Linguistic Typology at the Crossroads

ISSN 2785-0943

Volume 2 – issue 1 – 2022

Issue DOI: https://doi.org/10.6092/issn.2785-0943/v2-n1-2022

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

Responsible Editor

Caterina Mauri, University of Bologna
Department of Modern Languages, Literatures and Cultures, via Cartoleria 5, 40124 Bologna. Email: caterina.mauri@unibo.it

Linguistic Typology at the Crossroads

CONTENTS

Linguistic Typology at the Crossroads 2-1 (2022): 1-36 ISSN 2785-0943

Extending the typology: negative concord and connective negation in Persian

JOHAN VAN DER AUWERA[1], SEPIDEH KOOHKAN[2]
[1]UNIVERSITY OF ANTWERP, [2]TARBIAT MODARES UNIVERSITY

Submitted: 26/11/2021 Revised version: 9/02/2021
Accepted: 14/02/2021 Published: 14/04/2022

Abstract

This paper aims to advance the general understanding of negative concord (as in English *We don't need no education*) and connective negation (as in English *neither … nor'*) through an analysis of Persian. For negative concord with indefinites the analysis highlights differences between human vs. non-human and pronominal vs. nominal negative concord. It also deals with the problem that *hič*, the word that arguably marks negation in negative indefinites, also has a non-negative emphatic meaning in questions. For the relation between negative concord and connective negation the paper suggests the importance of two new parameters: (i) are the connective negator and the normal clausal negator similar? and (ii) can one of two negatively connected phrases precede the verb and the other follow it?

Keywords: negative concord; connective negation; negative indefiniteness; negative polarity; emphasis

1. Introduction

In this paper we intend to increase our understanding of negative concord and of its relation to connective negation through the lens of Persian. So we study how Persian expresses what corresponds to English *nobody* and the like and we focus on whether Persian has structures like substandard English *we don't need no education*. This clause contains two negative markers, viz. *don't* and *no education* and one takes them to agree with each other or 'to be in concord'. Languages can have obligatory negative concord, nowadays called 'strict', or not obligatory or 'non-strict' negative concord.

Persian is interesting for three reasons: (i) Persian has non-strict negative concord, but it is of a special kind, (ii) the strategies that Persian uses for *nobody*, *nothing* and *no education* are not identical, (iii) the marker that turns the indefinite into a negative indefinite has a non-negative emphatic use in questions. We also investigate how Persian connective negation, our term for 'neither … nor' structures, relates to negative concord.

The Persian examples are transliterated with the system described on http://www.eki.ee/wgrs/rom1_fa.htm. We use the same system in the references, except for the names of authors, where we keep the spelling that the authors themselves use. For glossing we rely on the Leipzig Glossing Rules (https://www.eva.mpg.de/lingua/resources/glossing-rules.php). For the sake of uniformity, we 'retransliterate' and 'regloss' examples from the literature according to the said systems. In one example from earlier Persian we use 'New Persian' for the language from 700 onwards, with first 'Early New Persian' and then 'Classical Persian' (Windfuhr 2009: 9; Windfuhr & Perry 2009: 533). 'New Persian' follows 'Middle Persian'. By itself, 'Persian' stands for the contemporary language, also known as 'Farsi', as spoken in Iran. The language has various dialects, e.g. Isfahani, Mashhadi, Shirazi, Yazdi, Qazvini, Tehrani, etc. The dialect that we discuss is dominant in the national media, which is very close to the dialect spoken in Tehran and which is the native variant of one of the authors. All our contemporary examples are in the informal (spoken) register.

2. Negative indefiniteness

2.1. 'Nobody'

How does one express negative indefiniteness in Persian? The matter is complex. (1) sketches at least the four main strategies to express what corresponds to English *nobody*.[1]

[1] In (1) the negative indefinite is subject. The syntactic function is not relevant: when the negative indefinite is object, for instance, we find the same four strategies.

(1) a. *hič-kas* *zang* *na-zad-Ø.*
NIND-HUM.SG ring NEG-hit.PST-3SG

b. *hič-kas-i*
NIND- HUM.SG-IND

c. *hič-ki*
NIND- INT.HUM.SG

d. *kas-i*
HUM.SG-IND
'Nobody called.'

The strategies in (1a) and (1d) are mentioned by Haspelmath (1997: 282-283) and Rasekh-Mahand (2015: 209). Haspelmath also includes (1b) and he adds that the pattern in (1d) can be preceded by *yek* 'one'. *Yek-i* may also occur on its own.

(2) a. *ye* *kas-i* *n-ist-Ø* *az* *in* *be-pors-e*
one HUM.SG-IND NEG-be-3SG from PROX SBJV-ask.PRS-3SG
dār-i *či-kār* *mi-kon-i!*
have.PRS-2SG INT.NHUM-task IPFV-do.PRS-2SG
'There isn't anybody to ask him what you are doing.'

b. *yek-i=o* *na-did-am* *ke* *az* *in* *xoš=eš*
one-IND=ACC NEG-see.PST-1SG REL from PROX like=3SG
bi-yā-d.
SBJV-come.PRS-3SG
'I didn't see somebody who likes him.'

The use of *yek* demands more research: we assume that the above constructions make the indefinite outscope negation, conveying that 'somebody to ask him what you are doing was not there' in (2a) and 'somebody who likes him was not seen' in (2b).

In the strategies in (1) we see two morphemes that diachronically relate to the numeral 'one'. The *-i* suffix derives from *aiva* 'one' (Briceño Villalobos 2019: 131) and it now functions as a marker of indefiniteness. *Hič* derives from *aiva* 'one', too, with

aiva followed by a free choice marker **-cid*, which originally was an additive and/or scalar marker 'also, even' (Briceño Villalobos 2019: 133). The current status of *hič* is controversial. We consider its appearance in (1) to mark negative indefiniteness, glossed as 'NIND', but the matter is complex and will be dealt in some detail in section 2.4. In (1c) we see *-ki*. We hypothesize that *-ki* is the interrogative human pronoun *ki* 'who', *pace* Mahootian & Gebhardt (1997: 211), who derive it from *kas.* We posit that *hič-ki,* a relatively new and informal construction, is built on the crosslinguistic frequent pattern of an indefinite formation with an interrogative pronoun (Haspelmath 1997: 26-27, 170-176; Van Alsenoy 2014: 26-27) with, in this case, the form *hič,* which serves negative indefiniteness, as it does with *-kas* and *-kas-i* in (1a) and (1b). The fourth morpheme we see in the indefinites in (1) is *kas*. This is historically an indefinite pronoun, ultimately based on an interrogative stem. At least since Middle Persian (Juan Briceño Villalobos, p.c.) *kas* is a hybrid form, advancing towards noun status on a degrammaticalization path (cf. Willis 2007; Norde 2009: 143-145). The progression is rather strong: at least for the contemporary language Lambton (1957: 33), Lazard (1992: 124), Haspelmath (1997: 283) and Yousef (2018: 71) all simply consider it to be a noun. It is not quite a normal noun yet, though. It cannot function in a sentence like (3).

(3) *ye* **kas=e* / *ādam=e* *jālebi=o* *did-am.*
one HUM.SG=EZ human=EZ interesting=ACC see.PST-1SG
'I saw an interesting person'.

Also, when we compare the expression of indefiniteness of *kas* with that of 'normal' nouns (in Table 1 below), we will see that *kas* does not behave like a normal noun. The intermediate pronoun – noun status of *kas* is reflected in the Boyle (1966) grammar. In the section on indefinite pronouns (Boyle 1966: 25) *kas* is first glossed as 'someone', but then Boyle goes on to call it a 'noun', but one that can 'serve as a pronoun'. The treatment in Mace (2003: 74) is similar: *kasi* is mentioned in the section of pronouns, but not as a pronoun but as an 'indefinite noun'. Cross-linguistically, the distinction between generic nouns and indefinite pronouns meaning 'somebody' or 'something' is often not a sharp one (see e.g. van der Auwera & Krasnoukhova 2021), to the extent even that Haspelmath (2005) includes both nominal and pronominal counterparts to e.g. *somebody* as 'indefinite pronouns'. In this paper we steer clear of this issue and gloss *kas* as 'human, singular' ('HUM.SG'). It is clear, however, that *kas*

is not indefinite by itself. In examples (1a), (1b) and (1d) it is made indefinite by either *hič-* or *-i* or by both.

(4) **kas* *zang* *na-zad-Ø.*
HUM.SG ring NEG-hit.PST-3SG
'Nobody called.'

2.2. 'Nothing'

In (5) we show how Persian expresses what corresponds to English *nothing*.

(5) a. *hič-či*
NIND-?
b. *hič-čiz-i*
NIND-NHUM.SG-IND
c. *čiz-i*
NHUM.SG-IND
'He said nothing.'
na-goft-Ø.
NEG-say.PST-3SG

The strategy in (5c) is listed by both Haspelmath (1997: 282-283) and Rasekh-Mahand 2015: 209) and the former also lists the one in (5b). Enriching or replacing the indefinite forms with *yek* seems possible under the same circumstances as for 'nobody' – see (2). Like for humans, we see an old interrogative stem turning into a noun (Boyle 1966: 25; Lazard 1992: 124). Neither Haspelmath (1997: 282-283) nor Rasekh-Mahand (2015: 209) list the form *hič-či* of (5a), but instead mention a *hič-čiz* form, which is not used in the variety we deal with. One could propose that *hič-či* is a short form of *hič-čiz,* but *hič-či* could also be a short form of *hič-čiz-i. Hič-či* could furthermore derive from *hič-če*, in which *če* is a non-human interrogative, and under this analysis *hič-či* is a counterpart of *hički* shown in (1c).

Interestingly, Boyle (1966: 25), Lazard (1992: 125) and Yousef (2018: 75) all mention a bare *hič* strategy. This is not accepted in the variety that we study, except in set phrases like *be hič* 'for nothing' or *hič šodan* 'be destroyed' (lit. 'nothing become').

(6) a. *ham=ro be hič foruxt-Ø.*
all=ACC to NIND.NHUM.SG sell.PST-3SG
'(S)he sold everything for nothing.'

b. *hame zahmat-ā=m hič šod-Ø.*
all effort-PL=1SG.POSS NIND.NHUM.SG become.PST-3SG
'All my efforts are destroyed.'

Earlier, bare *hič* did have the 'nothing' use.

(7) 10th century (Afshar & Afshari eds. 2006: 135)
tork=e bičāre hič ne-mi-dānest-Ø.
Turk=EZ poor NIND.NHUM.SG NEG-IPFV-know.PST-3SG
'Poor Turk knows nothing.'

For both human and non-human indefiniteness we see that negation can be expressed solely on the verb – (1d) and (5c). The negative verb then combines with the form that is also used for positive indefiniteness. The latter use is illustrated in (8).

(8) a. *kas-i zang zad-Ø.*
HUM.SG-IND ring call.PST-3SG
'Somebody rang the bell.'

b. *čiz-i goft-Ø.*
NHUM.SG-IND say.PST-3SG
'He/she said something.'

This means that *kas-i and čiz-i* are polarity-neutral. That specific and negative indefiniteness can use the same, polarity-neutral marking is cross-linguistically well-attested: it is arguably even the most frequent strategy world-wide (van der Auwera & Van Alsenoy 2016, 2018).

2.3. Negative noun phrases

Neither Haspelmath (1997) nor Rasekh-Mahand (2015) pay attention to nominal negative indefiniteness. As (9) shows, nominal negative indefiniteness differs from the pronominal one.

(9) a. **hič dānešāmuz* *emtehān=o qabul na-šod-Ø.*
NIND student exam=ACC pass NEG-become.PST-3SG

b. *hič dānešāmuz-i*
NIND student-IND

c. **dānešāmuz-i*
student-IND

d. **dānešāmuz*
student

'No student passed the exam.'

The nominal pattern with *hič* but without *-i* in (9a) corresponds to pronominal (1a) and, possibly, (5a), but whereas the pronominal strategies are grammatical, nominal (9a) is ungrammatical.[2] The two pronominal and the nominal b examples, with *hič* and *-i,* are all grammatical. The strategy with *-i* but without *hič,* i.e., (1d) and (5c) for the pronouns, are grammatical, but the nominal one, i.e., (9c), is not, at least in the intended 'no student' sense – it is grammatical when it means that a student didn't pass the exam. Then there is the bare nominal use in (9d). Bare pronoun uses with *kas* or *čiz* are ungrammatical – see (1d) and the discussion around (6) and (7). The bare nominal use is ungrammatical too, i.e., in the intended sense, for we get a definite sense – 'The student didn't pass the exam'. In the plural, however, an indefinite generic sense is possible.

(10) *dānešāmuz-ā emtehān=o qabul na-šod-an.*
student-PL exam=ACC pass NEG-become.PST-3PL
'Students didn't pass the exam.'

[2] Mahootian & Gebhardt (1997: 89, 160, 211) list three examples of *hič* in front of a bare noun. This divergence could be due to dialect or register differences.

This sentence needs a special context, for it is or was not a generic property of students to fail exams. A clearer use of the generic sense is illustrated in (11) and (12a), and in (12b) the generic sense does not require a plural.

(11) *ali ketāb ne-mi-xun-e.*
Ali book NEG-IPFV-read.PRS-3S
'Ali doesn't read books.'

(12) a. *gorbe nun ne-mi-xor-e.*
cat bread NEG-IPFV-eat.PRS-3SG
'A cat doesn't eat bread.'

b. *gorbe-hā nun ne-mi-xor-an*(*d*).
cat-PL bread NEG-IPFV-eat.PRS-3PL
'Cats don't eat bread.'

(11) is an example of Kwak's (2010: 62). Her translation is 'Ali doesn't read a book', which fails to bring out the generic sense. A study of the interaction of negation and generic readings goes beyond this study. Suffice it for us to note that a generic reading lifts the need for negative concord, though the negative concord version allows a generic reading too.

(12) c. *hič gorbe-i nun ne-mi-xor-e.*
NIND cat-IND bread NEG-IPFV-eat.PRS-3SG
'No cat eats bread.'

The expression of nominal negative indefiniteness thus contrasts with the expression of pronominal negative indefiniteness. We have already noted that the expression of negative indefiniteness is sensitive to the human - non-human parameter and we now see that the pronoun – noun parameter is even more important. We summarize the patterns in Table 1. In all cases the verbs are negative.

Pronominal		**Nominal**
human	**non-human**	
		N
hič-HUM.SG-*i*	*hič*-NHUM.SG-*i*	*hič*-N-*i*
hič-HUM.SG	*hič*-NHUM.SG	
hič-INT.HUM.SG		
HUM.SG-*i*	NHUM.SG-*i*	

Table 1: Negative indefiniteness in Persian

That Persian allows a nominal strategy without *hič* for a generic reading is interesting. That nominal and pronominal strategies can be different is not new. 'We the linguists' have been aware of this since the seminal work of Bernini & Ramat (1992: 201-226; 1996: 181-199). Without using the term 'negative concord', they discuss negative concord in European languages with pronouns and nouns in the same chapter. They analyse seventeen European languages as exhibiting obligatory or, in the current jargon, 'strict' negative concord for pronouns, but there is not a single one that they analyse as having strict negative concord for nouns. Bernini & Ramat (1992; 1996) also discuss the non-strict pronominal negative concord that depends on word order, as illustrated with Italian (13), which has negative concord when the negative indefinite follows the verb, but not when it precedes.

(13) Italian (Indo-European)

a. *(*Non*) *ho* *visto* *nessuno.*
NEG have.PRS.1SG see.PTCP.PST.MSG NIND.HUM
'I have not seen anybody.'

b. *Nessuno* *mi* (**non*) *ha* *visto.*
NIND.HUM 1SG.ACC NEG have.PRS.3SG see.PTCP.PST.MSG
'Nobody has seen me.'

They do not discuss this parameter for nouns, for lack of space, but it is clear from their example, shown in (14), that negative concord works in a different way.

(14) Italian (Indo-European; Bernini & Ramat 1992: 227; 1996: 199).

a. *Giovanni* *(*non*) *mangia* *nessun* *pesce.*
Giovanni NEG eat.PRS.3SG no fish
'Giovanni doesn't eat any fish.'

b. *Giovanni* *non* *mangia* *pesce.*
Giovanni NEG eat.PRS.3SG fish
'Giovanni doesn't eat fish.'

In (14) the indefinite nominal is postverbal, so if nominal negative concord functioned like the pronominal one, negative concord should be obligatory. But it is not, as is shown in (14b), and the versions with and without negative concord have a different meaning. What is more, Bernini & Ramat (1992: 227; 1996: 199) give (14b) a generic paraphrase, viz. 'not eating fish is a characteristic of John'. So in this respect the non-concordial nominal patterns of Farsi and Italian might be identical or, at least, similar. However, this needs more research. Schwarze (1995: 768), for example, describes the difference between (14a) and (14b) differently. He does not appeal to a +/- generic parameter but to +/- partitive and +/- emphatic parameters. In particular: the concordial pattern in (14a) goes with a partitive or an emphatic sense, absent in the non-concordial pattern in (14b). Is 'non-partitive non-emphatic' the same as 'generic'? If (14a) can be partitive, how is its partitivity different from the construction with the partitive *di* determiner, illustrated in (14c)?[3]

(14) Italian (Indo-European)

c. *Giovanni* *non* *mangia* *del* *pesce* *dal* *1997*
Giovanni NEG eat.PRS.3SG PART.DEF.M.SG fish from.DEF.M.SG 1997
'Giovanni hasn't been eating fish since 1997.'

It is clear that 30 years after Bernini & Ramat (1992) the presence or absence of nominal negative concord is still on the agenda (see also van der Auwera & De Lisser 2010; van der Auwera, to appear).

[3] On Italian partitives see Garzonio & Poletto (2020) and Cardinaletti & Giusti (2020) and the references therein.

2.4. The status of hič

Whether the indefinite is a noun or a pronoun, it can be preceded by *hič*. Judging by the current literature this constellation allows three analyses: either *hič* is negative and then the collocation with a negative verb is a matter of negative concord, as in substandard English (15a), or *hič* is a negative polarity item ('NPI'), as in standard English (15b), or it is an in-between item, which sometimes shows negative behavior and sometimes negatively polar behavior.

(15) a. *You ain't seen nobody yet.*
b. *You haven't seen anybody yet.*

Each of the three views is represented in the literature. Mansouri (2004) and Kwak (2010) advocate the first analysis, Taleghani (2008: 89) and, probably but implicitly, Erschler & Volk (2011: 135), are in line with the second one. Kahnemuyipour (2017: 7) as well as, implicitly, Haspelmath (1997: 282-284) adopt the third one. Interestingly, Turkish borrowed Persian *hič,* it functions in a similar (though not quite identical) way and we may find the same three approaches: a negative analysis in Jeretič (2018; Görgülü 2020), a negative polarity one in Kelepir (2001: 155-167; Görgülü 2017), and an (implicit) in-between view in Haspelmath (1997: 196, 286-287) (cf. also Suleymanova & Hoeksema 2018: 193, adopting this analysis for Azerbaijani *heç/hiç*).

The NPI analysis is *prima facie* supported by the fact that a form such as *hič-kas* can also occur in polar questions, as in (16).[4] We gloss this use of *hič* as 'NPIND' ('negatively polar indefinite').

(16) *hič-kas* *zang* *zad-Ø*?
NPIND-HUM.SG ring hit.PST-3SG
'Did anybody call?'

Interrogatives are NPI contexts. If the interrogative use shows that *hič-kas* is an NPI, it supports the view that it is an NPI in negative sentences too. We do, after all, see English *anybody* in questions too:

[4] What is to be discussed in this section is also valid for the other constructions with *hič*. For simplicity's sake the examples all use the *hič-kas* form.

(17) *Did anybody call*?

The parallel with *anybody* is limited, however. There is only one non-negative NPI context that *hič-kas* appears in and in this respect it differs from English *anybody*. The latter is allowed in the NPI contexts of conditionals and comparatives, as in (18a-b).

(18) a. *If anybody calls, let me know.*
b. *He can sing better than anybody I know.*

Perhaps *hič-kas* had a wider use earlier. Thus (19) shows a conditional use, no longer possible in present-day Persian (20).[5]

(19) Early New Persian (11th c.) (Ahmadi Givi 2001: 1586)

rasul	*goft-Ø:*	*agar*	*hič-kas*	*az*	*išān*	*tamiz*
prophet	say.PST-3SG	if	NPIND.HUM.SG	from	3PL	discernment
dārad-Ø	*ān*	*xodāvad=e*	*šotor*	*ast-Ø.*		
have.PRS-3SG	that	God=EZ	camel	be.PRS-3SG		

'The prophet said: "If any of them has the sense of discernment, that is the God of the camel".'

(20)

payāmbar	*goft-Ø*	*age*	*qarār*	*bāš-e*	*kas-i*
prophet	say.PST-3SG	if	arrangement	be.SBJV-3SG	HUM.SG-IND
**hič-kas*	*tašxis*	*be-d-e,*	*un*	*xodāy=e*	*šotor=e.*
NPIND-HUM.SG	discriminate	SBJV-give-3SG	DIST	God=EZ	camel=3SG

'The prophet said: "If anyone is going to judge, that is the God of the camel".'

But, independently of whether there was this wider use or independently of the extent of the wider use, for both the interrogative and the negative use one could adopt an NPI analysis.

This cannot be the whole story though. Kwak (2010) convincingly argues that

[5] It is noteworthy that the contemporary language has a special negative use of *age* 'if', as in (a) – a variation on an example in Anvari (2003: 520).

(a)

be	*xodā*	*age*	*hič-kas*	*hazer*	*bāš-e*	*in*	*kar=o*	*bo-kon-e.*
to	God	if	NIND-HUM.SG	ready	be.SBJV-3SG	PROX	task=ACC	SBJV-do.PRS-3SG

'I swear to God that nobody is ready to do that task.'

when *hič-kas* appears in a negative context, it is truly negative, and not just negatively polar. She appeals to tests that distinguish between the two types of elements. We illustrate these with English, and then come to Persian. First, English *nobody* is negative and *anybody* is negatively polar, and in a negative context, only the former can be modified by *almost.*

(21) a. *I called almost nobody.*
b. **I didn't call almost anybody*.

This test shows that *hič-kas* is negative.

(22) *man* *taqriban* *be* *hič-kas* *zang* *na-zad-am.*
1SG almost to NIND-HUM.SG ring NEG-hit.PST-1SG
'I called almost nobody.'

Second, in elliptic answers, the 'nobody' sense is conveyed with a bare *nobody*, but not with a bare *anybody.* The same holds for *hič-kas.*

(23) A: *Who came?* B: *Nobody / *Anybody*.

(24) A: *ki* *umad-Ø?* B: *hič-kas*.
INT.HUM.SG come.PST-3SG NIND-HUM.SG
'Who came?' 'Nobody.'

Of course, Persian *hič-kas* and English *nobody* are still different, for in (1) *hič-kas* shows negative concord and *nobody* does not.

If *hič-kas* is negative and if the non-elliptic clause requires a negative on the verb, then this pattern is a negative concord pattern. There are different types of negative concord. It is clear that the negative concord is 'non-strict' – (1d), (5c), (11) and (12a-b) are examples without *hič* – yet not in the 'classical' sense, illustrated with Italian (13). Classical non-strict NC depends on the position of the indefinite relative to the verb. This does not seem to be relevant, since in Persian the verb comes late.[6] There are many other types of non-strict negative concord, but what we find in Persian is

[6] The basic word order in Persian is now arguably moving towards SOV (Dabir Moghaddam 2013: 129), but this does not affect the negative marking on the indefinite.

not included in the typology of van der Auwera & Van Alsenoy (2018) nor in the extensions discussed in van der Auwera (2017). In Haspelmath's (1997) 40-language sample, there are parallels in Hindi (1997: 180-181) and Turkish (1997: 286-297). Hindi, like Persian, allows a polarity neutral pronoun (like *kas*), viz. *koi,* with a negative verb, but the pattern it alternates with, viz. *koi* followed by an originally additive-scalar particle *bhii,* has many more uses than the interrogative one found in Persian (Bhatia 1978: 60; Lahiri 1998; Kumar 2006). So Hindi is similar but not identical – and the same goes for Oriya, discussed by Van Alsenoy (2014: 493-496), and probably other Indo-Aryan languages (Bhatia 1978: 68). Turkish *hič* is also similar but different. Whereas there is alternation for *bir şey* 'something' (lit. 'one thing') in the example from Haspelmath (1997: 287), represented as in (25), for 'nobody', as in (26), *hič* cannot be dropped.[7]

(25) Turkish (Turkic; Şahin Beygu, Nisan Ece Gümüş p.c.)

a. *Hiç bir şey* *gör-me-di-m.*
NIND one thing see-NEG-PST-SG
b. *Bir şey*
one thing
'I didn't see anything.'

(26) Turkish (Turkic; Şahin Beygu, Nisan Ece Gümüş, p.c.)

a. *Hiç bir-i* *gel-me-di.*
NIND one.3SG.POSS come-NEG-PST
b. **Bir-i*
one-3SG.POSS
'Nobody came.'

So much for the use of negative *hič-kas* in negative contexts and the alternation with polarity neutral *kas-i*. We still have to account for the interrogative use, illustrated in (16). Kwak, the linguist that argues for negative concord, does not discuss the interrogative use. If we now consider both the interrogative and the negative use, we propose that *hič-kas* has a double status. In the negative context it is negative and in

[7] Just like in Persian there are other alternatives, and they have *hiç*.

the interrogative context it is not. And, given the prominence in the literature of the diachronic scenario from NPIs to negative elements, already sketched in the above, a good bet would be that in interrogative contexts *hič-kas* is an NPI element. Just like for the analysis that takes both the interrogative and the negative uses to involve an NPI, the fact that the only non-negative use is an interrogative one is not a problem.

This is still not the full story. Though both interrogative *anybody* and interrogative *hič-kas* are negative polarity items, they are different, and not only because *anybody* has a wider use. Interrogative *hič-kas* also differs from interrogative *anybody* in that the former has an emphatic meaning, characterized by Lazard (1992: 125) and Yousef (2018: 75) by 'at all' and by Lambton (1957: 33) by 'whatsoever'. (27) is Lambton's example.

(27) *hič* *nān*[8] *dār-id*?
NPIND bread have.PRS-2PL
'Do you have any bread whatsoever / at all?' / 'Do you have ANY bread?'

Thus the emphasis can be said to bring in an element of 'free choice' or, a better term, 'widening' (Kadmon & Landman 1993).[9] The *hič* question is furthermore – this is our observation – a biased one: it comes with the expectation of a negative answer and we hypothesize that this follows from the widening: by stressing that the hearer should do his/her best to consider every possible breadlike object, the speaker is preparing for a negative answer.

Interestingly, *hič* is also claimed to allow an adverbial widening sense – Lambton (1957: 34), Lazard (1992: 125) and Yousef (2018: 75). We think that this sense is associated with the positive use of a small number of verbs referring to cognitive processes, including 'know', 'think', 'understand', 'doubt' and 'see', and that it is

[8] In Lambton's example the noun does not carry the indefiniteness marker *-i*. In our variety of Persian the indefiniteness marker is necessary. A bare noun would be appropriate for a habitual or a generic reading.

[9] 'Widening' is a better term because 'free choice' is strongly associated with a use of *any* as in *Anybody can do this* and *hič* does not allow this use.

dialectal,[10] possibly influenced by Turkish or Kurdish.[11]

(28) a. *hič* *mi-dun-i* *age* *be-fahm-e* *či*
at.all IPFV-know.PRS-2SG if SBJV-understand.PRS-3SG INT.NHUM
mi-š-e?
IPFV-become.PRS-3SG
'Do you know at all what will happen if (s)he understands (the truth)?'

b. *hič* *mi-fahm-i* *dār-i* *či* *mi-g-i?*
at.all IPFV-understand.PRS-2SG have.PRS-2SG INT.NHUM IPFV-say.PRS-2SG
'Do you at all understand what you are saying?'

Negative (pro)nominal *hič* does not have the widening sense. Assuming that negative *hič* developed from negative polarity *hič,* the widening sense has bleached. Bleaching is frequently found in diachronies of negation. Thus in a classical Jespersen Cycle involving a minimal unit expression like French *pas* (originally) 'step', *point* (originally) 'point' or *mie* (originally) 'crumb', we see a 'not at all' sense, which comes from a 'not even a minimal thing' like a step, point or crumb. The 'not at all' sense then bleaches into the non-emphatic, i.e., non-widened 'not' (van der Auwera 2009). Something like this is a possible scenario for *hič.* The interrogative use is not merely a negative polarity sense, it is a negative polarity sense with widening. When this sense turned into a purely negative sense, it underwent two changes: it didn't only lose the negative polarity sense, but the widening as well. That emphasis or widening

[10] This needs more work. But at least we checked (28a) and b with 31 informants from different parts of Iran: four speakers from the North (Gilan province), 11 speakers from the West, (Kurdistan, Shahri-e Kord, Malayer, Tabriz), two speakers from the South (Jahrom and Sistan; two speakers), and 14 speakers from the center (Tehran, Irak, Qazvin, Yazd, and Isfahan). Except for the Kurdish informant and a Qazvini informant, whose native language is Turkish, they are all monolingual. Only three speakers accepted (28a), all from Shahri-e Kord in the West. (28b) fares better, with six speakers accepting it (four from Shahri-e Kord in the West, one from Jahrom in the South, and the bilingual Turkish Persian speaker from Qazvin in the center) and two speakers that are uncertain.

[11] For Turkish it is clear that the adverbial widening use is not restricted to verbs of cognition nor to an interrogative context (Kelepir 2001: 122; Görgülü 2017: 54). So adverbial *hič* has a wider use in Turkish than in Persian. Pronominal and nominal *hič*, however, seem more restricted in Turkish than in Persian. At least, native speakers disagree about the acceptability of the Turkish counterparts to (16) and (27) (Kelepir 2001: 164; Görgülü 2017: 54).

play a role in the development of negative indefinites is well-known (see e.g. Gianollo 2018:137-288; 2020: 554-55). It seems to us that the constellation in Persian, with an emphatic interrogative pronoun and a non-emphatic negative one, has not been documented yet.[12]

It is interesting to compare Persian with Dutch. In Dutch the negative 'nobody' pronoun is *niemand*. Its main use is negative, its morphology transparently says 'NEG.iemand', i.e., 'not somebody'.

(29) Dutch (Indo-European)
Ik heb niemand gezien.
1SG have.PRS.1SG NIND.HUM see.PTCP.PST
'I have seen nobody.'

But it also occurs in a question and it give the question a bias. This makes it similar to Persian *hič-kas*. The bias, however, when *niemand* is not stressed, is positive.[13] In that respect it is just like the Dutch clausal negator (Geerts et al. 1984: 1063).

(30) Dutch (Indo-European)

a. *Is dat niet de broer van Maria?*
be.PRS.3SG DIST.SG NEG DET.DEF brother.SG of Maria
'Isn't that Maria's brother?
b. *Heb je daar niemand gezien?*
have.PRS.SG 2SG there NIND.HUM.SG see.PTCP.PST
'Haven't you seen somebody there?'

The positive bias can be visualized if we rephrase the questions in (30) with English

[12] A much discussed pattern, prominent in the work of Giannakidou (e.g. 1998), of a link between emphasis and negative indefiniteness is what we find in Modern Greek. Modern Greek negative indefinite pronouns can be emphasized prosodically. The emphatic version is strongly associated with the negative use, the non-emphatic version much less so, for it is allowed in a wide variety of NPI contexts. In Persian it is the non-emphatic use that is associated with negation.

[13] When *niemand* is stressed, (30b) becomes a neutral question asking whether it is true or not that the hearer saw nobody.

tag questions that have a falling tone.[14] The tag question separates the assumption, expressed in the declarative part, from the question, expressed in the tag, and the assumption is each time positive.

(31) a. *That is the brother of Mary's, isn't it*?
b. *You have seen somebody there, haven't you?*

So in Dutch it is the negative meaning of *niemand* and *niet* that gives the question the positive bias. We see the same thing with the Persian clausal negator.

(32) *in* *barādar=e* *maryam* *n-ist-Ø?*
PROX brother=EZ Maryam NEG.be.PRS-3SG
'Isn't this Maryam's brother?'/ 'That is Maryam's brother, isn't it?'

In questions *hič-kas* is different from the Persian and Dutch clausal negators and from Dutch interrogative *niemand:* the latter three markers are negative in questions but the former, i.e., interrogative *hič-kas*, is not. Just why the presence of negation proper in a question can yield a positive bias is an issue that is beyond this paper – see Romero (2020) for a description of the state of the art. Beyond this paper is also the question how one should integrate the dimension of speaker bias on a semantic map or that of widening/emphasis, for that matter. These two issues were explicitly mentioned as tasks for later research in Haspelmath (1997: 82-86, 128) and they retain this status now – but see Van Alsenoy (2014: 321-346) for an attempt (cf. also Fălăuş 2013).

It is important to point out that our analysis of interrogative *niemand* does not imply that when an otherwise negative pronoun occurs in questions, it has to remain negative and bring along a bias. Negative pronouns can turn into NPIs. This kind of scenario can be the result of analogical pressure, with the negative pronoun adapting to the negative polarity pronouns in whose paradigm it entered. The change from negation to negative polarity can also be due to negative concord progressively locating the negative meaning solely on the clausal negator, thus allowing the negative pronoun to be reinterpreted as a negative polarity pronoun. The former

[14] With a rising tone, we get a neutral question. The difference between the rising tone and the falling tone in English is thus parallel to the difference between stressing and not stressing the negative pronoun in Dutch. We are grateful to a reviewer for this observation.

scenario has been argued by Haspelmath (1997: 213-233) for Romance pronouns like Spanish *ningun* (see also Breitbarth et al. 2020: 161-164), the second one for Jamaican Creole *nobadi* (van der Auwera & De Lisser 2019; van der Auwera, to appear).

2.5. Conclusion

We have argued that Persian negative pronominal and nominal indefinites do not pattern the same way and that there is a difference even in the two pronominal constellations, i.e., the one for 'nobody' and the one for 'nothing'. Each of them shows non-classical non-strict negative concord. For nominals, we have argued that negative concord is sensitive to genericity: a generic reading allows a non-concordial pattern. We have also suggested that *hič*, the morpheme that makes the pronoun or noun negative, is originally a negative polarity element with a widening component, which survives in questions. The *hič* that serves negative indefiniteness has narrowed the negative polarity to negation and it has lost the widening. These language-specific properties need to be accommodated in the general theory of negative concord and, more generally, negative indefiniteness.

Of course, many other things need to be investigated. Thus, we would want to find out whether the hypotheses can be extended to other negative indefinite words and phrases, like the counterparts to *never* or *nowhere*. One also needs a thorough study of 'negative spread', i.e., the co-occurrence of more than one negative indefinite. On this subject Haspelmath (1997: 221) claims that when a form with *hič* combines with a form without *hič*, the one with *hič* has to come first. Example (33) is Haspelmath's.

(33) a. *kas-I* *čiz-i* *na-šnid-Ø.*
HUM.SG-IND NHUM.SG-IND NEG-hear.PST-3SG

b. *hič-kas* *hič-či*
NIND-HUM.SG NIND-NHUM.SG

c. *hič-kas* *čiz-i*
NIND-HUM NHUM.SG-IND

d. * *kas-I* *hič-či*
HUM.SG-IND NIND-NHUM.SG
'Nobody heard anything.'

We found no support for this observation. In other words, the variant in (33d) is fine. Of course, there could be speaker variation and perhaps the variant in d is (for some speakers) less preferred.

We also need to study Haspelmath's 'indirect negation', i.e., the contexts in which the counterpart to a simple clausemate clausal negator is a superordinate negator, a negative verb like *lack* or a privative marker corresponding to English *without.* In the next section, we turn to another issue which has received little or no attention, viz., the expression of connective negation and its relation to negative concord. Last but not least, for the diachrony our cursory remarks invite serious corpus work on the older stages of the language.

3. Negative concord and connective negation

3.1. Cross-linguistic variation

Connective negation, also known as 'emphatic negative coordination' (Haspelmath 2007, Salaberri 2022) and 'correlative negation' (Briceño Villalobos 2019), is illustrated in (34) with *neither … nor.*

(34) *The man neither went nor came. He stayed right there.*

A working definition has connective negators as elements that both connect and negate structurally identical elements and contain no additional semantics. Thus, *neither* is different from *and* and *not* in (35) – *and* only connects and *not* only negates. Similarly, in (36) only *neither* is a connective negator, different from *either* in (37) and independently of the fact that the first part of (36) does not contain a connective negator and the connection is expressed twice (i.e., with *and* and with *neither*).

(35) *The man did not come and he did not go. He stayed right there.*
(36) *The man did not come and neither did he go. He stayed right there.*
(37) *The man did not come and he did not go either. He stayed right there.*

Lest is not a connective negator either, even though it connects and negates, but the connected elements are structurally different and the structure has, in this case,

additional apprehensional semantics.[15]

(38) *The man showed me the map, lest I get lost.*

How this pattern interacts with negative concord has not been the subject of much work. What we know is based on languages in which the connective negator is clearly different from the clausal negator, viz. the Balto-Slavic ones, Spanish and French (see van der Auwera 2021, van der Auwera et al. 2021, and the references therein) as well as Turkish (Şener & İşsever 2003; Jeretič, 2018). Thus Spanish, for instance, has a clausal negator *no*, but the main connective negator is *ni*.[16]

(39) Spanish (Indo-European)

a. *No he visto a nadie.*
NEG have.PRS.1SG see.PTCP.PST.M.SG to NIND.HUM
'I have seen nobody.'

b. *No somos (ni) de izquierda-s ni de derecha-s.*
NEG be.PRS.1PL CONEG of left-PL CONEG of right-PL
'We are neither from the left nor from the right.'

What we also know is that one must make a distinction between at least three connective negator uses, which may or may not have dedicated markers. In (40) the connective negators scope over a finite verb. Van der Auwera et al. (2021) call it the 'finite' type, different from the 'phrasal' and 'clausal' types.

(40) Spanish (Indo-European)
Ni puedo ni debo exponer-la a
CONEG can.PRS.1SG CONEG must.PRS.1SG espose.INF-3SG.ACC.F to
ciertos riesgos.
certain.M.PL risk.M.PL
'I neither can nor should expose her to certain risks.'

[15] This definition is similar to the one in Salaberri (2022), the main difference being that he follows Haspelmath (2007: 17-19) and requires the construction to encode emphasis. While we don't deny that connective negation may be emphatic, we remain to be convinced that emphasis should be part of the definition.

[16] All the examples in this section come from van der Auwera (2021) and van der Auwera et al. (2021).

(39b) illustrates the 'phrasal' type, the scoped phrases are *de izquierda-s* and *de derecha-s*. The third type is the 'clausal' type, and here the connective negators connect clauses. This is illustrated in Slovenian (41).

(41) Slovenian (Indo-European)
Niti programa ne zna napisati niti
CONEG program.GEN.SG.M NEG know.PRS.3SG write.INF.PFV CONEG
kongresa ne zna narediti.
congress.GEN.SG.M NEG know.IPFV.PRS.3SG make.INF.PFV
'Neither does he know how to write the program nor does he know how to organize a congress.'

For each type there is variation as to how connective negation combines with negative concord. Thus, for the phrasal type, one would expect that when the language has negative concord with negative indefinites, it would exhibit the same type of negative concord with negatively connected phrases. This holds true for Latvian, for instance. It has strict negative concord for negative indefinites, and we see the same for negatively connected phrases: the connected phrases are, of course, negative, and so is the verb, independently of the order of the negative phrases and the verb.

(42) Latvian (Indo-European)
Nedz Telma, nedz Jozefs nespēja pakustēties.
CONEG Telma.NOM.SG CONEG Jozefs.NOM.SG NEG.can.PST.3 PVB.move.INF.RFL
'Neither Telma nor Jozefs could move.'

In Spanish, just like in Italian, illustrated in (13), negative concord is 'classically' non-strict with the position of the indefinite relative to the verb determining whether or not there is negative concord. This holds true for connective negative phrases too, but there is nevertheless something special, something 'quirky' (van der Auwera 2021), also related to word order. When the connective negative phrases follow the verb, the first connective negator may remain absent. This is illustrated in (39b) with the bracketed '(*ni*)'.

Another quirky feature is that when the first connective negator is absent, French and Spanish allow the construction to scope over negative indefinites. This is illustrated with French (43).

(43) French (Indo-European)
Rien *ni* *personne* *changera* *ça.*
NIND.NHUM CONEG NIND.HUM change.FUT.3SG DIST.SG
'Nothing or nobody will change that.'

For the clausal type, one might expect there to be no need for a clausal negator. In Lithuanian, for instance, connective clausal negators and ordinary ones are incompatible, as illustrated in (44) (cf. Jeretič 2018 for Turkish), but in Slovenian they are compatible – see (41).

(44) Lithuanian (Indo-European)
Nei *aš* *jam* (**ne-*)*patinku* *nei* *jis*
CONEG 1.SG.NOM 3SG.DAT NEG-like.PRS.IPFV.1.SG CONEG 3SG.NOM
man (**ne-*)*patinka.*
1SG.DAT NEG-like.PRS.IPFV.3SG
'He does not like me and neither do I like him.'

In virtue of its clausal scope, one would furthermore expect a clausal connective negator to control negative concord inside the clause in the way an ordinary clausal negator does. But this is never the case – at least in the languages studied so far. Croatian, for instance, has strict negative concord with an ordinary negator, but not with the connective negator *niti.* In (45) *niti* goes with the Croatian 'anybody' and 'somebody' pronouns, but not with 'nobody'.

(45) Croatian (Indo-European)
Niti *je* **nikoga* / *ikoga* /
CONEG be.PRS.3.SG NIND.HUM.ACC.SG / NPIND.HUM.ACC.SG /
nekoga *vidio,* *niti* *ga*
IND.HUM.ACC.SG see.PTCP.PST.SG.M CONEG 3.SG.ACC.M
je *djevojka* *upozorila.*
be.PRS.3.SG girl.NOM.SG warn.PTCP.PST.SG.F
'He neither saw anybody/somebody nor did the girl warn him.'

The finite type behaves in a similar way. But there is one extra feature. If the clausal negator is expressed in a prefinite slot and the connective negator is of the finite type,

we have two contenders for the prefinite slot. The presence of the connective negator may make the presence of the ordinary superfluous (and even impossible), as in Spanish (40). But it may allow both, as with *ni ne* in French.

(46) French (Indo-European)

Je	*ne*	*veux,*	*ni*	*ne*	*peux*	*répondre.*
1SG	NEG	want.PRS.1SG	CONEG	NEG	can.PRS.1SG	answer.INF

'I don't want to answer and I can't either.'

From the earlier work we can formulate the generalizations in (47).

(47) a. A clausal connective negator may make the ordinary negator superfluous, and it cannot control negative concord.

b. A finite connective negator may make the ordinary negator superfluous, it may control negative concord in the same way as for negative indefinites, and a first connective negator may be absent.

c. A phrasal connective negator must be controlled by negative concord in the same way as negative indefinites, but a first connective negator may be absent, and in that case it may scope over negative indefinites.

It must be stressed that these generalizations are only based on the Balto-Slavic languages, Spanish, French and Turkish. They have to stand up to typological testing and in future work we should also find out whether a structure that is possible is either frequent or rare. The generalizations in (47) are thus highly tentative but, at least, they show some parameters of variation.[17] The generalizations are also not too forbidding – note the frequent use of the modal *may*. We will now check whether Persian stays within the bounds of this variation and we will see that it does not, at least not completely.[18]

[17] With the grounding of the generalizations in Balto-Slavic, Spanish, French and Turkish there is also a strong European bias. We thus do not advance much on the state of affairs described by Haspelmath (2007: 17) for the study of connective negation at that time. At least, we now know, thanks to Salaberri (2022), that connective negation is not an exclusively European phenomenon.

[18] Stilo (2004:321-322) already makes clear that the relation between connective negation and negative concord is interesting, but he does not go into detail.

3.2. Observations and hypotheses for Persian

In Persian the connective negator is *na*. It is a free-standing negator. It is similar to the ordinary negator *na.* The latter functions as a prefix when the predicate is verbal (and it has an allomorph *ne-*).[19] In (48) we illustrate the ordinary clausal negator for both verbal and adjectival predicates.

(48) a. (*unā*) *farār* *na-kard-an.*
3PL escape NEG-do.PST-3PL
'They didn't run away.'
b. (*unā*) *mariz* *na-bud-an.*
3PL sick NEG-be.PST-3PL
'They were not sick.'

Example (49) illustrates clausal connective negation.

(49) *na* *man* *umad-am* *piš=e* *to,* *na* *to* *montazer=e*
CONEG 1SG come.PST-1SG next.to=EZ 2SG CONEG 2SG waiting=EZ
man *mund-i.*
1SG stay.PST-2SG
'Neither did I come to you, nor did you wait for me.'

The meaning of (49) can also be expressed with connection marked with *=o* 'and' and negation with the non-connective prefixal *na.*

(50) *man* *na-yumad-am* *piš=e* *to=o* *to* *montazer=e*
1SG NEG-come.PST-1SG next.to=EZ 2SG=and 2SG wait=EZ
man *na-mund-i.*
1SG NEG-stay.PST-2SG
'I didn't come to you and you didn't wait for me.'

(49) shows that the clausal connective negator combines with a positive verb. This is in agreement with the generalization in (47a). We conjecture that the fact that the

[19] In his world-wide sample study Salaberri (2022) remarks that this formal similarity seems prevalent in the Indo-Iranian languages.

connective negator and the ordinary negator are formally close militates against having *na* twice.[20]

With clausal connective negation we get negative concord in the same way as with non-connective clausal negation. The example below is based on Mahootian & Gebhardt (1997: 76).

(51) *na man čiz-i āvord-am, na to čiz-i dār-i.*
CONEG 1SG NHUM.IND bring.PST-1SG CONEG 2SG NHUM.IND have.PRS-2SG
'Neither did I bring anything nor do you have anything.'

In (51) *čiz-i* can be replaced by *hič-čiz-i* and *hič-či* as one would expect from Table 1. Negative concord with nominal indefinites and with human pronominal ones similarly shows the options of Table 1. This way Persian does not obey the generalization formulated in (47a), which states that clausal connective negation does not control negative concord. We suspect that the fact that Persian clausal connective negation does control negative concord is again due to the fact that ordinary and connective clausal negators are formally similar.

Example (52) illustrates the finite use of the connective negator.

(52) *mard-e na raft-Ø na umad-Ø. hamunǰā vāysād-Ø.*
man-DEF CONEG go.PST-3SG CONEG come.PST-3SG there stay.PST-3SG
'The man neither went nor came. He stayed right there.'

Again, we see that the verb is positive. (53) shows what we find with indefinites.

(53) *na āvāz-i mi-xund-Ø, na čiz-i mi-goft-Ø.*
CONEG song-IND IPFV-sing.PST-3SG, CONEG NHUM.SG-IND IPFV-say.PST-3SG
'(S)he was neither singing nor saying anything.'

The rules for negative concord with finite connective negation are thus the same as with clausal connective negation. Persian conforms to the generalization in (47b).

[20] This does not rule out semantically double negation as in
(a) na man = o did-Ø, na na-did-Ø.
CONEG 1SG = ACC see.PST-3SG, CONEG NEG-see.PST-3SG
'He neither saw me nor didn't see me.'

Note also that the generalization in (47b) allows a language not to express a first finite connective negator, but does not require it. In Persian a first finite connective negator must be expressed. So in this respect too, Persian conforms to the generalization in (47b).

We now turn to the phrasal use.

(54) *na dānešāmuz-ā=ye bāhuš, na dānešāmuz-ā=ye tanbal*
CONEG student-PL=EZ intelligent CONEG student-PL=EZ lazy
dars=o (*na-*)*fahmid-an.*
lesson=ACC (NEG-) understand.PST-3PL
'Neither the intelligent students nor the lazy students understood the lesson.'

Example (54) has to contain both phrasal connective negators. In this respect, phrasal connective negators are like the finite and clausal ones and they conform to the generalization in (47c). But in two other respects they do not conform: we do not get the negative concord found with an ordinary clausal negator.[21] First, with (pro)nominal indefinites the indefinite need not be negative – see examples (1d), (5c), (11), (12a) and (12b). Negatively connected phrases, however, have to be negative. Second, with ordinary clausal negation the verb has to be negative. In (54), however, we see that the clausal negator may be absent, even preferably so. This way Persian resembles the non-strict negative concord of the Catalan type. In this language negative concord is obligatory for a postverbal negative indefinite, and optional for a preverbal one.

(55) Catalan (Indo-European)

a. *Ningú (no) ha vist Joan.*
NIND.HUM NEG have.PRS.3SG see.PTCP.PST.SG Joan
'Nobody has seen Joan.'

b. *Joan no ha ist ningú.*
Joan NEG have.PRS.3SG see.PTCP.SG NIND.HUM
'Joan has seen nobody.

[21] What is also special and needs more work is the intonational pattern. The negative verb seems to require a break after the first connective phrase, whereas the positive verb allows but does not require this break.

In Persian the negative concord with phrasal connective negation is similarly optional when the phrasal connective negators precede the verb. But there are differences, too. First, it is impossible for the phrasal connective negators to both follow the finite verb (be 'right-dislocated'). So there is no constellation resembling the obligatory negative concord that we see in Catalan (55b). Second, it is, however, possible for one of the two connective negators to follow the verb and then the verb has to be positive (Najafi 2013: 400-401).

(56) *na dānešāmuz-ā-ye bāhuš tu emtehān qabul*
CONEG student-PL=EZ intelligent in exam accept
(**na-*)*šod-an, na dānešāmuz-ā=ye tanbal.*
(NEG-)become.PST-3PL CONEG student-PL=EZ lazy
'Neither the intelligent students passed in the exam nor the lazy students.'

This constellation is difficult to explain. What we see in (54) invites an explanation in terms of a version of 'Neg Early' principle.[22] With respect to both Catalan (55a) and Persian (54) the Neg Early principle could be interpreted to imply that when the clause has special negative marking early in the sentence, the negation is sufficiently clear and it can dispense with a later general negative marker. In Persian (54) there are even two special negators and both come early, i.e., before the slot where the clausal negator could have come. In Persian (56) only one of the special negators precedes the slot for the clausal negator, so there should not be more tolerance for leaving the clausal negator unexpressed. But this is not the case. In fact, we see an obligation for leaving the clausal negator unexpressed. Perhaps a 'Neg Late' principle helps out. It could go as follows. In Persian a clausal negator needs an overt expression towards the end of the clause. In (54) the late exponent is the *na* on the verb, and the verb comes late. The Neg Late principle is relaxed by the Neg Early principle: prefinite but late *na-* is optional. In (56) the clausal negator is absent because of the early first connective negator. Neg Late does not come to 'rescue' to at least allow the clausal negator, for Neg Late is satisfied by the second connective negator. This has to be overt – Persian connective negators always are – and it comes later still than the

[22] This principle goes back to at least Jespersen (1917: 5), when he argued that '[T]here is a natural tendency, also for the sake of clearness, to place the negative first, or at any rate as soon as possible, very often immediately before the particular word to be negatived [*sic*] (generally the verb).' It was called 'Neg First' by Horn (1989).

clausal negator. Whether this explanatory attempt is successful, it appears that the 'split connective negation' pattern shown by Persian (56) has not attracted any attention in the literature. What has also gone unstudied is the pattern shown in (57). Here the connective negators are resumed by a 'neither of them' element. The latter contains a *hič* element and we see the normal negative concord.

(57) *na dānešāmuz-ā=ye bāhuš na dānešāmuz-ā=ye tanbal,*
CONEG student-PL=EZ intelligent CONEG student-PL=EZ lazy
hič-yek/hič-kodum tu emtehān qabul na-šod-an.
neither.of.them in exam accept NEG-become.PST-3PL
'Not the intelligent students and not the lazy students, none of them passed in the exam.'

Finally, it will be remembered that at least in French and Spanish connective negators can scope over negative indefinites and that the first connective negator has to be absent. In Persian, connective negators can scope over negative indefinites too, but, different from French and Spanish, the first connective negator has to be present. We again see that, like in (54), the verb may be negative or positive. (58) illustrates this point with one choice for both 'nothing' and 'nobody'. It seems that all choices are possible. The variants without *hič* seem most frequent, perhaps because they avoid the collocation of two negatives, viz. *na* and *hič*.

(58) *na hič-čiz-i na hič-kas-i harf-i beh=eš*
CONEG NIND-NHUM.SG-IND CONEG NIND-NHUM.SG-IND talk-IND to=3SG
(na-)zad-e.
(NEG-)hit.PST-3SG
'Nothing or nobody had told him/her anything'.

The meaning in (58) can also be expressed with non-connective negators. It is also possible to split the connective negators.

(59) a. *hič-čiz-i=o hič-kas-i na-yumad-e-Ø*
NIND-NHUM.SG-IND=and NIND-HUM.SG-IND NEG-come.PST-PTCP-3SG
b. *na hič-čiz-i umad-e-Ø, na*
CONEG NIND-NHUM.SG-IND come.PST-PTCP-3SG CONEG

hič-kas-i.
NIND-HUM.SG-IND
'Nothing or nobody has come.'

Again, the other alternatives for 'nothing' and 'nobody' are possible, and the 'best' ones seem to be the ones without *hič*. Still, double negative *na hič* structures are possible, and they are 'more possible' than their counterparts in French or Spanish, for in the latter the first connective negator has to be absent.

Like in the conclusion of section 2, we have by no means exhausted the topic. Like with negative indefinites, for instance, the issue of negative spread shows up. And it remains to be seen how connective negators interact with the various types of indefinite phrases discussed in section 2.

3.3. Conclusion

With respect to the existing generalizations on the relation between negative concord and connective negation, Persian is special for both the clausal and the phrasal use of connective negation. The clausal connective negator is interesting because it controls negative concord in the same way as the ordinary clausal negator. The phrasal connective negator is more interesting still, for it shows the relevance of a new parameter. Thus far the literature has shown the relevance of word order in terms of both of the negatively connected phrases either preceding or following the finite verb. Persian makes us aware of a third pattern: one of the two negatively connected phrases precedes the finite verb and the other one follows. Up to now the literature has appealed to a Neg Early principle. We tried to account for the resulting negative concord patterns in terms of a competition between a Neg Early and Neg Late principle. We have also shown that phrasal connective negator can connect negative indefinites, like in French and Spanish.

4. General conclusion

This paper offered an account of Persian negative indefiniteness and, more particularly, negative concord. The kind of negative concord shown in Persian is non-strict, but it is a subtype that has not been documented well, viz. a pattern that allows the negative verb to combine with either a negative or polarity neutral pronoun. We

have also seen that the negative concord for 'nobody' may be different from the one with 'nothing'. We have documented how both types of pronominal negative concord differ from the one with noun phrases. That this may be the case is well-known but there is not much research. A thorny point is the analysis of the marker that makes the indefinites negative, viz. the *hič* marker. It is intriguing because *hič* has an emphatic non-negative use in questions. We described this use of *hič* as a widening negative polarity element and we hypothesize that the negative use derives from this, with a loss of widening and a narrowing from negative polarity into negation. The general typology should provide for this and show how the Persian constellation is different from superficially similar constellations in languages like Dutch, Spanish or Jamaican Creole.

We have also offered a first account of the relation between Persian negative concord and connective negation. We see that Persian does not fall within the bounds of variation sketched in earlier work. The Persian observations and hypotheses suggest that at least two new parameters should be attended to in future work: (i) the question whether or not the ordinary and the connective negators are formally similar, and (ii) the question whether one of two connectively negated phrases can be put before the finite verb and the other behind it. A Neg Early principle is commonly appealed in the study of negative concord; we have argued that Persian could be seen as motivating the effect of a Neg Late principle.

Though we think that this study is the most comprehensive one on Persian negative concord and connective negation so far, there are many tasks undone. In the synchrony we abstained from the study of *yek* 'one', the interaction between negation and genericity, and negative spread, to recall just three examples. We have seen that that there is a lot of variation in the data. Crucially, negative concord is often non-strict, which invites a quantitative analysis of what steers the variation. It is also clear that even though we interpreted the synchronic data from a diachronic perspective and thus used some examples from the earlier stages of the language, a thorough diachronic study remains to be done.

Acknowledgments

Thanks are due to two reviewers and to Juan Briceño Villalobos (Santiago de Chile), Egbert Fortuin (Leiden), Chiara Gianollo (Bologna), Rajesh Kumar (Chennai),

Roohollah Mofidi (Qazvin), Masoud Mohammadirad (Cambridge) and Iker Salaberri (Bilbao).

Abbreviations

1 = first person
2 = second person
3 = third person
ACC = accusative
CONEG = connective negation
DAT = dative
DEF = definite
DET = determiner
DIST = distal
EZ = *ezafe*
F = feminine
FUT = future
GEN = genitive
HUM = human
IND = indefinite
INF = infinitive
INT = interrogative
IPFV = imperfective
M = masculine
N = noun
NEG = negation
NHUM = non-human
NIND = negative indefinite
NOM = nominative
NPI = negative polarity item
NPIND = negatively polar indefinite
NEG = negative
PART = partitive
PFV = perfective
PL = plural
POSS = possessive
PROX = proximal
PRS = present
PST = past
PTCP = participle
PVB = preverbal
REL = relative
RFL = reflexive
SBJV = subjunctive
SG = singular

References

Afshar, Iraj & Mehran Afshari (eds.). 2006. *Qesse Hossein Kord Shabestari bar asāse revāyate nāšenākhteye mosum be Hossein nāme* [The story of Hossein Kurd Shabestari, based on an anonymous story known as Hossein Name]. Tehran: Cheshme.

Ahmadi Givi, Hasan. 2001. *Dastoore tārixie fe'l.* [The historical grammar of the verb]. *Volume II.* Tehran: Ghatreh.

Anvari, Hasan. 2003. *Farhānge bozorge Soxan.* [The great dictionary of Sokhan]. Tehran: Sokhan.

Bernini, Giuliano & Paolo Ramat. 1992. *La frase negativa nelle lingue d'Europa.* Bologna: Il Mulino.

Bernini, Giuliano & Paolo Ramat. 1996. *Negative sentences in the languages of Europe.* Berlin: Mouton de Gruyter.

Bhatia, Tej Krishan. 1978. *A syntactic and semantic description of negation in South Asian languages.* PhD dissertation, University of Illinois at Urbana-Champaign.

Boyle, John Andrew. 1966. *Grammar of modern Persian.* Wiesbaden: Harrassowitz.

Breitbarth, Anne, Christopher Lucas & David Willis. 2020. *The history of negation in the languages of Europe and the Mediterranean. Volume II: Patterns and processes*. Oxford: Oxford University Press.

Briceño Villalobos, Juan Eugenio. 2019. *Negation, indefinites, and polarity in early Greek and Indo-Iranian: a typological and comparative approach*. PhD dissertation, Complutense University Madrid.

Cardinaletti, Anna & Giuliana Giusti. 2020. Indefinite determiners in informal Italian: A preliminary analysis. *Linguistics* 58(3). 679-712.

Dabir-Moghaddam, Mohammad. 2013. *Radešenāsie zabānhāye Irani*. [Typology of Iranian Languages]. Tehran: Samt.

Erschler, David & Volk, Vitaly. 2011. On negation, negative concord, and negative imperatives in Digor Ossetic. In Agnes Korn, Geoffrey Haig, Simin Karimi & Pollet Samvelian, (eds.), *Topics in Iranian linguistics*, 135-150. Wiesbaden: Reichert.

Fălăuş, Anamaria. 2013. Broaden your views, but try to stay focused: a missing piece in the polarity system. In Ivano Caponigro & Carlo Cechetto (eds.), *From grammar to meaning. The spontaneous logicality of language,* 81-107. Cambridge: Cambridge University Press.

Garzonio, Jacopo & Cecilia Poletto. 2020. Partitive objects in negative contexts in Northern Italian dialects. *Linguistics* 58(3). 621-650.

Geerts, G., W. Haeseryn, J. de Rooij & M.C. van den Toorn (eds.). 1984. *Algemene Nederlandse spraakkunst.* Groningen: Wolters Noordhoff.

Giannakidou, Anastasia. 1998. *Polarity sensitivity as (non)veridical sensitivity*. Amsterdam: Benjamins.

Gianollo, Chiara. 2018. *Indefinites between Latin and Romance.* Oxford: Oxford University Press.

Gianollo, Chiara. 2020. Evolution of negative dependencies. In Viviane Déprez & M. Teresa Espinal (eds.) *The Oxford handbook on negation,* 546-562. Oxford: Oxford University Press.

Görgülü, Emrah. 2017. Negative polarity in Turkish: from negation to nonveridicality. *Macrolinguistics* 5: 51-69.

Görgülü, Emrah. 2020. Negative sensitive items in Turkish: Negative polarity or negative concord? *RumeliDE Dil ve Edebiyat Araştımaları Dergisi* 21: 724-749.

Haspelmath, Martin. 1997. *Indefinite pronouns.* Oxford: Oxford University Press.

Haspelmath, Martin. 2005. Indefinite pronouns, In Martin Haspelmath, Matthew S. Dryer, David Gil & Bernard Comrie (eds.), *The world atlas of language structures*, 190-183. Oxford: Oxford University Press. (Also at http://wals.info/chapter/46)

Haspelmath, Martin. 2007. Coordination. In Timothy Shopen (ed.), *Language typology and syntactic description. vol II.,*1–51. Cambridge: Cambridge University Press.

Horn, Laurence R. 1989. *A natural history of negation*. Chicago: The University of Chicago Press.

Jeretič, Paloma. 2018. Hybrid and optional Negative Concord in Turkish: A unified analysis. *Indiana University Linguistics Club Working Papers* 18

Jespersen, Otto. 1917. *Negation in English and Other Languages*. Copenhagen: Høst & Søn.

Kadmon, Nirit & Fred Landman. 1993. Any. *Linguistics and Philosophy* 16(4). 353-422.

Kahnemuyipour, Arsalan. 2017. Negation is low in Persian: evidence from nominalization. *Proceedings of the 2017 conference of the Canadian Linguistic Association*.

Kelepir, Meltem. 2001. *Topics in Turkish syntax: Clausal structure and scope*. PhD dissertation, Massachusetts Institute of Technology.

Kumar, Rajesh. 2006. *Negation and licensing of negative polarity items in Hindi syntax.* New York: Routledge.

Kwak, Saera. 2010. Negation in Persian. *Iranian studies* 43(5). 621-636.

Lahiri, Utpal. 1998. Focus and negative polarity in Hindi. *Natural Language Semantics* 6(1). 57-123.

Lambton, Ann K. 1957. *Persian grammar*. Cambridge: University Press.

Lazard, Gilbert. 1992. *A grammar of contemporary Persian*. Cosa Mesa: Mazda Publishers.

Mace, John. 2003. *Persian Grammar: for reference and revision*. London: Routledge.

Mahootian, Shahrzad & Lewis Gebhardt. 1997. *Persian.* London: Routledge.

Mansouri, Mehrzad. 2004. Nafy dar zabāne Farsi az manzare radešenāsie zabāni [Negation in Persian from a typological perspective]. *Proceedings of Allameh Tabatabaie University*, *vol* 186: 339-350.

Najafi, Abolhasan. 2013 [1988]. *Qalat nanevisim* [Let's not write wrong]. Tehran: Nashre Daneshgahi.

Norde, Muriel. 2009. *Degrammaticalization.* Oxford: Oxford University Press.

Oryan, Saeid. 1997. Nahve zabānhāye Irāni-e miyāne qarbi [translation of Brunner's the Syntax of Middle West Iranian]. Tehran: Reseach Institutefor Islamic Culture and Art Language and Literature.

Rasekh-Mahand, Mohammed. 2015. Kārbodhāye moxtalefe zamāyere nakare dar barxi zabāhāye Iranie no. [Semantic maps of indefinite pronouns in Iranian languages] In Mahmood Ja'fari Dehaghi & Nazanin Khalilipoor (eds.), *Majmoo'e maghalate dovvomin hamyeshe beinolmelaliye zabanha va guyeshhaye Irani (gozashte va hal)* [Proceedings of the Second International Conference on Iranian Languages and Dialects (Past and Present)], 203-226. Tehran: Centre of the Great Islamic Encyclopedia.

Report on the current status of United Nations Romanization systems for geographical names - Persian, http://www.eki.ee/wgrs/rom1_fa.htm

Romero, Maribel. 2020. Form and function of negative, tag, and rhetorical questions. In Viviane Déprez & M. Teresa Espinal (eds.), *The Oxford handbook of negation*, 235-254. Oxford: Oxford University Press.

Salaberri, Iker. 2022. A cross-linguistic study of emphatic negative coordination (ENC). *Studies in Language*

Schwarze, Christophe. 1995[2]. *Grammatik der italienischen Sprache.* Tübingen: Niemeyer.

Suleymanova, Venera & Jack Hoeksema. 2018. Minimizers in Azerbaijani from a comparative perspective. *Folia Linguistica* 52(1). 177-211.

Şener, Serkan & Selçuk İşsever. 2003. The interaction of negation with focus: *ne … ne …* phrases in Turkish. *Lingua* 113(11). 1089-1117.

Stilo, Donald. 2004. Coordination in three West Iranian languages: Vafsi, Persian and Gilaki. In Martin Haspelmath (ed.), *Coordinating constructions,* 269-330. Amsterdam: Benjamins.

Taleghani, Azita H. 2008. *Modality, aspect and negation in Persian.* Amsterdam: Benjamins.

Van Alsenoy, Lauren. 2014. *A new typology of indefinite pronouns, with a focus on negative indefinites.* PhD dissertation, University of Antwerp.

van der Auwera, Johan. 2009. The Jespersen cycles, In Elly van Gelderen (ed.), *Cyclical Change*, 35-71. Amsterdam: Benjamins.

van der Auwera, Johan. 2017. Englishes, English Creoles and their negative indefinites, In Debra Ziegeler & Zhiming Bao (eds.), *Negation and contact: with special reference to Singapore English,* 115-149. Amsterdam: Benjamins.

van der Auwera, Johan. 2021. Quirky negative concord: Croatian, Spanish and French *ni*'s. *Jezikoslovlje* 22(2). 195-225.

van der Auwera, Johan. To appear. Nominal and pronominal negative concord, through the lens of Belizean and Jamaican Creole, *Linguistics*

van der Auwera, Johan & Tamirand Nnena De Lisser. 2019. Negative concord in Jamaican, Ampersand 6

van der Auwera, Johan & Olga Krasnoukhova. 2021. How to do words with "things" – Multiple grammaticalization from 'thing' in Tupi-Guarani, *Italian Journal of Linguistics* 33(2). 69-98.

van der Auwera, Johan, Motoki Nomachi & Olga Krasnoukhova. 2021. Connective negation and negative concord in Balto-Slavic. In Peter Arkadiev, Jurgis Pakerys, Inesa Šeškauskienė & Vaiva Žeimantienė (eds.), *Studies in Baltic and other languages. A Festschrift for Axel Holvoet on the occasion of his 65th birthday,* 44-64. Vilnius: Vilnius University Press.

van der Auwera, Johan & Lauren Van Alsenoy. 2016. On the typology of negative concord, *Studies in Language* 40(3). 473-512.

van der Auwera, Johan & Lauren Van Alsenoy. 2018. More ado about *nothing*: on the typology of negative indefinites. In Ken Turner & Laurence R. Horn (eds.), *Pragmatics, truth and underspecification: towards an atlas of meaning,* 107-146. Leiden: Brill.

Willis, David. 2007. Syntactic lexicalization as a new type of degrammaticalization, *Linguistics* 45(2). 271–310.

Windfuhr, Gernot. 2009. Dialectology and topics. In Gernot Windfuhr (ed.), *The Iranian languages.* 5-42. London: Routledge.

Windfuhr, Gernot & John R. Perry. 2009. Persian and Tajik. In Gernot Windfuhr (ed.), *The Iranian languages.* 416-544. London: Routledge.

Yousef, Saheed (with the assistance of Hayede Torabi). 2018. *Persian. A comprehensive grammar*. London: Routledge.

CONTACT

johan.vanderauwera@uantwerpen.be
sepideh.koohkan@gmail.com

Linguistic Typology at the Crossroads 2-1 (2022): 37-64 ISSN 2785-0943

Deictic marking in adpositions in Ap Ma and Waran

DON KILLIAN[1], RUSSELL BARLOW[2]
[1]UNIVERSITY OF HELSINKI, [2]MAX PLANCK INSTITUTE FOR EVOLUTIONARY ANTHROPOLOGY

Submitted: 21/09/2021 Revised version: 01/12/2021
Accepted: 03/12/2021 Published: 14/04/2022

Abstract

Two neighboring languages of the Sepik region of Papua New Guinea, Ap Ma (Keram) and Waran (Ramu), exhibit an unusual structure in adpositional phrases. In both languages, all postpositions that govern nouns are obligatorily marked with deictic morphemes that indicate the position of the referent of the NP relative to the deictic center. Both languages employ deictic morphemes that index whether a referent is near, medial, or far. In addition to having other crosslinguistically common and expected applications, these morphemes occur as obligatory elements in adpositional phrases. This article examines the details of these unusual deictic-marked adpositional constructions, placing them in a typological context. We conclude that diachronic changes in Ap Ma phonology were likely the historical impetus for these constructions, which may have subsequently spread to Waran through contact.

Keywords: adpositions; postpositions; deixis; demonstratives; Papuan; Keram-Ramu.

1. Introduction

Demonstratives are words that prototypically indicate spatial deictic reference, indicating the location of a referent relative to the deictic center (or "origo", following Bühler 1934) (cf. Lyons 1977); as in, for example: "I would like to purchase *this* book".[1] Although they are often stereotypically thought of as comprising pronouns

[1] Demonstrative systems can also include other types of demonstratives, such as those that are deictically neutral as well as those that make reference to geographical features. Although languages that make reference to geographical features are outside the scope of this paper, it is worth mentioning the possibility of having a spatially neutral or unmarked term that contrasts with other terms with a clear anchorage and reference zone (cf. Levinson 2018).

and determiners, demonstratives belonging to several other lexical categories have also been identified in the world's languages. In his typological work, Diessel (1999: 57-58) recognizes four basic types of demonstratives, based on their syntactic position (1).

(1) Basic types of demonstratives
(i) pronoun: a pronoun being used to replace a noun, e.g., *I like this*.
(ii) determiner: a determiner used in apposition to a noun, e.g., *I like this book*.
(iii) identifier: a non-verbal predicator used in non-verbal clauses, e.g., *This is John*.
(iv) adverb: an adverb that modifies a verb, e.g., *He read the book here*.

Dixon (2003; 2010: 225), on the other hand, recognizes three types of demonstratives: nominal demonstratives (encompassing both Diessel's pronoun and determiner categories), local adverbial demonstratives (corresponding to Diessel's adverb category), and verbal demonstratives, which have a meaning of 'do like this/that'. Demonstrative verbs have also been investigated in more recent works, such as Gruzdeva (2013), Guérin (2015), and Breunesse (2019).

Killian (2021) expands considerably on the syntax and semantics of non-verbal predicating demonstratives (corresponding to Diessel's identifier category). Additional and more fine-grained distinctions may also be needed for the categories of determiners and adverbs, categories which contain under-researched semantic types such as manner, quality, quantity, and degree (for adverbs; cf. König & Umbach 2018), and definiteness and specificity (for determiners). Some of these categories may end up being shown to behave syntactically more like adjectives than determiners for instance, adding yet another category.

However, one category of demonstratives that has not yet been discussed in the literature, with perhaps the sole exception of a single sentence in Breunesse (2019: 197),[2] is that of adpositional demonstratives. The term *adposition* is used here to refer to "words that combine with noun phrases and that indicate the semantic relationship of that noun phrase to the verb" (Dryer 2007: 81-82). Hagège (2010: 175) acknowledges spatial deixis marking as a possible inflectional category of adpositions; however, he states that this constitutes "deixis agreement", rather than existing as a

[2] "… a paradigm of demonstratives described as prepositions exists in Begak (see Goudswaard 2005: 90-92), Buru (see Grimes 1991: 255-256), and Semelai (see Kruspe 1999: 359 ff.)."

category in its own right. Thus, although adpositions commonly possess spatial or temporal meanings, they do not normally include a deictic component.

Indeed, deictic distinctions in adpositions are extremely rare.[3] However, they do occur in a small number of Austronesian and Atlantic languages, such as Wolof (Atlantic, Senegal), where the paradigm of locative prepositions *ci* and *ca* indexes either proximal (2) or distal (3) relations.[4]

(2) Wolof (Atlantic; Robert 2006: 158)
ci *néeg* *bi*
in.PROX room the.PROX
'in the room (close to me)'[5]

(3) Wolof (Atlantic; Robert 2006: 158)
ca *néeg* *ba*
in.DIST room the.DIST
'in the room (far away from me)'

Similarly, in Begak (Austronesian, Malaysia), the two prepositions *nong* and *di'* include a deictic component to their meaning. Whereas *nong* refers to locations near the speaker or deictic center (4), *di'* is used for locations removed from the speaker or deictic center (5).

[3] By "deictic distinctions" we refer to relative proximity to the origo, rather than default viewpoints of relative frames of reference that originate from a deictic center. Although a preposition like English *behind* (as in *the man is behind the tree*) may contain a deictic component to its meaning, this has little to do with the deixis of relative proximity.

[4] Although the Wolof examples may appear to be exhibiting agreement between the preposition and the deictic article, such an analysis would not be correct (Stéphane Robert, p.c.). It is, for example, possible for the deictic prepositions *ci* and *ca* to occur with proper nouns, which do not take articles.

[5] We have maintained the original orthographies of the various sources in our examples, including the authors' use of hyphens, equal signs, and spaces. In a few instances, however, where morpheme boundaries were ignored in the original, we have added hyphens. We do not, however, wish to argue for the syntactic status of any morphs (i.e., whether they be affixes, clitics, or words). Also, although we have tried to be faithful in reproducing these orthographies, we have made some changes to the interlineal glossing, primarily to make them conform to the abbreviations of the Leipzig Glossing Rules, but occasionally to reflect a difference in analysis.

(4) Begak (Austronesian; Goudswaard 2005: 246)

kəmmi	*gərə-i-dagang*	*kaset*	***di'***	*Indonesia*
1PL.EXCL.NOM	AV.DSTP-COMPL-buy	cassette	**PREP.DIST**	Indonesia

'We bought cassettes in Indonesia.'

(5) Begak (Austronesian; Goudswaard 2005: 85)

-u-rənna'	*kat*	*nupi*	*key*	***nong***	*monay*
-DEP-descend	CDM	dream	FOC	**PREP.PROX**	young.man

'A dream came down to Young Man.'

A far more elaborate and unusual construction occurs in two neighboring and distantly related Papuan languages of Papua New Guinea. In Ap Ma (also known as Kambot or Botin, Keram family) and Waran (also known as Banaro, Ramu family), all adpositions governing nouns obligatorily mark spatial deixis (proximal, distal, or medial).

Although spoken in close proximity, the two languages are only distantly related. Figure 1 provides a map of the region where both languages are spoken. Figure 2 provides a tentative subgrouping of the Keram-Ramu family, to which both languages belong, based largely on the work of John Z'graggen (1971: 73-92) and Timothy Usher (p.c.) (cf. Barlow 2020: 14-17).

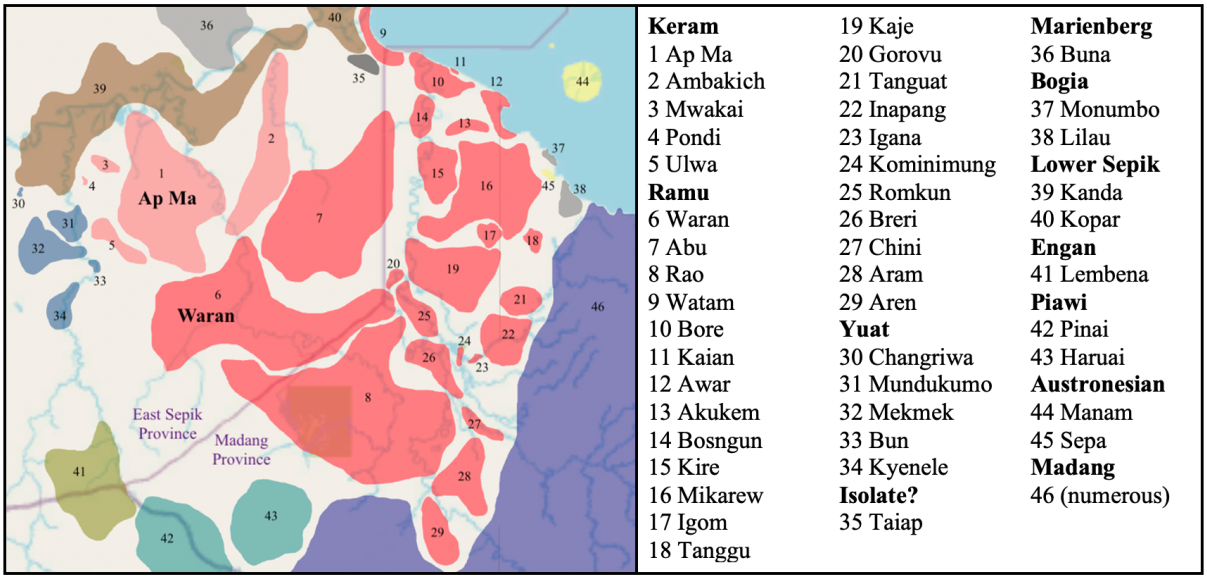


Figure 1: Map of northeastern New Guinea, depicting where Ap Ma, Waran, and some other neighboring languages are spoken.

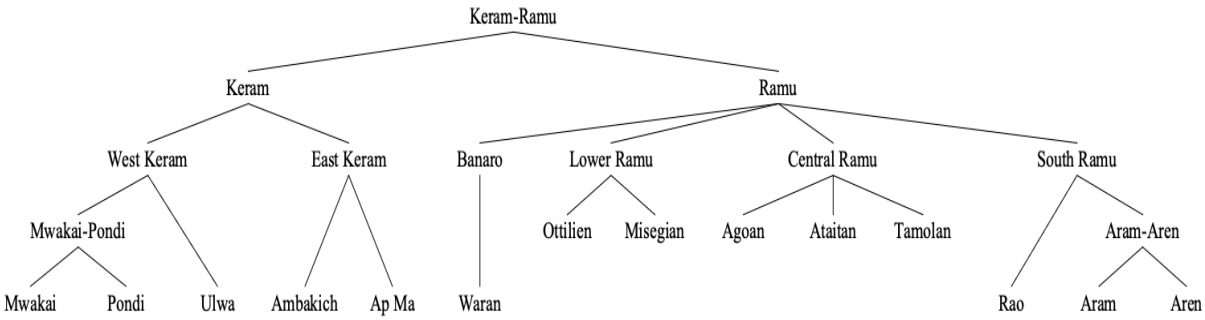


Figure 2: The Keram-Ramu family.[6]

As in Wolof and Begak, adpositions in Ap Ma and Waran not only relate the nominal argument to the clause but also carry deictic information, relating the entire adpositional phrase to the speaker or deictic center. Unlike in Wolof and Begak, the deictic element in such constructions in Ap Ma and Waran is a morphologically distinct (although bound) component, as illustrated for Ap Ma in (6).

(6) Ap Ma (Keram; Barlow, unpublished field notes 2018)
ni-ŋga *nindan* ***ape*** ***nda-nɨn*** *ta-p*
1SG-PROX yesterday **father** **DIST-with** walk-PFV
'Yesterday I walked around with father.'[7]

The phrase *ape ndanɨn* not only relates the noun phrase 'father' to the clause as a whole, but also indicates (by means of the morpheme *nda* 'DIST') that the object of the postposition ('father') is or was distant from the current location of the speaker, either because father is not nearby at the time of the speech event, or because the event of walking itself occurred at a distance from the speaker's current location. In other words, the adposition conveys deictic information on the position of the referent of the NP relative to the speaker or deictic center.

[6] Ottilien consists of Watam, Bore, Kaian, Awar, and Bosngun. Misegian consists of Akukem, Kire, and Mikarew. Agoan consists of Abu and Gorovu. Ataitan consists of Igom, Tanggu, Kaje, and Tanguat. Tamolan consists of Romkun, Breri, Kominimung, Igana, Inapang, and Chini.

[7] The grapheme <ɨ> is used in the transcriptions by Barlow (2021) and Pryor (1986, 1990) to represent the high central unrounded vowel in Ap Ma, which has phonemic status in the language (Barlow 2021: 36). Wade's (1982, 1983a, 1983b, 1984) transcriptions, however, do not use this grapheme. What is written elsewhere as <ɨ> corresponds to other vowels in her transcriptions.

Such constructions – in Ap Ma at least – are only possible if the argument governed by the adposition is nominal; if the argument is pronominal, then the argument occurs in the same position that the deictic marker otherwise would occupy, and no deictic reference is indicated (7).

(7) Ap Ma (Keram; Barlow, unpublished field notes 2018)
ape *ndɨ* ***ni-nɨn*** *ta-p*
father FOC **1SG-with** walk-PFV
'Father walked around with me.'

In the following sections we aim to explain in greater detail the phenomenon of adpositional spatial deixis in Ap Ma (Section 2) and in Waran (Section 3), before proposing a diachronic explanation for this unusual feature (Section 4), thereby placing it in a broader typological context (Section 5).

2. Ap Ma

There are three (singular) forms for Ap Ma deictics, which ostensibly signal proximal (*gɨ-*), medial (*mɨ-*),[8] or distal (*dɨ-*) distances. Deictics appear to have a very high functional load in Ap Ma: not only do they mark physical distance, but they are also used to track referents in discourse, as well as to indicate the attitude of the speaker towards the referent (Pryor 1990). Although such functions of deictics are not uncommon among the world's languages (cf. Himmelmann 1996), the degree to which Ap Ma recruits deictics for such purposes is remarkable. Indeed, Pryor & Farr (1989: 116) acknowledge deixis as "a major part" of the language. According to Pryor (1990: 22), "demonstratives serve as the major cohesive device in the language."

In example (8), the proximal form is used with the postposition *-ak* (glossed by Wade as 'ORG', i.e., 'origin'). However, the story was not actually told in or near the garden mentioned in the story. Rather, this use of a proximal form instantiates a deictic shift and cognitive framing, the proximal form being used to set the scene and

[8] Levinson (2018: 25) writes that "... there are grounds for suspicion that most if not all terms that might be described as 'medial' in descriptive grammars are in fact neutral or unmarked distal, and pick up their medial usage through privative opposition with proximal or far-distal terms." It is possible that this scenario would apply to Ap Ma as well, although we do not know for certain.

the deictic center for the story. The deictic center has been transposed to the garden, and future deictic references in the story occur accordingly.

(8) Ap Ma (Keram; Wade 1982: 9)
wuluk *ma-n* ***sak*** ***g-ak*** *ma-n* *nage* *l-ul-ul-me*
man MED-OBL **garden** **PROX-from** MED-OBL banana 3PL-cut-IPFV-EPIS
'(She) cut the man's garden's bananas.'[9]

Deictics in Ap Ma have a somewhat unusual distribution, attaching to verbs (9) and to postpositions (10), as well as to various bound grammatical morphemes, including case markers, emphatic markers, and topic markers (11), amongst other possible morphemes whose functions are not yet understood. Additionally, all deictics may also fuse with personal pronouns or with other deictics. Such deictic-deictic combinations create meaningful words that serve several functions (Wade 1982: 25-34), such as that of a copula (12) (see also (41) for an example of the same combination of deictics being used as a focus marker).

(9) Ap Ma (Keram; Wade 1984: 145)
nyi-ga *no* *pa-**di-m**-a-me*
1SG-PROX own house-**DIST-go**-IRR-EPIS
'I will certainly go to my own house.'

(10) Ap Ma (Keram; Pryor 1986: 32)
nyimi-ba *Astilelia* ***d-okol*** *wis*
some-NT Australia **DIST-through** come.PFV
'Some come from Australia.'

(11) Ap Ma (Keram; Wade 1982: 23)
73 ***da-yen*** *o* *Mis* *Rop* *ba-lup* *nyi-tin-a*
73 **DIST-TOP** INJ Miss Rop turn-PFV 1SG-get-IRR
'In 1973 Miss Rop was again coming to get me …'

[9] The gloss 'EPIS' is used for the Ap Ma forms *ma* and *me* (both always occurring clause-finally), since – although their functions are not fully understood – they seem to have epistemic force. The former occurs in perfective contexts, whereas the latter occurs elsewhere (i.e., imperfective and irrealis contexts). They may derive from suffixed versions of the medial deictic *mi-*.

(12) Ap Ma (Keram; Wade 1982: 16)

neg	*da-n*	*dinat*	***da-ma***
woman	DIST-OBL	heavy	**DIST-MED**

'It is that woman's problem.'

Deictics do not appear in their bare forms, nor do they fuse directly with nouns, unless the fusion additionally includes verbs, postpositions, topic or case markers, or other deictics. More crosslinguistically familiar deictic categories, such as demonstrative pronouns and demonstrative adverbs, are extremely rare (if not totally absent) in Ap Ma. Table 1 provides a list of some Ap Ma postpositions with their fused deictic markers.

Gloss	**Root**	**Proximal**	**Medial**	**Distal**
		g(i̱)-	m(i̱)-	d(i̱)-
'with' (comitative)	-ni̱n	gani̱n	mani̱n	dani̱n
'with' (inanimate comitative)	-pa	gi̱pa	mi̱pa	di̱pa
'with' (instrumental)	-n	gan	man	dan
'for' (beneficiary)	-ayak	gayak	mayak	dayak
'at, from'	-aak	gaak	maak	daak
'from'	-uk	guk	muk	duk
'in'	-in	gin	min	din
'to'	-ada	gada	mada	dada
'under'	-abe	gabe	mabe	dabe
'on'	-aal	gaal	maal	daal
'near'	-el	gel	mel	del
'through'	-okol	gokol	mokol	dokol
'above'	-atiyel	gatiyel	matiyel	datiyel

Table 1: Ap Ma postpositions (adapted from Pryor 1990: 5).

Note that, in addition to this paradigm of (singular) deictic markers, there are two plural deictic forms: *w(i̱)-* 'PL.PROX' and *l(i̱)-* 'PL.DIST'. Also, what is treated here as "medial" appears to be ambiguous between marking a third person referent and marking spatial deixis. This colexification of 3SG pronominal and MED deictic forms is common among languages of the Keram-Ramu family. Finally, based on comparative data from other Keram-Ramu languages, the form *-n*, presented above as a postposition ('with'), could alternatively be considered an "oblique marker"; however,

it appears to have nearly (if not exactly) the same morphosyntactic distribution as (other) postpositions in Ap Ma. This morpheme is discussed further in Section 4.2.

The origin of this rather curious phenomenon appears to lie in phonology. Postpositions in Ap Ma belong to a closed set of morphs (along with deictic enclitics) that must form "tightly-knit" noun phrases (Wade 1983b). Another way of viewing these forms is to consider them to be enclitics, which – quite remarkably – require proclitics as their phonological hosts. Only when these enclitics join with proclitics can they make well-formed words. The group of proclitics that can function in this way consists of personal pronouns (13, 14) and deictics (14, 15).

(13) Ap Ma (Keram; Wade 1984: 302)
ny-e ***wu-nin*** *y-e-w-o*
1SG-? **2SG-with** ground-PRF-1.2-STAT
'I am here with you.'

(14) Ap Ma (Keram; Wade 1984: 55)
nyi-ba *lo-sala* ***li-nin*** *pasta* *woksap* ***m-in*** *s-ap-ma*
1SG-NT QUANT-three **3PL-with** pastor workshop **MED-in** go-PFV-EPIS
'It's a fact that I went there three times with them to the pastor workshop.'

(15) Ap Ma (Keram; Wade 1984: 293)
*mi-**d-aak*** *lol* *da-ma*
water-**DIST-from** man DIST-MED
'He was a man from the water.'

Note that the form *mi* 'water' in (15) belongs to a class of "semi-free nouns", which are discussed further in Section 4.3.

There is thus a phonologically based functional motivation for postpositions to index spatial deixis. Since they belong to a class of enclitics that require proclitics as hosts, they can only permissibly follow pronouns or deictics. When the object of the postposition is a pronoun like *wu-* 'you' (13), then no deictic is "needed". However, it would not be permitted for a postposition immediately to follow a noun, since nouns do not participate in this peculiar proclitic-enclitic "co-hosting" relationship that postpositions require. By fusing with a deictic, the postposition becomes capable of taking any NP as its object. (Based on its usage in texts, it is assumed here that, when

no particular deictic distinction is intended, a speaker has recourse to use the medial form in a somewhat semantically bleached fashion as a generic referent.)

Table 2 summarizes the possible combinations of postpositions and other parts of speech in Ap Ma.

Object of the postposition	**Deictic marking?**	**Postpositional phrase**
(Independent) noun	Obligatory	NOUN DEIC = PP
Semi-free noun	Obligatory	SF.NOUN = DEIC = PP
Pronoun	Prohibited	PRON = PP

Table 2: Postpositions and deictic marking in Ap Ma.

We should note that, although we refer to "marking" here and elsewhere in this article, we do not wish to make any theoretical claims about the locus of marking in these constructions in Ap Ma (or in Waran). That is, we are not concerned with whether spatial deixis is an inflectional category of adpositions, or whether deictic forms are being inflected for case-like categories (if these postpositions may be considered as such). The simple facts of this fusion of deictic forms and adpositions are fascinating in their own right and, we believe, deserving of closer examination.

3. Waran

Less is known about Waran compared to Ap Ma, and nearly all information concerning the language originates from Butler's (1981a) sketch grammar. However, we may begin by noting that there are three deictics in the language, which signal proximal (*ga-*), medial (*na-*), or distal (*da-*) distances. The semantics of this contrast may be somewhat different from the three-way contrast found in Ap Ma, since what is treated here as a "medial" category in Waran is described by Butler (1981a: 38) as indicating referents that are near the hearer (i.e., addressee-proximal). Moreover, unlike in Ap Ma, deictics in Waran may appear in their bare forms (i.e., free) when they are used to indicate a location, in crosslinguistically expected fashion for demonstrative adverbs (16, 17).

(16) Waran (Ramu; Butler 1981a: 38)
gu ***ga*** *pe-Ø*
1SG **PROX** exist-N3SG
'I live here.'

(17) Waran (Ramu; Butler 1981a: 16)
a-na *kai* *erake* ***da*** *se-t*
1PL.EXCL-POSS banana garden **DIST** exist-3SG
'Our banana garden is over there.'

Deictics suffixed with the "demonstrative marker" *-ŋ* <ng> (Butler 1981a: 39) or with the 3SG pronoun *ma* both may function as modifiers of nouns (18, 19, 20) as well as pronominally (21). It is uncertain what the difference in usage between the two suffixes in such contexts is. It may be noted that the Waran 3SG pronoun *ma* is cognate with the Ap Ma medial deictic form *mɨ-*, although synchronically Waran uses *na-* as its medial form.

(18) Waran (Ramu; Butler 1981a: 15)
sin *da-dɨk* *u* *erake* ***ga-ng*** *tigo-pa-magɨn*
long.ago DIST-INS 2SG garden **PROX-DEM** cut-COMPL-FP
'Long ago you cleared this garden.'

(19) Waran (Ramu; Butler 1981a: 15)
maraksong ***ga-ma*** *enabu* *ra-se-t*
child **PROX-3SG** big become-PRS-3SG
'This child is becoming big (i.e., is growing).'

(20) Waran (Ramu; Butler 1981a: 37)
kabop *gu-ng* *kañe-ñe* *wa* *eñabu* ***ga-ng*** *kañe-ñe*
little PROX-DEM get-2.FUT or big **PROX-DEM** get-2.FUT
'Will you take this little (one) or this big (one)?'

(21) Waran (Ramu; Butler 1981a: 20)
ga-ma *kabop-bu*
PROX-3SG little-CMPR
'This one is smaller.'

Additionally, deictics may occur with the "specifier" suffix *-(a)p* (22), described as functioning "to make specific the item to which it is attached" (Butler 1981a: 44).

(22) Waran (Ramu; Butler 1981a: 10)

nga-na	*nga-may*	***da-p***	*Rikan*	*da-wa*	*kaka-ra-ñ*
1SG-POSS	1SG-wife	**DIST-SP**	Likan	DIST-at	leave-NP-N3

'I left my wife at Likan [village].'

Finally, deictics attach to all postpositions, in much the same way as in Ap Ma. Paradigms for adpositions in Waran are presented in Table 3.

Gloss	**Root**	**Proximal**	**Medial**	**Distal**
		ga-	na-	da-
'with' (instrumental)	-dɨk	gadɨk	nadɨk	dadɨk
'on' (temporal)	-jin	gajin	najin	dajin
'at'	-wa	gawa	–	dawa
'to'	-na	gana	–	dana
'from'	-ka	gaka	–	daka
'from' (animate source)	-gɨn	?	?	dagɨn
[unknown meaning]	-kɨn	?	?	dakɨn

Table 3: Waran postpositions (adapted from Butler 1981a: 49).

The meaning of *-kɨn* in Table 3 is uncertain. Additionally, Butler (1981a: 40) notes that he has not observed medial deictic forms occurring with "location relators" (i.e., 'at', 'to', 'from'), but we do not know whether such constructions would be ill-formed. It is also uncertain whether *-gɨn* 'from' (animate source) can combine with proximal or medial forms in addition to distal forms.

The following examples (23, 24, 25) illustrate the contrast among the three deictic markers in Waran in combination with the postposition *dɨk* 'with'. Note that deictic reference is not made with a pronoun but rather with the postposition.

(23) Waran (Ramu; Butler 1981a: 40)

kamin	***ga-dɨk***	*gora*	*na-ng*	*ra-pa-se-Ø*
bow	**PROX-with**	bird	MED-DEM	shoot-COMPL-PRS-N3SG

'With this bow I shot a bird.' (The bow is close to the speaker.)

(24) Waran (Ramu; Butler 1981a: 40)

kamin	***na-dɨk***	*gora*	*na-ng*	*ra-pa-se-Ø*
bow	**MED-with**	bird	MED-DEM	shoot-COMPL-PRS-N3SG

'With that bow I shot a bird.' (The bow is close to the hearer.)

(25) Waran (Ramu; Butler 1981a: 40)

kamin	***da-dik***	*gora*	*na-ng*	*ra-pa-se-Ø*
bow	**DIST-with**	bird	MED-DEM	shoot-COMPL-PRS-N3SG

'With that bow I shot a bird.' (The bow is not close to the speaker or the hearer.)

Other postpositions behave similarly, immediately following deictic markers (as opposed to the nouns that they govern), as illustrated in the following examples (26, 27, 28).

(26) Waran (Ramu; Butler 1981a: 46)

kas	*pep*	*nginang*	*enebu*	*Angoram*	***da-wa***	*kaje-Ø*
dog	black	two	big	Angoram	**DIST-at**	get.PRS-N3SG

'I bought two big black dogs at Angoram.'

(27) Waran (Ramu; Butler 1981a: 32)

gu	*krorowom*	***da-ka***	*paye-Ø*
1SG	Klorowom	**DIST-from**	come-N3SG

'I came from Klorowom.'

(28) Waran (Ramu; Butler 1981a: 40)

ñaka	*gu*	*krung*	***ga-na***	*yi-riyu-ñ*
yesterday	1SG	village	**PROX-to**	come-PST-N3

'Yesterday I came to the village.'

In unmarked situations where the location relationship is unimportant, the distal deictic form *da-* is used (Butler 1981a: 39). Notably, Waran differs in this regard from Ap Ma, which instead employs the medial deictic as the unmarked (or neutral) form in such circumstances. It is also worth noting here that, although the Waran proximal (*ga-*) and distal (*da-*) forms are both cognate with their Ap Ma equivalents (proximal *g(i)-* and distal *d(i)-*), the Waran medial form (*na-*) is innovative. In Ap Ma – as in many Keram-Ramu languages – the medial form (*m(i)-*) is clearly related to the 3SG pronominal form (*mi- ~ ma-*). In Waran, however, the medial form (*na-*) is, as Butler (1981a: 38) notes, very similar to the 2SG pronominal form (*na*) in that language; this accords with the form's addressee-proximal meaning.

No information is given on whether anything comparable to Ap Ma's semi-free nouns exists in Waran, nor are there any examples showing how postpositions interact with personal pronouns.

Waran also employs deictics with a temporal postposition *-jin*, which Butler (1981a: 41) describes as a "temporal relator". When following the proximal marker, this form is used to refer to an immediate time ('right now') (29); following the addressee-proximal marker, it refers to a time further off but still close (30); and following the distal marker, it refers to a time that is far off, at least a day in the future or past (31). Note that constructions with the temporal relator do not always appear to have an argument, such as in examples (29) and (30).

(29) Waran (Ramu; Butler 1981a: 41)
ga-jin *krung* *da-na* *me-ñe*
PROX-TR village DIST-to go-2.FUT
'Are you going to the village right now?'

(30) Waran (Ramu; Butler 1981a: 41)
na-jin *gu* *ma-bi-ta-mas*
MED-TR 1SG 3SG-see-?-1.FUT
'I'll look at it in a minute.'

(31) Waran (Ramu; Butler 1981a: 41)
waran *imanong* ***da-jin*** *me-mas*
no tomorrow **DIST-TR** go-1.FUT
'No, we'll go tomorrow.'

There is fairly clear evidence that deictics and adpositions interact in Waran in much the same way that they do in Ap Ma. However, unlike Ap Ma, which seems to contain rather nuanced phonological requirements for various parts of speech, there is no clear language-internal motivation for the development of such adpositional deictic marking in Waran. However, the two language areas are in close proximity and show a historical relationship, so one plausible explanation would involve contact-induced change. Therefore, we suggest that the Waran deictic adposition system was at least partially influenced by Ap Ma. Section 4.4 addresses this in greater detail.

4. A possible diachronic explanation

The origin of this construction likely stems from some particular aspects of Ap Ma phonology. Postpositions in Ap Ma belong to a closed set of enclitic morphs that require proclitics (such as deictics) as phonological hosts in order to form well-formed words. Before detailing the peculiar behavior of these clitics in Ap Ma in Section 4.3, we discuss two potentially relevant grammatical features found in the Keram-Ramu family that may have encouraged the development of deictic-marked adpositions in both Ap Ma and Waran. These are a semantically versatile class of articles (Section 4.1) and an oblique-marking morpheme that commonly follows deictic forms (Section 4.2). Finally, in Section 4.4, we consider the possibility that Waran was influenced by Ap Ma in the development of its similar postpositional constructions.

4.1. "Articles"

Throughout the Keram-Ramu family there are examples of NP-final determiners that are historically related to medial deictic demonstratives and/or third person personal pronouns. These determiners, which will be referred to here broadly as *articles*, seem to vary widely in their particular functions, both across and even within individual languages. They may indicate definiteness or specificity, but also may serve as topic markers or focus markers, and may index number and (to a more limited extent) case. They also vary in the degree to which they are used in particular Keram-Ramu languages: some members have no attestations of their use, some make occasional use of them, and some employ them rather frequently (Ap Ma sits at one extreme of this spectrum, using either these articles or deictic determiners for essentially every NP). In the languages that exhibit these articles, they are generally in complementary distribution with other determiners, namely proximal (Proto-Keram-Ramu *ŋga) and distal (Proto-Keram-Ramu *anda) deictic demonstratives and (if present) an indefinite marker. They may follow either common or proper nouns, but they do not cooccur with personal pronouns.

Since these frequently occurring articles may have paved the way for the proliferation of Ap Ma deictic forms (Section 4.3), we take a moment here to describe their behavior in some other languages belonging to the Keram-Ramu family.

These articles have been described, variously, as "subject markers" (and "object markers") in Ulwa (Barlow 2018: 166-195, 2019b: 4-7) and in Pondi (Barlow 2020:

87), as “echo pronouns” in Waran (Butler 1981a: 51-52), and as pronouns in “appositional noun phrases” in Rao (Christensen 1977: 23-24).

The following examples illustrate various uses of these articles. In Pondi they may index referent number (32, 33, 34). In Ulwa they may disambiguate grammatical relations (35, 36). In languages with flexible constituent order, such as Rao, they may provide essential case information (37). In Ambakich (as in Ap Ma), they have also developed predicative functions (cf. Li & Thompson 1977; Diessel 1999: 143-148), occurring in non-verbal clauses (38, 39).

(32) Pondi (Keram; Barlow 2020: 110)
kula-m ***mï***
boy-NPL **3SG.SUBJ**
‘the boy’[10]

(33) Pondi (Keram; Barlow 2020: 111)
kula-m ***min***
boy-NPL **3DU**
‘the (two) boys’

(34) Pondi (Keram; Barlow 2020: 111)
kula-wi *ndïn*
boy-PL **3PL.SUBJ**
‘the (more than two) boys’

(35) Ulwa (Keram; Barlow 2018: 170)
inom *manji* *ay* *ma = ama-p*
mother 3SG.POSS sago 3SG.OBL = eat-PFV
(a) ‘Mother ate her sago.’
(b) ‘(Someone) ate mother’s sago.’

(36) Ulwa (Keram; Barlow 2018: 170)
inom ***mï*** *manji* *ay* *ma = ama-p*
mother **3SG.SUBJ** 3SG.POSS sago 3SG.OBL = eat-PFV
(a) ‘Mother ate her sago.’
*(b) ‘(Someone) ate mother’s sago.’

[10] The grapheme <ï> in the Pondi and Ulwa data represents /ɨ/.

(37) Rao (Ramu; Christensen 1978: 32)

frendɨ	***me***	*ma*	*ole-lɨ*
man	**3PL.SUBJ**	3SG.OBL	drag-COND

'If the men dragged it …'

(38) Ambakich (Keram; Barlow, unpublished field notes 2018)

Lucy	*anetʃo*	***man***
[name]	woman	**NPL.MED**

'Lucy is a woman.'

(39) Ambakich (Keram; Barlow, unpublished field notes 2018)

anɨ	*erer*	***arin***
1PL	dog	**PL.MED**

'We have dogs.'

Although, as shown above, these Keram-Ramu articles can serve grammatical functions, their overall use seems largely determined by pragmatic factors – in particular, topic marking. Diessel & Breunesse (2020: 331) mention several Papuan languages in which topic markers have developed historically from demonstratives.

The presence of articles in a Papuan family deserves comment. Foley (2000: 383) notes that articles are uncommon in Papuan languages, and he (Foley 2000: 393) even attributes the loss of articles in some Austronesian languages to contact with neighboring Papuan languages. It is not inconceivable that contact has had the opposite effect in the case of the Papuan Keram-Ramu family – that is, that the evolution of the Keram-Ramu articles was influenced by contact with nearby (unrelated) Austronesian languages (specifically, Oceanic languages of the Schouten Chain; see Ross 1988: 122-132). There are several well-known examples of metatypy between Papuan and Oceanic languages in this region of New Guinea (cf. Ross 1996). It is even possible that the Oceanic influence here was not only structural but also formal: the nearby Oceanic language Manam, for example, has a "specifier" suffix *-ma*, which, among other things, follows NPs to "identify an object or a person as being the same as the one that has been spoken of before" (Lichtenberk 1983: 207-208, 364). Similarly, Kairiru (closely related to Manam) has a postnominal topic marker *mai*, which may be used either anaphorically or cataphorically (Wivell 1981:

187-190). Juillerat (2000: 50) even mentions Waran foundation legends of their ancestors having originally come by canoe from Manam island.

Regardless of the origin of these articles, their versatility and frequency in Keram-Ramu languages such as Ap Ma may have encouraged the proliferation of deictic forms in that language, due to the formal and semantic similarities between the two types of morphs.

*4.2. The oblique marker *-n*

Another morphosyntactic feature that may have encouraged the development of the Ap Ma and Waran systems of deictic marking is a postposed oblique marker that occurs in many languages of the family (Proto-Keram-Ramu *-n). This oblique marker, which commonly immediately follows deictic forms, may have served as a model for semantically similar postpositions to require immediately preceding deictic markers.

In some Keram-Ramu languages, this marker may attach to any non-subject NP (including direct objects and objects of postpositions), whereas in other languages it is restricted to non-core arguments, especially non-core arguments that do not contain postpositions. Indeed, the oblique marker *-n, which often indicates instrumental, genitive, or beneficiary roles, behaves semantically very much like a postposition; however, it is considered separately here, since – at least in some languages – it patterns differently from postpositions morphosyntactically. For example, in Pondi, the oblique marker *=n* is restricted to determiners, whereas postpositions can follow any nominal element in that language. Examples of this oblique marker are given below for Pondi (40), Ap Ma (41), Ulwa (42), Waran (43), Akukem (44), Mikarew (45), and Kire (46).

(40) Pondi (Keram; Barlow 2020: 129)

tatï	*sanglama*	*ma=**n***	*kondiyam*	*oli-ï*
papa	axe	3SG.OBL=**OBL**	palm.sp	cut-IPFV

'Papa cuts a palm with an axe.'

(41) Ap Ma (Keram; Barlow, unpublished field notes 2018)

ape	*da-ma*	*ŋgai*	*nda-**n***	*su-ndi-e-p*	*ma*
father	DIST-MED	axe	DIST-**OBL**	pig-DIST-hit-PFV	EPIS

'Father killed the pig with an axe.'

(42) Ulwa (Keram; Barlow 2019a: 1031)
nungol = ***nï*** *ama-p-e* *lam*
child = **OBL** eat-PFV-DEP meat
'The meat was eaten by the child.'

(43) Waran (Ramu; Butler 1981a: 41)
waran *gu* *u-**na*** *kas* *a-i-Ø-rang*
no 1SG 2SG-**POSS** dog NEG-hit.PRS-N3SG-NEG
'No, I didn't hit your dog.'

(44) Akukem (Ramu; Daniels 2010: 121)
a *gua* *utuav-**in***
3SG go path-**OBL**
'He goes along the path.'

(45) Mikarew (Ramu; Capell 1951: 140)
'kø *'pokon* *uŋi'pamu-**ni*** *i'tji*
1SG yesterday village-**LOC** was
'I was in the village yesterday.'

(46) Kire (Ramu; Stanhope 1972: 60)
ana *Kire-ttə-**n*** *iki*
3SG Giri-upper-**LOC** be.at
'He is at [Upper] Giri [village].'

The preceding examples illustrate the broad use of this postposition-like oblique marker in the Keram-Ramu family. In some languages of the family, especially those in and around the Keram branch to the west, it seems to have developed a "preference" for pronominal or deictic (as opposed to nominal) hosts (this may have especially been the case in those languages that frequently employ postnominal articles). In Pondi, there developed an outright prohibition against *-n immediately following nominal arguments. Since this Keram-Ramu morpheme is semantically very much like a postposition, it is reasonable to imagine that it provided a model for all postpositions in Ap Ma to disfavor nominal hosts.

4.3. The proliferation of Ap Ma deictics

Although the presence in Keram-Ramu both of topic-marking articles and of a postposed oblique marker may have played a role in the genesis of the Ap Ma and Waran deictic systems, these facts alone certainly do not suffice to explain the extreme degree to which deictics pervade Ap Ma grammar. Part of the solution may lie in diachronic phonology – namely, the fact that almost every multisyllabic Ap Ma noun lost its initial syllable. Since many roots were originally disyllabic, this resulted in a great many monosyllabic nouns, including numerous homophones. For example, earlier forms like *mbapa 'wing' and *kapa 'house' both ended up as *pa* in Ap Ma. Following this great shedding of initial syllables, there seems to have developed a strong aversion to ending prosodic units with (unbound) vowel-final monosyllabic words. Notably, when giving citation forms of such words (e.g., *pa* 'house'), speakers almost invariably provide bound versions thereof (commonly by adding the focus marker *ndɨma* – thus: [*pandɨma*]).[11]

Wade (1983b: 3-4) refers to words such as *pa* 'house' in Ap Ma as "semi-free nouns", a class of vowel-final monosyllabic words that are "grammatically free" but are "phonologically bound to the following word or suffix and can never occur in isolation." The set of semi-free nouns is closed but includes a great many high-frequency concepts, such as 'house', 'village', 'jungle', 'sky', 'spear', 'water', 'fire', 'talk', 'name', 'dog', and 'pig'.

Semi-free nouns can potentially attach to a variety of hosts, so long as the host is a phonological word, such as a verb (47) or an adjective (48). Also, a semi-free noun may immediately precede a deictic, provided the deictic is itself a proclitic attaching to a following form (49), which may itself be an enclitic (50). This suggests that proclitic-enclitic pairings can themselves constitute phonological words.

(47) Ap Ma (Keram; Wade 1984: 303)

me-t	***pi-k**-ap*
MED-take	**fire-put**-PFV

'(He) caused it to stay on the fire.'

[11] Even based on very limited data, Laycock & Z'graggen (1975: 759) make a similar observation: "[Pronouns] take a range of suffixes … Nouns also show similar suffixes, including a prevalent *-dama* which occurs on all monosyllabic nouns in isolated, final and predicative positions, but which may be omitted elsewhere."

(48) Ap Ma (Keram; Wade 1984: 295)
li-nyaam *da-ma*
mosquito.net-big DIST-MED
'It is a big mosquito net.'

(49) Ap Ma (Keram; Wade 1984: 54)
l-e *alak* *no* ***pa-di-m****-a*
3PL-? also own **house-DIST-go**-IRR
'… they also will go to their own house.'

(50) Ap Ma (Keram; Wade 1984: 239)
su-du-ba *w-odu-p*
pig-DIST-NT 2SG-see-PFV
'… the pig saw you …'

Thus, it seems as though rules of prosody have strongly influenced the proliferation of postnominal deictic markers in Ap Ma, since they provide a semantically weak means of addressing a prosodic problem. Since a postposition is not a phonological word in Ap Ma, it cannot by itself serve as a host for a semi-free noun. However, by adding a deictic proclitic and thereby creating a proclitic-enclitic pairing, a speaker may permissibly use a semi-free noun in an adpositional phrase. Considering the great number of high-frequency items constituting this class of semi-free nouns, it is not unlikely that the [(semi-free) noun = deictic = postposition] structure was generalized to encompass all nouns, including those that are phonologically free.

4.4. Diffusion to Waran?

While this historical sound change and these (unusual) prosodic requirements may help to explain the situation in Ap Ma, they do not necessarily account for the similar (albeit somewhat less extreme) situation in Waran. Given the proximity of the two language areas, as well as their longstanding relationship, however, we consider it plausible that the Waran deictic system was influenced at least in part by that of Ap Ma.

Several Waran villages lie along the well-traversed Keram River, just upstream (southeast) from Ap Ma villages lying along the same river. Today, the closest Ap Ma

and Waran settlements on the Keram are within 10 kilometers of each other: they are Bopten (Ap Ma) and Yar (Waran). Almost exactly equidistant between the two is Yaut, one of the seven villages where Ambakich (Keram family) is spoken. However, the Ambakich language area mainly lies farther to the east, along the Porapora River. The village of Yaut was only built in the middle of the 20th century by the transposed Ambakich-speaking population of the former Angang village. Thus, traditionally, the Ap Ma and Waran people occupied contiguous stretches of land along the important conduit that is the Keram River. Furthermore, to the west of the Keram River, in the "Grass" area, sit several Ap Ma and Waran villages in considerable proximity, perhaps the closest two of which are Kekten (Ap Ma) and Pushyten (Waran), positioned some three kilometers apart from each other, along the same bayou. Figure 3 presents a map of these villages.

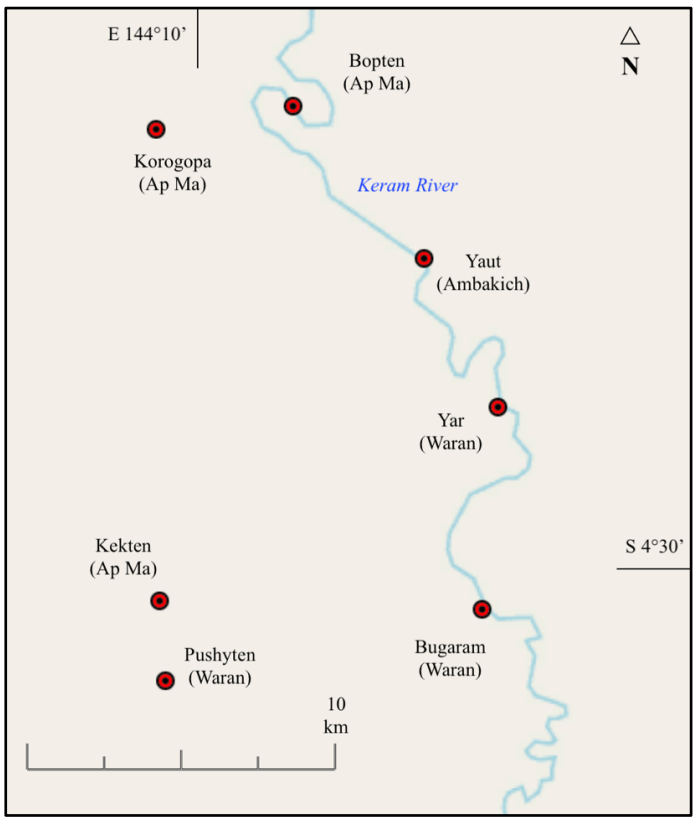


Figure 3: Neighboring Ap Ma and Waran villages.

Thurnwald (1916: 254) also notes some shared cultural traits of this region, including the use of bow and arrow (as opposed to dart) and pottery making, both of which he says are unknown in the Lower Sepik area. Thurnwald (1916: 281-282) also states that Waran kinship and marriage customs (the subject of his work) are shared by the

Ap Ma ("Kambót") people. Finally, linguistic contact is evident from the presence of lexical loans, occurring apparently bidirectionally between the two languages. Some likely borrowings from Waran into Ap Ma are given in Table 4, and some likely borrowings from Ap Ma into Waran are given in Table 5.

Gloss	**Waran**	>	**Ap Ma**
'woman'	miŋak	>	neŋ
'basket'	uman	>	wumɨn
'spirit sp.'	koŋim	>	koŋim
'spirit sp.'	mɨrom	>	molombi

Table 4: Likely loans from Waran into Ap Ma.[12]

Gloss	**Ap Ma**	>	**Waran**
'arrow'	peu	>	peʌp
'paddle'	napuŋ	>	nap
'betel nut'	au	>	o
'ground'	mbɨn	>	mbɨŋ

Table 5: Likely loans from Ap Ma into Waran.[13]

Therefore, in light of the known cultural and linguistic interactions between the Ap Ma- and Waran-speaking peoples, and in the absence of any known language-internal motivation for this highly unusual deictic behavior in Waran, we consider it plausible that deictic-marked adpositions in Waran developed at least in part thanks to influence from the neighboring Ap Ma language.

[12] Sources: <miñɑk> 'woman' (Butler 1981b: 31), <neg> 'woman; wife' (Wade 1983a: 68), <umɑn> 'woven basket' (Butler 1981b: 7), <wumɨn> 'basket' (Barlow 2021: 85), <koñim> 'spirits of nature' (Z'graggen 1972: 20), <konyim> 'spirit' (Wade 1983a: 48), <mʉrɔm> 'spirits of ancestors' (Z'graggen 1972: 21), <molobi> 'spirit whose face is found on garamut etc.' (Wade 1983a: 63).

[13] Sources: <peu> 'arrowhead' (Barlow 2021: 82), <peʌp> 'arrow' (Z'graggen 1972: 196b), <napug> 'paddle' (Wade 1983a: 67), <nɑp> 'paddle' (Butler 1981b: 18), <au> 'betel nut' (Wade 1983a: 5), <o> 'betelnut' (Butler 1981b: 6), <mbɨn> 'land, ground' (Barlow 2021: 79), <bʉŋ> 'ground' (Z'graggen 1972: 168).

5. Conclusion

This article has examined the synchronic and diachronic aspects of obligatory spatial deixis marking in adpositions, an extremely rare phenomenon crosslinguistically. This construction seems, at least in part, to have arisen due to a phonological requirement in Ap Ma. Namely, certain enclitics (among them, postpositions) require proclitics as phonological hosts in order to form well-formed words. This proclitic-enclitic co-hosting relationship is itself a remarkable phonological phenomenon and warrants future research to better understand its nature and, especially, its origin. At any rate, the particular prosodic demands of the language are likely what have encouraged the general spread of deictic markers in the language and the particular pattern of obligatory deixis marking in adpositional phrases discussed here.

While it is possible that phonology may have also played a role in the development of this phenomenon in Waran, it is reasonable to assume that the marking of spatial deixis in adpositional phrases was borrowed from Ap Ma, or at the very least was influenced by it.

Looking further back into the history of the Keram-Ramu family, to which both languages belong, we see some of the structural conditions that may have helped give rise to deictic-marked adpositions in Ap Ma and Waran. First, the frequent use of semantically versatile articles may have provided a model for the proliferation Ap Ma deictic forms, which share some semantic and distributional features with these articles. Likewise, the fact that the Keram-Ramu oblique marker *-n commonly follows deictic forms may have encouraged semantically similar postpositions to require immediately preceding deictic markers. Thus, even though areal diffusion likely played a role in the development of deictic marking in Waran, this language as well would have possessed some of the morphosyntactic conditions facilitating the advent of this unusual construction.

Regardless of its origin, however, the obligatory combination of deictic forms with adpositions is very rare in the world's languages. Languages such as Ap Ma and Waran add further attestation to the diversity and degree to which languages can encode spatial deixis in their grammars.

Acknowledgements

We are grateful to Martha Wade for her valuable comments and help. We also wish to thank the audience of the Linguistic Society of Papua New Guinea 2021 annual conference as well as two anonymous reviewers for their helpful questions and suggestions.

Abbreviations

? = morpheme of unknown meaning
1 = 1st person
2 = 2nd person
3 = 3rd person
AV = actor voice
CDM = core development marker
CMPR = comparative
COMPL = completive
COND = conditional
DEM = demonstrative
DEP = dependent
DIST = distal
DSTP = distant past
DU = dual
EPIS = epistemic marker
EXCL = exclusive
FOC = focus
FP = far past
FUT = future
INJ = interjection
INS = instrumental
IPFV = imperfective
IRR = irrealis
LOC = locative
MED = medial
N3 = non-3rd person
N3SG = non-3SG
NEG = negative
NOM = nominative
NP = near past
NPL = nonplural
NT = neutral topic
OBL = oblique
PFV = perfective
PL = plural
POSS = possessive
PREP = preposition
PRF = perfect
PROX = proximal
PRS = present
PST = past
QUANT = quantity
SG = singular
SP = specifier
STAT = stative
SUBJ = subject
TOP = topic
TR = time relator

References

Barlow, Russell. 2018. *A grammar of Ulwa*. PhD Dissertation, University of Hawai'i at Mānoa.

Barlow, Russell. 2019a. Agent demotion through inverted word order: Syntactic passives in Ulwa. *Studies in Language* 43(4). 1015-1037.

Barlow, Russell. 2019b. A syntactic motivation for valency reduction: Antipassive constructions in Ulwa. *Oceanic Linguistics* 58(1). 1-30.

Barlow, Russell. 2020. *A sketch grammar of Pondi*. Canberra: Asia-Pacific Linguistics (College of Asia and the Pacific, the Australian National University).

Barlow, Russell. 2021. Ambakich phonological sketch. *Language and Linguistics in Oceania* 13. 32-88.

Breunesse, Merlijn. 2019. *Demonstratives in space and discourse: A synchronic and diachronic analysis*. PhD Dissertation, University of Jena.

Bühler, Karl. 1934. *Sprachtheorie: Die Darstellungsfunktion der Sprache*. Jena: Gustav Fischer Verlag.

Butler, William. 1981a. Preliminary notes on Banaro grammar. Pioneer Bible Translators. Unpublished MS.

Butler, William. 1981b. A tentative phonology of Banaro. Pioneer Bible Translators. Unpublished MS.

Capell, Arthur. 1951. Languages of Bogia District, New Guinea. *Oceania* 22(2). 130-147.

Christensen, Fay. 1977. Rao grammar (from stem to clause): A preliminary study. Unpublished MS.

Christensen, Fay. 1978. Combinations of predications in Rao. Unpublished MS.

Daniels, Don. 2010. Akukem 1 notebook. Papuan languages collection: A collection of various materials gathered about several Papuan languages spoken in Madang Province, Papua New Guinea. Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC).

Diessel, Holger. 1999. *Demonstratives: Form, function, and grammaticalization*. Amsterdam: John Benjamins.

Diessel, Holger & Breunesse, Merlijn. 2020. A typology of demonstrative clause linkers. In Åshild Næss, Anna Margetts & Yvonne Treis (eds.), *Demonstratives in discourse*, 305-341. Berlin: Language Science Press.

Dixon, R. M. W. 2003. Demonstratives: A cross-linguistic typology. *Studies in Language* 27(1). 61-112.

Dixon, R. M. W. 2010. *Basic linguistic theory, volume 2: Grammatical topics*. Oxford: Oxford University Press.

Dryer, Matthew S. 2007. Word order. In Timothy Shopen (ed.), *Language typology and syntactic description, volume 1: Clause structure*, 61-131. Cambridge: Cambridge University Press.

Foley, William A. 2000. The languages of New Guinea. *Annual Review of Anthropology* 29. 357-404.

Goudswaard, Nelleke. 2005. *The Begak (Ida'an) language of Sabah*. Utrecht: LOT (PhD Dissertation, Vrije Universiteit Amsterdam).

Grimes, Charles E. 1991. *The Buru language of eastern Indonesia*. PhD Dissertation, Australian National University.

Gruzdeva, Ekaterina. 2013. Демонстративные глаголы в языках мира: проблемы идентификации и анализа [Demonstrative verbs in the languages of the world: Problems of identification and analysis]. *Acta Linguistica Petropolitana* 9(3). 51-80.

Guérin, Valérie 2015. Demonstrative verbs: A typology of verbal manner deixis. *Linguistic Typology* 19(2). 141-199.

Hagège, Claude. 2010. *Adpositions*. Oxford: Oxford University Press.

Himmelmann, Nikolaus P. 1996. Demonstratives in narrative discourse: A taxonomy of universal uses. In Barbara Fox (ed.), *Studies in anaphora*, 205-254. Amsterdam / Philadelphia: John Benjamins.

Juillerat, Bernard. 2000. Do the Banaro really exist? Going back after Richard Thurnwald. *Oceania* 71(1). 46-66.

Killian, Don. 2021. Towards a typology of predicative demonstratives. *Linguistic Typology*. https://doi.org/10.1515/lingty-2021-2078.

König, Ekkehard & Umbach, Carla. 2018. Demonstratives of manner, of quality and of degree. In Marco Coniglio, Andrew Murphy, Eva Schlachter & Tonjes Veenstra (eds.), *Atypical demonstratives: Syntax, semantics and pragmatics*, 285-328. Berlin: De Gruyter.

Kruspe, Nicole D. 1999. *Semelai*. PhD Dissertation, University of Melbourne.

Laycock, Donald C. & Z'graggen, John A. 1975. The Sepik-Ramu phylum. In Stephen A. Wurm (ed.), *New Guinea area languages and language study, volume 1: Papuan languages and the New Guinea linguistic scene*, 731-764. Canberra: Research School of Pacific and Asian Studies, the Australian National University.

Levinson, Stephen C. 2018. Introduction: Demonstratives: Patterns in diversity. In Stephen C. Levinson, Sarah Cutfield, Michael J. Dunn, N. J. Enfield & Sérgio Meira (eds.), *Demonstratives in cross-linguistic perspective*, 1-42. Cambridge: Cambridge University Press.

Li, Charles & Thompson, Sandra. 1977. A mechanism for the development of copula morphemes. In Charles Li (ed.), *Mechanisms of syntactic change*, 419-444. Austin: University of Texas Press.

Lichtenberk, Frantisek. 1983. *A grammar of Manam*. Honolulu: University of Hawai'i Press.

Lyons, John. 1977. *Semantics, volume 2*. Cambridge: Cambridge University Press.

Pryor, John. 1986. A provisional semantic-structure analysis of Botin hortatory and descriptive discourse. Pioneer Bible Translators. Unpublished MS.

Pryor, John. 1990. Deixis and participant tracking in Botin. *Language and Linguistics in Melanesia* 21. 1-29.

Pryor, Bonita & Farr, Cindi. 1989. Botin deictics: Go and come. *Language and Linguistics in Melanesia* 20. 115-145.

Robert, Stéphane. 2006. Deictic space in Wolof: Discourse, syntax and the importance of absence. In Maya Hickmann & Stéphane Robert (eds.), *Space in languages: Linguistic systems and cognitive categories*, 155-174. Amsterdam / Philadelphia: John Benjamins.

Ross, Malcolm. 1988. *Proto-Oceanic and the Austronesian languages of western Melanesia*. Canberra: Pacific Linguistics, the Australian National University.

Ross, Malcolm. 1996. Contact-induced change and the comparative method: Cases from Papua New Guinea. In Mark Durie & Malcolm Ross (eds.), *The comparative method reviewed: Regularity and irregularity in language change*, 180-217. New York: Oxford University Press.

Stanhope, John M. 1972. The language of the Kire people, Bogia, Madang District, New Guinea. *Anthropos* 67. 49-71.

Thurnwald, Richard. 1916. Bánaro society: Social organization and kinship system of a tribe in the interior of New Guinea. *Memoirs of the American Anthropological Association* 3. 251-391.

Wade, Martha. 1982. Pronouns and deictics in Botin. Pioneer Bible Translators. Unpublished MS.

Wade, Martha. 1983a. Botin/English dictionary. Pioneer Bible Translators. Unpublished MS.

Wade, Martha. 1983b. A tentative analysis of Botin morphophonemics in a generative format. Pioneer Bible Translators. Unpublished MS.

Wade, Martha Lynn. 1984. *Some stratificational insights concerning Botin (Kambot), a Papuan language*. MA Thesis, University of Texas at Arlington.

Wivell, Richard. 1981. *Kairiru grammar*. MA Thesis, University of Auckland.

Z'graggen, John A. 1971. *Classificatory and typological studies in languages of the Madang district*. Canberra: Pacific Linguistics, the Australian National University.

Z'graggen, John A. 1972. Comparative wordlist of the Ramu language group, New Guinea. Alexishafen: Anthropos Institute. Unpublished MS.

CONTACT

donald.killian@helsinki.fi

russell_barlow@eva.mpg.de

Linguistic Typology at the Crossroads 2-1 (2022): 65-93 ISSN 2785-0943

Comparative constructions in Suansu and the languages of northeastern India

JESSICA K. IVANI
UNIVERSITY OF ZÜRICH

Submitted: 28/10/2021 Revised version: 27/01/2022
Accepted: 09/02/2022 Published: 14/04/2022

Abstract

This paper provides a first description of comparative constructions in Suansu, an unreported Tibeto-Burman language spoken in northeastern India, and frames the characteristics of Suansu comparative constructions from a typological perspective (following Stassen's 1985 classification). To this purpose, comparative constructions from a sample of 25 Tibeto-Burman languages of the area are collected in an ad-hoc designed database and typologically discussed. Results reveal the presence of two main types that cluster geographically in the region, as well as high internal variation with respect to the subtypes. Based on the classification, Suansu is assigned to the Exceed comparative type, the only representative of this type in the sample.

Keywords: Suansu; Tibeto-Burman; typology; comparative constructions; language documentation.

1. Introduction

Comparison can be defined as a "mental act where two elements occupy a symmetric or asymmetric position on a specific property predicative scale" (Treis 2018: I): this mental act is linguistically encoded by comparison constructions. Several types of comparison structures have been distinguished in the literature. Based on the subdivision proposed by Fuchs (2014) and implemented in Treis (2018), this study focusses on comparison of inequality, and specifically on comparison of relative superiority. The default template for constructions of relative superiority (hence, comparative constructions) is exemplified by structures such as *X is bigger than Y* or *X is more intelligent than Y*.

The linguistic literature has proposed several terms to define the elements of prototypical comparative constructions. The terminology adopted in this study relies on Treis (2018: I), where the following elements are distinguished: COMPAREE (the entity being compared: *X*); STANDARD (what the compared is being compared to: *Y*); standard marker (grammatical function of the standard: *than*); PARAMETER (the property of comparison: *tall, intelligent*); DEGREE (the degree of presence of a property in the comparee: *-er, more*).

This aim of the present study is two-fold. Firstly, to describe the comparative constructions in Suansu, a virtually undescribed Tibeto-Burman language from northeastern India. The second objective is to explore the structural variation of comparative constructions of other 24[1] Tibeto-Burman languages spoken in the same area, to frame their typology, and outline the geographical distribution of the identified types.

The paper is structured as follows. Section 1.1. summarizes and discusses the classification of comparative constructions outlined in Stassen (1985), that is the typology adopted in this study. In Section 2, I briefly introduce Suansu language, and report a first profile of Suansu comparative constructions (2.1.). The second part of the study (Section 3) is dedicated to the description of comparative constructions across 24 Tibeto-Burman languages spoken in northeastern India. In Section 3.1., the methodological approach used for the cross-linguistic data exploration and the database design are briefly illustrated. Section 3.2. describes the expression of comparative constructions in the Tibeto-Burman sample. Section 4 discusses the types attested in the area, and Suansu comparative constructions are included in the typology. The paper concludes (Section 5) with a summary of the typological findings and sets the ground for further typological research on the topic.

1.1. Background

There is broad and rich literature available on the expression of comparative constructions, from theoretical approaches to language specific overviews, as well as typological research. Typological classifications of comparative constructions have a long tradition and include the works of Ultan (1972), Heine (1997), Stassen (1985), and Dixon (2008), among others (see Stolz 2013 & Treis 2018 for a review).

[1] In an earlier version of this study, the sample included three more languages from the Tani subgroup (Tawrã, Upper Adi and Mising), which were later excluded because the cited grammars have been contested as unreliable.

The present study follows the typology of comparative constructions outlined by Stassen (1985), with additional input from Stolz (2013). The types identified in Stassen's typology are summarized in Table 1.

Taxonomy	**Types**
Fixed-case comparatives	a) Exceed b) Adverbial • Allative • Locational • Separative
Derived-case comparatives	a) Conjoined b) Particle

Table 1: Taxonomy of Stassen's typology.

Stassen's typology (as well as several other classifications, e.g., Dixon 2008) is mostly grounded on the etymological and semantic properties of the constructions found on the standard of comparison. Based on these properties, Stassen distinguishes between two main types: derived-case and fixed-case comparatives, with the main criterion behind the partition being whether the case of the standard is dependent or independent from the comparee.

Within the fixed-case comparatives, Stassen differentiates further in Exceed and Adverbial comparatives (the latter relabeled Locational in Stassen 2013). Derived-case comparatives include Conjoined and Particle comparatives subtypes.

Exceed comparatives are characterized by the presence of a transitive verb with the meaning of 'exceed' and 'surpass', with the comparee and the standard of comparison being the subject and the object, respectively. Conjoined comparatives, on the other hand, are formed by two independent clauses, one containing the comparee and the other the standard, with the parameter present in both; the clauses are in adversative relation.

Most of the languages identified in this study fall into Stassen's Adverbial/Locational (henceforth, Adverbial) and Particle types. The Adverbial comparative type includes three subtypes: the Separative comparatives (Stassen 1985: 114–135), where the standard of comparison is marked through a morpheme of source and origin. Stassen provides as examples of standard markers ablative forms

such as 'from'.[2] Comparatives of this subtype are found in Kambaata (Afroasiatic, Cushitic, Treis 2018) and several Athabascan languages (Koyukon, Tanana and Athan among others, Tuttle 2018).

The second subtype within the Adverbial type is the Allative comparative. In Stassen's classification, Allative comparatives mark the standard of comparison as a goal, benefactive or direct object (Stassen 1985: 136–145).

The third subtype within the Adverbial type is the Locative comparative (Stassen 1985: 146–152), and the standard is marked with an element that indicates contact or static location ('on', 'at', 'beside'), usually derived from locational adverbs.

Derived-case comparatives include the above-mentioned Conjoined comparatives and Particle comparatives. The Particle type, according to Stassen's classification, includes a comparative marker that "does not influence the case marking of the standard" (Treis 2018: II). These heterogeneous particles include, among others, disjunctive forms, negators and coordinators.

The above typology leaves some questions unresolved (c.f. Stassen 1985; Heine 1997; Stolz 2013). One potential source of ambiguity pertains to the heterogeneity of the types identified in the typology: Stassen explicitly mentions "sources of indeterminacy" (Stassen 1985: 36) with respect to the Adverbial comparatives, stressing the semantic variation of the standard markers categorized under this type, and which further segmentation would lead to "a proliferation of other subclasses" (Stassen 1985: 36). The Particle type encounters similar limitations, being "not a homogeneous class" (Stassen 1985: 46), and often characterized by "etymologically nontransparent forms" (Heine 1997: 120), which at times "cannot be associated with any co-existing functional element of the language in which they occur" (Stolz 2013: 21). The heterogeneous nature of the Particle type, combined with a general opaqueness of the constructions included to the type, prompted the introduction of further subclassifications. Additional subtypes have indeed been proposed, and only two types are relevant to this study. The first is the Pure comparative type introduced by Stolz (2013: 22), that includes the constructions in which the standard is marked by a dedicated comparative marker. The second subtype is the Companion Schema by Heine (1997: 93–94), that comprises constructions with comitative and instrumental forms marked on the standard of comparison. Comitative marking on the standard, considered typologically rare, is tentatively included by Stassen (1985: 37) under the Separative comparatives, but it is not discussed further.

[2] In Stassen's terminology, marking is intended in broad morphological and non-morphological sense.

Both Heine and Stassen, as well as most of the typological classifications, do not include in their respective typologies the forms found outside the standard (one exception being Bobaljik 2012). Parameters such as the presence or the morphological characteristics of the degree markers are often discarded (Treis 2018: IX). This study follows Stassen's typology, and therefore the degree markers will not be discussed in detail nor incorporated in the classification. However, considering potential future updates in the typology of comparative constructions, the variables and properties related to the degree, comparee, and parameter have been collected in a detailed database (Section 3.1).

2. Suansu language

Suansu is an endangered Tibeto-Burman language spoken in a small cluster of villages in Manipur, northeastern India. Suansu has approximately 2200[3] speakers, located in the Ukhrul district of Manipur, not far from the Myanmar border.

Suansu is currently virtually undescribed in the literature. Suansu features comprise strict verb-final word order, a rich case marking system, and ergative - absolutive alignment. Ergative marking appears to be motivated by pragmatic and/or semantic factors, consistent with several other Tibeto-Burman languages (DeLancey 2011). Further features align Suansu to the scarce typology available on the languages of the eastern border area (Burling 2003b: 173). These include the presence of three lexical tones, noun compounding strategies and frequent "frozen prefixes" (Marrison 1967: 108) attached to verbs, nouns and adjectives (with unclear semantic distribution). Other common traits within the languages of the area and attested in Suansu include a rich verb morphology, with several affixes (mostly suffixes), and the absence of verb agreement altogether.

The linguistic data on Suansu used in this study has been collected between 2017 and 2019, during several fieldwork trips to Pune, Maharashtra, where a heterogeneous Suansu speaking community works and lives. Part of the data has been time-aligned, transcribed and annotated; specific examples here reported are drawn from a corpus that includes narratives, folk tales and elicited examples from three different native speakers.

[3] According to the most recent public Census available, the 2011 Indian Census: http://censusindia.gov.in.

2.1. Comparative constructions in Suansu

The present Section outlines a first report on comparative constructions in Suansu. Being a first report on the language, the description that follows should be treated as preliminary, and has no claim of exhaustivity.

The word order template for Suansu comparative constructions is structured as follows: COMPAREE - STANDARD - PARAMETER - DEGREE MARKER. An example is illustrated in (1).

(1) *ha-ʃi-ne* *nɔ-ʃi-di* *tʰazu* *mɛn-le*
1PL-house-ERG 2SG-house-ABS beautiful more-be.PRS
'Our house is more beautiful than your house.'

The comparee is morphologically marked by the suffix *ne*, which is used in Suansu to mark ergative case and A in general (2). The standard immediately follows the comparee and is suffixed by the absolutive marker *di*, which covers P (2), S (3), and experiencer (4).

(2) *ba-ne* *klui-di* *kətərum-no* *huamsuɛ*
3SG-ERG rope-ABS tree-LOC tie.PST
'She/He tied the rope to the tree.'

(3) *hai* *həma-di* *mari-də* *samhai*
DET pot-ABS iron-INS make.PST
'This pot is made of iron.'

(4) *gəpʰem-di* *miŋə-le*
ice-ABS melt-PRS
'Ice melts.'

The word order template illustrated above is not strict, and the standard can be found preceding the comparee, as illustrated in (5). Further research is needed for a better understanding of the word order distributions in Suansu.

(5) *pjəs-di* *kaminta-ne* *am:etok^{h}* *mɛn-le* *ʔasə*
onion-ABS tomato-ERG expensive more-be.PRS today
'Tomatoes are more expensive than onions today.'

Overt marking on the comparee is not obligatory in Suansu comparative constructions (6), as ergative marking can be optional in general (7); more data is required to account for the distributional trends and occurrences of this marker, although first insights from the data available seem to suggest that the presence of agent marking is motivated by pragmatic contexts.

(6) *ha-ʃi* *nɔ-ʃi-di* *tʃutʃu* *mɛn-le*
1PL-house 2SG-house-ABS small more-be.PRS
'Our house is smaller than yours.'

(7) *ba* *tatʃu-di* *dukan-də* *lu-le*
3SG rice-ABS market-ABL buy-PRS
'She/He buys rice at (from) the market.'

The parameter is followed by the particle *mɛn*: the meaning can be linked, based on the occurrences of the form found elsewhere in the grammar, to 'more' (8).

(8) *ha* *ɬo* *tʃõ* *mɛn* *laŋe* *dorgat^{h}e* *le*
1PL field CLF more cultivate need be.PRS
'We need to cultivate more land.'

At the state of the art, the origin of *mɛn* is unclear: it is not linked to any spatial nor locational particle found in the data and it does not find correspondences in Tangkhul, the predominant linguistic neighbor. The particle *mɛn*, used in combination with the verb *le*, 'to be', acquires the meaning of 'to be more', 'exceed', as shown in the following non-comparative example (9).

(9) *asserikom-va* *taciu-di* *assokom-va-ne* mɛn-*le*
last.year-GEN grain-ABS present.year-GEN-ERG more-be.PRS
'This year's harvest surpassed last year's harvest.'

Similar particles, possibly related, are attested in the neighboring languages spoken in the Ukhrul district. Examples include Huishu (*kʰə-mu* 'more than', Mortensen 2013), Kachai (*kʰə-mi* 'more than', Mortensen 2013), Tusom (*cũ-kʰə-ma* 'more than', Mortensen 2013), Ukhrul (*kʰə-mɒj* 'to be more', Mortensen 2013). These forms suggest a hypothesized Proto-Tangkhulic *mej 'more' (Mortensen 2013: 402). Data on these languages is limited to wordlists and does not contain grammatical information on these forms.

In Suansu comparative constructions, the standard is constructed as the direct object, with the comparee as the subject. The transitive predicate suggests a meaning related to 'exceed', 'surpass', construed through the particle *mɛn* and the verb form *le,* 'to be'.

Thus, Suansu comparative constructions can be assigned to the Exceed comparative type defined in Stassen's typology. This assignment confirms Stassen's findings on the geographical distribution of this type, whose presence seems restricted to two geographical areas, sub-Saharan Africa and South-East Asia. In the following Section, I turn to other languages of the eastern border area and the respective comparative constructions.

3. Comparative constructions in the languages of northeastern India

3.1. Methodology

In what follows, I present data on comparative constructions from 24 Tibeto-Burman languages (25 including Suansu). With a few exceptions, the languages are spoken in northeastern India: the sample includes languages spoken in the fringes of this region, such as Myanmar and the Himalayan range.

The languages of the sample belong to different branches within the Tibeto-Burman sub-family.[4] Kuki-Chin-Naga and Brahmaputran are the most represented subgroups in the sample, followed by Mruic. Other genealogical subgroups, such as Bodic, Kiranti, Burmo-Qiangic, Karenic, Macro-Tani, and Raji, are also represented, although with fewer languages. The main criteria for language selection include their location (northeastern India) and the availability of the sources. The detailed sample, grouped by linguistic subgroup, is illustrated in Table 2.

[4] Genealogical affiliations follow the classification reported on Glottolog (Hammarström et al. 2020).

Subgroup	**Languages**
Bodic (1)	Bunan (gahr1239)
Brahmaputran (7)	Atong (aton1241), Chothe (chot1239), Garo (garo1247), Kadu (kado1242), Konyak (kony1248), Turung (sing1264), Rabha (rabh1238)
Kiranti (2)	Chintang (chhi1245), Yakkha (yakk1236)
Karenic (2)	Geba Karen (geba1237), Kayah Monu (kaya1316)
Kuki-Chin-Naga (6)	Karbi (karb1241), Mao Naga (maon1238), Mongsen Ao (aona1235), Moyon (moyo1238), Suansu (suan1234), Tangkhul (tang1336)
Macro-Tani (3)	Apatani (apat1240), Galo (galo1242), Tangam (tang1377)
Mruic (2)	Hkongso (anuu1241), Mru (mruu1242)
Burmo-Qiangic (1)	Burmese (nucl1310)
Raji-Raute (1)	Raji (rawa1264)

Table 2: Language sample by genealogical subgroup.

The linguistic data is collected in a multivariate typological database. The primary goal of the database is to collect information on comparative constructions at the most refined level of detail.[5] The methodology adopted in designing the database relies on the autotypologizing method and the late aggregation principles. The autotypologizing method (Bickel & Nichols 2002) describes a bottom-up, data-driven approach that starts at the earliest stages of data collection. Instead of "fitting" the structures found in the languages of the sample in a "conceptual grid" of traits established a priori, the constructions are collected dynamically, along with the data collection process (in a multivariate approach fashion, cf. Bickel 2010a; Bickel 2010b). Results from the data are eventually framed within the types identified in the literature at a later phase (following the principle of late aggregation, see Bickel et al. 2016). These methods have shown their potential in several typological databases (Bickel et al. 2017; Ivani & Zakharko 2019); in addition, the granularity of the data collected enables its reusability for further research.

The data collection procedure is structured as follows. For each language, I describe in detail the structures that characterize the respective comparative construction. The forms and types illustrated in the database encompass both morphological and non-

[5] The database is stored and freely accessible on GitHub (https://github.com/jkivani/coi-neils). The dataset version for the present study (Ivani 2020) is available on the public access Zenodo repository (https://doi.org/10.5281/zenodo.4274488).

morphological means used to define the relations among the comparee, the standard and the parameter.

Each construction is stored with a set of metadata information: these include an identifier, the respective language name (with the related Glottocode), and the language genealogical subgroup. The identifier is unique, but it can be shared by several constructions in different languages in cases of attested cognacy relationships. Linguistic information includes the language specific form, the type of marker, the locus, and the function or meaning (when available), associated to the form in the grammar. Table 3 exemplifies the coding sheet for Atong language. The forms individuated during the data collection process are then assigned to the pre-existing typology described in Section 1.1. and then discussed further.

ID	Glottocode	Language	Subgroup	Form	Type	Locus	Function	Source
ID003	aton1241	Atong	Brahmaputran	*-na*	suffix	standard	goal	Van Breugel 2014
ID004	aton1241	Atong	Brahmaputran	*-khal*	suffix	parameter	more	Van Breugel 2014

Table 3: Coding sheet for Atong language.

3.2. Data overview

The data available on comparative constructions in the languages of the sample reveals interesting structural variation. In what follows, I present cases from individual languages by genealogical subgroup, and describe the strategies used to express the respective comparative constructions.[6]

The Tani languages of the sample, spoken mostly in Arunachal Pradesh, are Apatani, Tangam, and Galo. All the Macro-Tani languages included in the sample share a degree marker, *ya* in Apatani, *yaŋ* in Tangam, and *jaa* in Galo. The meaning of this form seems to be 'more', as reported in the linguistic sources, and it is linked to the Proto Tani form *jaŋ (Sun 1993: 122). All three languages show the same

[6] The examples from the individual languages are reported *verbatim*, with the original glosses described in the respective sources, without any relabeling, except for minor normalization adjustments.

behavior in expressing comparison. In addition to the degree marker, Apatani comparative constructions have a standard marker, *mi* (10), and it is linked to functions such as non-agentive and direct object, among others.[7]

(10) Apatani (Macro-Tani; Abraham 1985: 136)
mado *rinyo-mi* *ka-pyo* *ya-do*
mado rynio-ACC see-good COMP-exist
'Mado is more beautiful than Rynio.'

Similarly, the standard of comparison in both Tangam and Galo comparative constructions is marked by a non-agentive relational marker: the form corresponds to the postposition *me* in Tangam and the clitic *nè* in Galo, related in form and function to the marker found in Apatani. In both languages, the distribution of the non-agentive and the accusative is semantically controlled by the definiteness and the animacy of the referent (Post 2017: 102).[8] This restriction is reflected in the respective comparative constructions. In example (11), the only sentence available from Galo sources where all the elements of the comparative construction are expressed overtly, the accusative clitic *əəm* is used on the standard, being the standard of comparison a non-human noun. The non-agentive postposition is found on the human referent in the example from Tangam (12).

(11) Galo (Macro-Tani; Post 2007: 548)
…takə́ = əəm *dór-tə̀-jàa-dó(o)-nà = əə = na*
…squirrel = ACC CLF:HIGH.ANIMAL-big-COMP-STAT-NZR:SUB = COP.IPFV = DECL
'…you know, they're bigger than squirrels.'

(12) Tangam (Macro-Tani; Post 2017: 128)
nodɨ *ŋo = me* *abəŋ-yaŋ-du(ŋ)*
3SG 1SG = NAGT mature-COMP-IPFV
'He is elder to me.'

[7] A reviewer suggests caution in using Abraham (1985) as a source and recommends consulting in parallel other sources and descriptions treating related languages. The functions listed here rely on Simon (1972: 5) and through comparison with Galo and Tangam.

[8] Both languages include dative and locational forms in their respective referential marking systems. See Post (2007: 58) for Galo and Post (2017: 108) for Tangam language.

The Brahmaputran languages Atong, Garo and Rabha employ the marker *na* (cliticized in Atong) which covers, in the three languages, goal, indirect object, and beneficiary role. In addition, the languages show variation with respect to the marking combination in the respective comparative constructions. In Atong, the dative form is suffixed to the standard (13). The dative suffix in Garo is used in combination with *bate* (14), whose meaning is glossed 'as compared to'. This marker operates as a dedicated form used in Garo comparative constructions (Burling 1961: 46). Furthermore, Atong presents the suffix *khal*, that occurs on the parameter. The form, used for generic intensification and meaning 'more', is a degree marker (Burling 2003a: 99).

(13) Atong (Brahmaputran; Van Breugel 2014: 278)
aŋ *naŋʔ=na* *cuŋ-khal=a*
1SG 2SG=GOAL big-COMP=CUST
'I am bigger than you.'

(14) Garo (Brahmaputran; Burling 1961: 19)
acak-na-bate *moiyr* *da'r-bate*
dog-GOAL-COMP elephant big-COMP
'An elephant is bigger than a dog.'

A suffixal compounding strategy analogous to the one illustrated for Garo, is found in Rabha (15). Rabha comprises affixal particles and postpositions that are used in combination with case markers to accomplish a more specific meaning or function. In Rabha comparative constructions, the dative *na* on the standard is followed by the particle *kára*, 'above, over', which appears to be partially grammaticalized (Joseph 2007: 762).

(15) Rabha (Brahmaputran; Joseph 2007: 360)
e-kai *pan* *o-kai* *pan-na* *kára* *cuh-a*
this-ATTR tree that-ATTR tree-DAT more big-PRS
'This tree is bigger than that tree.'

The other Brahmaputran languages of the sample, Kadu and Konyak, show a range of strategies in the respective comparative constructions that differ from the cases

illustrated above. In Kadu, the clitic *athá* follows obligatorily the standard and is of unclear origin: its use seems to be restricted to the marking of comparison. A topic marker, *ká*, is optionally found on the comparee (Sangdong 2012: 331).

Konyak (16) uses the form *phəy* (Nagaraja 2010: 59) suffixed to the standard; *phəy* is used elsewhere in the grammar to signal the "inanimate force or object casually involved in the action or state identified by the verb" (Nagaraja 2010: 62). In addition, *phəy* covers sociative and instrumental marking (excluding causal, goal and source functions, Nagaraja 2010: 69). The marker *phəy* is homophonous with the adverb *phəy*, 'behind', used also in combination with the locative marker *me* such as in *nòkphəyme*, 'behind the house' (Nagaraja 2010: 68), for which I assume that it is the etymological source of the standard marker in Konyak. In addition, the form *si*, possibly a degree marker, is prefixed to the parameter.

(16) Konyak (Brahmaputran; Nagaraja 2010:155)
kùy-ə *əmi-phəy* *si-yòŋ*
dog-NOM cat-with COMP-big
'The dog is bigger than the cat'

In Turung, the standard is followed by the particle *ngga* (or *nloh*), that corresponds to 'more'. The comparee is marked by the agentive/ergative marker *î* (17).

(17) Turung (Brahmaputran; Morey 2010:296)
Kon Kham î Kon Seng ngga coh
Kon Kham AG Kon Seng COMP tall
'Kon Kham is taller than Kon Seng.'

No further information nor additional occurrences are available on the form *tre*, the standard marker in Chothe (Singh 2000: 271). The suffix *he* on the parameter is reported with the meaning of 'excessive' (Singh 2000: 215). The comparee carries the nominative marker *na*.

In Bunan (18), the standard is marked by both the dative clitic *tok* and the ablative *tɕi*. The two forms combined indicate "a motion away from a generic location" (Widmer 2014: 237). A similar strategy is found in Chintang (Kiranti). Chintang (19) has an extensive case marking system (Paudyal 2015: 42), and the standard in the comparative constructions is suffixed by *ʔ*, that is the marker of focussed, specific

location, followed by *ya*, that expresses instrumental force (Paudyal 2015: 44). In Yakkha, the other Kiranti language of the sample, comparative constructions are expressed through the particles *haʔniŋ* and *haksaŋ* (in free distribution). Schackow (2015: 145) discusses possible meanings of the form, the most likely candidate being the verbal stem *haks* ‘send, send up’, but also ‘weigh’.

(18) Bunan (Bodic; Widmer 2014: 328)
tedzi=tsuk *tete* *gi=tok=tɕi* *tedzi jen*
big=REL grandfather 1SG=DAT=ABL big EQ.CJ
‘The older grandfather (of yours) is older than me.’

(19) Chintang (Kiranti; Paudyal 2015: 49)
hani-ʔ-yã *the=kha*
2s-LOC-ABL big=NZR
‘Bigger than you.’

The Mruic languages Hkongso (20) and Mru (21), closely related to each other, employ the particles *luki* and *lake*, that follow the standard in their respective comparative constructions. These forms have been linked to *luk* (Wright 2009: 66), that has the temporal and locative meaning of ‘side’. The particle has the additional meaning ‘more’, as attested in expressions such as *la luk* lit. ‘month more’, “next month” (Wright 2009: 34). Besides, Mru uses the particle *lang* meaning ‘different’ (Ebersole 1996: 9).

(20) Hkongso (Mruic; Wright 2009: 119)
dai˥ *cəʔ˩* *koko˧* *luk˩* *r̥ʰau˧*
dai TOP koko COMP tall
‘Dai is taller than Koko.’

(21) Mru (Mruic; Ebersole 1996: 17)
enning kim *lake* *anging kim* *lang* *iuk*
2PL house COMP 1PL house difference big
‘Our house is bigger than yours.’

Plural markers can be involved in comparative constructions: Raji-Raute language Raji, spoken in Uttarkand, uses the plural particle *jəmma* (Krishan 2001: 84), followed by the parameter (22).

(22) Raji (Raji-Raute; Krishan 2001: 84)
pəhare *ti* *gədde* *ti* *jəmma* *thənda* *hwā*
mountain water river water COMP cold COP
'Mountain water is colder than river water.'

The languages belonging to the Kuki-Chin-Naga subgroup show higher internal diversity, but scarce information is available on the respective comparative constructions in general. In Moyon, the particle *hak* is postponed to the parameter (e.g., *irun hək*, 'bigger', Devi 2010: 174). The particle can be also suffixed to the verbal root, where it covers an associative/collective function (*enno ensaʔhəkne* 'they eat together', Devi 2010: 84). The data available does not provide further information regarding the presence of additional markers in comparative constructions.

In Mao Naga, comparative constructions show the presence of the marker *zhü*, with the meaning of 'rather' (Giridhar 1994: 206). Mao Naga uses different particles to express an increasing level of intensity (Giridhar 1994: 373). The comparee is marked by *ko*; no further information is available on the structure of comparative constructions in Mao Naga.

Mongsen Ao (23) uses the comitative marker *thən* on the standard of comparison (Coupe 2007: 183). The marker *phān* found on the standard in Karbi (see example 24) is glossed as 'non subject': it is mainly used to mark O arguments, recipients and oblique participants (Konnerth 2014: 480).

(23) Mongsen Ao (Kuki-Chin-Naga; Coupe 2007: 293)
nì *sənti-pàʔ* *thən* *la* *tə-hláŋ-pàʔ*
1SG PN-M COM TOP NZP-be.long-NR
'I am taller than Sentiba.'

(24) Karbi (Kuki-Chin-Naga; Konnerth 2014: 454)
methān *a-phān-te* *ingnàr* *thè-mū*
dog POSS-NSUBJ-COND? elephant be.big-COMP
'Elephants are bigger than dogs.'

Tangkhul (25) marks the standard NP of comparative constructions with the suffix *ki*, which corresponds to the locative form meaning ‘on’ (Stassen 1985: 147).

(25) Tangkhul Naga (Kuki-Chin-Naga; Stassen 1985: 147)
Themma hau lu-ki vi-we
man this that-on good-COP
‘This man is better than that man’

Data on Karenic languages is limited. In Kayah Monu, comparative constructions display the particle *khlu* after the verb and before the standard (Wai 2013: 23). The particle is attested in Kayan Pekon with an adverbial function and translated with the meaning ‘more’ (Wai 2013: 35). No other means seem to be used in the expression of comparison of inequality, as shown in (26).

(26) Kayah Monu (Karenic; Wai 2013: 23)
hè sáplá khlù phàlú
1SG dejected ADV phalu
‘I am more dejected than Phalu.’

In Burmese, the only Burmo-Qiangic language of the sample, the standard is suffixed by the marker *hte?*, meaning ‘over’ (27).

(27) Burmese (Burmo-Qiangic: Stassen 1985: 126)
Thu-hte? pein-te
him-over be.thin-NONFUT
‘She is thinner than him.’

4. Data discussion

Following Stassen’s classification, the languages discussed in Section 3.2 can be assigned to the Adverbial and to the Particle types. Suansu is the only language among the ones described in this study that can be linked to the Exceed type. Table 4 illustrates the types and the related subtypes individuated for the languages of the sample.

Type (No. of languages)	**Languages**
Exceed (1)	Suansu
Adverbial: Allative (6)	Apatani, Atong, Galo, Garo, Karbi, Tangam
Adverbial: Locative (5)	Burmese, Hkongso, Mru, Rabha, Tangkhul
Adverbial: Separative (6)	Bunan, Chintang, Konyak, Mongsen Ao, Moyon, Yakkha
Particle (7)	Chothe, Geba Karen, Kayah Monu, Kadu, Mao Naga, Raji, Turung

Table 4: Typology and respective languages.

The most common type identified in the sample corresponds to the Adverbial type, which can be further distinguished in Allative, Separative and Locative subtypes.

Six languages belong to the Allative subtype. These include the Brahmaputran languages Atong and Garo, characterized by goal and benefactive forms marked on the standard of comparison. According to Stassen's typology, direct object markers are also classified within the Allative subtype, adding Tangam, Apatani, Galo and Karbi (Kuki-Chin-Naga) to the subtype.

Six languages from the sample belong to the Separative subtype within the Adverbial type. The Separative subtype comprises standard markers of source and origin. The Kiranti languages Yakkha and Chintang, as well as Konyak and Bunan belong to the Separative subtype. Comitative markers are also included within the Separative subtype in Stassen's typology, and they are attested in Mongsen Ao and Moyon through the forms *thən* and *hak* respectively. The marker *phəy* in Konyak has also a comitative meaning, and it is included in the Separative subtype. However, *phəy* somehow challenges the typology, since its original meaning is `behind', opening the interpretation of the marker as a locative particle and thus linking the form to the Locative subtype. Comitative markers used as standard markers in comparative constructions are found cross-linguistically, for example in Nuer (Nilotic, Ultan 1972), and Muna (Austronesian, Van Den Berg 2018), and they are considered rare.

Five languages are assigned to the Locative subtype. These include Rabha (Brahmaputran), Tangkhul (Kuki-Chin-Naga), Burmese,[9] and the Mruic languages Hkgonso and Mru. Within the Locative subtype, the markers on the standard have the

[9] Stassen (1985:40) includes Burmese comparative constructions under the Separative type. Since the form *hte?* indicates a spatial location ('over'), rather than "a motion away from a location" prototypical of the Separative subtype, I have reinterpreted Burmese comparative constructions under the Locative subtype.

function of a static spatial location 'on' and 'over' (*ki* in Tangkhul, *kára* in Rabha and *hte?* in Burmese), and 'side' (*lake* in Mruic).

The distribution of the types (and the related subtypes) reveals interesting aspects. Comparatives of the Allative subtype are found in 6 languages of the sample. This subtype is claimed to be rare in typological investigations (Stolz 2013: 19; Ultan 1972: 140). Stassen (1985: 40) reports the presence of the Allative comparative subtype in 7 languages out of the 110 that compose his sample. None of the languages listed by Stassen under the subtype is a Tibeto-Burman variety.[10] Based on his results, Stassen hypothesizes that the Allative subtype occurs in languages with basic verb-initial word order (Stassen 1985: 41). The cases illustrated above, from verb-final Tibeto-Burman languages, suggests that word order in general does not constrain the assignment to a specific comparative subtype, such as Allative.

Seven of the Tibeto-Burman languages explored in this study are assigned to the Particle comparative type. The Particle type is characterized by broad definitory criteria in Stassen's classification. The formal heterogeneity of the Particle type, combined with the scarce data available on the functions and meanings of the particles, makes this classification and the assignment of comparative constructions to this type blurred and tentative at times. Only a few languages of the Particle type provide detailed information on the respective comparative constructions. In some cases, the markers found in the languages assigned to this type do not seem to appear elsewhere in the grammar, suggesting the presence of dedicated comparative markers. Dedicated markers are referred to as Pure comparatives in Stolz's terms (see Section 1.1.). The particle *zhii*, glossed as 'rather', found in Mao Naga comparative constructions, appears to be a dedicated comparative marker. A similar specific comparative function can be hypothesized for the standard marker *tre* in Chothe.

Finally, the standard clitic *atha* found in Kadu comparative constructions, appears to be a dedicated comparative marker (Sangdong 2012: 331).

Degree markers do not constitute a parameter in Stassen's classification, and they are in general neglected from comparative constructions typologies (e.g., Dixon 2008). Degree markers are often hard to identify in linguistic sources. This aspect has been pointed out in the literature by Ultan (1972: 127), who has stressed the difficulties in distinguishing between standard and degree markers within languages.

[10] The languages that are listed by Stassen under the Allative subtype are Breton, Jacaltec, Kanuri, Maasai, Nuer, Siuslawan, and Tarascan.

Raji comparative constructions offer an example of this ambiguity. The form *jəmma* identified in Raji corresponds to the plural marker and is the only instance attested in the sample of a synchronic fully functional plural marker used in comparative constructions. This form is assigned to the Particle comparative type. Based on the shared meaning of 'moreness' and increased quantity in general that is conveyed both in plurality and asymmetric relations of comparison, I hypothesize a developing multipurpose function of this marker. From this perspective, the form *jəmma* acquires the meaning and functional properties of a degree marker rather than a standard marker, making its assignment to the Particle type more ambiguous.

Other markers from the languages assigned to the Particle type have a dubious status. In Turung, the post standard particle is glossed as 'more', and thus interpreted as a degree marker. Both Karenic languages of the sample, Geba Karen and Kayah Monu, appear to express comparison exclusively through a degree marker, *doli* and *khlù* respectively. These languages are tentatively assigned to the Particle type.

Degree markers are widespread across the sample and they often co-occur with other markers in comparative constructions. It is outside the scope of this study to propose a novel typology that would comprise the presence and the properties of degree markers, and this enterprise is left to future research. However, I report for exhaustivity the distribution of degree markers in the sample. Table 5 shows the number of languages with degree markers distributed over the identified types.

Type (No. of languages)	**No. of languages with degree markers**
Exceed (1)	0
Adverbial: Allative (6)	6
Adverbial: Locative (5)	0
Adverbial: Separative (6)	0
Particle (7)	7

Table 5: Typology and presence of degree markers.

Degree markers are found in half of the languages of the sample and appear to be absent in comparative constructions of the Locative and Separative subtypes. In these subtypes, the asymmetry between the comparee and the standard is made explicit by 'separative' and ablative morphemes in general. The Allative type, on the other hand, comprises standard markers related to goal, direct object and comitative forms, 'allowing' the presence of markers that overtly express the degree of comparison. The

locus of degree markers is usually on the parameter. Degree markers are also present in the languages assigned to the Particle type, either as standalone forms or in addition to other markers.

The presence of overt marking on the comparee is another parameter usually left unexplored in typological classifications of comparative constructions. Markers on the comparee are found on a small subset of 7 languages of the sample. Half of these languages belong to the Brahmaputran subgroup: the forms include topic markers, such as *ká* in Kadu (Sangdong 2012: 315) and *cə?* in Hkgonso (Wright 2009: 119). Other languages include Konyak, Chothe (Singh 2000: 271), and Suansu.

As shown in Section 2.1., Suansu comparative constructions (of the Exceed type) consist of a transitive predicate that takes the comparee as its subject and the standard as object. Ergative marking on the comparee outside a transitive predicative unit in a comparative construction is considered a typological oddity. It is found in Turung, where the ergative/agentive marker *î* is suffixed to the comparee, while the respective standard of comparison is followed by a particle meaning 'more'. A similar use of the ergative marker in a comparative construction has been reported by Jacques (2016) in Japhug (a Qiangic language spoken in Sichuan), where the functions and the possible diachronic pathways of evolution of this marker are presented and discussed extensively.

Given the types, it is possible to explore their genealogical and geographic distribution. Table 6 illustrates the breakdown of the types for each linguistic subgroup in the sample.

Subgroup	**Allative**	**Locative**	**Separative**	**Particle**	**Exceed**	**Total**
Bodic			1			1
Brahmaputran	2	1	1	3		7
Burmo-Qiangic		1				1
Karenic				2		2
Kiranti			2			2
Kuki-Chin-Naga	1	1	2	1	1	6
Macro-Tani	3					3
Mruic		2				2
Raji-Raute				1		1
						25

Table 6: Types per linguistic subgroup.

The sample is not genealogically balanced and is not fully representative of each linguistic subgroup; however, it allows for some qualitative considerations. Data from the Brahmaputran stock suggests high heterogeneity in the types (and subtypes) found within a genealogical stock. In the Brahmaputran stock, both the Adverbial and the Particle types are found, and the Adverbial type is represented by each of the respective subtypes. The same scenario is observed in Kuki-Chin-Naga, where all the main types -and subtypes- are found. The Mruic linguistic subgroup consists of 2 languages, Mru and Hkgonso, both included in the sample: they follow the Locative subtype. The Macro-Tani stock includes 12 languages (according to Glottolog, Hammarström et al. 2020), and three are included in the sample. These languages follow the Allative type.

Comparative structures tend to spread areally (Stassen: 1985; Dixon 2008: 813). The three main types, Adverbial, Particle and Exceed comparatives are plotted in Figure 2 to illustrate their geographical distribution.

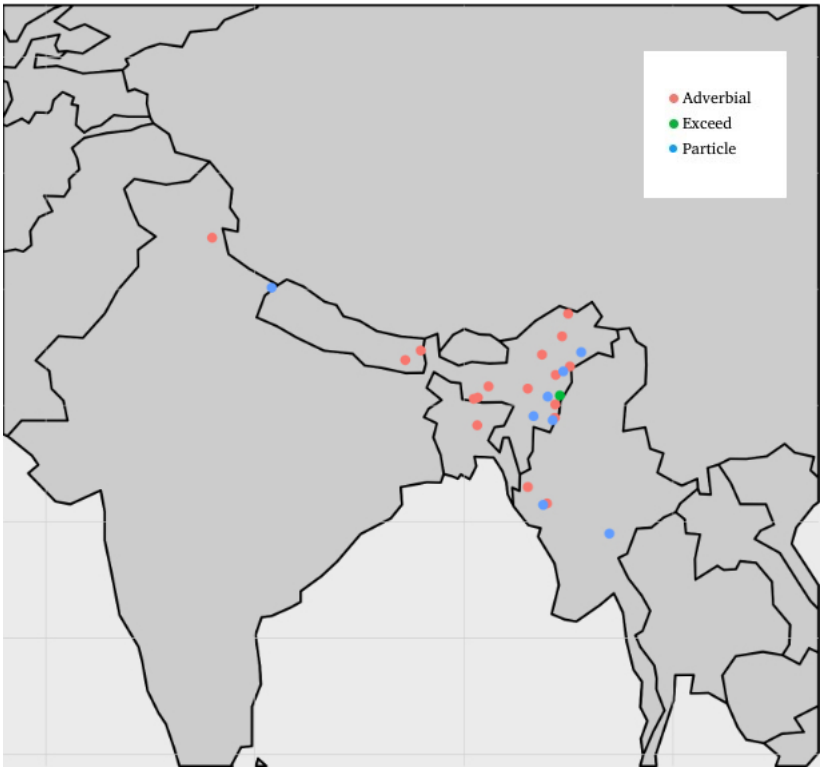


Figure 2: Geographical distribution of the types.

At a visual inspection, the geographical distribution of the main types suggests the presence of two clusters for the Adverbial and Particle types.[11] To assess the presence of the clusters, I test the plot through a specific visual statistical inference inspection process (Wickham et al. 2010), followed by a "Line Up" protocol (Kerman et al. 2008).

Visual statistical inference has the main purpose of bringing statistical support in quantifying the significance of structure in plots of data. It relies on the human capacity to spot visual trends in data, avoiding at the same time the risks of over-interpreting the patterns. Through this procedure, the plot representing the true data is placed among a set of randomly generated decoy data plots that support the null hypothesis (in the present case, no clustering). If the true data plot stands out from the rest in terms of representativity of the current hypothesis (here, the presence of two areal clusters), based on the judgment of a group of impartial and unbiased observers, one could consider this result as a rejection of the null hypothesis. This procedure is known as "Line-up", and the generated plot is found in the Appendix. The plot is generated using the R package `nullabor` (Buja et al. 2009). The true data plot is 12, and it passed the "Line-up" protocol test.

The Adverbial type clusters in the western part of the region. Comparative constructions of the Adverbial type are also widespread in the languages and language families of the Indian subcontinent (Stassen 2013). They are found, for example, in Hindi, Marathi and Bagri (Indo-Aryan), in the Dravidian languages Tamil and Telugu, and in Mundari and Santali (both belonging to the Munda branch of Austroasiatic). The languages of the sample located in the western part of the area contribute to this Adverbial comparative constructions *continuum*.

The Particle type clusters in the eastern part of the region and is assigned mainly to the Tibeto-Burman languages geographically contiguous to the Myanmar border, except for Raji, spoken in the western fringes of the targeted area and whose assignment to the Particle type is dubious. There is no detailed cross-linguistic information on the comparative types found in the linguistic stocks contiguous to the languages belonging to the Particle type.

Suansu is the only representative of the Exceed comparative in the sample. Outside the sample, comparatives of the Exceed type seem to be restricted to two geographical areas: South-East Asia and Sub-Saharan Africa (Stassen 2013). Within South-East Asia, the Exceed type is attested in Mandarin, Thai, Khmer, and Vietnamese, among others. Further studies on lesser-known languages are required to assess the presence

[11] No significant clustering can be observed in the geographical distribution of the subtypes.

of this type in contiguous geographical areas such as southern China or northern Myanmar.

5. Concluding remarks

The main contributions of this paper are summarized in the following. Suansu comparative constructions, described for the first time in this study, can be typologically assigned to the Exceed comparative type in the classification proposed by Stassen.

Comparatives of the Exceed type are not found in Suansu genealogical and geographical linguistic neighbors. The cross-linguistic exploration of comparative constructions from a sample of Tibeto-Burman languages of northeastern India rather shows the presence of two main types: Adverbial and Particle types.

These types cluster geographically. The Adverbial type is found in the languages spoken in the western part of the region. This distribution is consistent with the extensive presence of comparative constructions of the Adverbial type found in the genealogically diverse languages spoken in South Asia (Indo-Aryan, Dravidian, Munda) and attested in the literature.

Comparative constructions of the Particle type are rather found in the eastern part of the region, and specifically in the languages spoken on the geographical and cultural border between South Asia and South-East Asia.

Comparative constructions of the Exceed type are not found outside Suansu; however, the presence of this type aligns with the assumptions in the cross-linguistic literature that describe the distribution of this type as geographically restricted to South-East Asia (and sub-Saharan Africa). The presence of the Exceed type comparative in Suansu might suggest a larger spread of this type, not limited to the 'core' South-East Asia, but also extended to the western fringes of the region.

A more general contribution to the typology of comparative constructions pertains to the presence of comparative constructions of the Allative subtype. This type is described as rare in many typologies and linked to verb-initial word order. The relative high occurrences of the Allative type found in several verb-final Tibeto-Burman languages seem to suggest that the Allative subtype is not constrained by any specific word order frame.

Several aspects that surfaced in the data discussion call for an updated approach in the typology of comparative constructions. These aspects include the high internal

diversity attested within the linguistic subgroups, the blurred boundaries among the types (for example, the double interpretation of the marker *phəy* in Konyak), and the tentative assignment of several comparative constructions to the Particle type. A typology of comparative constructions based exclusively on the marking on the standard of comparison may suffice to outline broad characteristics of comparative constructions, but fails in describing the full linguistic diversity of these structures. As a result, different comparative structures are lumped together in broad types, such as the Particle type, or several subtypes proliferate with arbitrary definitory criteria.

A proposed typology, whose implementation is left for future research, supports a multivariate approach to comparative constructions. In this framework, each structural variable is considered: all the structures found in a comparative construction are captured, such as degree markers, parameters markers, word order, and so on. The premature labelling of the variables should be also avoided, to prevent potential ambiguities. No variable is discarded *a priori,* and the set of potential variables is defined during the data collection process. These variables should then be described in a granular and systematic way, to ensure a detailed overview of their diversity. The definition of the types can then be outlined in a latter phase, in order to capture the systematic similarity of the constructions that are assigned to a specific type and, at the same time, to facilitate the cross-linguistic viability of the types.

Acknowledgements

My thanks go to the Suansu speaking community and especially to J. V. for their constant support in explaining and clarifying several aspects of Suansu grammar to me. I am also grateful to Taras Zakharko for his helpful comments, and to the two anonymous reviewers for their insightful feedback on an earlier draft of this paper.

Abbreviations*

1 = 1st person	CUST = customary aspect	NR = general nominalizer
2 = 2nd person	DAT = dative	NSUBJ = non-subject

* The abbreviations used in the Suansu examples found in Section 2 follows the Leipzig Glossing Rules. The examples from the other Tibeto-Burman languages described in this study are reported with the original glosses found in the respective sources, with some minor adjustments for normalization purposes.

3 = 3rd person
ABS = absolutive
ABL = ablative
ACC = accusative
ADV = adverb
AG = agentive
ATTR = attributive
CJ = conjunct
CLF = classifier
COM = comitative
COMP = comparative
COND = conditional
COP = copula
DECL = declarative
DET = determiner
EQ = equative copula
ERG = ergative
GEN = genitive
GOAL = goal
INS = instrumental
IPFV = imperfective
LOC = locative
M = masculine gender
NOM = nominative
NAGT = non-agentive
NONFUT = non-future tense
NZP = nominalizing prefix
NZR = nominalizer
PL = plural
PN = personal name
POSS = possessive
PRS = present
PST = past
PL = plural
REL = relativizing subject
SG = singular
STAT = stative
SUB = subject
TOP = topic

References

Abraham, P. T. 1985. *Apatani grammar*. Vol. 12. Central Institute of Indian Languages: Grammar Series. Mysore: Central Institute of Indian Languages.

Bickel, Balthasar & Nichols, Johanna. 2002. Autotypologizing databases and their use in fieldwork. In Peter Austin & Helen Dry & Peter Wittenburg (eds.), *Proceedings of the International LREC Workshop on Resources and Tools in Field Linguistics.* Nijmegen: Max Planck Institute for Psycholinguistics.

Bickel, Balthasar. 2010a. Capturing particulars and universals in clause linkage. Isabelle Bril (ed.), *Clause-Hierarchy and Clause-Linking: The Syntax and Pragmatics Interface*, 51–101. Amsterdam/Philadelphia: John Benjamins.

Bickel Balthasar. 2010b. Towards a multivariate typology of reference tracking. Presentation given at the Meeting of the Research Group 742, Leipzig, 21 April.

Bickel, Balthasar & Zakharko, Taras & Nichols, Johanna. 2016. Better data with late aggregation: Autotyp and beyond. Paper presented at Poznan Linguistics Meeting, Poznań, 15-17 September.

Bickel, Balthasar & Nichols, Johanna & Zakharko, Taras & Witzlack-Makarevich, Alena & Hildebrandt, Kristine & Rießler, Michael & Bierkandt, Lennart & Zúñiga, Fernando & Lowe, John B. 2017. The AUTOTYP typological databases. *Version 0.1.0.* (Available online at https://github.com/autotyp/autotyp-data/tree/0.1.0, Accessed on 2020-11-10).

Bobaljik, Jonathan D. 2012. *Universals in comparative morphology: suppletion, superlatives, and the structure of words*. Vol. 50. Cambridge, MA: MIT Press.

Buja, Andreas & Cook, Dianne & Hofmann, Heike & Lawrence, Michael & Lee, Eun-kyung & Swayne, Deborah F. & Wickham, Hadley. 2009. Statistical inference for exploratory data analysis and model diagnostics. *Royal Society Philosophical Transactions.* 367(1906). 4361–4383.

Burling, Robbins. 1961. *A Garo grammar*. Vol. 25. Deccan College Monograph Series. Poona: Deccan College.

Burling, Robbins. 2003a. *The language of the Modhupur Mandi, Garo: Vol. II: The Lexicon*. New Delhi: Bibliophile South Asia.

Burling, Robbins. 2003b. The Tibeto-Burman Languages of Northeastern India. In Graham Thurgood & Randy J. LaPolla (eds.), *The Sino-Tibetan Languages*, 169-192. London/New York: Routledge.

Coupe, Alexander R. 2007. *A grammar of Mongsen Ao*. Berlin/New York: Mouton de Gruyter.

DeLancey, Scott. 2011. 'Optional' 'ergativity' in Tibeto-Burman languages. *Linguistics of the Tibeto-Burman Area* 34(2). 9-20.

Devi, Kongkham Hemabati. 2010. *A descriptive grammar of Moyon*. PhD Dissertation, Manipur University.

Dixon, Robert. 2008. Comparative constructions: a cross-linguistic typology. *Studies in Language.* 32(4). 787–817.

Ebersole, Harold. 1996. *The Mru language: a preliminary grammatical sketch*. Unpublished manuscript.

Fuchs, Catherine. 2014. *La comparaison et son expression en Français*. Paris: Éditions Ophrys.

Giridhar, P. P. 1994. *Mao Naga grammar*. Central Institute of Indian Languages: Grammar Series. Mysore: Central Institute of Indian Languages.

Hammarström, Harald & Forkel, Robert & Haspelmath, Martin & Bank, Sebastian. 2020. Glottolog 4.2.1. Jena: Max Planck Institute for the Science of Human History. (Available online at https://glottolog.org/, Accessed on 2020-11-10)

Heine, Bernd. 1997. *Cognitive foundations of grammar*. Oxford: Oxford University Press.

Ivani, Jessica K., & Zakharko, Taras. 2019. Tymber. Database on nominal number marking constructions. Version 0.1.0. (Available online at https://github.com/jkivani/tymber/, Accessed on 2020-10-10)

Ivani, Jessica K. 2020. Comparative constructions in Tibeto-Burman languages of northeastern India [Data set]. (Available online at https://doi.org/10.5281/zenodo.4274488, Accessed on 2020-10-10).

Jacques, Guillaume. 2016. From ergative to comparee marker: multiple reanalyses and polyfunctionality. *Diachronica* 33(1). 1–30.

Joseph, U. V. 2007. *Rabha*. Vol. 1. Languages of the Greater Himalayan Region. Leiden: Brill.

Kerman, Jouni & Gelman, Andrew & Zheng, Tian & Ding, Yuejing. 2008. Visualization in Bayesian data analysis. In Chun-houh Chen & Wolfgang Hardle & Antony Unwin (eds.), *Handbook of Data Visualization*, 709–24. Berlin: Springer.

Konnerth, Linda. 2014. *A grammar of Karbi*. PhD dissertation, University of Oregon.

Krishan, Shree. 2001. A sketch of Raji grammar. In Yasuhiko Nagano and Randy J. Lapolla (eds.), *New Research on Zhangzhung and Related Himalayan Languages*, 449–497. Bon Studies. Osaka: National Museum of Ethnology.

Marrison, Geoffrey Edward. 1967. *The classification of the Naga languages of North-East India*. PhD dissertation, University of London.

Morey, Stephen. 2010. *Turung: A variety of Singpho language spoken in Assam*. Vol. 614. Pacific Linguistics. Canberra: Research School of Pacific; Asian Studies, Australian National University.

Mortensen, David. 2013. Database of Tangkhulic Languages. Unpublished manuscript. Accessed via the STEDT Database (*https://stedt.berkeley.edu/)* on 2020-11-10.

Nagaraja, K. S. 2010. *Konyak grammar*. Vol. 594. Central Institute of Indian Languages: Publication. Mysore: Central Institute of Indian Languages.

Paudyal, Netra Prasad. 2015. *Aspects of Chintang syntax*. PhD dissertation, Universität Zürich.

Post, Mark W. 2007. *A grammar of Galo*. PhD dissertation, LaTrobe University.

Post, Mark W. 2017. *The Tangam language: grammar, dictionary and texts*. Leiden: Brill.

Sangdong, David. 2012. *A grammar of the Kadu (Asak) language*. PhD dissertation, La Trobe University.

Schackow, Diana. 2015. *A grammar of Yakkha*. Studies in Diversity Linguistics 7. Berlin: Language Science Press.

Singh, Hidam Brojen. 2000. *A descriptive grammar of Chothe*. PhD dissertation, Manipur University.

Simon, Ivan Martin. 1972. *An introduction to Apatani*. Gangtok: Government of India Press.

Stassen, Leon. 1985. *Comparison and universal grammar*. Oxford: Blackwell.

Stassen, Leon. 2013. Comparative 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 https://wals.info/chapter/121, Accessed on 2020-11-10).

Stolz, Thomas. 2013. *Competing comparative constructions in Europe*. Berlin/New York: Mouton de Gruyter.

Sun, Tianshin. 1993. *A historical-comparative study of the Tani (Mirish) branch in Tibeto-Burman*. Linguistics of the Tibeto-Burman Area. 16. Ann Arbor: UMI.

Treis, Yvonne. 2018. Comparative constructions: An Introduction. *Linguistic Discovery* 16(1). i-xxvi.

Tuttle, Siri G. 2018. Comparative and superlative constructions in Alaskan Athabascan languages. *Linguistic Discovery* 16(1). 100-116.

Ultan, Russell. 1972. Some features of basic comparative constructions. *Working Papers on Language Universals (Stanford)* 9. 117–162.

Van Breugel, Seino. 2014. *A grammar of Atong*. Leiden: Brill.

Van Den Berg, René. 2018. More than most: comparative constructions in Muna. *Linguistic Discovery* 16(1). 117-140.

Wai, Lin Aung. 2013. *A descriptive grammar of Kayah Monu*. Master's thesis, Payap University.

Wickham, Hadley & Cook, Dianne & Hofmann, Heike & Buja, Andreas. 2010. Graphical inference for infovis. *IEEE Transactions on Visualization and Computer Graphics* 16(6). 973–979.

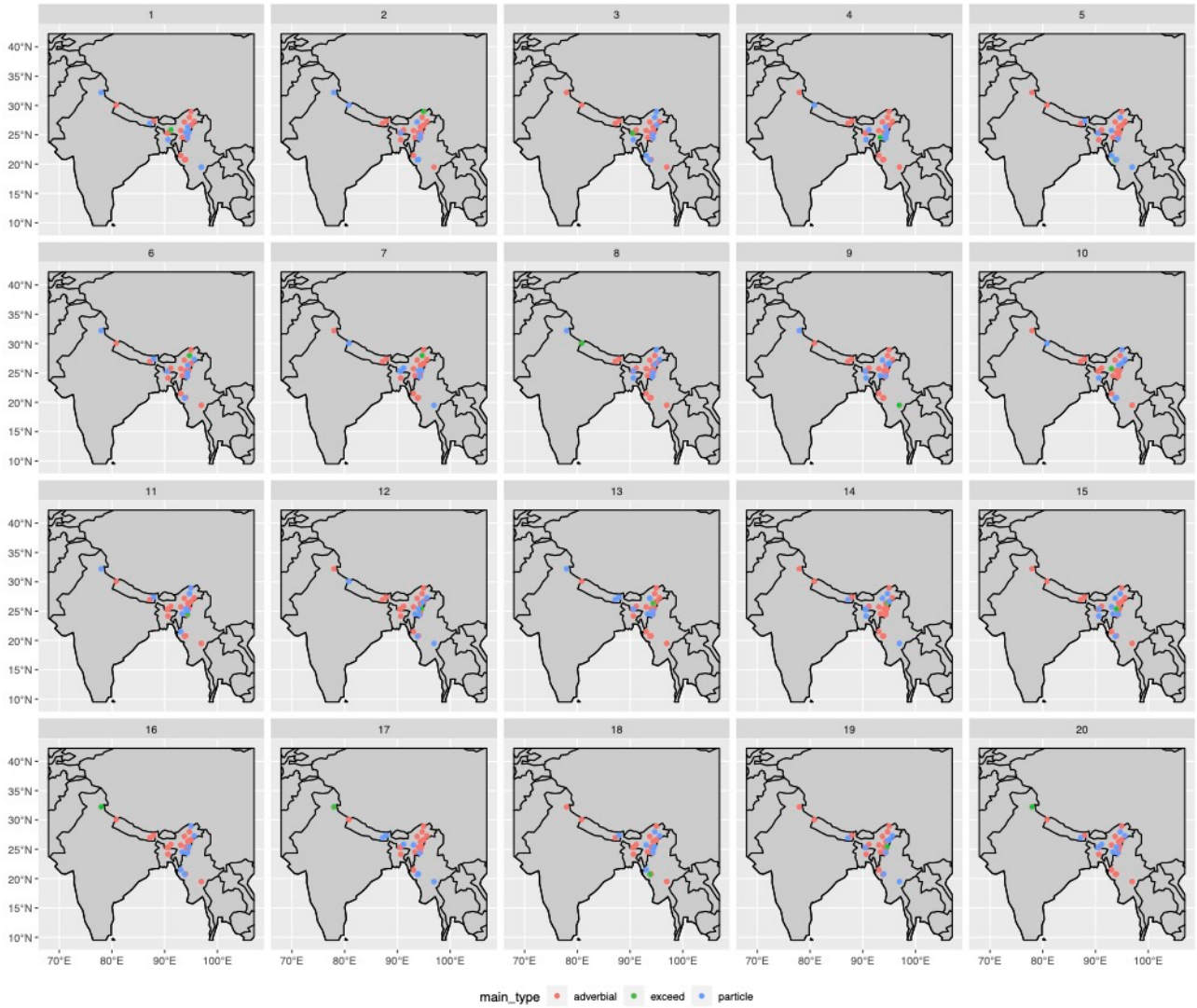
Widmer, Manuel. 2014. *A descriptive grammar of Bunan*. PhD dissertation, Universität Bern.

Wright, Jonathan Michael. 2009. *Hkongso grammar sketch*. PhD dissertation, Graduate Institute of Applied Linguistics.

CONTACT

jessica.ivani@uzh.ch

Appendix


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
40°N
35°N
30°N
25°N
20°N
15°N
10°N
70°E
80°E
90°E
100°E
main_type
adverbial
exceed
particle

Linguistic Typology at the Crossroads 2-1 (2021): 94-120 ISSN 2785-0943

Verb-argument lability and its correlations with other typological parameters: a quantitative corpus-based study

NATALIA LEVSHINA[1], JOHN A. HAWKINS[2,3]

[1] MAX PLANCK INSTITUTE FOR PSYCHOLINGUISTICS, NIJMEGEN,

[2]UNIVERSITY OF CALIFORNIA, DAVIS, [3]CAMBRIDGE UNIVERSITY

Submitted: 24/11/2021 Revised version: 28/02/2022

Accepted: 10/03/2022 Published: 14/04/2022

Abstract

We investigate the correlations between lability for verbal arguments with other typological parameters using large, syntactically annotated corpora of online news in 28 languages. We focus on A-lability, when the A argument alternates with S (e.g., *She is singing* vs. *She is singing a song*), and P-lability, when the P-argument alternates with S (e.g., *She opened the door* vs. *The door opened*). To estimate how much lability is observed in a language, we measure associations between Verbs or Verb + Noun combinations and the alternating constructions in which they occur. Our correlational analyses show that high P-lability scores correlate strongly with the following parameters: little or no case marking; weaker associations between lexemes and the grammatical roles A and P; rigid order of Subject and Object; and a high proportion of verb-medial clauses (SVO). Low P-lability correlates with the presence of case marking, stronger associations between nouns and grammatical roles, relatively flexible ordering of Subject and Object, and verb-final order. As for A-lability, it is not correlated with any other parameters. A possible reason is that A-lability is a result of more universal discourse processes, such as deprofiling of the object, and also exhibits numerous lexical and semantic idiosyncrasies. The fact that P-lability is strongly correlated with other parameters can be interpreted as evidence for a more general typology of languages, in which some have highly informative morphosyntactic and lexical cues, whereas others rely predominantly on contextual environment, which is enabled by fixed word order. We also find that P-lability is more strongly correlated with the other parameters than any of these parameters are with each other, which means that it can be a very useful typological variable.

Keywords: verb-argument lability; corpora; Universal Dependencies; word order; case marking; tight-fit and loose-fit languages.

1. Theoretical background

The usefulness of a typological parameter depends on how many other parameters it helps us to predict. Greenberg's (1963) word order correlations have been such a major achievement in linguistics because they connected many diverse and seemingly unrelated word order patterns. In this paper we demonstrate that the strength of attraction of verbs (as well as their arguments) to specific subcategorization frames, which can be defined in terms of verb-argument lability, can be a useful parameter, because it is strongly correlated with many others.

The attraction of verbs to specific subcategorization frames has been argued to be a part of the typology of tight-fit versus loose-fit languages. The terms were coined by Hawkins (1986: 121–127, 1995; see also Müller-Gotama 1994). Generally speaking, tight-fit languages have unique surface forms that map onto more constrained meanings, whereas loose-fit languages have vaguer forms with less constrained meanings and they rely on word-external and contextual features as defined in Hawkins (2019) for the assignment of meanings. For example, grammatical roles in tight-fit languages have a narrower semantic range than grammatical roles in loose-fit languages. The languages Jakaltek and Halkomelem strictly exclude inanimate subjects in transitive clauses (Aissen 2003), while English and Swedish merely strongly disprefer them (Dahl 2000). There are also more gradient distinctions. Both English and German allow for different kinds of subjects, but English is looser than German in the semantic roles that can fill the subject position (Hawkins 1986, 2019), and also than Russian and Korean.

The strength of the associations with grammatical roles is correlated with other linguistic parameters, including more explicit grammatical coding (e.g., formal case marking and use of complementizers and relativizers), avoidance of raisings and long distance WH-movements. Tight-fit languages have fewer instances of syntactic category ambiguity. For example, the English word *book* can be both a noun and a verb, while in German the corresponding noun and verb have different forms, *Buch* – *buchen*. Moreover, verb-final languages are often semantically tight (see also Levshina 2020).

If these parameters change, they often change together. English is a well-known case (Hawkins 1986, 2012). The loss of morphology correlated, in particular, with the emergence of SVO order, long distance movement and raising, greater category ambiguity and other features, including fewer restrictions on the semantics of

syntactic arguments. In contrast, German is more conservative. It preserves case marking, verb-final order (for all verbs in subordinate clauses and for non-finite verbs in main clauses), and it still has some variability in the order of Subject and Object. In addition, German has fewer instances of category ambiguity, tighter associations between semantics and roles, and very limited examples of raising. Generally speaking, English is more structurally ambiguous than German. For example, raising and control constructions are not distinguished formally in surface structure. Compare *Sue happened to win the lottery* (raising) and *Sue hoped to win the lottery* (control). In German, these are distinguished by formally different constructions. In Hawkins' terminology (2019), English relies more on word-external properties to derive meanings from ambiguous or vague surface forms, whereas German relies more on distinct grammatical and lexical patterns and on word-internal properties.

Importantly for our study, verbs in loose-fit languages have a broader set of subcategorization frames than in tight-fit languages. For example, the English verb *open* can be both transitive (e.g, *I opened the door*) and intransitive (*The door opened*), while German distinguishes formally between the transitive *öffnen* "open (tr.)" and the reflexive verb *sich öffnen* "open (intr.)".

The term 'lability' in the context of verb-argument structure is used in the typological literature to refer to alternations between the arguments of one and the same verb (as e.g., P or S with a verb like *open*) or between the arguments of alternating verbs in pairs such as *teach* and *learn* that describe a common event from different perspectives. The precise definitions proposed by different authors have varied in terms of how general or restrictive they are with respect to the alternations in question (compare, e.g., Nichols 1986; Nichols et al. 2004; Dixon 1994; Haspelmath 1993; Letuchiy 2009; Creissels 2014). In the present context we focus on a set of core patterns that most definitions subsume, which can be readily identified in our corpora and in which one and the same verb can be used transitively or intransitively without any formal change while the arguments alternate in their status as A or P or S. Examples from English are *break, open, eat* and *sing*. We are agnostic with regard to the question of which use (transitive or intransitive) is basic, and which is a result of valency derivation (but see Nichols et al. 2004 for insightful discussion of this issue and typological comparison). Syntactically annotated corpora enable us to track the arguments of individual verbs in many different languages and so permit us to quantify the degree of lability within and across languages. The resulting lability

scores can then be tested against other typological variables derived from the same corpora.[1]

More precisely, we consider two types of verb-argument lability, referred to as A-lability and P-lability. A-lability is defined in (1):

(1) A-lability occurs when the A argument of a verb in a transitive clause remains the same in a corresponding intransitive clause from which the P argument has been removed without any formal change to the verb.

In other words, with the same verb, the A-argument can turn into an S-argument: A = S (Dixon 1994).[2] Examples are the unspecified object alternation (2a), the understood body-part alternation (2b) and the characteristic property alternation (2c) (Levin 1993).

(2) a. Unspecified object alternation
Jack ate the cake. - Jack ate.
b. Understood body-part alternation
The Queen waved her hand at the crowd. - The Queen waved at the crowd.
c. Characteristic property alternation
The dog bites strangers. - The dog bites.

P-lability is defined in (3):

(3) P-lability occurs when the same argument can be used as the subject of a verb in an intransitive clause (S) and as the direct object (P) of a corresponding transitive clause without any formal change to the verb.

[1] Notice that the large-scale empirical study described here makes it straightforward to identify alternating arguments for one and the same verb, but does not lend itself readily to identifying alternating arguments among semantically related pairs of verbs such as *teach* and *learn* or *like* and *please*, in the absence of semantic features or tags that can identify the relevant pairs of verbs. Our definition is accordingly more restricted in this respect than that proposed in, e.g., Letuchiy's (2009) summary paper of the typology of lability, but less restricted in other respects (see fn.2).

[2] Note that this type of lability is not considered as such by Letuchiy (2009), for whom lability necessarily involves a change in the semantic role of the subject. For him, verbs like *sing* and *drink* occurring with and without a direct object are 'pseudo-labile'. Letuchiy's definition of lability proper is accordingly more restrictive in this respect than ours (compare fn.1).

In other words, with the same verb S = P (Dixon 1994). Examples are the causative-inchoative alternation (4a), the middle alternation (4b) and the induced action alternation (4c) (Levin 1993).

(4) a. Causative-inchoative alternation
The boy broke the vase. - The vase broke.
b. Middle alternation
The publisher sells the book. - The book sells well.
c. Induced action alternation
She jumped the horse over the fence. - The horse jumped over the fence.

The above-mentioned contrast between English *open* (both transitive and intransitive) and German *öffnen* "open" (transitive only) with its reflexive verb counterpart in the corresponding intransitive *sich öffnen* "open (intr.)" suggests that English has more P-lability (causative-inchoative alternations, in particular) than German. However, this has not yet been examined in corpora and using quantitative measures.

In this paper we fill this gap, measuring A- and P-lability in languages with the help of large corpora, which are described in Section 2. We compute the Mutual Information between verbs, or combinations of verbs and nouns, and the alternating constructions in which they occur. The procedure and the scores are presented in Section 3. Then, we test the correlations between different measures of A- and P-lability and four other variables which have been used in the literature on tight-fit and loose-fit languages and more generally: word order rigidity; the position of the verb in the sentence; case marking; and the strength of associations between nouns and the grammatical roles of Subject and Object (Section 4). Finally, in Section 5 we discuss our findings and conclusions.

2. Data and method

We used the Leipzig Corpus Collection (Goldhahn et al. 2012)[3]. We first selected 30 online news corpora with 1M sentences in each of the following languages: Arabic, Bulgarian, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek (modern), Hindi, Hungarian, Indonesian, Italian, Japanese, Korean,

[3] http://wortschatz.uni leipzig.de/en/download/ (accessed 2022.02.27).

Latvian, Lithuanian, Persian, Portuguese, Romanian, Russian, Slovenian, Spanish, Swedish, Tamil, Turkish and Vietnamese. The corpora were annotated with the Universal Dependencies pipeline *udpipe* (Wijffels, Straka & Straková 2018), which allowed us to extract the subject, the direct object, the predicate, as well as their lemma, part of speech and morphological features, and other useful information. Due to our doubts about the quality and consistency of verb lemmatization in the data from Tamil and Turkish, these languages were later excluded. This is why we had 28 languages in the final sample.

In order to find patterns of A-lability, we extracted the frequencies of all verb lemmas with the same noun in subject position (represented by the Universal Dependency 'nsubj') with and without any kind of nominal or pronominal direct object (the Universal Dependency 'obj'). Consider the examples in Table 1.

A-lability Frequencies			
Verb	**Subject**	**Transitive**	**Intransitive**
be	idea	0	140
learn	student	21	35
play	team	55	47

Table 1: Examples of frequencies relevant for A-lability.

The table shows that the verb *be* with the noun *idea* as subject occurs 140 times (e.g., *the idea was…*), only in intransitive clauses. This is not surprising. The combination *student* + *learn* occurs 21 times with a direct object (e.g., *the students learn languages*) and 35 times without (e.g., *the students learn*). This is an example of A-lability. In this paper, we measure the degree of A-lability by estimating the degree of the skew of a Verb + Noun combination towards the Transitive construction, where the noun is the A-argument, or the Intransitive use, where the noun is the S-argument. The stronger the bias towards one or the other use across different Verb + Noun combinations in a corpus, the weaker the A-lability. If many combinations behave like *play* + *team*, the A-lability will be high. If more combinations behave like *be* + *idea*, the A-lability will be low.

In order to identify examples of P-lability, we extracted the frequencies of all verb lemmas (only predicates of main clauses) with the same noun occurring as direct object and as intransitive subject. Consider the examples in Table 2.

P-lability Frequencies			
Verb	**Noun**	**Intr. subject + Verb**	**Verb + Object**
die	people	64	0
open	door	36	149
begin	work	35	33

Table 2: Examples of frequencies relevant for P-lability.

The numbers should be read as follows. The verb *die* occurs with the noun *people* only as an intransitive subject (64 times), and never as an object. The verb *open* with the noun *door* as intransitive subject (*The door opened*) occurs 36 times, and as a direct object (*I opened the door*) 149 times. This is an example of P-lability. The stronger the skew of the Verb + Noun combinations towards one or the other use, the weaker the P-lability in a given corpus. If many rows contain the frequencies of *begin* + *work,* the P-lability will be high. If more Verb + Noun combinations behave like *die* + *people*, the P-lability will be low.

If we simply counted intransitive and transitive uses of verbs, it would be impossible to distinguish A-lability from P-lability. As will be shown below, making this distinction is crucial, and it is why it was necessary to control for the nouns as A, P or S.

Note that we only selected the verbs that served as predicates of main clauses. Particle verbs and verbs with separable prefixes were treated as one lemma (e.g., *break+out*, *um+leiten*). We also excluded verbs with reflexive, passive, antipassive, middle morphology or auxiliaries because of the substantial cross-linguistic differences in their semantics, formal properties and annotation. One consequence of this decision is that we are primarily measuring looseness vs. non-looseness (the formal marking of which can be quite variable across languages). We also excluded ditransitive clauses, in which one and the same verb had dependencies labelled as 'iobj' (indirect object) and 'obj' (direct object). The measures of lability presented below are based only on combinations of verbs and nouns that occur ten times or more in a corpus.

Lability measures were computed using two methods. According to the first, we controlled for both the verb and the noun, which means that our measures took into account not only the flexibility of the verb with regard to the alternation variants, but also the flexibility of the noun with regard to the roles of A or S (in cases of A-lability)

and S or P (in cases of P-lability). In the second method, we took into account the verbs only, adding up the frequencies of all nouns occurring as A and S, or as S and O with a given verb. The formulas and the scores are discussed in the next section.

3. Measures of lability

3.1. Mutual Information related to A-lability

Using the kinds of frequencies shown in Table 1, we computed Mutual Information (MI) related to A-lability for twenty-eight languages. For Verb + Noun combinations, the formula was as follows:

(5) $$I\,(V\&N;\ Cx) = \sum_{i,\,j} p\left(V\&N_i,\ Cx_j\right) log_2\ \frac{p\left(V\&N_i, Cx_j\right)}{p\left(V\&N_i\right) p\,(Cx_j)}$$

where *V&N* represents Verb + Noun combinations, and *Cx* stands for the constructional alternation, which includes the transitive construction ('nsubj' + Verb + some object) and the intransitive construction ('nsubj' + Verb). The higher MI, the stronger on average the associations between the Verb + Noun combinations and the particular constructions. Therefore, high MI scores indicate weak lability, characteristic of a tight-fit language, and low MI scores correspond to strong lability, characteristic of a loose-fit language.

For verbs only, the formula was as follows, where V stands for a verb:

(6) $$I\,(V;\ Cx) = \sum_{i,\,j} p\left(V_i,\ Cx_j\right) log_2\ \frac{p\left(V_i, Cx_j\right)}{p\left(V_i\right) p\,(Cx_j)}$$

Both types of scores are shown in Figure 1. The languages are ordered by their MI scores based on Verb + Noun combinations, but the two types of scores are strongly correlated: Spearman's rank-based correlation coefficient is 0.97, and the p-value $<$ 0.0001 (but see a more precise measure with genetic dependencies taken into account in Section 4). This means that the measures represent very similar information. The scores based on verbs only are lower in all languages, but the ordering is more or less the same, as the high correlation coefficient suggests. The highest scores are found in Portuguese, followed by Italian, Hindi, English and Slovene. The lowest score belongs to Lithuanian, followed by Vietnamese, Korean, Arabic and Persian. This ranking is

not predictable from any typological, genealogical or areal properties of the languages.

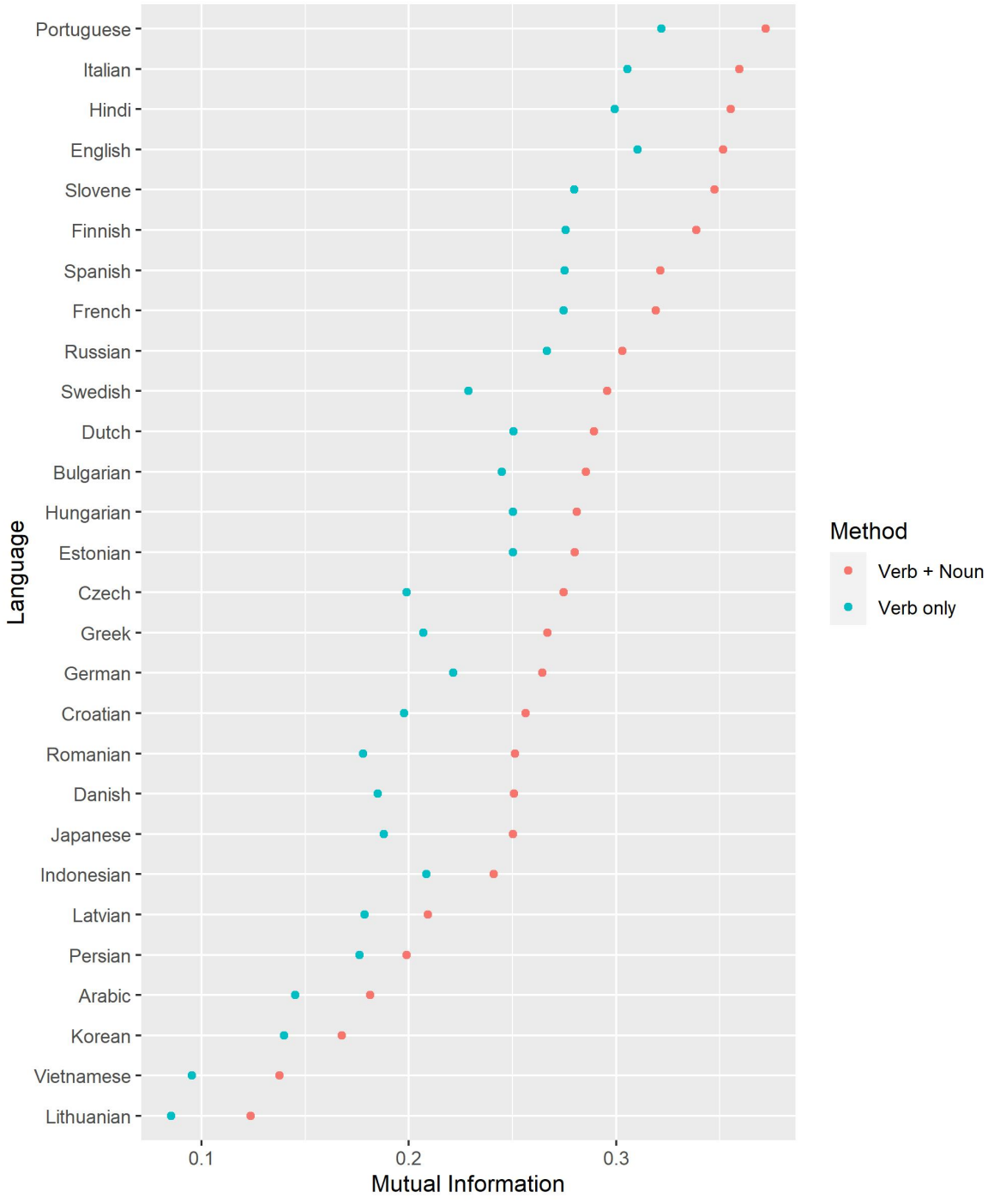


Figure 1: Distribution of MI scores representing A-lability. The higher the score, the weaker this type of lability in a language.

3.2. Mutual Information related to P-lability

To compute MI related to P-lability, we used the same approach as for A-lability, but took the frequencies of verbs and nouns in the construction 'nsubj' + Verb without object and the construction Verb + 'obj' (regardless of the presence or absence of any subject). The two methods, Verb + Noun (as 'nsubj' or 'obj') and Verb only, yield scores that are highly correlated: Spearman's correlation coefficient rho is 0.96, with the p-value < 0.001.

Figure 2 displays both types of MI scores. The top scores belong to Hungarian, Russian, Estonian, Latvian, Korean and Finnish. The high scores mean that these languages have strong associations between the Verb + Noun combinations and the constructions in which they appear as 'nsubj' or 'obj' respectively, characteristic of tight-fit languages. These languages also have formal case marking and relatively free word order of the core arguments. Many of the languages at the top are verb-final, or at least allow for the V-final order. The two languages at the bottom are Indonesian and Vietnamese, followed by English, French and Romanian. These have weaker associations between the Verb + Noun combinations and the constructions in which they appear as 'nsubj' or 'obj'. So they display stronger P-lability characteristic of loose-fit languages. They also have fairly rigid SVO order and no case morphology.

If we compare the range of values in Figure 1 and Figure 2, we see that the MI scores related to A-lability are on average lower than the MI scores related to P-lability. This impression is supported by paired Wilcoxon tests. The difference between the A- and P-lability scores is significant for both methods ($p = 0.028$ for verbs only, and $p < 0.001$ for Verb + Noun combinations). This means that languages are more tolerant with regard to A-lability in general. In addition, the spread of the P-lability scores is greater, which suggests more substantial cross-linguistic differences.

3.3. How variable is lability across individual verbs and arguments?

An important question is whether low MI scores, which characterize high-lability languages, are due to most verbs being weakly associated with a specific construction, or whether they are primarily influenced by a handful of very frequent idiosyncratic verbs with high verb-argument lability, e.g., English *break* or *learn.*[4]

[4] We thank Thomas Hörberg for pointing out this important distinction to us.

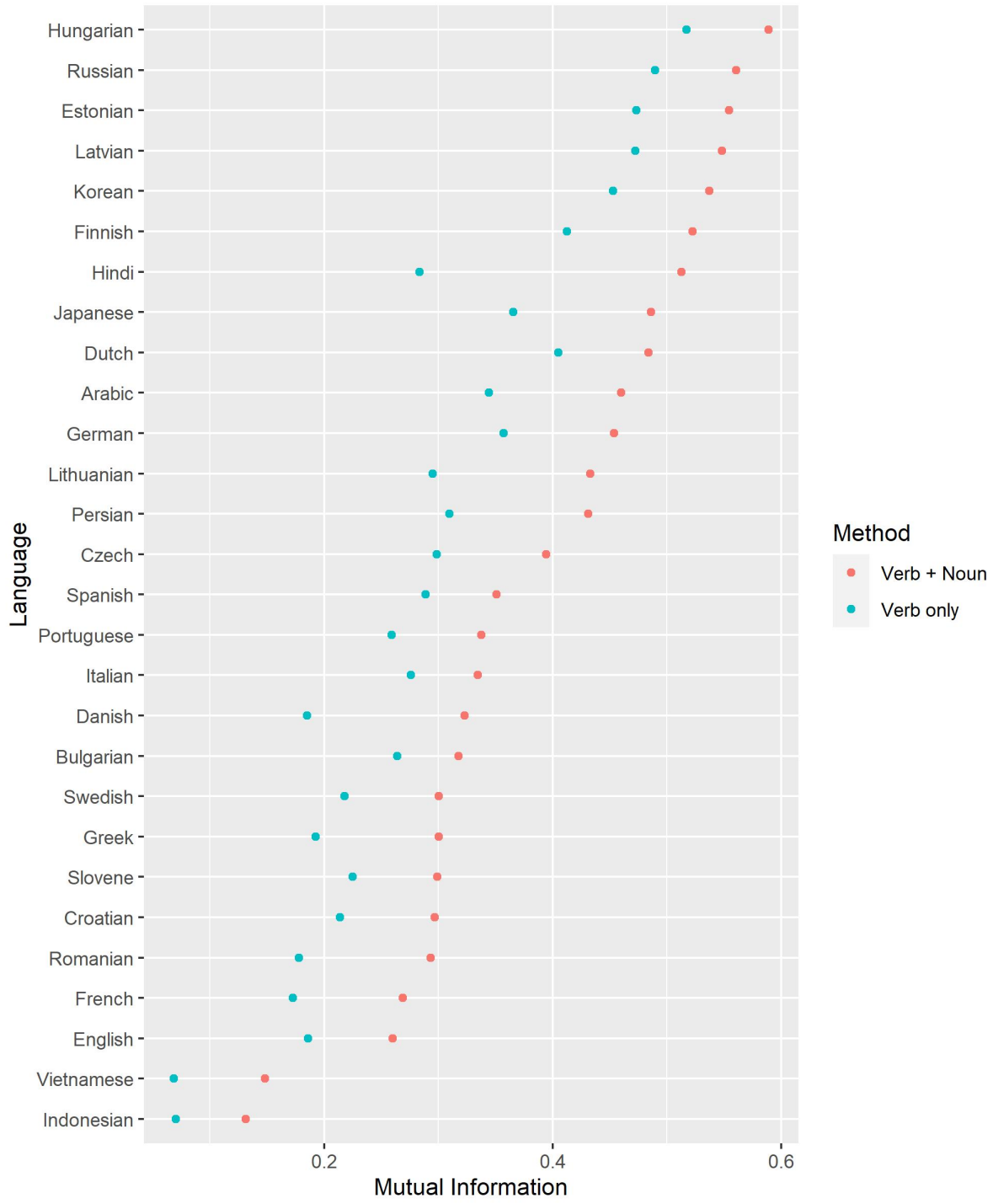


Figure 2: Distribution of MI scores representing P-lability. The higher the score, the weaker this type of lability in a language.

In order to answer this question, we computed how strongly the frequencies of a Verb + Noun combination (or of a Verb only) deviated from the expected proportion based on the total frequencies for the alternating constructions in the datasets. This measure is called Kullback—Leibler Divergence (D_{KL}) in information theory. It is also known as

relative entropy. The formula for a Verb + Noun combination *V&N* and Construction *Cx* is as follows:

$$(7) \quad D_{KL}\,(P||Q) = \sum_j P\left(V\&N,\ Cx_j\right) log_2\ \frac{P\left(V\&N, Cx_j\right)}{Q\ (V\&N, Cx_j)}$$

where P stands for the observed proportions of the specific Verb + Noun combination in each of the constructional variants, whereas Q represents the probability of this constructional variant for all Verb + Noun combinations, derived from the corpus frequencies. The higher the measure, the more this Verb + Noun combination contributes to MI. For verbs only, the procedure is identical, but instead of the observed and expected proportions of constructional variants in all Verb + Noun combinations we compute the proportions of constructional variants for Verbs only.

As a result, we obtained D_{KL} scores for individual combinations of Verb + Noun and for Verbs only. For example, the combinations *drive* + *people*, *open* + *room*, *begin* + *work*, *show* + *video* and *ask* + *woman*, had very low D_{KL} scores for P-lability because the Noun in these combinations occurred as S and P of the Verb with comparable frequencies. In contrast, the combinations *have* + *opportunity*, *do* + *job*, *play* + *role*, *score* + *point* and *make* + *sense* had very high D_{KL} scores for P-lability because the Noun was only used as P with these verbs.

The average D_{KL} scores were strongly correlated with the analogous MI scores presented in Sections 3.1 and 3.2. When computed on the A-lability data, Pearson's correlations were $r = 0.89$ (Verb + Noun combinations) and $r = 0.67$ (Verbs only), both $p < 0.001$. As for P-lability, the correlations were even stronger: $r = 0.98$ for Verb + Noun combinations and $r = 0.86$ for verbs, both $p < 0.001$. This means that both measures reflect similar information.

In order to estimate how strongly the individual Verbs or Verb + Noun combinations vary, we computed standard deviations of D_{KL} for every language. The higher the standard deviations, the more variability there is between the Verb + Noun combinations or individual verbs with regard to their association to one or the other construction.

Figures 3a and 3b display the MI scores and the standard deviations of the D_{KL} scores for A-lability. Figure 3a shows the scores for Verb + Noun combinations, and Figure 3b displays the scores for Verbs only. There are no obvious correlations between the measures in either plot.

Now let us turn to P-lability. Figure 4a shows the MI scores and the standard deviations of the D_{KL} scores for Verb + Noun combinations. There is a clear negative correlation: lower MI (and higher P-lability) scores correspond to higher D_{KL}, although there is also quite a lot of variability in the middle of the plot. Figure 4b displays the same data, but for Verbs only. In this case, the negative correlation is even more obvious.

How can we interpret these results? The looser languages on the left-hand side of the plot are loose due to certain individual Verb + Noun combinations which are highly labile, while other verbs are less or not labile. By contrast, if a language is generally tight, as are the languages on the right-hand side, the individual verbs in such a language are quite uniformly tight.

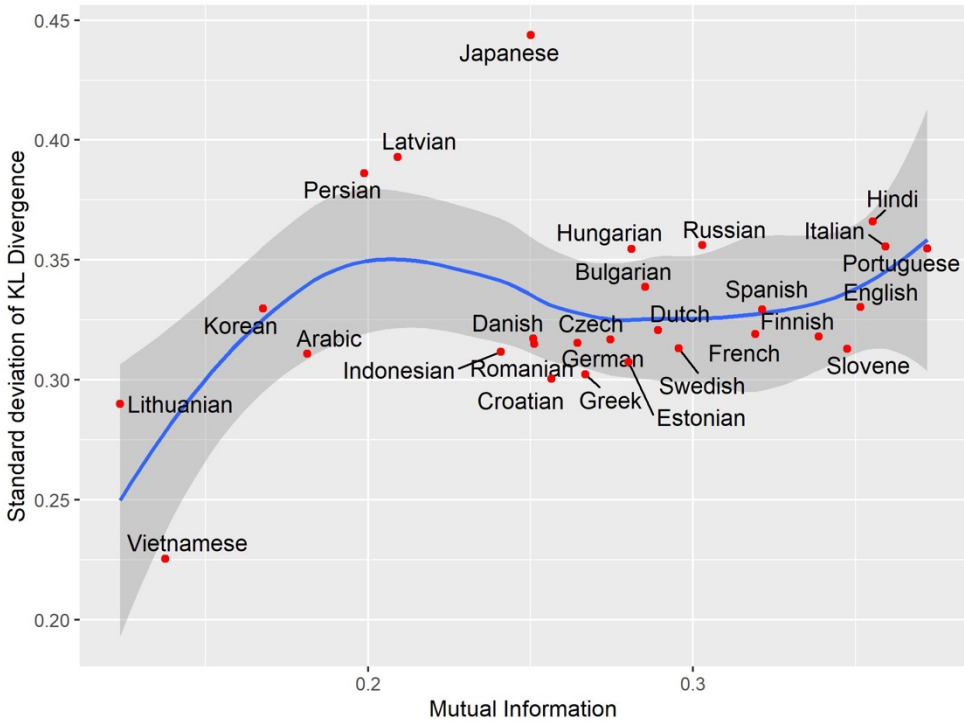


Figure 3a: MI scores representing A-lability (horizontal axis) and the standard deviations of the D_{KL} scores for individual Verb + Noun combinations (vertical axis). The curve is based on the LOESS smoothing method.

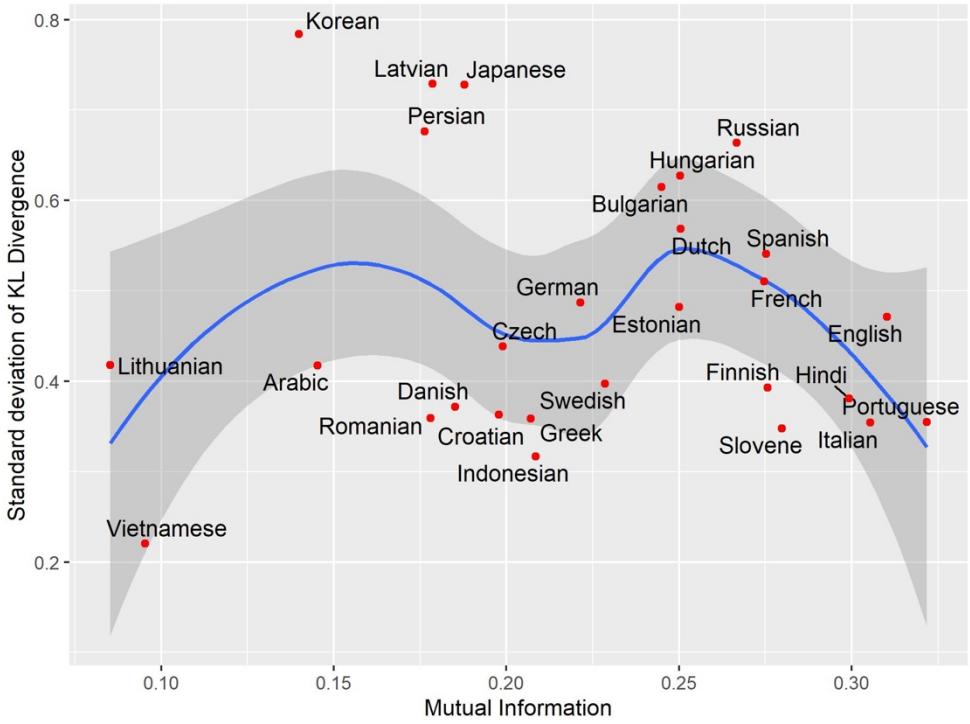


Figure 3b: MI scores representing A-lability (horizontal axis) and the standard deviations of the D_{KL} scores for individual Verbs (vertical axis). The curve is based on the LOESS smoothing method.

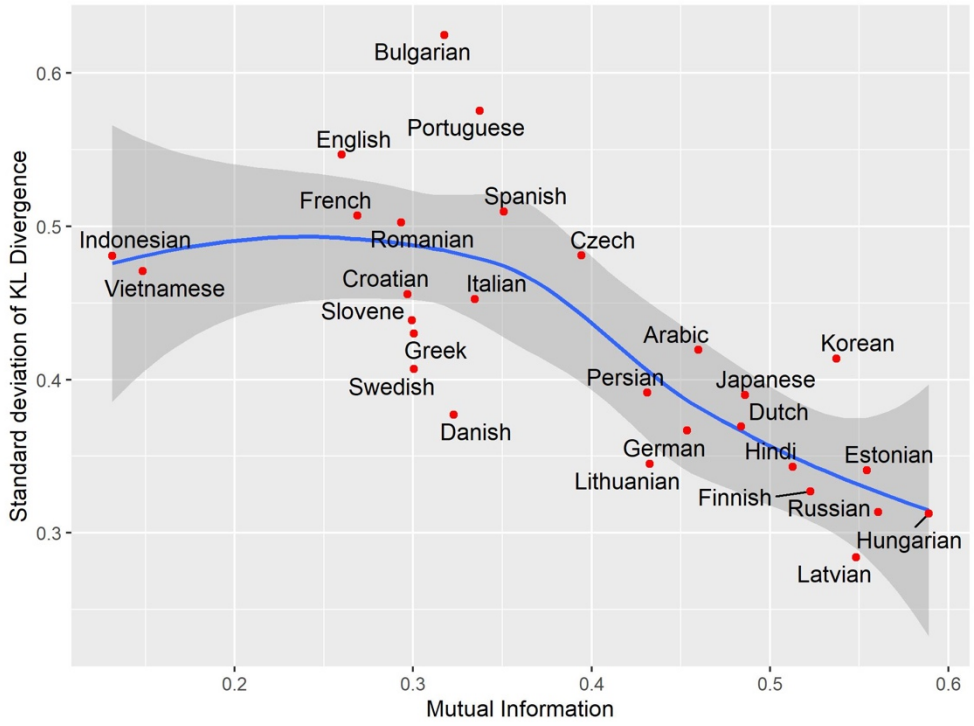


Figure 4a: MI scores representing P-lability (horizontal axis) and the standard deviations of the D_{KL} scores for individual Verb + Noun combinations (vertical axis). The curve is based on the LOESS smoothing method.

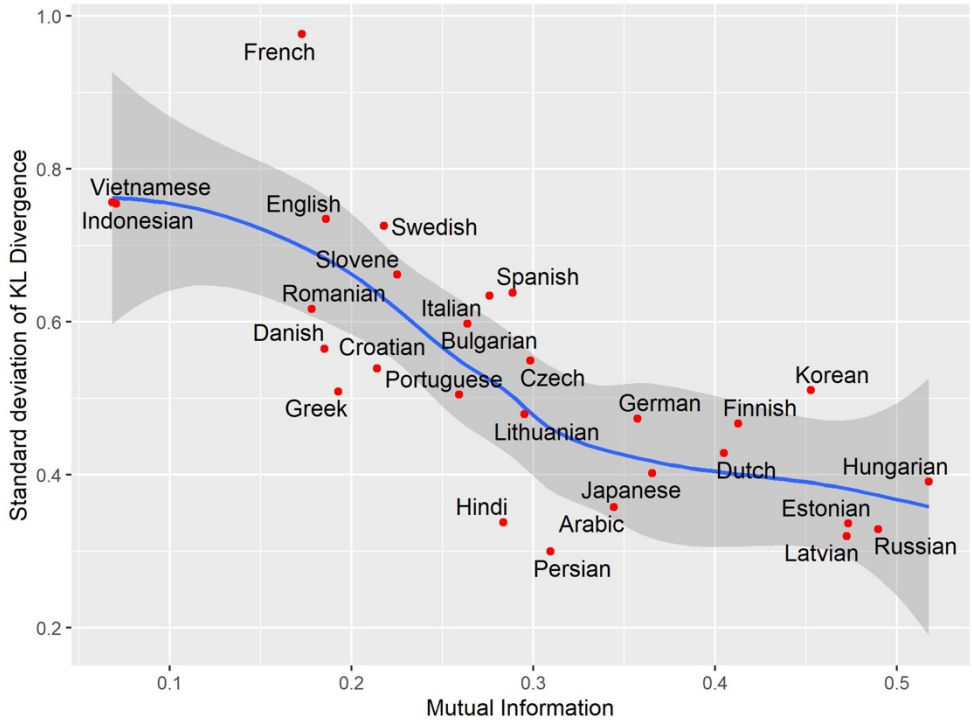


Figure 4b: MI scores representing P-lability (horizontal axis) and the standard deviations of the D_{KL} scores for individual Verbs (vertical axis). The curve is based on the LOESS smoothing method.

4. Correlations with other typological parameters

In this section we test for correlations between the A-lability and P-lability scores based on Mutual Information and the following parameters: rigidity of Subject and Object order; position of the lexical verb in the clause; case marking; and associations between lexemes and grammatical roles, which serves as a proxy for semantic tightness. We recycle the data from Levshina (2021), where the parameters were estimated by using the same online news corpora. More specifically, rigidity of Subject and Object order was computed as 1 minus entropy of SO and OS orders. To compute entropy, we used Shannon's (1948) famous formula, as shown below:

(8) $H(X) = -(P(SO)\,log_2\,P(SO) + P(OS)\,log_2\,P(OS))$

The proportions of SO and OS orders in transitive clauses were computed first based on the corpora, and then these entropy scores were computed (see Levshina 2019). If the proportions of SO and OS orders are equal (0.5), this leads to an entropy score of 1. If only one of the orders is used (either SO or OS), this leads to zero entropy. Since

entropy represents word order variability, we subtracted the entropy scores from 1 in order to obtain measures of word order rigidity. Lithuanian, Hungarian, Latvian, Czech and Estonian had the lowest scores and therefore the most variable orders, and Indonesian, French, English, Danish and Swedish had the highest scores and thus the most rigid orders. Note that in all languages, the SO order was the more frequent one. So we can speak about the rigidity of SO order. This variable was called 'Rigid Order (SO)'.

Another measure was the proportion of main clauses with a lexical verb between the Subject and Object. As expected, it was near-zero in verb-final languages, such as Japanese, Korean, Persian and Hindi, and close to one in Indonesian, English, French, Vietnamese and Portuguese. This variable was labelled 'Verb between Subj and Obj'.

We also took into account how much case marking was present to help in identifying the Subject and the Object. In Levshina (2021), the scores represented Mutual Information between case and the corresponding grammatical roles (A and P). For languages with adpositional case marking, the data were extracted automatically. As an illustration, consider the frequencies for Spanish in Table 3.

Case	**Transitive Subject**	**Direct Object**
Zero marking	126,736	569,252
Preposition *a*	0	55,442

Table 3: Frequencies of zero case marking and the direct object marker *a* for Subject and Object in Spanish.

For languages with case morphology, random samples were drawn and analyzed manually in order to take account of case syncretism. Next, the counts were extrapolated to all occurrences of Subjects and Objects in transitive clauses in a corpus. Consider an illustration in Table 4, which contains frequencies for Russian. In languages with distinct forms for Subject and Object and also forms with case syncretism, as in Russian, these three situations were represented by separate rows.

Case	**Transitive Subject**	**Direct Object**
Nominative	47,521	0
Accusative	0	93,520
Nominative/Accusative (case syncretism)	42,884	246,361

Table 4: Frequencies of Nominative, Accusative and case syncretism forms in Russian

German was a special case, where all feminine, neuter and plural forms were treated as ambiguous, since their Nominative and Accusative forms are formally indistinguishable, whereas masculine nouns were analyzed as Nominative or Accusative only in the presence of determiners or adjectives, which normally carry the distinct marking in combination with the noun. See more details about the procedure in Levshina (2021).

Based on numbers like those displayed in Tables 3 and 4, we computed the Mutual Information between cases (*C*) and grammatical roles (*R*) for each language, using the formula in (9).

(9) $$I\,(C;\,R) = \sum_{i,\,j} p\,(C,\,R_j)\,log_2\,\frac{p\,(C_i, R_j)}{p\,(C)\,p\,(R_j)}$$

The higher the Mutual Information, the more strongly the case forms are associated with the grammatical roles in question. Languages with zero scores had no case marking on Subject and Object (Danish, Dutch, English, Indonesian, Swedish and Vietnamese). Languages with the highest scores were those with rich morphological case marking (Lithuanian, Hungarian, Latvian, Estonian and Japanese). Languages with some type of differential, lexically restricted or optional marking were in-between (the Slavic languages, Hindi, Korean, German, Persian and Turkish). The variable with these scores was called 'Case Marking'.

Finally, we took the Mutual Information between Nouns only and the grammatical role of Subject and Object as a proxy for semantic tightness. If the proportions with which a Noun is found as a transitive Subject and Object are similar to the baseline proportions of Subject and Object, this contributes to the semantic looseness of a language. If a Noun is strongly biased towards one of these roles, this increases its semantic tightness (see Levshina 2021 for more details). The higher the Mutual Information, the tighter the language. The languages with the highest scores were Hindi, Korean, Russian, Hungarian and Japanese. They are known as tight-fit languages in the literature with regard to the relationships between arguments and their semantics. Indonesian had the lowest score, followed by English and Spanish. These were the loosest languages in our sample. This tightness measure was labelled as 'MI Nouns'.

The correlation analyses were based on Spearman's rank-based correlations. In order to control for the genealogical dependencies in our data (i.e., the fact that many languages come from one and the same genus), we used a sampling procedure

whereby we created 1,000 samples. For every sample, we drew randomly only one language per genus and computed the correlation coefficient (ρ). For null hypothesis significance testing, the test statistic was first computed and logged for the original pairs of scores in every simulation. Next, we also ran 1,000 permutations, in which the original scores of the second variable were randomly reshuffled. The permutation scores helped us to capture the distribution of the test statistic under the null hypothesis. Next, we counted the number of cases out of 1,000 permutations where the permuted scores were equal to or more extreme than the original test. After we had these data for all samples, we averaged the coefficients and the *p*-values.

Figure 5 represents the correlation coefficients between the parameters, with the p-values being shown in Figure 6. One can see in Figure 5 that both types of P-lability scores (Verbs and Verb + Noun) are correlated with the other typological parameters.

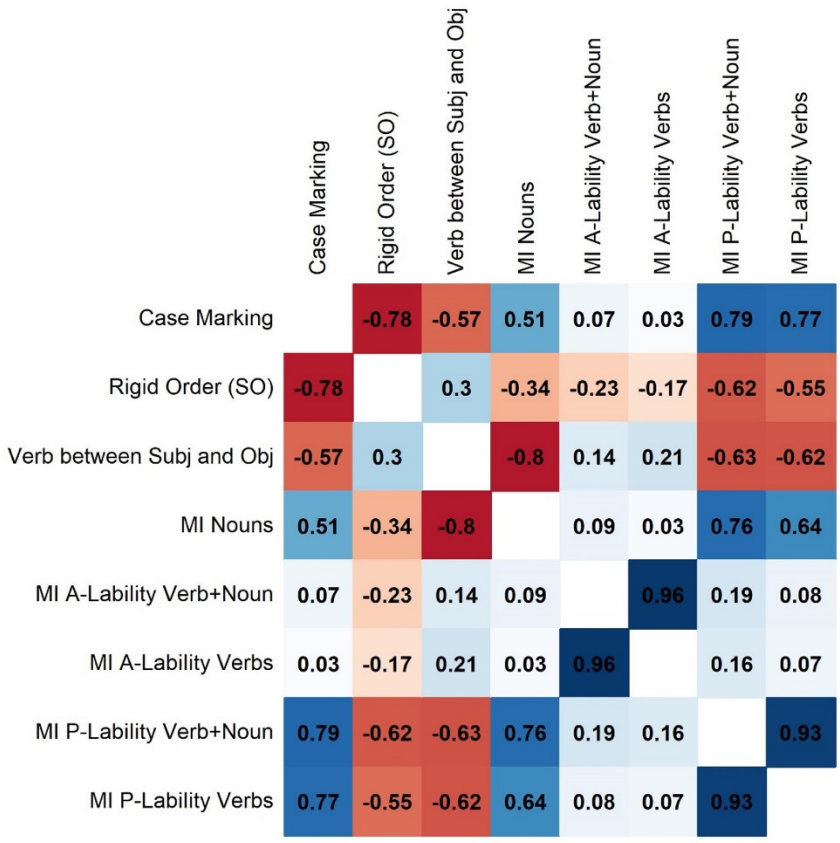


Figure 5: Correlations between the typological parameters and lability scores. The colour intensity represents the strength of the correlation. Blue cells stand for positive correlations. Red cells display negative correlations.

	Case Marking	Rigid Order (SO)	Verb between Subj and Obj	MI Nouns	MI A-Lability Verb+Noun	MI A-Lability Verbs	MI P-Lability Verb+Noun	MI P-Lability Verbs
Case Marking		**0.001**	**0.025**	**0.052**	**0.352**	**0.351**	**0.002**	**0.004**
Rigid Order (SO)	**0.001**		**0.161**	**0.141**	**0.233**	**0.278**	**0.02**	**0.037**
Verb between Subj and Obj	**0.025**	**0.161**		**0.001**	**0.314**	**0.245**	**0.015**	**0.017**
MI Nouns	**0.052**	**0.141**	**0.001**		**0.336**	**0.34**	**0.002**	**0.01**
MI A-Lability Verb+Noun	**0.352**	**0.233**	**0.314**	**0.336**		**0**	**0.27**	**0.375**
MI A-Lability Verbs	**0.351**	**0.278**	**0.245**	**0.34**	**0**		**0.295**	**0.37**
MI P-Lability Verb+Noun	**0.002**	**0.02**	**0.015**	**0.002**	**0.27**	**0.295**		**0**
MI P-Lability Verbs	**0.004**	**0.037**	**0.017**	**0.01**	**0.375**	**0.37**	**0**	

Figure 6: Average p-values based on resampling and permutation

The correlation between both types of MI scores for P-lability and case marking is strong and positive. This means that languages with systematic case marking have low P-lability. There is also a positive correlation between MI related to P-lability and MI based on Nouns only. At the same time, P-lability scores are negatively correlated with verb-medialness and word order rigidity. This means, in turn, that languages with SVO and rigid SO order have more P-lability. Judging from the magnitude of the coefficients, we can also see that the P-lability scores based on verbs only are overall less strongly correlated with the other typological parameters than the P-lability scores based on Verb + Noun combinations.

We also observe significant negative correlations between rigid SO order and case marking, and between verb-medialness and case marking. In addition, there is a strong and significant negative correlation between Mutual Information based on associations between nouns and grammatical roles, and verb-medialness.

The A-lability scores do not participate in any significant correlations, however. They are only strongly correlated between themselves. This means that we do not find evidence that A-lability is correlated with any of the typological parameters.

5. Conclusions

Our quantitative analyses reveal that P-lability scores are systematically correlated with the other parameters related to tight and loose fit. Languages with low P-lability (and high MI scores) tend to have case marking, stronger associations between nouns and grammatical roles, relatively flexible order of Subject and Object, and verb-final order. These features are associated with tight-fit languages. In contrast, languages with high P-lability (and low MI scores) tend to have little or no case marking, quite rigid SVO order, and weaker associations between nouns and grammatical roles. These features are associated with loose-fit languages. Therefore, our data support Hawkins' (1986, 1995) observation that verbs in loose-fit languages are used in more diverse subcategorization frames, while the argument co-occurrences of verbs in tight-fit languages are more constrained. More generally, our analysis provides further evidence that the classification of languages into loose- and tight-fit is a theoretically useful one.

In order to explain why these and other grammatical features of this general typology correlate in the ways they do Hawkins (1995, 2014, 2019) appealed to language processing, and especially to the crucial role of the verb in online processing. At issue is: does the verb precede its co-occurring arguments (as in VSO and SVO languages), or do these latter precede the verb (i.e. SOV)? In the former case there is early activation in processing of the verb's co-occurrence possibilities (with temporary ambiguities and possible garden paths when parsing English-type languages as the intended frame is gradually selected). In SOV languages, co-occurrence frame activation and selection based on preceding material will be almost instantaneous.

Hawkins (1995, 2014, 2019) formulated predictions for correlating properties based on this difference. Languages with verb-final structures (like Japanese and Korean) should exhibit what he called greater 'predicate frame differentiation' and 'argument differentiation'. Predicate frame differentiation refers to the degree to which a verb is distinctive from others by virtue of its unique selectional restrictions or syntactic co-occurrence possibilities. A verb that is uniquely transitive is more differentiated, and less labile in the terminology of the present paper, than one that is ambiguously transitive or intransitive. A verb that selects restricted direct objects for 'putting on clothing' according to the body part and the type of clothing in question (hats on the head, a coat over the rest of the body, etc.) is more differentiated than one (like English *put on*) that is compatible with many different types of body parts

and clothing activities (Planck 1984, Hawkins 1986). Argument differentiation refers to the degree to which arguments are, e.g., differentially case-marked versus ambiguous as to case, and the degree to which they are assigned a narrow set of thematic roles like Agent and Patient rather than the broader set of English (which permits e.g., Locative and Instrumental subjects as in *This tent sleeps four people* and *The key opened the door*).

When the verb is the last constituent in the clause and the very next item to be parsed belongs in an altogether different clause, the parser must succeed instantly in selecting the correct frame and its arguments. The grammar and lexicon must, Hawkins hypothesized, help the parser of a verb-final language by ensuring that predicate frame selection is immediately successful. More differentiated predicate frames and arguments can accomplish this, and in a number of observable ways.

First, subcategorization restrictions can be made tighter in SOV languages. For example, it is desiderable to avoid transitive/intransitive ambiguities with the result that arguments can be paired with their predicates more uniquely and more easily and with less variability and lability. Second, additional selectional restrictions can be imposed so that certain verb-NP pairings are more constrained, frequently co-occurring and easily recognizable. Third, subjects and objects can be made less semantically diverse in SOV languages with the result that there are fewer co-occurrence possibilities to choose from and more constrained assignments of thematic roles to NPs (transitive subjects are agents, transitive objects are patients, etc.). We expect fewer assignments of thematic roles to a transitive subject such as Location and Instrument in these languages. Fourth, surface coding devices can be grammaticalized for arguments that permit immediate thematic role recognition. This is what case marking generally does. It constrains the thematic roles that can be assigned to surface NPs, making them less semantically diverse (and also making them less labile) compared with case-less languages. This has the dual advantage of making thematic role information available early on-line, prior to the verb, and of facilitating argument-predicate assignments once the verb is encountered. And indeed it has long been known that there is a strong correlation between SOV and case marking, as confirmed in the correlations of the present paper.

More generally we expect to see verb-final languages with a more constrained set of verb co-occurrence possibilities. We expect more predicate frame differentiation and more argument differentiation, but less 'argument trespassing', which Hawkins (1995, 2014, 2019) defined to mean less movement of NP arguments into clauses in

which they contract no semantic relations with their most immediate predicates. For verb-early (SVO, VSO, VOS) languages, however, no such constraints are predicted. These languages do not need to conventionalize devices that permit immediate and correct predicate frame recognition at the verb, because the parser still has the remainder of the clause in which to complete its predicate frame selection and argument recognition, and because aspects of the verb's processing (e.g. its precise interpretation and disambiguation) that depend on subsequent arguments cannot be identified at the verb anyway. As a consequence, the need for immediate and correct decision-making at the verb, and the resulting need for clear predicate-frame differentiation, argument differentiation, and for local argument-predicate matching will impose much weaker requirements on the grammars and lexicons of such languages. This is why verb-early languages are quite variable: Indonesian (SVO) reveals many similarities with English, for example, whereas Hebrew (SVO) has far fewer (Müller-Gotama 1994). Verb-final languages should be more constrained in these respects, therefore, whereas languages with earlier verbs in the clause are predicted to be more variable. Verb-final languages are accordingly generally tight-fit, whereas verb-early and verb-medial languages can be tighter (like Hebrew) or looser (like English), as we have seen in this paper.

What ultimately underlies the tight vs loose typology, as explained in Section 1, is the mapping between forms and meanings. Tight-fit languages have richer, more complex and more unique surface forms that map onto less ambiguous and more constrained meanings, i.e. there is more of a one-to-one correspondence between form and meaning. This simplifies the mapping between them, but at the expense of processing more complex forms, for example case-marked nouns. Loose fit languages involve simpler processing of their more minimal and semantically general forms like caseless nouns, but they require more complex contextual processing and disambiguation of meaning assignments through 'word-external properties', see Hawkins (2019). This trade-off can be measured and made more precise in terms of the different MI and lability scores that we have seen in this paper for different languages, in the assignments of different nouns to grammatical roles, and in the further correlations with case marking, SOV, verb-mediality and rigid order. For further exploration of the processing basis for these correlations in terms of competing efficiencies and the general theory of efficiency in processing (as laid out in Gibson et al. 2019; and Levshina 2022), and for suggestions for psycholinguistic experiments

that need to be conducted on different language types from this perspective, the reader is referred to Hawkins (2019).

We have found further that P-lability scores based on verbs only are overall less strongly correlated with the other typological parameters than the P-lability scores based on Verb + Noun combinations. This is not surprising, because these latter scores also include the attraction of nouns to different grammatical roles. At the same time, both of these scores are more strongly correlated with word order rigidity and case marking than semantic tightness scores based on nouns only, but the latter has a stronger correlation with verb-final order (since the languages in the sample, except for Arabic, are either verb-medial, or verb-final). This may have to do with the fact that the attraction of nouns to one or the other role helps to avoid costly reanalysis when the verb comes last. Whether or not the verb has special marking depending on the roles of its arguments is less important for that purpose.

It is remarkable that our P-lability scores are more strongly correlated with the other typological parameters than the latter are among themselves. This is an unexpected finding, but it can be explained by the fact that P-lability scores convey information not only about the verbs, but also (explicitly or implicitly) about the nouns in different roles. These scores can thus be a useful indicator of the word-external or word-internal orientation of the language in question (Hawkins 2019).

We should also mention that high P-lability for a language in general seems to be due to some highly labile verbs, which behave 'promiscuously', appearing in both constructional frames. At the same time, there are also lexemes with low promiscuity. In contrast, the verbs or combinations of verbs and nouns in tight languages with low P-lability tend to be uniformly faithful to one or the other constructional frame. These conclusions are based on our analysis of lexical variability based on Kullback—Leibler Divergence scores (also known as relative entropy).

In contrast to P-lability, the A-lability scores are not correlated with any of the typological properties we have examined here. A-lability is also found more frequently in our corpora than P-lability, as we see from the lower MI values in the former. A possible explanation for this is that A-lability is often driven by general pragmatic factors. For example, the object can be omitted due to its high accessibility, e.g., *And Italy wins [the final]!* Many objects are omitted due to specific conventionalized inferences, *e.g., He drinks again [liquor].* Object omission is also possible if the focus is on the action, while the object has low discourse prominence, e.g., *She chopped and chopped [e.g., meat]* (Goldberg 2005). Other reasons are cultural.

For example, the object can be omitted when it is taboo, e.g., *Pat sneezed [mucus] onto the computer screen*, or for feelings of tact, *I contributed [$1,000] to UNICEF* (Goldberg 2005). In addition, many rules allowing for object omission are also lexically and semantically specific (Fillmore 1986). All these pragmatic factors and lexical idiosyncrasies explain the lack of systematic correlations between A-lability and the other typological properties of the languages in question. It would be very interesting to investigate if there exist cross-linguistic regularities at the level of individual verbs and verb classes, and if so which semantic roles alternate (cf. Letuchiy 2009). As mentioned above, it would not be realistic to annotate manually all verbs for their semantics in the twenty-eight languages, so we have to leave this question for future research.

We also hope to have demonstrated that information-theoretic measures, such as Mutual Information, entropy and Kullback-Leibler Divergence, can be fruitfully used for language comparison. As initiatives such as the Universal Dependencies progress and typologically diverse large corpora are made available it becomes increasingly possible to infer typological variables directly from texts and to build on, and refine, earlier typological patterns derived largely from grammars, as in the present study. We hope that new corpora and tools will allow for further testing of the findings of this study using a larger and more diverse sample of languages and genres. A further question is then whether there are causal relationships between the parameters themselves, and what they would look like exactly. A causal analysis in Levshina (2021) showed that case marking is more likely to be affected by other typological parameters (word order and associations between lexemes and syntactic roles) than the other way round. We need a larger sample of languages in order to answer this question and to test all possible causal links. It would also be interesting to add the other parameters of Hawkins (1995, 2014, 2019), such as the frequency of long-distance syntactic dependencies or of categorial ambiguity, and to test their relationships with the parameters examined here.

Data availability statement

The dataset with the Mutual Information and other values for every language in the sample is available as supplementary materials on the journal website.

Acknowledgements

The first co-author's work on this paper was funded by the Netherlands Organisation for Scientific Research (NWO) under Gravitation grant "Language in Interaction", grant number 024.001.006. We thank the reviewers for their insightful and constructive feedback, which has helped us improve the paper. All remaining errors are only ours.

References

Aissen, Judith. 2003. Differential object marking: Iconicity vs. economy. *Natural Language and Linguistic Theory* 21. 435-483.

Creissels, Denis. 2014. P-lability and radical P-alignment. *Linguistics* 52(4). 911-944.

Dahl, Östen. 2000. Egophoricity in discourse and syntax. *Functions of Language* 7(1). 37-77.

Dixon, R.M.W. 1994. *Ergativity*. Cambridge: Cambridge University Press.

Fillmore, Charles J. 1986. Pragmatically Controlled Zero Anaphora. *Proceedings of the Berkeley Linguistics Society* 12. 95-107.

Gibson, Edward, Richard Futrell, Steven T. Piantadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen, & Richard Levy. 2019. How efficiency shapes human language. *Trends In Cognitive Sciences* 23(5). 389-407.

Goldberg, Adele E. 2005. Argument realization: the role of constructions, lexical semantics and discourse factors. In Jan-Ola Östman & Miriam Fried (eds.), *Construction Grammars: Cognitive grounding and theoretical extensions*, 17-44. Amsterdam: John Benjamins.

Goldhahn, Dirk, Thomas Eckart & Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig Corpora Collection: From 100 to 200 languages. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck et al. (eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, 759-765. Istanbul: ELRA.

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.

Haspelmath, Martin. 1993. More on the typology of inchoative/causative alternations. In Bernard Comrie & Maria Polinsky (eds.), *Causatives and transitivity*, 87–120. Amsterdam: John Benjamins.

Hawkins, John A. 1986. *A Comparative Typology of English and German: Unifying the Contrasts*. London: Croom Helm.

Hawkins, John A. 1995. Argument-predicate structure in grammar and performance: A comparison of English and German. In Irmengard Rauch & Gerald F. Carr (eds.), *Insights in Germanic Linguistics, Vol. 1 Methodology in Transition*, 127-44. Berlin: Mouton de Gruyter.

Hawkins, John A. 2012. The drift of English towards invariable word order from a typological and Germanic perspective. In Terttu Nevalainen & Elizabeth C. Traugott (eds.), *The Oxford Handbook of the History of English*, 622-632. Oxford: Oxford University Press.

Hawkins, John A. 2014. *Cross-linguistic Variation and Efficiency*. Oxford: Oxford University Press.

Hawkins, John A. 2019. Word external properties in a typology of Modern English: A comparison with German. *English Language and Linguistics* 23(3). 701-723.

Letuchiy, Alexander. 2009. Towards a typology of labile verbs: Lability vs. derivation. In Patience Epps & Alexandre Arkhipov (eds.), *New Challenges in Typology: Transcending the Borders and Refining the Distinctions*, 247-268. Berlin / New York: De Gruyter Mouton.

Levin, Beth. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago: University of Chicago Press.

Levshina, Natalia. 2019. Token-based typology and word order entropy. *Linguistic Typology* 23(3). 533-572.

Levshina, Natalia. 2020. How tight is your language? A semantic typology based on Mutual Information. In *Proceedings of the 19th International Workshop on Treebanks and Linguistic Theories*, 70-78. Düsseldorf: ACL.

Levshina, Natalia. 2021. Cross-linguistic trade-offs and causal relationships between cues to grammatical subject and object, and the problem of efficiency-related explanations. *Frontiers in Psychology* 12. 648200.

Levshina, Natalia. 2022. *Communicative Efficiency: Language Structure and Use*. Cambridge: Cambridge University Press.

Müller-Gotama, Franz. 1994. *Grammatical Relations: A Cross-Linguistic Perspective on Their Syntax and Semantics*. Berlin: Mouton de Gruyter.

Nichols, Johanna. 1986. On form and content in typology. In Winfred P. Lehmann (ed.), *Language Typology 1985: Papers from the Linguistic Typology Symposium, Moscow, 9-13 Dec. 1985*, 141 - 162. Amsterdam: John Benjamins.

Nichols, Joanna, David A. Peterson, & Jonathon Barnes. 2004. Transitivizing and detransitivizing languages. *Linguistic Typology* 8. 149-211.

Plank, Frans. 1984. Verbs and objects in semantic agreement: Minor differences between English and German that might suggest a major one. *Journal of Semantics* 3(4). 305-360.

Shannon, Claude E. 1948. A Mathematical Theory of Communication. *Bell System Technical Journal* 27. 379-423, 623-656.

Wijffels, Jan, Milon Straka & Jana Straková. 2018. udpipe: Tokenization, parts of speech tagging, lemmatization and dependency parsing with the UDPipe NLP Toolkit. R package version 0.7.

Corpora

Leipzig Corpora Collection

https://wortschatz.uni-leipzig.de/en/download

CONTACT

natalia.levshina@mpi.nl