mouton. <https://doi.org/10.1515/9783110864823>
- Miyaoka, Osahito. 2012. *A Grammar of Central Alaskan Yupik (CAY)*. Berlin, Boston: De Gruyter Mouton. <https://doi.org/10.1515/9783110278576>
- Montgomery-Anderson, Brad. 2008. *A Reference Grammar of Oklahoma Cherokee*. University of Kansas. (Doctoral dissertation.)
- Moroz, George. 2017. *lingtypology: easy mapping for Linguistic Typology*. <https://CRAN.R-project.org/package=lingtypology>
- Murane, Elizabeth. 1974. *Daga Grammar: From Morpheme to Discourse*. Norman: Dallas, Texas: The Summer Institute of Linguistics and the University of Texas at Arlington.
- Mushin, Ilana. 2012. *A Grammar of (Western) Garrwa*. Berlin, Boston: De Gruyter Mouton. <https://doi.org/10.1515/9781614512417>

- Müller, Neele. 2014. Language Internal and External Factors in the Development of the Desiderative in South American Indigenous Languages. In Loretta O'Connor & Pieter Muysken (eds). *The Native Languages of South America: Origins, Development, Typology*, 203–222. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9781107360105.012>
- Nassenstein, Nico & Helma M. Pasch. 2021. Phasal polarity in Lingala and Sango. In Raija L. Kramer (ed.), *The Expression of Phasal Polarity in African Languages*, 93–128. Berlin: De Gruyter Mouton.
- Nassenstein, Nico & Helma Pasch. Phasal polarity in Lingala and Sango. In Raija L. Kramer (ed.), *The Expression of Phasal Polarity in African Languages*, 93–128. Berlin, Boston: De Gruyter Mouton. <https://doi.org/10.1515/9783110646290-006>
- Nau, Nicole. 2019. The Latvian continuative construction *runāt vienā runāšanā* ‘talk in one talking’ = ‘keep talking’. *Baltic Linguistics* 10, 21–63. <https://doi.org/10.32798/bl.360>
- Nichols, Johanna. 2003. Diversity and stability in language. In Brian D. Joseph & Janda, Richard D. (eds.), *Handbook of historical linguistics*, 283–310. Oxford: Blackwell. <https://doi.org/10.1002/9780470756393.ch5>
- Nordlinger, Rachel. (1998). *A Grammar of Wambaya, Northern Territory (Australia)*. Canberra: Research School of Pacific; Asian Studies, Australian National University.
- Nurse, Derek. 2008. *Tense and aspect in Bantu*. Oxford: Oxford University Press.
- Onishi, Masayuki. 1994. *A Grammar of Motuna (Bougainville, Papua New Guinea)*. Canberra: Australian National University. (Doctoral dissertation.)
- Ostrowski, Norbert. 2011. Pochodzenie litewskiego afiksu duratywnego *teb(e)-*. *Folia Scandinavica Posnaniensia* 12. 205–210.
- Ostrowski, Norbert. 2016. Lithuanian discontinuatives *nebe-* / *jau nebe-* ‘no more, no longer’ and German-Lithuanian language contacts. *Folia Scandinavica Posnaniensia* 20 (1). 175–179. <https://doi.org/10.1515/fsp-2016-0035>
- Oxford, Will. 2007. Towards a grammar of Innu-aimun particles. Memorial University of Newfoundland. (MA thesis.)
- Persohn, Bastian. 2017. *The Verb in Nyakyusa: A Focus on Tense, Aspect, and Modality*. Berlin: Language Science Press. <https://doi.org/10.5281/zenodo.926408>
- Persohn, Bastian. 2021. Phasal polarity in Nyakyusa (Bantu, M31). In Raija L. Kramer (ed.), *The Expression of Phasal Polarity in African Languages*, 129–160. Berlin, Boston: De Gruyter Mouton. <https://doi.org/10.1515/9783110646290-007>

- Pickett, Velma. 2007. *Vocabulario zapoteco del Istmo: español-zapoteco y zapoteco-español*. México: Instituto Lingüístico de Verano.
- Plungian, Vladimir A. 1999. A Typology of Phasal Meanings. In Abraham, Werner & Kulikov, Leonid (eds.), *Tense-Aspect, Transitivity and Causativity: Essays in Honour of Vladimir Nedjalkov*, 311–22. Amsterdam, Philadelphia: John Benjamins.  
<https://doi.org/10.1075/slcs.50.21plu>
- Priestley, Carol. 2008. *A grammar of Koromu (Kesawai), a Trans New Guinea language of Papua New Guinea*. Canberra: Australian National University. (Doctoral dissertation.)
- Pjurbeev, Grigorij C. (ed.). 2001. *Bol'shoj akademičeskij rusko-mongol'skij slovar'* [The Big Academic Russian-Mongolian Dictionary]. Moscow: Academia.
- Rongier, Jacques. 2004. *Parlons Éwé: Langue du Togo*. Paris: L'Harmattan.
- Ross, Claudia & Jing-heng Sheng Ma. 2014. *Modern Mandarin Chinese Grammar: A Practical Guide*. London: Routledge.
- Samarin, William J. 1970. *Sango: Langue de L'Afrique Centrale*. Leiden: Brill.
- Schapper, Antoinette. 2022. *A Grammar of Bunaq*. Berlin, Boston: De Gruyter Mouton.  
<https://doi.org/10.1515/9783110761146>
- Smeets, Ineke. 2008. *A Grammar of Mapuche*. Berlin, New York: Mouton de Gruyter.  
<https://doi.org/10.1515/9783110211795>
- Sneddon, James Neil & Alexander Adelaar & Dwi Noverini Djenar & Michael C. Ewing. 2010. *Indonesian Reference Grammar. 2nd edition*. London: Allen & Unwin.
- Stoynova, Natalia M. 2013. *Pokazateli refractiva* [Refractive markers]. Moscow: AST-Press.
- Trudgill, Peter. 2011. *Sociolinguistic typology: Social determinants of linguistic complexity*. Oxford: Oxford University Press.
- Ullrich, Jan. 2018. *Modification, Secondary Predication and Multi-Verb Constructions in Lakota*. Heinrich-Heine-Universität Düsseldorf. (Doctoral dissertation.)
- van Baar, Theodorus Martinus. 1997. *Phasal Polarity*. University of Amsterdam. (Doctoral dissertation.)
- van der Auwera, Johan. 1993. 'Already' and 'still': Beyond duality. *Linguistics and Philosophy* 16. 613–653. <https://doi.org/10.1007/BF00985436>
- van der Auwera, Johan. 1998. Phasal adverbials in the languages of Europe. In Johan van der Auwera (ed.), *Adverbial Constructions in the Languages of Europe*, 25–145. Berlin, New York: De Gruyter Mouton.  
<https://doi.org/10.1515/9783110802610.25>

- van der Auwera, Johan. Phasal polarity – warnings from earlier research. In Raija L. Kramer (ed.), *The Expression of Phasal Polarity in African Languages*, 25–38. Berlin, Boston: De Gruyter Mouton. <https://doi.org/10.1515/9783110646290-003>
- Veselinova, Ljuba. 2015. Not-yet expressions in the languages of the world: a special negator or a separate cross-linguistic category? Paper presented at Diversity Linguistics: Retrospect and Prospect. Leipzig, Max Planck Institute for Evolutionary Anthropology, 1-3 May 2015.
- Veselinova, Ljuba & Maud Devos. 2021. NOT YET expressions as a lexico-grammatical category in Bantu languages. In Raija L. Kramer (ed.), *The Expression of Phasal Polarity in African Languages*. Berlin, Boston: De Gruyter Mouton, 445-496. <https://doi.org/10.1515/9783110646290-019>
- Veselinova, Ljuba & Leif Asplund & Jozina Vander Klok. To appear. Phasal polarity in Malayo-Polynesian languages of South East Asia. In Adelaar, Alexander & Schapper, Antoinette (eds.), *The Oxford Guide to the Malayo-Polynesian Languages of South East Asia*. Oxford: Oxford University Press.
- Vuillermet, Marine. 2012. *A Grammar of Ese Ejja, a Takanan language of the Bolivian Amazon*. Université Lumière Lyon 2. (Doctoral dissertation.)
- Visser, Eline. 2020. *A grammar of Kalamang: The Papuan language of the Karas islands*. University of Lund. (Doctoral dissertation.)
- Wilson, Jennifer. 2017. *A grammar of Yeri: A Torricelli language of Papua New Guinea*. State University of New York at Buffalo. (Doctoral dissertation.)
- Xiao, Richard & Tony McEnery. 2004. *Aspect in Mandarin Chinese. A corpus-based study*. Amsterdam, Philadelphia: John Benjamins. <https://doi.org/10.1075/slcs.73>
- Yumitani, Yukihiro. 1998. *A phonology and morphology of Jemez Towa*. University of Kansas. (Doctoral dissertation.)
- Zahran, Aron & Eva-Marie Bloom Ström. 2022. Against expectations – the rise of adverbs in Swahili phasal polarity. *Studies in African Linguistics* 51 (2). 295–323. <https://doi.org/10.32473/sal.51.2.129687>
- Ziervogel, Dirk & Jacobus A. Louw & P. Taljaard 1967. *A Handbook of the Zulu Language*. Pretoria: J.L. van Schaik.

#### CONTACT

anastasia.b.panova@gmail.com