


Linguistic Typology

at the Crossroads



ISSN 2785-0943

Volume 2 – issue 2 – 2022

Issue DOI: <https://doi.org/10.6092/issn.2785-0943/v2-n2-2022>

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

<https://typologyatcrossroads.unibo.it>

Editors

Nicola Grandi (University of Bologna, Editor in chief)

Caterina Mauri (University of Bologna, Editor in chief)

Francesca Di Garbo (University of Aix-Marseille)

Andrea Sansò (University of Insubria)

Publisher

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

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

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



Linguistic Typology

at the Crossroads



Editorial board

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

Scientific Board

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

Responsible Editor

Caterina Mauri, University of Bologna

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

Production editors

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

Assistant production editors

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



Linguistic Typology

at the Crossroads



CONTENTS

Expletive negation and related problems

Paolo Ramat ----- 1-38

Person alignment in reported speech and thought: the distribution and typology of participant roles (based on six Finno-Ugric languages)

Denys Teptiuk----- 39-92

Predicting grammatical gender in Nakh languages: Three methods compared

Jesse Wichers Schreur, Marc Allasonnière-Tang, Kate Bellamy, Neige Rochant----- 93-126

Cross-linguistic sources of anticausative markers

Guglielmo Inglese----- 127-186

Reports from the field

The rise of unemphatic negation: two standard negation constructions in Oji-Cree and their patterns of use

Matthew Windsor----- 187-221



Expletive negation and related problems

PAOLO RAMAT

ACCADEMIA NAZIONALE DEI LINCEI - ROMA

Submitted: 06/07/2022 Revised version: 31/10/2022

Accepted: 31/10/2022 Published: 22/12/2022

There are more things in Heaven and Earth, Horatio, / than are dreamt of in your Philosophy

W. Shakespeare, *Hamlet* 1.5.167-8.

‘Ci sono più cose in cielo e in terra, Orazio, / che **non** ne sogni la tua scienza’

Não sou nada. /Nunca serei nada.

‘Non sono niente. / I will never be anything.’

Fernando Pessoa, *Tabacaria*

Abstract

This paper makes two main claims: the presence of two or more negative elements in Negative Concord, Negative Comparison, and Expletive Negation basically rests upon the pragmatic need of intensifying the negative import of the sentence. Secondly, the paper aims at identifying a possible path of functional expansion that may account for the use of the Expletive Negation in the constructions under scrutiny. Expletive Negation is originally tied to the core concept of inequality comparison (x is more/has more y than z, where y refers to a state, a quality or property), as well as to temporal comparison (x before y) – where the second member is implicitly negated; it may then expand to other constructions such as Negative Concord and the construction of fear verbs.

Keywords: Double Negation, Equative and Negative comparison, Expletive negation, Negative concord, Fear verbs, Pragmatic strategy.

1. Introduction and Summary

More than 20 years ago, taking advantage of a stimulating article by Carlotta Viti on the equatives in RgVeda (Viti 2002), I wrote a note on the negative comparison (Ramat 2002). Since then, the literature on negative comparison (henceforth

NegComp) has experienced a real boom, including a thread of LINGTYP¹ (January 2022) on ‘Negation marks adverbial clauses’ with interesting examples from many different languages. Moreover, Johan van der Auwera and Chiara Gianollo, at the 55th Annual Meeting of the Societas Linguistica Europaea (Bucharest 2022), organized a workshop on “A hundred years of negative concord”. Therefore, I felt myself pushed to reconsider the issue based on the many publications that offer different viewpoints, and consequently different answers.

NegComp (for instance, *Paris is **not** as big as Tokyo*) cannot be considered in isolation since it reveals deep connections with other phenomena concerning negative constructs, such as negative concord (NegConc: for instance, *I did **not never** hear such a song*) and expletive negation (EN: for instance, It. *questo discorso è più pericoloso di quel che tu **non** creda* ‘this speech is more dangerous than you (***not**) believe’). Starting with the statement that “[n]egation is one of the few truly universal grammatical categories: every language seems to have some grammaticalized means to deny the truth of an ordinary declarative sentence” (Willis et al. 2013: 1), I agree with Garzonio and Poletto’s conclusions (2015: 147) that “negation is not a simple element [...] but a complex one, formed as the result of the interaction of several abstract processes” that have to do not only with grammatical categories but also with the cognitive and pragmatic behaviour of the speakers.

The received opinion is that negative concord (NegConc) is a separate phenomenon from expletive negation (EN), while NegComp and negation with fear verbs are traditionally assumed to be subtypes of EN. In the following sections, I shall consider the four negative constructs separately to uncover the ‘red thread’ uniting not only EN, NegComp, and negation with ‘fear verbs’, but also NegConc.

The question of the so-called ‘Negative Concord’, though apparently different from NegComp, has several contact points with the latter, at least from the theoretical point of view.

The main points of the negative expressions that do not concern the standard negation marked by \neg (i.e., the so-called polarity NEG), namely, the negation of the truth of an affirmation (p : *Paul likes ice cream* \rightarrow $\neg p$: *Paul does not like ice cream*), are Negative Concord, Negative Comparison, Expletive Negation, and Expletive Negation with Fear Verbs. The aim of the following sections is to discuss each of these points and to provide a key to understanding their interconnections. I will start from Expletive Negation as a useful point of departure.

¹ lingtyp@listserv.linguistlist.org

Expletive negation is dealt with in Section 2, followed by the discussion of Negative Concord in Section 3 and of Negative Comparison in Section 4. Negation with fear verbs is treated in Section 5. Finally, Section 6 discusses whether the expletive negation is always, or has always been, expletive and suggests an explanatory hypothesis for its diffusion.

2. Expletive Negation

A useful starting point to treat the problems alluded to is the so-called ‘expletive negation’, which, as we will see in many of the following examples, occurs quite often, particularly in non-standard language varieties. EN and its equivalents like ‘vacuous’, ‘pleonastic’, ‘abusive NEG’, commonly refer to the presence of a negative marker that does not give a negative sense to the utterance. For a more detailed and restrictive definition of EN, see below Jin & Koenig (2021: 41). Not much attention was traditionally paid to EN: in descriptive/prescriptive grammars, it was even disapproved of as useless and disturbing to the logic of the sentence.² Even recent comprehensive descriptions of Negation and its diachronic evolutions do not pay much attention to this largely and crosslinguistically diffused phenomenon; for instance, in Larrivé & Ingham’s collective volume on the evolution of NEG (2011) there is no ‘expletive (or pleonastic, vacuous, abusive) negation’ entry in the subject index. The same holds for Mosegaard Hansen & Visconti (2014), and Miestamo (2017). On the contrary, in the ten languages considered as case studies by Willis et al. (2013), the term EN appears with reference to the Italo-Romance, (High) German,

² Delfitto et al. (2019: 58) are right in stating that EN has received less attention in the literature concerning negation. However, already in 2002, Nocentini (not quoted in Delfitto et al.) had studied the following cases in connection with the ‘so-called expletive negation’ in Italian: dubitative interrogatives (*mi domando se non ci siamo-IND/SBJ.PRS comportati male* ‘I ask myself whether we behaved badly’); exclamatives (*che cosa non darei-COND.PRS per vederla contenta!* ‘what I would (not) give in order to see her be happy!’); comparatives (*la strada è più pericolosa di quel che tu non creda-SBJ.PRS* ‘the road is more dangerous than you believe’); ‘verba timendi’ constructions (*temo che non sia-SBJ.PRS vero* when meaning ‘I’m afraid it is true’); and some particular constructions such as the temporal one using *finché* (*rimasero a giocare finché non si stufarono-IND.PST* ‘they kept playing until they got tired’ or *finché non telefoni-SBJ.PRS qualcuno* ‘until someone will call’) and the avertives (*c’è mancato poco che non cadessi-SBJ.IMPF* ‘I almost fell (but I didn’t)’). Some of these constructions will be discussed in the following pages. To my knowledge, the most exhaustive description of the EN properties is Delfitto’s chapter in the *Oxford Handbook of Negation* (Delfitto 2020).

Low German and Dutch, Brythonic Celtic, Ancient and Modern Greek, and Slavonic languages. Jin (2021: 49) states that EN

occurs rather widely. In Jin's 1,142 language sample it occurred in 128 languages, on all continents, and in 63 genera. [...] Out of the 45 languages for which both research papers and reference grammars were consulted, expletive negation was mentioned in research papers but not grammars in 27 languages, suggesting that expletive negation most likely occurs in many more languages in Jin's sample: expletive negation is a relatively widespread phenomenon and certainly not an oddity of Romance languages.

Furthermore, Jin & Koenig's survey of 722 languages reports (2021: 40) that 74 languages show examples of EN. Their list of EN triggers goes beyond the traditional list considered in the present paper and considers not only temporal operators (see (10)-(12) below) or verbs such as 'regret', 'fear' etc. (see (49)-(53) below), but also logical operator triggers such as 'impossible', 'unless', 'without'. As can be seen in the references below, some valuable monographs have been recently dedicated to this topic (see fn. 2).

However, even languages that have EN can do without it (see (4) and the Italian translation of (8)). Significantly, van der Auwera (2009) is right in considering EN a universal potentiality. He collects examples where present-day French *ne*, *ne pas* and *pas* can all be considered as expletive; as *pas* has inherent negative meaning in all contexts and registers, independently of the presence of *ne* (cf. Mosegaard Hansen & Visconti 2012: 454f.), it can be expletive, i.e., pleonastic, in non-standard sentences such as ex. (1a), instead of the standard (1b):

(1) French (Indo-European)

- a. *J'imagine que je désire plus que j'peux pas obtenir*
I imagine that I desire morethan I can NEG obtain
'I imagine that I want more than I can obtain.'
- b. *J'imagine que je désire plus que je ne peux obtenir.*

Here are some examples where EN appears in utterances having different values: comparative, as in (1), (2) and (4), fear verbs in (3) and (5), and temporal in (6) and (7). By no means is this list exhaustive.³

From the following examples (2)-(7), we see that EN is a NEG that does not have the function of expressing negative polarity (see, among others, Dobrushina 2021: 121):

(2) French (Indo-European)

Il est plus grand que vous ne pensez.
He is bigger than you NEG think
'He's bigger than you think.'

(3) French (Indo-European)

J'ai peur qu' il ne pleuve demain
I have fear that it NEG rains tomorrow
'I am afraid that it will rain tomorrow.'⁴

(4) Italian (Indo-European)

Maria ha mangiato più biscotti che (non) Piero.
Maria has eaten more cookies than (NEG) Piero
'Mary has eaten more cookies than Peter.'

(5) Russian (Indo-European)

Boj-u-š, čto syn ne zaboje-l
Fear-1SG-RFL COMPL son NEG fall.ill-PST
'I am afraid that my son might get ill.'

³ For a general survey on EN across the languages of the world, see Jin & Koenig (2021) with reference to the previous literature.

⁴ As can be seen in the examples reported in footnote 2, the question of mood associated with EN in the complement clause is very intricate, especially in the Romance languages; however, it does not impact the very nature of EN. On the usage of indicative vs. subjunctive in the Italian historical tradition, see Nocentini 2003. As Jin & Koenig (2021: 71) have noted, only a subset of EN triggers selects the subjunctive mood, and the subjunctive can be optional. Moreover, the nature of the verb (accomplishment / resultative / event, state, etc.) plays a role in the choice of the verbal mood. In the following, I shall not deal with the moods that appear in the examples, except for a short discussion in Section 5.

(6) Neo-Aramaic (Afro-Asiatic; Khan 2016: 499)

ʾé-⁺dān léla ⁺vāra, |jəxcəla. |
 that-time NEG.COP.3FSG enter.PROG laugh.PROG.3FSG
 ‘Just as she enters, she laughs.’

(7) Danish (Indo-European)

Ikke så snart Amsterdam- traktaten trådte i kraft,
not as soon Amsterdam treaty-the came into force,
begyndte Europa-Kommissionen ...
 began EU-Commission
 ‘No sooner had the Treaty of Amsterdam entered into force than the European Commission began...’

Helmut Haberland (LINGTYP, 12 Jan. 2022) comments (7) as follows:

One could translate this [scil. *Ikke så snart*] into English with ‘no sooner’ and one might ask if a difference between “no sooner” (with negation) and ‘as soon as’ (without negation) always is observed by all speakers, as subtle it may be.

When speaking of translation, it is relevant to note that EN can be used in a target language, whereas it is not present in the source language, as in the famous Shakespearean verse quoted in the exergon and repeated here as (8):

(8)

a. English (Indo-European)

There are more things in heaven and earth, Horatio, than are dreamt of in your Philosophy.

b. Italian (Indo-European)

*Ci sono più cose in cielo e in terra, Orazio, che **non** ne sogni la tua scienza.*

In English it would be impossible to say *than aren't dreamt*.

To note that the EN is not obligatory in the Italian translation: ...*che ne sogni la tua filosofia* would be quite correct as well. Interestingly, Old Church Slavonic knew the same optionality of the negative particle *ne*, whereas later, in the 15th-17th centuries, the presence of *ne* with preverbal N-words became the preferred variant. Finally, the

ne accompanying N-words became almost obligatory from the 17th century on: Russian is nowadays a strict NegConc language (see Garzonio 2019, Section 5, with the corresponding diachronic examples).

This hint regarding the historical development and the uncertainty of the speaker about the presence / absence of NEG is very relevant. Generally speaking, we agree with Delfitto et al. (2019: 86) conclusion that EN has to be considered in the context of the speaker's presuppositions and implicatures (we will come back to this also in the following sections). A more detailed and restrictive definition of EN, intended to eliminate phenomena that, strictly speaking, do not belong to the EN domain, such as rhetorical questions, concessives, polite requests, etc., is offered by Jin & Koenig (2021: 41):

The occurrence of a negator is an instance of expletive negation if (i) it is included in a syntactic dependent of a lexical item (verb, adposition, adverb, or collocation), (ii) it is triggered by the meaning of that lexical item, but (iii) it does not contribute a (logical) negation to the proposition that the syntactic dependent denotes.

The French, Italian, and Russian examples above do fit the three conditions: *J'ai peur qu' il ne pleuve demain* is fine, but **Je souhaite qu'il ne pleuve demain*, intended as 'I wish that it will rain tomorrow', does not work because of condition (ii). The real negative sentence for 'I wish that tomorrow it will **not** rain' would be *Je souhaite qu'il ne pleuve pas demain*: with fear verbs as *avoir peur* EN entails the omission of *pas* (or, more precisely, the omission of *pas* marks *ne* as EN).

As we have just seen in (3) and (5), complement clauses of verbs of fear often contain an EN. I shall come back to this topic later. There are different contexts that invite EN. Though Jin & Koenig (2021) do not include exclamatives in ENS, it is worth paying attention to sentences like (9). Greco (2019, § 2.1) notes that EN rejects 'strong' Negative Polarity Items (NPIS) such as *affatto* 'not at all':

(9) Italian (Indo-European)

*Che cosa non ha (*affatto) capito Gianni!*

what not has (*not at all) understood John!

'What has John understood!' (= John understood everything!).

The *non* in (9) has an emphasizing effect, and the sentence without *non* (i.e. *che cosa ha capito Gianni!*) could even mean that Gianni has totally misunderstood what has been previously talked about (on the exclamative *non*, cf. Parry 2013: 99. Delfitto et al. 2019: 66ff. speak of (9) as capturing “the universal flavor of wh-exclamatives featuring EN”). The distinction between ‘strong’ and ‘weak’ NPIS (see below, fn. 13) does not impact the very nature of EN, which is a unitary phenomenon. Moreover, the intrinsic semantics of the verb can or cannot favour an EN construction, as in the case of *souhaiter*, or Russian *starat’sja* ‘to try’. Consequently, the decision to consider the presence / absence of NEG in a sentence is not a black or white question to be solved simply on the basis of syntactic criteria: the context implicatures play a decisive role in interpreting EN sentences.

Another aspect of the multifarious EN has been alluded to by Ludwig Paul (LINGTYP 12 Jan.2022) with the German sentence:

(10) German (Indo-European)

*Bevor ich das Geld nicht gespart habe, kann ich mir kein
 before I the money NEG saved have can I to.me no
 Auto kaufen.
 car buy*

‘Before I have saved money, I can’t buy a car for myself.’

This *nicht* is superfluous from a logical point of view, since the negative indefinite pronoun *kein* in the main clause already assures the global negative sense of the sentence, but in this context *nicht* has a hybrid temporal / conditional nature: ‘Before / If I do not save money enough, I can’t buy a car’. Hence, its possibility in the spoken language to extend the value of EN to different domains.

Among the EN constructions, Wälchli (2018: 149) introduces what he calls the ‘opportunity.before’ type, used for taking advantage of an opportunity which will not be possible anymore at some later point in time. He quotes the following Russian sentence from Iordanskaja & Mel’čuk (2009):

(11) Russian (Indo-European)

Poka (ja) *ne* *zabył*, *sxodi* *za*
 as.long.as/until/before (I) not forget:PST.MSG go.after:IPFV behind

xlebom!

bread:INS.SG

‘Before I forget: go buy some bread.’⁵

Japanese has the same construction (see Kuno 1973: 155 also cited in Wälchli 2018: 150):

(12) Japanese (Japonic)

Wasure-nai uti ni henzi o kakimasyoo

forget-NEG inside to answer ACC let’s.write

‘I will write an answer before I forget it.’⁶

In conclusion, we have seen thus far in examples (1)-(7) different usages of EN in different contexts (by no means all the possible contexts we can find in languages).

The following sections will briefly discuss some negative constructs that are dealt with in the literature and examine the relations of EN with these constructs to find a unifying ratio, the ‘red thread’ I allude to in the introduction.

3. Negative Concord and Embracing Construction

As is well-known, the term NegConc refers to the phenomenon according to which a negative element (a so-called ‘N-word’, or ‘Negative Concord Item’, NCI) occurs in a sentence with a negative marker ‘no(t)’. The negative component is expressed twice or more, if more than one N-word is present, but the sentence is interpreted as being negated only once.⁷

Romance languages have NegConc as almost obligatory. For instance, in Rumanian a sentence such as:

⁵ *before I don’t forget would be impossible, but *avant que je n’oublie (pas)* would be fine.

⁶ In French we might have *avant que je n’oublie pas*.

⁷ There exists a vast literature on NegConc. For recent reports on the state of the art, see van der Auwera & Van Alsenoy (2016), Giannakidou & Zeijlstra (2017), and Breitbarth et al. (2020). For a wider, philosophical approach, see Kuhn (2022). Moser (2019) gives an exhaustive overview of NegConc in the German dialects, with a focus on Upper German. As for the cognitive and processing operations tied to NegConc, with particular attention to its acquisition by children, see Tagliani et al. (2022), with a useful summary of the previous literature.

(13) Rumanian (Indo-European)

**Niciun student a citit cartea*
 No student has read book.the
 ‘No student has read the book.’

is not accepted. You have to insert the negator *nu* and say, instead

(14) Rumanian (Indo-European; see Moscati, forthcom.)

Niciun student nu a citit cartea
 ‘No student has read the book.’

In a sense, Rum. *nu* in (14) is logically superfluous since *niciun* already gives the sentence a negative meaning, as Germ. *kein* in (10). This fact shows the (diachronic) ties between NegConc and EN, the difference being that Rum. *nu* belongs now to NegConc as a grammaticalized element, while EN is ‘per se’ optional (on the discussion of the logic of NegConc, see below in this section).

The tendency to accompany negative quantifiers (often an indefinite PRO) by a predicate negator is widespread (see examples below). For instance, Russian shows NegConc in

(15) Russian (Indo-European)

Nikto ničego ne skazal.
nobody nothing not said
 ‘Nobody said anything.’

which corresponds to Italian (with triple NEG as in Russian)

(16) Italian (Indo-European)

Nessuno non disse niente
nobody not said **nothing**
 ‘Nobody said anything.’

However, (17) sounds rather emphatic, with the more standard form being

(17) Italian (Indo-European)

Nessuno disse niente
'Nobody said nothing.'

In Russian a sentence without NegConc (18a) is ungrammatical; the correct form is instead (18b):

(18) Russian (Indo-European)

- a. **Nikto zdes' menja znaiet*
nobody here me knows.
- b. *Nikto zdes' menja ne znaiet*
nobody here me **not** knows
'Nobody knows me here.'

NegConc appears also in dependent clauses:

(19) Portuguese (Indo-European; Pessoa, Tabacaria).

Como não fiz propósito nenhum, talvez tudo fosse
since not I.did purpose not.any perhaps everything would be
nada.
nothing
'Since I acted out no purpose, perhaps then everything was nothing.'

German is a language that does not have NegConc, and a possible translation of (19) would be, with two coordinated sentences:

(20) German (Indo-European)

Allerdings hatte ich mir nichts vorgenommen, und so war
however had I to.me nothing proposed, and thus was
dies alles vielleicht nichts
this all perhaps nothing

Logicians have maintained that the meaning of (15) should be 'Everyone said something' and that there is a form-meaning mismatch. Hence, the consideration of

NEG in (15) and (16) as ‘vacuous’, ‘pleonastic’, i.e. ‘expletive’. However, it is a fact that (16) and (17) are both usually meant as negative, and all NEG elements contribute to the negative (emphatic) sense of the sentence with a NegConc strategy.⁸

One hundred years ago, Jespersen noted that in Old English it was the regular idiom to say: *nan man nyste nan þing*, ‘no man not-knew nothing’ (Jespersen 1922: 352). This construction still survives in non-standard English; Jespersen quotes from George Eliot: *there was niver nobody else, [...] gen* (‘given’) *me nothin*’ (see ‘Negative Concord’ in Glottopedia). The construct where negative indefinite pronouns such as *nobody*, Rus. *nikto*, Span. *nadie* etc. co-occur with another indefinite negative (e.g., *nothing*, Rus. *ničego*, Span. *nada.*, etc., and also *never*, Rus. *nikogda*, Span. *nunca*, etc.)

⁸ Exceptions such as the colloquial French type *C’est pas rien* ‘It is not nothing’ (=‘It is quite something’), where two NEG words give a positive meaning, are rare. See de Swart (2010: 252). In the present case, the indefinite negative *rien* is in the scope of *pas*, and consequently the two negatives produce a positive outcome according to the mentioned principle of logic that such a statement is equivalent to the denial of the second negative *rien*, as in

- i. *it is not the case that John is not here* which means ‘John is here’.

See the classical example of Lat. *nemo non* (Gianollo 2018: 145; Moscati, forthcom. § 4.1):

- ii. *aperte enim adulantem nemo non videt* (Cic., *Lael.*,99)
 blatantly in.fact flattering:ACC no.one:NOM not see:3SG
 ‘no one does not recognize someone who is blatantly flattering.’ (i.e., everyone recognizes someone who...)
- iii. **Nemo** ergo **non** miser Prorsus *nemo* (Cic., *Tusc.* 1.9)
nobody then **not** unhappy.
 ‘No one then is not unhappy. Absolutely nobody.’, i.e. ‘Everyone is unhappy’

Remember that Classical Latin is basically not a NegConc language but a ‘double NEG language’, where two negations lead to a positive statement: each additional negative element / operator contributes to the meaning of the sentence. As stated by Gianollo (2019: 245), in ‘double negation languages’ (like German, English, Latin, and also Homeric Greek, Gianollo 2021) two negative indefinites do not produce a NegConc but represent two effective negations whose sum results in a positive sense:

- iv. German (Indo-European)
Hans gab keinem Teilnehmer nichts
 Hans gave to.none participant nothing
 ‘Hans gave something to every participant.’

is largely diffused in languages around the world, particularly in non-standard or non-standardized varieties (see de Swart 2010; Moser 2019). Old Italian knew NegConc: [...] *e comandò a' baroni che nessuno non li insegnasse spendere questo oro* (Novellino, 7, 33.10-11) '...and he ordered the barons that nobody should instruct them how to spend this gold'. Van der Auwera and Neuckermans (2004: 462) quote the following examples from the Flemish dialects:

(21) Flemish (Indo-European)

'k ben niemand ni tegengekomen

I am nobody NEG met

'I have not met anyone.'

with *ni* as NEG, along with the standard Dutch strategy, that is, without NegConc:

(22) Dutch (Indo-European)

'k ben niemand tegengekomen

I am nobody met

'I have not met anyone.'⁹

Indefinite pronouns are usually \neg , when used as elements of the so-called 'non-canonical NEG', i.e., in negation strategies that are pragmatically marked (see Schwenter 2006; cf. Ballarè 2019: 211).¹⁰

Therefore, in English we have both

(23) *Nobody has said anything*

as well as

(24) *Nobody has said nothing*

⁹ See also the ten case studies presented in Willis et al. (2013), with the interaction between NEG and indefinites that raise NegConc.

¹⁰ On the Romance indefinite 'N-forms' in NegConc, as in It. *Non ho visto nessun bambino* 'I haven't see any child', Span. *No tengo ningún comentario oficial sobre eso* 'I don't have any official comment on it', Port. *Não vou assistir a nenhum espetáculo hoje à noite*. 'This night I'm not going to see any show', etc., see, among others, Gianollo (2018).

and both have the same first-order logic representation (see de Swart 2010: 248):

(25) $\neg\exists x\neg\exists y \text{ Say } (x,y)$

As an anonymous reviewer has noted, (25) contains two negative operators, that is, it describes the double negation reading found in the standard language. However, (24) is actually used in spoken language as being equivalent to (23). Accordingly, it has to be taken as a non-standard NegConc structure. Such a meaning equivalence shows the optionality and the speaker's possible uncertainty between NegConc and double NEG: *any* is not a NEG-expression per se, cf. (26).

(26) *Is there anything I can do to help?*

The optionality and uncertainty between NegConc and double NEG reading in certain environments – as is the case in Russian for holophrastic answers to negative questions (Garzonio 2019: 177) – is very important from the diachronic point of view since, as we will see below, it can give a hint in explaining the presence of EN.

However, first we must have a look at the so-called ‘embracing (or discontinuous) negation’, such as in Catalan *Joan no menja pas peix*, or French *Jean ne mange pas de poisson*, ‘John does not eat fish’ (see Bernini & Ramat 1996: 44;49). As the following examples (29)-(35) show, embracing negation, with its two negative items, may lead to understanding why one negative element may become superfluous and hence an EN.

Embracing constructions as well as NegConc have double negative lexemes, and both are interpreted as having just one negative meaning. Note that the second negative element of the embracing NEG and NegConc is usually located in the focus position, which in short sentences such as (27) and (28) coincides with the end of the sentence. Zanuttini (1997), and subsequently Poletto (2008), have collected a large sample of this kind in Italian dialects, with the indefinite PRO closing the sentence in focus position like the NPIs:

(27) Emilian dialect (Indo-European)

E' n m' a vest entsun
 SCL NEG me has seen nobody
 ‘Nobody saw me.’

(28) Piedmontese (Indo-European)

*A parla **nen** cun **gnun***
he speaks **not** with **nobody**
'He does not speak with anybody.'

Such a trend in underlining the negativity of the sentence at its very end is psychologically and cognitively on a par with the historical development of the embracing NEG-type, as in Fr. *ne...pas* or in Malt. *ma....-x*:¹¹

(29) Maltese (Afro-Asiatic)

Marika ma rat-x lit-tifel
Marika NEG saw-NEG the-boy
'M. did not see the boy.'

¹¹ The Malt suffixes *-x*, Arab. *-š* derive from *šay?* 'thing' (cf. Fr. *rien* < Lat. *rem*-ACC 'thing'). Thus, originally they were NPIS; see Bernini & Ramat 1996: 235, fn.54. Later, they became obligatory parts of the NEG construction, i.e., they grammaticalized, according to the so-called Jespersen's cycle. I do not think it necessary to explain once more the well-known 'Jespersen's Cycle' with its evolution illustrated by the classical French example: 1. *Ne* + VB → 2. *Ne* + VB + *pas* → VB + *pas*. It is enough to note that Ahern's Dissertation (2015), by using mathematical and statistical methods, has reconstructed the same evolution from Middle English to Modern English: from pre-verbal *ne* to an initially emphatic embracing *ne...not*; from embracing *ne...not* to post-verbal *not*. As for the French cycle, it must be remembered that French-based creoles use NEG + VB (= back to stage 1 of the cycle and hence ready to start a new cycle!); *mo pa kup* lit. 'I NEG cut', i.e. 'I don't cut' (Louisiana); *nu pa vle* lit. 'we NEG want', i.e. 'We don't want' (Guadeloupe), along with the more French type that has postverbal NEG: *mo kup pa*; *nu vle pa* (see Posner 1985: 180-83; Ramat 2006; for a critical assessment of Jespersen's cycle, see Larrivé 2011). Preverbal *nu* is found also in the Spanish-based creole of Palenque (see further below):

i. Palenquero (Indo-European creole; Schwegler 1991a: 173, from Friedemann & Patiño Rosselli 1983)

Pero kumu nu ta yobé juu!, sé mori toito
pero como **no** TENSE/ASP llover uu TENSE/ASP morir todito
but if **not** rain uu die everything
'But since it doesn't rain, uuh! everything dies.'

Note that the term 'negative polarity item' used by Booij & Bettelou & Rem (2006) has nothing to do with the NPIS alluded to in this footnote. *Ghe-* is a prefix that in Middle Nederland transforms the following verb (mostly *kunnen*) into a negative polarity item.

Both *pas* and the suffix *-x* are (or were) negative polarity items whose function was originally to enhance, as postverbal minimizer objects, the negative sense of the sentence. And both can be subsumed under the label of ‘multiple negation’, which intensifies the negation, such as the double NEG and the NegConc.

On the other hand, Siller-Runggaldier (1985) has studied the Central Ladin (near the Austrian-Italian border) negation in sentences like

(30) Central Ladin (Indo-European)

- a. Livinallongo and Fassano
La no 'veŋ ('nɔ)
- b. Gardonese
La ne 'veŋ ('nɔ)
- c. Badiotto
Ara ne vegn (no)
‘She’s not coming (at all).’

She has observed (1985: 74) that the second *no* is always sentence-final (= NEG#) and does not refer to any particular element of the sentence, but resumes the negative value of the entire sentence. Further examples from Northern Italian dialects can be found in Poletto (2008; see, e.g., her ex. 32: **No ghe so ndà no**, lit. **Not** there I.am gone **not**, i.e. ‘I didn’t go there’ with the second *no* in focus position).

Similarly, in Afrikaans, a V2 language like Dutch and German, every negative sentence must end with the NEG *nie* (or another negative morpheme):¹²

(31) Afrikaans (Indo-European)

- Jan het nie geëet nie*
J. has **not** eaten **not**
‘John has not eaten.’

versus Dutch *Jan heft niet gegeten.*

¹² According to Bernini (1994, § 4.2) the Afrikaans negative construction is a calque from the colonial Portuguese, once again a non-cultivated language variety. Bernini reports Valkhoff’s (1972: 94) distinction between ‘High’ and ‘Low Portuguese’: “a pidginized, creolized or simplified Portuguese” used in South Africa. (On Brazilian Portuguese, see immediately below.)

If a negative sentence has a subordinated clause, the NEG *nie* is repeated at the end of the subordinated clause and refers to the main sentence, according to the general rule of *nie* in final position.

(32) Afrikaans (Indo-European; Bernini 1994, ex. 13)

Jan het nie gedink [dat iets met hom sou gebeur] nie
J. has NEG thought [that something with him would happen] NEG
'John didn't think that something could happen to him.'

In (Brazilian) Portuguese, the NEG *não* is frequently repeated at the end of the sentence and the same holds for the creoles of Palenque (Colombia) and Chocó-Spanish (Northwestern Colombia; see Schwegler 1990: 170):

(33) Brazilian Portuguese (Indo-European)

Ele não fala português não
he NEG1 speaks Portuguese NEG1
'He does not speak P.'

(34) Palenquero (Indo-European creole)

I nu ta kandá nu
I NEG1 TENSE/ASPECT sing NEG1
'I'm not singing (at all).'

(35) Chocó-Spanish (Indo-European)

Yo no sé no
I NEG1 know NEG1
'I do not know.'

The Afrikaans ex. (32) follows exactly the same scheme (hence, the calque hypothesis referred to in fn. 12): 'NEG1...NEG1#' is valid also when a subordinate clause is inserted:

(36) Brazilian Portuguese (Indo-European)

Ele não sabe que o pai chegou não
he NEG1 knows that the father arrived NEG1#
'He does not know that his father arrived.'

where the final NEG1# refers to the main sentence ‘He doesn’t know’ and not to the subordinate clause (see Schwegler 1991a: 199). Schwegler (ibid.) has shown that there may be constructions with double NEG as in (33) or final NEG only as in

(37) Palenquero (Indo-European creole; Schwegler 1991b: 172)

E kelé fruta nu
 he want fruit **not**
 ‘He doesn’t want any fruit.’

There are other examples of sentence-final NEG with the Scheme NEG1 + VB + NEG1#, which is rather widespread among the languages of the world (cf. Bernini 1994 for the African languages; Ramat 2006).¹³

In examples (33)-(36), the second NEG is the repetition of the first one (NEG1...NEG1(#)), and therefore its history is different from that of NPIs such as *pas* or *goutte*, *mica*, etc., which originally represented the (emphatically minimizing) object of the sentence (NEG1...NEG2(#)): *non bibo guttam* ‘I don’t drink a drop’, used in ancient French as NEG2 (see *Jeu d’Adam*, 12th cent.: *Par Dieu, je n’ai goute d’argent* ‘By God, I don’t have a penny’); *non vado passum* ‘I don’t move a step’. As is well-known, (parts of these) NPIs were finally grammaticalized as a NEG marker: *Je ne vais pas, je ne vois pas* (see fn. 11)¹⁴.

According to the Optimality Theory, there is a contrast between the overall tendency to place the verb in the scope position of NEG (i.e., to have preverbal NEG)

¹³ A good history of the French negation, endowed with fine examples, can be read in Grieve-Smith’s Dissertation (2009). From the acquisition point of view, we may refer to Tagliani (2019; reported by Tagliani et al. 2022: 116): “double negation, which is extremely marked from both a pragmatic, prosodic and syntactic perspective, is learned by children only at a much later age, at around 7, when they start to correctly interpret and produce these marked constructions”. As for the creoles, among whom the already-mentioned Palenquero, see the accurate description in Bernini (1994, Section 3.2). Note that the second NEG does not necessarily coincide with the first one. In the Portuguese-based creole of São Tomé, for instance, we find *na* (=NEG₁)...*fa* (=NEG₂) and in angular (São Tomé and Príncipe) *na* (=NEG₁)...*wa* (=NEG₂). It has been suggested that *fa* and *wa* derive from the aboriginal languages of West Africa. But the problem is too complex and too far from the topic of the present paper. The main point remains that creoles do have embracing NEG.

¹⁴ Greco’s (2019) twofold distinction between ‘strong’ and ‘weak’ NPIs, based almost exclusively on syntactic properties, does not help very much in defining what NPIs properly are: *It alzare un dito* ‘to lift a finger’ or *avere la più pallida idea* ‘to have the faintest idea’ are considered as ‘weak-NPI’. But this makes the concept of ‘NPI’ very vague and potentially unlimited. For a more exhaustive definition, see Bernini & Ramat (1996: 30-34).

and to place the postverbal NEG in the focus position¹⁵. This contrast has caused the emergence of discontinuous NEG (= NEG... NEG; cf. de Swart 2010: 247), both via the repetition of NEG1 or the grammaticalization of NPIs.

However, what we finally get in both cases is an embracing NEG, which is really not necessary for the negative sense of the sentence, so that one negative element can disappear: *Je vois pas* ‘I don’t see’. If both NEGs are maintained, this may open one of the ways leading to EN.

The repetition of NEG1 in the sentence-final position represents an afterthought strategy, the aim of which is (or was, before its grammaticalization) to enhance the negative force of the entire utterance, as in:

(38) French (Indo-European)

Je (ne) sais pas moi, non.

‘I don’t know, **nah**.’¹⁶

These *no*, *non*, *nah*, etc. may be considered as a kind of comment on the previous utterance; we have already seen that NEG reinforcement, both via negative PROs (i.e., NegConc) or quantifiers and via NPIs, appears frequently in non-standard varieties.¹⁷

This fact can easily be considered as the speaker’s intention to underline the negative content of her/his utterance when s/he speaks in an informal, spontaneous but pragmatically efficient way (‘emphatic negation’). And this strategy is not limited to the spontaneous face-to-face conversation. Even modern literary texts make current use of this strategy: Bernini (1994) quotes an example from Jorge Amado:

(39) Brazilian Portuguese (Indo-European)

“*Quem vem lá?*” – disse Ferreira. “*É homem. Não é bicho não*”

who comes there? - said Ferreira. is man. **Not** is animal **not**

‘Who’s coming?’ –said F. ‘It is a man, not an animal.’

¹⁵ On the focus properties of comparison (a topic we shall come back to in a moment) see Gawron 1995.

¹⁶ Jespersen (1917: 72) had already spoken for the Port. construction *não* VB *não* of “incorporation of a post-sentential, afterthought like ‘resumptive negation’”; see further, Schwegler (1990: 170) and (1991a); Ramat (2006).

¹⁷ See further examples in Google s.v. Negative Concord, with reference to the *Yale Grammatical Diversity Project*, which offers many instances from non-standard varieties of American English, as *I don't never* have **no** problems ‘I don't ever have any problems’ (African American English).

To sum up what we have said thus far: whatever the origin of the doubling negative element – i.e., via the repetition of NEG1 as in Port. *não*, or via NPIs as in Fr. *pas* – the final step in the construction is more or less the obligatory grammaticalization of postverbal or sentence final NEG (i.e., ‘...NEG#’), as we have seen in the Afrikaans ex. (31); see also *Ek het dit nie gesien nie* lit. I have this **not** seen **not**, ‘I don’t have seen this’. This entails, particularly for the NPIs, the well-known progressive bleaching of their original meaning. However, this holds also for the repetition of ‘NEG1#’ as in Braz. Port. *não... não*. Moreover, the final NEG, which can have a holophrastic value, particularly in answers¹⁸, has finally become a simple marker of NEG. In other words, ‘...NEG#’ is the final grammaticalized and no longer emphatic step in the ‘NEG....NEG#’ emphatic construction: cf. Dutch

(40) Dutch (Indo-European)

<i>Die man verstaán</i>	<i>die werk eenvoudig</i>	<i>nie!</i>
that man understand:PRS.3SG	the work simply	not!
‘That man simply does not understand the work!’		

4. Negative comparison

Some of the constructions that express (negative) comparison belong to the same psychological strategy as illustrated in the previous section (see examples (26)-(29)), though the outcome is not a double NEG. Stassen (1985: 218), based on a sample of 110 languages, says there exists an “intimate relationship of the comparative construction with negation” in a large part of his language sample and that the comparison may include a negative pronoun.¹⁹ Holthausen (1913), in a short note entitled ‘Negation statt Vergleichungspartikel beim Komparativ’, had already observed that comparisons such as

¹⁸ On the holophrastic negative forms, see Floricic (2016). Already present in Old French: *Nel feras? – Non!* (*Jeu d’Adam*, 171) ‘Won’t you do it? – No’, and continued in today’s spoken French: *T’as rien vu hier soir? – Rien* ‘Did you see anything yesterday evening? - Nothing’.

¹⁹ In her PhD thesis, Ballarè quotes (2019: 34) the following sentence twitted by the famous goalkeeper Gigi Buffon for the footballer Paulo Dybala upon his winning the ‘Gazzetta dello Sport’ Award: *Meriti questo premio più di nessun altro @gianluigibuffon !!!*. lit. You deserve such a prize more than **nobody** else, i.e., ‘Nobody would deserve such a prize more than you!’

(41) Middle English (*Leg.Schott.* III1103, anno 1400)

Sonare na onyman cuth thynke

‘eher als jemand denken konnte’

‘sooner than anybody could [*not!] think.’

have an implicitly negative sense, which explains the presence of *na*. Holthausen refers to a previous article by Fraenkel (1911) that mentions the same construction in Sanskrit, Ancient Greek, the Slavonic languages, and Lithuanian. Holthausen quotes further English dialectal comparative forms using *nor* (< OEngl. *náwðer* < *náhwæðer* < I.-E. NE-K^{WO}-TERO- with the negative element *ne* followed by the typical I.-E comparative suffix *-tero-*):

(42) *He’s older nor I/me*

by Holthausen literally translated as ‘er ist älter, und nicht ich’ meaning ‘he’s older than me’.

In Modern English, *nor* is a conjunction used before the second or last in a set of negative possibilities: its etymologically negative comparative meaning has been preserved. Similarly, Klein (1980) considers EN as arising from an underlying negation. But NEG can even surface in negative comparisons, as we have seen in (42). Some languages have grammaticalized the negative strategy: “(a BIG) & (b not-BIG)” meaning ‘a is bigger than b’. See Classic Nahuatl (examples from Stolz & Stolz 2001: 39; cf. Andrews 1975: 350ff.):

(43) Classic Nahuatl (Aztecan)

a. *Oc achi ni-yec-tli in amo tehua-tl*
still some 1SG-beautiful-ABS DET NEG 2SG-ABS

‘I am more beautiful than you.’ (lit. I’m beautiful, you not)

b. *Tlapanahuia ic tilitic in cacalo-tl in amo huexolo-tl*
surpass with black DET raven-ABS DET NEG peacock-ABS

‘The raven is blacker than the peacock.’ (lit. the raven surpass with blackness not the peacock).

Twenty languages in Stassen's sample (110 languages) adopt this strategy (called by Stassen 1985: 44ff. 'conjoined comparative [with] a positive-negative polarity'):

(44) Hixkaryána (Cariban)

kaw-ohra naha Waraka, kaw naha Kaywerye
 tall-not he:is W. tall he:is K.
 'K. is taller than W.'

(45) Telugu (Dravidian)

I-pandu a-pandu-kanna tipi-ga undi
 this-fruit that-fruit-not.be:PRS.PTCP sweet-this is
 'This fruit is sweeter than that one.' (Stassen 1985: 310)²⁰

Consider now the French sentence

(46) French (Indo-European)

Je suis parti avant qu'il ne soit arrivé
 'I left before he arrived.'

As we have already seen in examples (10) and (11), *avant* ('before') establishes a temporal comparison between two states of affairs (A before B), and the sentence refers in its second part to a non-real state of affairs: at the moment of my departure he was **not yet** arrived: "il s'agit toujours d'une comparaison établie [...] entre deux événements dont l'un n'existait pas encore quand l'autre s'est produit" ('we have to do with a comparison between two events, and the second one did not yet exist when the first happened', Vendryes 1950: 5; cf. Delfitto 2020, referring to Krifka 2010: "For every time *t* preceding A, $\neg B(t)$ "). Forest (1993: 115) quotes Welsh and Berber parallels.²¹

²⁰ On the contrary, the non-comparative negative sentence has

(i) *I-pandu tipi-di ka-du*
 this-fruit sweet-this not-is (Stassen 1985: 51)

²¹ See also Lithuanian (Indo-European; Wälchli 2018:194)

(i) [Šlovės Dievas apsireiškė mūsų tėvui Abraomui Mesopotamijoje,]
kai jis dar nebuvo persikėlęs į Charaną,
 as he yet NEG.be:PST.3SG move:RFL.PST.PA.NOM.SG.M into Haran

As Vendryes states, (47a) sharply contrasts with (47b):

(47) French (Indo-European)

- a. *il a bien changé depuis qu'il a été malade*
'he has really changed since he has been ill.'
- b. **depuis qu'il n'ait été malade*
'after he was not ill.'

(47b) is impossible, since it is a fact that he's been ill.

As we will see below in Sections 5 and 6, what might appear as a series of uncorrelated observations can be reduced to a common denominator: namely, the presence of negative elements: via Negative Polarity Items in embracing negation (as in Fr. *il ne mange pas*), in Negative Concord (as in (14): *Niciun student nu a citit cartea*), in the Expletive Negation (as in (46): *Je suis parti avant qu'il ne soit arrive*), and in Negative Comparison (as in (41): *sonare na ony man cuth thynke*).

An anonymous reviewer has noted that there is a difference between NegConc and embracing negation on one side and comparatives on the other since the latter do not seem to have a pragmatic need for enhancement nor to have a negative meaning. This

'[The God of glory appeared to our father Abraham while he was in Mesopotamia,] before he settled in Haran.'

Old Church Slavonic uses 'before' also in negative hortatives that refer to two actions where one has to occur before the other, exactly as in the French *avant* example (43):

- (ii) *Sъnidi* *prěžde* *daže* *ne* *oumьretь* *otročę* *moe*
descend:IMP.2SG before HORT.SUB NEG die:(PFV)PRS.3SG. son my
'...come down before my child dies!' (Wälchli, *ibid.*)

Cf. Old Italian (Bono Giamboni, *Libro*, Chap. 11, para 28):

- (iii) *...e da te non si partiranno giammai (...) infino che non t'hanno data la vittoria*
'...and from you they will never separate **until/before** they will [***not**] give you the victory.'

On the Italian before-clauses with EN of the type

- (iv) *Avvertila prima che non succeda qualche guaio!*
'Warn her before some trouble does (***?not**) happen!'

see the discussion in Delfitto et al. (2019: 61ff.). See also Wälchli's proposals in Section 6 below.

is correct as far as the pragmatic enhancement is concerned, but it does not change the fact that we have seen lot of examples, drawn from very different languages, where comparisons are or were expressed via a negative strategy (see, particularly, (42)). Therefore, we are entitled to consider negative comparative constructs along with other negative constructs, particularly when we find sentences containing temporal comparison or sentences with fear verbs, which will be discussed below (cf. examples (61) and (50), respectively).

As we have said, one of the causes that triggers the insertion of a negative form or the doubling of NEG is the pragmatic need to enhance the negative meaning of the sentence. Jin & Koenig's (2021; see Section 2) conclusion seems to be valid for all NEG types considered in this paper: the presence of affirmative sentences *p* (*John is heavier than Dik*) may lead speakers to produce $\neg p$ (*John is heavier than **not** Dik*, scil. is): Sanskrit *mṛtam śreyo ná jīvitam*” lit. ‘death better **not** life’, i.e., ‘death is better than life’, where the negative particle *ná* functions as a comparative particle (cf. Viti 2002: 71). In Italian we can have two possible translations: *la morte è meglio che **non** la vita*, vs. *la morte è meglio che la vita*. In the first sentence *non* is clearly EN.

Another interesting aspect of EN has to be considered. Kuteva (1998: 124ff.) cites the following Bulgarian example of ‘action nearly averted’:

(48) Bulgarian (Indo-European)

<i>Bez / Za</i>	<i>malko</i>	<i>ne</i>	<i>padnax</i>
without / for	few	NEG	fall1:SG.AOR
‘I almost fell.’			

which corresponds exactly to Ital. “per poco **non** cadevo”, along with the equally possible “per poco cadevo” (that implies ...*but I didn't*).

Queffélec (1988: 422ff.) has a specific section on the facultative presence of EN vs. Ø in Old French. See for instance, *Mais moct crienment qu'ele leur faille* ‘but they much fear that she could fail them’ vs. *Et por crienme que il n'i faillent / S'esvertuent de lor poeirs* ‘and fearing that they [***not**] fail there, they strengthen their troops’, Benoît de Sainte Maure *Chronique*, v. 34830 and 35814f., respectively.²²

²² However, utterances expressing the idea of ‘failing to’ do not necessarily belong to the domain of EN, pace Queffélec (1988), who quotes Marie de France, *Lais. Guigemar* 735f.: *Quant el l'oi, si suspira / Par un petit ne se pasma* ‘As she heard him, she sighed so strong / that she failed to faint’: actually, the dame did not faint, and consequently the NEG is not superfluous.

5. Negation with fear verbs

The above observations suggest some reflections on a particular kind of EN we have already introduced in Section 2. In fact, EN concerns not only fear verbs but also verbs such as ‘to think’, ‘to doubt’, ‘to be uncertain’, ‘to avoid’, etc. In other words when used with these kinds of verbs, EN “is a means for the speakers not to commit to the truth of what they are saying” Dobrushina (2021: 126, and cf. Grieve-Smith 2009: 166). Verbs like ‘think, fear, avoid’ etc., refer to a state of affairs which is possible but not considered as real by the speaker. This is the reason EN mostly co-occurs with non-indicative moods.

Moreover, an intrinsically negative meaning of the verb can make NEG quite superfluous. Jin & Koenig (2021: 56ff.) notice that many English speakers can say ‘I miss not seeing you’ when they really mean ‘I miss seeing you’: “A speaker intends to say p , but because the meaning of a trigger [as Fr. *regretter* ‘to regret’ or Engl. *miss*: P.Rt.] strongly activates $\neg p$, $\neg p$ is produced instead”.²³

As for the non-indicative moods used with fear verbs, notice that EN has been found in Greek since Homeric times, with imperatives, interrogatives, conditionals, optatives, called using the hypernym ‘nonveridical’ or ‘antiveridical’:

A veridical operator entails the truth of p in all worlds in the model, while a nonveridical operator expresses uncertainty [...] the class of nonveridical operators is the class of antiveridical ones, among which is negation; antiveridical operators entail the falsity of p . In other words, veridical operators reflect the speaker’s certainty and commitment to the truth of the proposition which is uttered, whereas nonveridical operators reflect uncertainty and lack of commitment, Chatzopoulou (2014, § 3.1), and see Dobrushina’s statement reported above.²⁴

²³ Delfitto et al. (2019: 72) rightly quote Yoon (2011: 24), who states that EN triggers a likelihood scale based on the speaker’s presuppositions of uncertainty or unlikelihood (see also the concept of ‘veridical’ discussed immediately below). In the present context, the threefold division of fear verbs introduced by Dobrushina (2021) is not relevant. She also considers structures of fear verbs that do not show NEG. On the contrary, the main topic of my paper is precisely the presence of NEG in sentences which logically would not request the presence of NEG.

²⁴ As for the concept of ‘nonveridical’, Chatzopoulou refers, among others, to Giannakidou (1998). Exceptions such as Japanese, Korean, and Russian using the indicative mood in fear-complements are mentioned by Dobrushina (2021, § 3.1).

(49) Ancient Greek (Indo-European; Il.20, v.30)

Δεῖδω **μὴ** καὶ τεῖχος ὑπὲρ μόρον ἐξαλαπάξῃ

I.fear NEG and wall against destiny destroy:3SG.AOR.SBJ.

‘I am afraid that he will destroy the wall even against the destiny.’²⁵

And in standard Modern Greek:

(50) Modern Greek (Indo-European; Chatzopoulou 2014).

Ο Γιάννης φοβάται μὴν αρρωστήσει.

O Jánis fováte min arostísi

the Janis fear:IND.3SG NEG get.sick:3SG.PRS.SBJ

‘John is afraid that he may get sick.’

The same situation occurs in Latin:

(51) Latin (Indo-European)

Timeo ne id [...] astute fecerint

I.fear NEG it astutely do:3PL.PERF.SBJ.

‘I fear that they did (***not**) it astutely.’ (Cic. *Caec.* 4)

This type of construction with the ‘fear verbs’ (‘*verba timendi*’) was maintained in the Romance languages up to the present times (see Grieve–Smith 2009: 20):

(52) Old French (Indo-European; J. Bodel, *Jeu de St. Nicolas*, ca. 1200)

J’ai paour qu’ele ne t’escape

‘I’m afraid she will get away from you.’

(53)

a. French (Indo-European)

Jean a peur qu’elle ne soit malade

b. Italian (Indo-European)

Giovanni teme che lei non sia malata

c. Portuguese (Indo-European)

Juan teme que ela não esteja doente

‘John is afraid that she be (***not**) ill.’

²⁵ Humbert (1986, §184) calls this construct ‘*subjonctif d’appréhension*’ and gives examples from Homer, Herodotus, and Plato.

6. Expletive (?) Negation

The question mark alludes to the fact that not all (apparent) ENs are truly expletive, i.e., pleonastic or vacuous. Note first that there are contexts which implicitly invite inserting NEG: if I say ‘the bar is open from nine until as long as there are clients’ (translation of the Lithuanian example (38) in Wälchli 2018: 186), I am implicitly saying that that bar remains open until there are **no longer** clients: in Italian it would be *Il bar resta aperto finché non ci sono più clienti*, and not **...finché ci sono più clienti*; in Spanish *el bar está abierto hasta que no hayan más clientes* and not **...hasta que hayan más clientes*.

For their part, Kehayov & Boye (2016) find that in complements of verbs of fear, the Russian complementation marker system encodes the distinction between a remotely possible threat and a more real threat: *kak* refers to a feared state of affairs portrayed as only remotely possible, while *čto* conveys a real threat. Many other fine-grained observations have been made in different studies, mainly those dedicated to a single language or a restricted language family. Thus, for instance, Greco (2020: 2 and passim) has studied an unnoticed usage of EN in Italian, namely the ‘Surprise Negation Sentences’ such as:

(54) Italian (Indo-European)

E non mi è scesa dal treno Maria ?!
And NEG to.me is got off.the train Mary ?!
‘Mary (to my surprise) got off the train!’

meaning ‘Mary (to my surprise) got off the train!’, with strong pragmatic effects.

However, the crucial point is that, concerning negative comparison (Section 3), we have already seen that EN is very frequently used in comparative constructions (cf. Stassen’s statement quoted above and the examples (43)-(45)). In Old French we find (see Grieve-Smith 2009: 20):

(55) Old French (Indo-European; *Ordo Representacionis Ade*, ca.1160)

Es plus fresche que n’ est rose
‘You are fresher than a rose.’

(56) Old French (Indo-European; *Graal* 107, 25)

Et en cele biauté sans faille m'enorguilli un poi plus que je ne deusse

‘Je fus plus orgueilleuse de ma beauté que je n’aurais dû’ (Queffélec 1988: 421)

‘I have been prouder of my beauty than I should (*not) have been.’

This construction is still used in Modern French, though it is possible to cancel the *ne*:

(57) French (Indo-European)

Ce travail est plus difficile que je (n’) avais imaginé

‘This work is more difficult than I (*NEG) had thought.’

In this kind of inequality comparison, the second member (e.g., *the rose*) refers to a state of affairs which is negated (i.e., it has a nonveridical status; cf. Delfitto 2020): ‘the rose is **not** fresher than the dame’; or ‘the dame is fresh, the rose is **not**’; and similarly ‘I did **not** imagine this work be so difficult’. In this kind of comparison, NEG is not expletive/vacuous in the same sense that ‘expletive’ is used in the examples examined in the previous sections (e.g., Russ. *Poka ja ne zabył*, in (11)): ‘before I (*not) forget’. These constructs of inequality comparison may provide the reason the NEG marker may become the comparison marker tout court as in Sanskrit: *mṛtam śreyo na jīvitam* (see above, Section 4; Viti 2002). According to Wälchli (LINGTYP, 12 Jan. 2022), “typical comparative examples with ‘expletive’ negation are all comparison of inequality (which strengthens my point that ‘expletive’ negation can have something to do with non-identity)”.

In (56) from *Graal* 107, the speaker thinks for sure that ‘I should **not** have been too much proud of my beauty’. Compare this example with the following (from Jin & Koenig 2021: 61):

(58) French (Indo-European)

Je regrette qu’il ne faille souvent attendre des

I regret that.it NEG should:3SG.SBJ often wait ART.INDF

années avant que l’histoire ne juge les tyrans.

years before that the.history NEG judge:3SG.SBJ the tyrants.

‘I regret that it often should take years before history judges tyrants.’

The difference lies in the ‘nonveridical’ status of *je regrette qu’il ne*, etc., while *Graal* 107 refers to a statement (*m’énourguilli* ‘I became proud’) that reports a situation which the speaker considers to have been true in the past.

The same holds for:

(59) French (Indo-European; Jin & Koenig 2021)

Niez-vous qu’il ne soit-3SG.SBJ un grand artiste?

deny-you that.he NEG be a great artist

‘Do you deny that he is (*not) a great artist?’

The speaker is sure that the person he’s speaking about is a great artist, but he thinks that his/her conversation partner has a different opinion; therefore, he uses an interrogative construction which entails a subjunctive (*soit*), typical of a nonveridical status. The speaker intends to say *p*, but the verb ‘deny’ triggers a $\neg p$.

The conclusion is that the presence / absence of EN depends (particularly in subordinated clauses) on the semantics of the individual lexical items (cf. Jin & Koenig 2021: 59): see the different constructs of *avoir peur* (‘to be afraid’) and *souhaiter* ‘to wish’ mentioned in Section 2. Moreover, grammatical descriptions (particularly of French) do not provide a consistent definition, particularly regarding the correctness of the usages. Some grammarians consider EN as not belonging to standard French. In fact, its usages vary in spoken language and extend to various construction types. In standard, prototypical NegConc, two or more negative elements yield a single semantic negation (see Section 3), but often a NEG element can be inserted emphasizing the negativity value of the utterance. Consequently, NegConc acquires an EN as in the Russian example quoted above as (15), *Nikto ničego ne skazal* lit. **Nobody nothing not** said i.e. ‘Nobody said anything’.

As Wälchli (2018: 152) writes,

a range of similar, but different, meanings are expressed by the same form and are treated as if they were the same thing although they are slightly different meanings, [and] cross-linguistically recurrent identity of form reflects similarity in meaning.

It seems that this applies to EN as well.

Conclusion: we can imagine that EN started in (inequality) comparisons where the second member of the comparison is implicitly negated as in the Sanskrit NEG *ná* (*mṛtām śreyo ná jīvitam*) and in (55) *es plus fresche que n'est rose*; i.e., 'la rose *n'est pas* si fraîche comme tu l'es'. As already noted, Old French is particularly rich in this type (see Vendryes 1950: 5ff.):

(60) Old French (Indo-European; *Perceval* 5443)

Ele vaut miauz que vos ne faites
 'She is worthier than you (* NEG) think.'

EN extended further to the other constructs triggered by operators such as 'deny', 'unless', or 'before' that implicitly invited a (negative) comparison, as we have discussed in the previous sections, until it finally expanded to other domains such as NegConc and fear verb constructions, under the pragmatic need of the speaker to enhance the negative meaning of the sentence. Compare the Russian sentences:

(61) Russian (Indo-European; Wälchli 2018: 151)

Poka stanet temno, ja eščë porobotaju
 as.long.as/**until** become:PFV.PRS.3SG dark:N.SG I yet DLM:work:PFV
 'Until it has become dark, I'll still (manage to) work a bit.'

(62) Russian (Indo-European; Wälchli 2018: 149)

Japorobotaju, poka ne stalo temno
 I DLM:work:PFV as.long.as/**until/before** **not** become:PST.N.SG dark:N.SG
 'I'll work a bit, before it (***not**) gets dark.'

In the second sentence, NEG is semantically unnecessary since the fact that it will get dark is a matter of fact, but NEG is inserted according the 'opportunity.before' scheme. Wälchli (2018) has based his study on a temporal comparison considering how a sample of languages (mainly the Baltic languages) comes to grips with the concepts 'AS LONG AS', 'UNTIL' and 'BEFORE'. He concludes there has been a gradual expansion of NEG and EN beyond the temporal domain. My article concludes that the origin of EN lies in inequality comparisons, but we have seen above that concepts such as 'UNTIL' and 'BEFORE' implicitly entail a comparison between two states of affairs:

consider, for example, the Latin *ante quam* ('before') where *quam* is the typical comparative form:

(63) Latin (Indo-European; Cic., *Cato Maior*, 50)

Sex annis antequam ego natus sum

'Six years before I was born.'

(64) Latin (Indo-European; Cic., *Cat.*, 4,20)

nunc antequam [...] ad sententiam redeo, de me pauca dicam

'now before [...] I return to the sentence, I'll say a few things about me.'

Whether we start from the temporal dimension or the predicative one (*plus fresche que*), both underline the progressive expansion of the EN from the core concept of comparison,²⁶ where the second member is (more or less) implicitly negated:

(65) French (Indo-European)

Je suis parti avant qu'il ne soit arrivé

'I left before he arrived.'

7. Final summary

This paper maintains that various negative constructions, already alluded to in the abstract, crucially obey the psychological, and hence pragmatic, need of the speaker to enhance and intensify the negative import of the sentence. Thus, we have Negative Concord where two or more negative elements yield a single semantic negation. Consequently, some negative elements in the sentence may be redundant, 'expletive'. Expletive Negation is mostly and originally found in inequality comparisons – including temporal comparisons – where the second member of the comparison is implicitly negated. Expletive Negation is also found with fear verbs that implicitly invite a negative comparison.

²⁶ In 1974, Seuren had already proposed that comparative clauses implicitly involve semantic negation, as clearly surfaces in languages such as Nahuatl (43), Hixkaryana (44), or Telugu (45). More precisely, we have to speak of comparative clauses expressing inequality: *Mary is taller than John* = *John is not as tall as Mary*.

Acknowledgments

Many thanks are due to Giuliano Bernini, who carefully read and commented on a first draft of this article. Thanks are due also to two anonymous reviewers for their very useful suggestions. Obviously, all remaining errors are mine.²⁷

Abbreviations

1 = 1 st person	NCI(s) = Negative Concord Item(s)
3 = 3 rd person	NEG = Negation
ABS = Absolute	NEG# = NEG's final position
ACC = Accusative	NEG1 = NEG's first member
AOR = Aorist	NEG2 = NEG's second member
ART = Article	NegComp = Negative Comparison
COMPL = Complementizer	NegConc = Negative Concord
COND = Conditional	NOM = Nominative
COP = Copula	NPI(s) = Negative Polarity Item(s)
DET = Determinative	PA = Active Participle
DLM = Delimitative	PERF = Perfect
EN = Expletive Negation	PFV = Perfective
F = Feminine	PRO = Pronoun
HORT = Hortative	PROG = Progressive
IMP = Imperative	PRS = Present
IMPF = Imperfect	PST = Past
IND = Indicative	PTCP = Participle
INDF = Indefinite	RFL = Reflexive
INS = Instrumental	SBJ = Subjunctive
IPFV = Imperfective	SCL = Clitic Subject
M = Masculine	SG = Singular
N = Neuter	SUB = Co-subordinating Particle

References

Ahern, Christopher A. 2015. *Cycles and Stability in Linguistic Signaling*. University of Pennsylvania (Dissertation).

²⁷ Unfortunately, I had the opportunity to look in the important book *Intorno alla negazione. Analisi di contesti negativi dalle lingue antiche al Romanzo*. Atti della giornata di studi, Roma 26 febbraio 2009, a cura di Mauro Lasagna, Anna Orlandini, Paolo Poccetti. Pisa – Roma: Serra Editore 2012, only when this article was finished and the proofs were ready.

- Andrews, Richard. 1975. *Introduction to Classical Nahuatl*. Austin & London: Texas Univ. Press.
- Ballarè, Silvia. 2019. *La negazione di frase: forme e funzioni. Studi di caso nel dominio italo-romanzo*. Universities of Bergamo & Pavia (Doctoral Dissertation).
- Bernini, Giuliano. 1994. La negazione finale di frase in afrikaans e il suo retroterra portoghese. *Quaderni del Dipart.^{to} di Linguistica e Letterature Comparete. Univ. di Bergamo* 10. 287-325.
- Bernini, Giuliano & Paolo Ramat. 1996. *Negative Sentences in the Languages of Europe. A Typological Approach*. Berlin / New York: Mouton de Gruyter.
- Breitbarth, Anne & Christopher Lucas & David Willis (eds.). 2020. *The History of Negation in the Languages of Europe and the Mediterranean*. Volume II: *Patterns and Processes*. Oxford: Oxford University Press.
- Booij, Geert & Los Bettelou & Margit Rem. 2006. De oorsprong van *ghe-* als negatief-polar element in het Middelnederlands. *Taal en Tongval* 58. 3-23.
- Chatzopoulou, Katerina. 2014. The Greek Jespersen's Cycle: Renewal, stability and structural microelevation. In Chiara Gianollo & Agnes Jäger & Doris Penka (eds.), *Language Change at the Syntax-Semantics Interface*. 323-354. Berlin / New York: Mouton De Gruyter.
- Delfitto, Denis. 2020. Expletive Negation. In Viviane Déprez & M. Teresa Espinal (eds.), *The Oxford Handbook of Negation*. 255-268. Oxford: Oxford University Press.
- Delfitto, Denis & Chiara Melloni & Maria Vender. 2019. The (en)rich(ed) meaning of expletive negation. *Evolutionary Linguistic Theory* 1(1). 57-89.
- De Swart, Henriëtte. 2010. *Expression and interpretation of negation: An OT typology*. Berlin: Springer.
- Dobrushina, Nina. 2021. Negation in complement clauses of fear-verbs. *Functions of language* 28(2). 121-152. <https://doi.org/10.1075/fo1.18056.dob>.
- Florici, Franck. 2016. À propos de la négation holophrastique dans les langues romanes. In Emilia Hilgert & Silvia Palma & Pierre Frath & René Daval (eds.), *Négation et référence*. 377-399. Éditions et presses universitaires de Reims.
- Forest, Robert. 1993. *Négations. Essai de syntaxe et de typologie linguistique*. Collection de la Soc. de Linguist. de Paris, vol. 77. Paris: Klincksieck
- Fraenkel, Ernst. 1911. Grammatisches und Syntaktisches, IV: οὐδέ nach Komparativ im Sinne von ἤ. *Indogermanische Forschungen* 28. 236-239.
- Friedemann, Nina S. de & Carlos Patiño Rosselli. 1983. *Lengua y sociedad en el Palenque de San Basilio*. Bogotá: Instituto Caro y Cuervo.

- Garzonio, Jacopo. 2019. Negative Concord in Russian. An Overview. In Iliyana Krapova & Svetlana Nistratova & Luisa Ruvoletto (eds.), *Studi di linguistica slava. Nuove prospettive e metodologie di ricerca*. 175-190. Venezia: Ediz. Ca' Foscari.
- Garzonio, Jacopo & Cecilia Poletto. 2015. On Preverbal Negation in Sicilian and Syntactic Parasitism. *Isogloss 18* (Special Issue). 133-149.
- Gawron, Jean Mark. 1995. Comparatives, Superlatives, and Resolution. *Linguistics and Philosophy* 18. 333-380.
- Giannakidou, Anastasia. 1998. *Polarity Sensitivity as (Non)Veridical Dependency*. Amsterdam / Philadelphia: Benjamins.
- Giannakidou, Anastasia & Hedde Zeijlstra. 2017. The Landscape of Negative Dependencies: Negative Concord and N-Words. In Martin Everaert & Henk van Riemsdijk (eds.), *The Wiley Blackwell Companion to Syntax*, 2nd Edit. 1-38. Hoboken (NJ): Wiley & Sons. <https://doi.org/10.1111/1467-968X.00059>.
- Gianollo, Chiara. 2018. *Indefinites between Latin and Romance*. Oxford: Oxford University Press.
- Gianollo, Chiara. 2019. Quanta variazione è possibile nella concordanza negativa? I dati del greco classico. *CLUB. Working Papers in Linguistics* 3. 244-256.
- Gianollo, Chiara. 2021. Indefinites and negation in Ancient Greek. *Journal of Historical Syntax* 5. 1-38.
- Greco, Matteo. 2019. Is expletive negation a unitary phenomenon? *Lingue e linguaggio* 18(1). 25-58.
- Greco, Matteo. 2020. *The syntax of surprise: expletive negation and the left periphery*. Cambridge: Cambridge Scholars Publishing.
- Grieve-Smith, Angus B. 2009. *The Spread of Change in French Negation*. Albuquerque, The University of New Mexico (PhD. Dissertation).
- Holthausen, Ferdinand. 1913. Negation als Vergleichungspartikel beim Komparativ. *Indogermanische Forschungen* 32. 339f.
- Humbert, Jean. 1986. *Syntaxe grecque*. 3^{me} éd., Paris: Klincksieck.
- Jespersen, Otto. 1917. *Negation in English and other languages*. Copenhagen: I.F. Høst & So.
- Jespersen, Otto. 1922. *Language: Its Nature, Development, and Origin*. London: Allen & Unwin.
- Jin, Yanwei. 2021. *Negation on your mind: A cross-linguistic and psycholinguistic study of expletive negation*. Buffalo, NY: University at Buffalo (Dissertation).

- Jin, Yanwei & Jean-Pierre Koenig. 2021. A cross-linguistic study of expletive negation. *Linguistic Typology* 25(1). 39-78
- Iordanskaja, Lidija & Igor Mel'čuk. 2009. Semantics of the Russian conjunction *poka* 'while, before, until'. In Tilman Berger & Markus Giger & Sybille Kurt & Imke Mendoza (eds.), *Von grammatischen Kategorien und sprachlichen Weltbildern — Die Slavia von der Sprachgeschichte bis zur Politsprache. Wiener Slawistischer Almanach. Sonderband 73.* 253-262. München / Wien: Kubon & Sagner.
- Kehayov, Petar & Kasper Boye. 2016. Complementizer semantics in European languages: Overview and generalizations. In Kasper Boye & Petar Kehayov (eds.), *Complementizer Semantics in European Languages*, 793-878. Berlin / New York: Mouton de Gruyter.
- Khan, Geoffrey. 2016. *The Neo-Aramaic Dialect of the Assyrian Christians of Urmi.* Leiden/Boston: Brill.
- Klein, Wolfgang. 1980. Some remarks on Sander's typology of elliptical coordinations. *Linguistics* 18, 871-876.
- Krifka, Manfred. 2010. How to interpret "expletive" negation under *bevor* in German. In Thomas Hanneforth & Gisbert Fanselow (eds.), *Language and Logos. Studies in Theoretical and Computational Linguistics*, 214–236. Berlin: Akademie Verlag.
- Kuhn, Jeremy. 2022. The dynamics of Negative Concord, *Linguistics and Philosophy* 45. 153-198.
- Kuno, Susumu. 1973. *The Structure of the Japanese Language.* Cambridge, MA: MIT Press.
- Kuteva, Tania. 1998. On identifying an evasive gram: Action Nearly Averted. *Studies in Language* 22. 113-160.
- Larrivé, Pierre. 2011. Is there a Jespersen cycle?. In Pierre Larrivé & Richard P. Ingham (eds.), *The evolution of Negation. Beyond the Jespersen Cycle.* 165-178. Berlin / New York: Mouton de Gruyter.
- Lasagna, Mauro & Anna Orlandini & Paolo Poccetti (eds.). 2012. *Intorno alla negazione: analisi di contesti negativi, dalle lingue classiche al romanzo. Atti della giornata di studi, Roma, 26 febbraio 2009.* Pisa / Roma: Fabrizio Serra editore.
- Miestamo, Matti. 2017. Negation. In Alexandra Aikhenvald & Robert M. W. Dixon (eds.), *The Cambridge Handbook of Linguistic Typology.* 405-439.
- Moscato, Vincenzo (forthcom.). Negation and Polarity in the Romance languages. In *The Oxford Encyclopaedia of Romance Linguistics.*

- Mosegaard Hansen & Maj-Britt & Jacqueline Visconti. 2012. The evolution of negation in French and Italian: Similarities and differences. *Folia Linguistica* 46. 453-482.
- Mosegaard Hansen & Maj-Britt & Jacqueline Visconti (eds.). 2014. *The Diachrony of Negation*, Amsterdam/ Philadelphia: Benjamins.
- Moser, Ann-Marie. 2019. Form und Funktion der doppelten Negation in deutschen Dialekten, mit einem Schwerpunkt im Oberdeutschen. *Linguistik Online* 98(5). 179–195. <https://doi.org/10.13092/lo.98.5935>.
- Nocentini, Alberto. 2003. La cosiddetta negazione espletiva in italiano. *Archivio Glottologico Italiano* 88. 72-90.
- Parry, Mair. 2013. Negation in the history of Italo-Romance. In David Willis & Christopher Lucas & Anne Breitbarth (eds.), 2013. *The History of Negation in the Languages of Europe and the Mediterranean*. Volume I: *Case Studies*. 77-118.
- Poletto, Cecilia. 2008. On negative doubling. *Quaderni di lavoro ASIT* 8. 57-84.
- Posner, Rebecca. 1985. Post-verbal negation in Non-standard French: A historical and comparative view. *Romance Philology* 39. 170-197.
- Queffélec, Ambroise. 1988. La négation ‘explétive’ en ancien français: une approche psycho-mécanique. In André Joly (ed.), *La linguistique génétique. Histoire et théories*, 419-443. Lille: Presses Univ. de Lille.
- Ramat, Paolo. 2002. La comparazione negativa. *Archivio Glottologico Italiano* 87. 223-229.
- Ramat, Paolo. 2006. Italian Negatives from a typological/areal point of view. In Nicola Grandi & Gabriele Iannàccaro (eds.), *Zhì. Scritti in onore di Emanuele Banfi in occasione del suo 60° compleanno*, 355-370. Cesena / Roma: Caissa Italia editore.
- Schwegler, Armin. 1990. *Analyticity and Syntheticity. A Diachronic Perspective with Special Reference to Romance Languages*. Berlin / New York: Mouton de Gruyter.
- Schwegler, Armin. 1991a. Predicate negation in contemporary Brazilian Portuguese – A change in progress. *Orbis* 34 [1985-1987 but 1991]. 186-214.
- Schwegler, Armin. 1991b. Negation in Palenquero: synchrony. *Journal of Pidgin and Creole Languages* 6. 165-214.
- Schwenter, Scott A. 2006. Fine-tuning Jespersen’s Cycle. In Betty J. Birner & Gregory Ward (eds.), *Drawing the boundaries of meaning: Neo-Gricean studies in pragmatics and semantics in honor of Laurence R. Horn*. 327-344. Amsterdam / Philadelphia: Benjamins.

- Seuren, Pieter A. M. 1974. Negative's travels. In Pieter A. M. Seuren (ed.), *Semantic syntax*. 96–122. Oxford: Oxford University Press.
- Siller-Runggaldier, Heidi. 1985. La negazione nel ladino centrale. *Revue de Linguist. Romane* 48. 71-85.
- Stassen, Leon. 1985. *Comparison and Universal Grammar*. London: Blackwell.
- Stolz, Christel & Thomas Stolz. 2001. Hispanicized comparative constructions in indigenous languages of Austronesia and the Americas. In Klaus Zimmermann & Thomas Stolz (eds.), *Lo propio y lo ajeno en las lenguas austronésicas y amerindias*, 35-56. Frankfurt: Vervuert.
- Tagliani, Marta. 2019. The Acquisition of Double Negation in Italian. In *Language Use and Linguistic Structure. Proceedings of the Olomouc Linguistics Colloquium 2018*. 109–126. Olomouc: Palacký University.
- Tagliani, Marta & Maria Verder & Chiara Melloni. 2022. The Acquisition of Negation in Italian. *Languages* 7. 116-149.
- Valkhoff, Marius F. 1972. *New light on Afrikaans and 'Malayo-Portuguese'*. Louvain, Peeters.
- Van der Auwera, Johan. 2009. The Jespersen cycles. In Elly van Gelderen (ed.), *Cyclical change*. 35-71. Amsterdam / Philadelphia: Benjamins.
- Van der Auwera, Johan & Annemie Neuckermans. 2004. Jespersen's cycle and the interaction of predicate and quantifier negation in Flemish. In Bernd Kortmann (ed.), *Dialectology meets Typology. Dialect Grammar from a Cross-Linguistic Perspective*. 453-478. Amsterdam / Philadelphia: Benjamins.
- Van der Auwera, Johan & Lauren van Alsenoy. 2016. On the typology of negative concord. *Studies in Language* 40. 473–512
- Vendryes, Joseph. 1950. Sur la négation abusive, *Bulletin de la Soc. de Linguist. de Paris* 46. 1-18.
- Viti, Carlotta. 2002. Comparazione e individuazione: uno studio sugli equativi *ŗgvedici iva e ná*. *Archivio Glottologico Italiano* 87. 46-87.
- Wälchli, Bernhard. 2018. 'As long as', 'until' and 'before' clauses: Zooming in on linguistic diversity, *Baltic Linguistics* 9. 141-236.
- Willis, David & Christopher Lucas & Anne Breitbarth (eds.). 2013. *The History of Negation in the Languages of Europe and the Mediterranean*. Volume I: *Case Studies*. Oxford, Oxford University Press
- Yoon, Suwon. 2011. "Not" in the mood: *The syntax, semantics, and pragmatics of evaluative negation*. University of Chicago (Doctoral dissertation).

Zanuttini, Raffaella. 1997. *Negation and Clausal Structure: A Comparative Study of Romance Languages*. New York: Oxford University Press.

CONTACT

paoram@unipv.it

Person alignment in reported speech and thought: the distribution and typology of participant roles (based on six Finno-Ugric languages)

DENYS TEPTIUK

UNIVERSITY OF TARTU

Submitted: 04/05/2022 Revised version: 03/10/2022

Accepted: 06/10/2022 Published: 22/12/2022

Abstract

This paper investigates how person alignment is arranged in discourse reporting. I focus on participant roles appearing in narrated and speech events (Jakobson [1957] 1971) and how they are linguistically encoded in (re)presentations of speech and thought. Besides the (re)presentations of speech and thought attributed to other speakers, I include two other types of report: self-quotations (Reported Speaker = Reporter) and quotations with an unknown source (Reported Speaker = ?). For illustrative purposes, I use data from internet communications of six Finno-Ugric languages: Hungarian, Estonian, Finnish, Erzya, Udmurt, Komi. The results show that three types of reports exhibit idiosyncrasies regarding the participant distribution in the narrated event. These idiosyncrasies affect how the linguistic encoding of participants is arranged and how different perspectives are highlighted in reported speech and thought. In addition to two canonical perspectives, i.e. Reported Speaker's and Reporter's, there are some ambiguous cases where person marking does not index only one type of perspective. Such ambiguity is characterized by the overlap between different roles carried out by one participant or subsumption of participants from different events under one formal reference. Furthermore, ambiguous cases often contain a generic reference equally suitable for participants in the narrated and current speech event.

Keywords: reported speech; reported thought; reported evidence; person alignment; Finno-Ugric.

1. Introduction

Cross-linguistically, reported speech (RS) and reported thought (RT) (often abbreviated as RST ‘reported speech and thought’) exhibit many similarities on the level of construction. They are both formed by a binary structure: report and the unit introducing the report, i.e. *Matrix*,¹ cf. (1). In many languages, the same unit can introduce both speech and thought, making these types of report often indistinguishable without an appropriate context, as e.g. in (2) from Komi (Uralic; Russia) and (3) from Ungarinyin (Worrorran; Australia).

(1) English (Indo-European; enTenTen20)²

a. [“Maybe tomorrow...”]_{REPORT} **she said**_{MATRIX}.

b. [‘Probably just a stupid mailing list again...’]_{REPORT} **she thought**_{MATRIX}.

(2) Komi-Zyrian (Uralic; KoZSmC, komimy1)

Medvodž jurö voisny koz da požöm. Miša,
 first head:ILL come:PST:3PL fir and pine QUOT.SELF
parmayd mijan pom ni dor.
 taiga:2SG 1PL.GEN end nor edge

‘First, fir and pine came to my mind. **I said/thought**, our taiga is never-ending.’

(3) Ungarinyin (Worrorran; Spronck 2015: 71, emphasis added)

[[ngurrba nyungiminda] **ama jirri**]
 [[ngurrba nyunga₂-iy-minda] *a₁-ma-ø jirri*]
 [[hit.ITRV 3SG.F.O:1SG.S-FUT-take] 3SG.M-do-PRS M.ANAPH]

‘**He says/thinks**: “I will hit her.”’

¹ Another term used vastly in the literature is *quotative index*, coined by Güldemann (2001; also see Güldemann 2008: 11 for an overview of different terms used). Güldemann’s (2008) framework views the quotative index (alias Matrix) as an optional element in the reported discourse (alias RST) construction. In contrast, Spronck & Nikitina (2019) argue for a stable syntactic relation between Matrix and Report. They show that RST tends to preserve its syntax even in the Matrix-less RST constructions (cf. Spronck & Nikitina 2019: 126-129). In this study, I follow the latter consideration and adopt the terms Report and Matrix as they are used in Spronck & Nikitina (2019).

² See list of data sources at the end of the paper.

Despite the close connection between RS and RT, there is a semantic difference between them. Speech as a dialogic phenomenon and a tool for communication presupposes an interlocutor to whom it was/is/will be addressed. Even a monologue presupposes some addressee: either the speaker herself (e.g. thoughts uttered out loud in solitude) or an imaginary audience (e.g. in theatre) (cf. Clark 2016; also see Bakhtin 1981; Goffman 1981; Jakobson 1990). Thinking as an individual process does not require any interlocutor to be involved but has its addressee analogically to speech. Thought is always self-addressed or egocentric in its original manifestation (see Vygotsky [1934] 1986). In contrast, speech has another speaker as an addressee, although it can also be egocentric, e.g. *I said to myself*.³

Such differences in the dialogic nature of speech and thought shall be preserved and reflected in reports of these processes. This study explores how these differences are encoded linguistically. It aims to increase the understanding of speech and thought and how their formal representation is constructed in RS and RT and construed in the current speech. At the same time, differences in formal representation shall point to how RS can be distinguished from RT, especially when speech and thought can be introduced identically (see (2) and (3)).

The paper builds upon the previous observations on idiosyncrasies of RS and RT (Teptiuk forth.) and explores how differences in the dialogic nature of speech and thought affect person indexing. Other *shifters* (Jakobson [1957] 1971), i.e. temporal, modal and evidential categories changing in the report to match the current speech (see Section 2), may also index such differences. To limit the scope of this investigation, I will only concentrate on person indexing. By investigating this parameter, I aim to contribute to the typology of person alignment in reported speech (Nikitina 2012a). Since the typology in Nikitina (2012a) did not include RT, this study attempts to cover this gap and investigate RT in comparison to RS in a relatively small set of typologically similar languages.

This study pursues descriptive goals and does not provide any quantitative outcomes. Empirical data from six Finno-Ugric languages (Hungarian, Estonian, Finnish, Erzya, Udmurt, Komi) support theoretical discussion and illustrate tendencies common to these languages. The languages represent four branches (Ugric, Finnic, Mordvinic, and Permic) and three linguistic areas (Central and Northeastern Europe

³ Interestingly, constructions pointing out egocentricity of speech often introduce RT and may conventionalize as RT-introducers in self-quotations (see e.g. Teptiuk 2021a on Hungarian *mondom/mondok* 'I say').

and Russia) of the Uralic language family. The data are derived from social media corpora available online for all languages (for more details, see Section 3).

The selection of the languages is based merely on the author's familiarity with their structure and different aspects regarding RST therein. That said, they are not expected to reflect unique patterns in person alignment in RST, not found elsewhere among the world's languages. Conversely, I expect to find patterns shared by other languages, stemming from the foundational mechanisms of speech and thought reporting, and peculiarities of participant distribution therein. Having a relatively small sample of typologically similar languages shall allow comparability between them and stimulate generation of hypotheses beyond the selected sample. Furthermore, it shall be possible to point out language-specific patterns of organizing discourse participants in RST and check for possible alternative strategies among the selected sample.

The paper is organized as follows. Section 2 starts with a theoretical discussion on discourse reporting with the focus on discourse participants, perspectivization, and types of report according to their author. Method and data are introduced in Section 3. In Section 4, I illustrate the theoretical points outlined in Section 2 with the corpus data. I present a short summary of results and discuss the implications of this study in Section 5. A short conclusion closes the paper in Section 6.

2. Discourse reporting: typology and participant roles

Voloshinov's ([1931] 1973: 115) definition of reported speech states that it is "speech within speech, utterance within utterance, and at the same time speech about speech and utterance about utterance."⁴ This definition highlights the relationship between two different discourse situations where one is embedded into another. Jakobson ([1957] 1971) further developed this idea in his concept of evidential meaning. According to Jakobson (1971: 135), reported speech, among other evidential strategies,⁵ reflects the interaction between three events: Speech event (E_S), Narrated event (E_N), and Narrated Speech event (E_{NS}).⁶ Figure 1 demonstrates this interaction

⁴ Note that Voloshinov's (1973) term 'reported speech' also includes reported thought, viewed as 'inner speech.'

⁵ Cf. "[t]he speaker reports an event on the basis of someone's report (quotative, i.e. hearsay evidence), of a dream (relative evidence), of a guess (presumptive evidence), or of his own previous experience (memory evidence)" (Jakobson 1971: 135).

⁶ I use capital letters and subscripts to define the event since they are more visible in the text than small letters and superscripts in Jakobson (1971).

and highlights the borders between three events with color. E_S highlighted with red corresponds to the current speech situation. In turn, E_N highlighted with blue is the situation during which reported speech (or thought) is assumed to occur,⁷ embedded in E_S . Finally, E_{NS} highlighted with green is “the alleged source of information about the narrated event” (Jakobson 1971: 135).

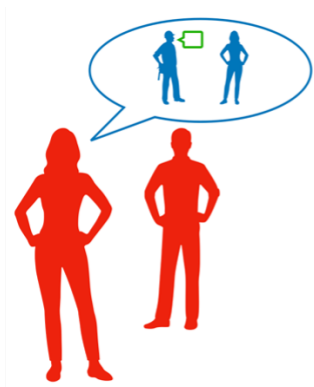


Figure 1: Visual interpretation of Jakobson’s (1971) conceptualization of the evidential meaning and reported speech [E_S – speech event (red), E_N – narrated event (blue); E_{NS} – narrated speech event (green)]

According to Jakobson (1971: 136), personal and temporal deictics, as well as modal and evidential categories may change in E_N to refer to E_S or its participants. Jakobson (1971: 136) coined the term ‘shifter’ for such elements. To illustrate such shifters, consider two types of RS in (4). In (4a), temporal (*saw*, *yesterday*) and personal (*I*) deictics correspond to E_N . In (4b), they shift to match E_S . Temporal deictics align with E_S and highlight the sequence of events: *had seen (the previous day) > said*. Furthermore, the personal deictic *he* is coreferential with Reported Speaker, John.

- (4) a. *John said: ‘I saw Fred yesterday.’*
b. *John_i said (that) he_i had seen Fred the previous day.*
(adapted from Aikhenvald 2011: 238)

Shifters are not selected arbitrarily: their presence or absence highlight difference in perspectives. Perspective, or in other words, referential orientation, is a location of the anchor for deictic and expressive elements. In (4a), all referential elements

⁷ RS and RT do not always entail factuality of their content and can demonstrate the occurrence of speech or thought in hypothetical situations.

correspond to E_N and demonstrate the perspective of Speaker in E_N , i.e. Reported Speaker. In contrast, (4b) illustrates the perspective of Speaker in E_S , i.e. Reporter.

The two perspectives are traditionally discussed in connection with two formal manifestations labeled as ‘direct’ (shifters-) and ‘indirect’ (shifters+) speech. Recent studies in languages outside Europe have shown discrepancies between the marking of tense, modality, and evidentiality, personal pronouns, honorifics, and vocatives in RST (see Spronck 2012; Evans 2013; Spronck & Nikitina 2019 and references therein). They observe that the traditional opposition between direct and indirect report involves a range of intermediate types (see e.g. Aikhenvald 2008; Evans 2013; Nikitina & Bugaeva 2021). These findings bring forth problems of the dichotomy and indicate that it does not hold for many languages outside Western Europe (see e.g. ex. (5) below). That considered, I will not use the labels ‘direct’ and ‘indirect’. Instead, I will explore the selection of shifting elements in the report as a manifestation of the two perspectives, Reporter’s vs. Reported Speaker’s.

Although the two perspectives are usually split in RST, some languages allow the combination thereof (see e.g. Aikhenvald 2008; Evans 2013; Knyazev 2022). Consider (5) from Golin (Chimbu-Wahgi; PNG) where the subject argument in RS corresponds to E_N , but the object shifts and corresponds to E_S , leading to a mixture of perspectives.⁸

(5) Golin (Chimbu-Wahgi; Evans 2013: 85)

I_x [na_y si- \emptyset_x -w-a] di-n_x-g-e
 you 1SG hit-1SG.S-RPRT-DIST say-2SG-ASS-PROX
 ‘You_x said you_x hit me_y.’ [Lit. ‘you_x “I_x hit me_y” you_x-said’]

Languages, where such mixture is not that prominent, may still exhibit it in certain pragmatic environments.⁹ For instance, in (6) from Colloquial British English Reporter presents the command initially addressed to her to Current Addressee (or Addressee in E_S).¹⁰ Mood remains unchanged, while the personal deictic shifts.

⁸ Also, note that the participants in E_N acquire the opposite roles in E_S : Addressee in E_N > Speaker/Reporter in E_S ; Speaker in E_N > Addressee in E_S .

⁹ To a certain degree, genre can also be important. Consider cases of *free indirect speech* in European literature, reflecting the mixture of two perspectives (see e.g. Pascal 1977; Roncador 1988; Vandelanotte 2021).

¹⁰ NB: Current Addressee \neq Reported Speaker.

- (6) I_{CS} rang Paul_p, and Paul_i said ‘Come_{CS} and see him_i’
(Aikhenvald 2011: 354; _{CS}: current speaker)¹¹

As becomes apparent from the illustration of different perspectives in RS, it is tightly connected to two participants: Reported Speaker (Speaker in E_N) and Reporter (Speaker in E_S). The choice between two perspectives affects how other participants are encoded linguistically but largely depends on their status in E_N and E_S .

Based on a few previous studies discussing perspectivization in RS (and less so in RT),¹² I propose the following typology of participants present in the RS situation (P_N):

- 1) *Reported Speaker*, i.e. Speaker in E_N whose utterance is reported;
- 2) *Reported Addressee*, i.e. Participant in E_N to whom the reported utterance is directed;
- 3) *Reported Interlocutor*, i.e. ‘bystander’¹³ in E_N ;
- 4) *Reported Other*, i.e. Participant absent in E_N .

Since every speech situation presupposes speaker and addressee (see Section 1), only these two participants are necessary to establish a speech situation in E_N . Other participants are optional and therefore may be absent.

As discussed in Section 1, RT differs from RS in dialogic nature. These differences shall affect the distribution of participants and partial overlap thereof in RT. Reported Speaker in RT shall coincide with Reported Addressee due to egocentricity of thoughts. Interlocutor and Other are also relevant participants in the RT situation. Speaker may think of Interlocutor(s) surrounding her in E_N and bring thoughts about participants absent in E_N . However, thought can be directly accessible only to their author (see Section 1), and it does not require an interlocutor to be involved. That said, Interlocutor and Other shall be equally excluded from the Reported Speaker’s mental space. Due to this characteristic, these participants can be viewed as indistinguishable in RT. Consequently, Interlocutor and Other shall be encoded identically as far as person indexing is concerned, unless they shift referring to participants in E_S . Otherwise, they can be distinguished when referred to with social roles (‘daughter’, ‘mom’) or proper nouns (John, Fred).

¹¹ Another case discussed by Aikhenvald (2011: 356) *Mummy says: ‘Sam_{CS}, wash_{CS=ADDRESSEE} my_{CS} hands’* reflects a toddler’s (2; 9) speech.

¹² To name just a few inspirations, see Li (1986), Aikhenvald (2008, 2011), Nikitina (2012a, 2012b).

¹³ For more details, see Goffman (1979, 1981).

The participants in E_S (P_S) can be easily derived from their counterparts in E_N :

- 1) *Current Speaker*, i.e. Reporter¹⁴ in E_S ;
- 2) *Current Addressee*, i.e. Participant in E_S to whom the report is presented;
- 3) *Current Interlocutor*, ‘bystander’ in E_S ;
- 4) *Current Other*, i.e. Participant absent in E_S .

In practice, P_S may overlap with P_N , and thus one participant will fulfill different roles in both events. For instance, Current Speaker may report something said to her and hence also be Reported Addressee. In some cases, this condition is governed on the pragmatic level, i.e. dependent on the situation and presence of the participant with a distinct role in both events (e.g. Current Speaker reports speech addressed to her). In others, such an overlap of roles is realized due to characteristic features of report type (see below) and hence can be considered semantic.

Before discussing the method and data in the following section, one more issue concerning the author of report shall be touched upon. Since there is no linguistic restriction regarding to whom RS and RT can be attributed,¹⁵ Reported Speaker can have many manifestations. Speech and thought can be attributed to other speakers specified in E_S or E_N , or to Reporter herself. Furthermore, the original author of report, as well as the time and circumstances of the utterance or thought, may remain covert.

I propose adding one more parameter to the investigation and classify RS and RT according to who Reported Speaker is. Similarly to the effect produced by the differences in the dialogic nature between RS and RT (see above), I expect this parameter to influence the distribution of participant roles and their linguistic encoding (see below). I distinguish three types of report: (i) *self-quotations* (i.e. Reported Speaker = Reporter), (ii) *quotations* (i.e. Reported Speaker \neq Reporter), and (iii) *quotations with an unknown source* (i.e. Reported Speaker = ?).¹⁶ Quotations

¹⁴ I neglect other roles that can be assigned for Speaker in E_S and focus only on the one relevant when the report is produced.

¹⁵ Cultural restrictions are, of course, possible. For instance, some cultures disfavor or even prohibit attributing thoughts to other speakers or their reports even if they somehow became available to Reporter (see e.g. Besnier 1993; Michael 2015).

¹⁶ I use the term *quotation*, typical for formal semantics (although see Clark & Gerrig 1990; Evans 2013, i.a.) to avoid unnecessary recycling of the term *reported speech/thought* or *report*. Despite using the term *quotation*, this study does not adopt formal approaches to reported speech and focuses on it from functional-typological perspective.

with an unknown source are usually discussed in linguistic literature under the label *reported evidentiality* restricted to the presentation of spoken material (and sometimes to grammatical means of expression, e.g. in Aikhenvald 2004).¹⁷ However, one can assume that thoughts like utterances can be attributed to an unknown cognizant (e.g. generic ‘people’, universal ‘all’ and existential ‘some’).

What makes quotations with an unknown source different from other reports is that they usually do not contain reference to the author, time, and circumstances of the original utterance (or thought) (Holvoet 2018: 248; also see Aikhenvald 2004, i.a.). A quotation of this type is illustrated in (7) consisting of a proverb and introduced merely by the reported evidential marker *állítólag* ‘allegedly’.

(7) Hungarian (Uralic; MNSz)

Bár *állítólag* *akit* *utálnak* *az*
although **allegedly** who:ACC hate:PRS.3PL DEM.DIST
sokáig *él.*
much:TERM live.PRS.3SG

‘Although **it is said** that those who are hated live long.’

The label *unknown* is far from ideal because sometimes these characteristics may be deliberately unspecified by Reporter. However, I stick to this label since it seems to be more accurate and covers more cases of reported evidence than the label *unspecified* does. The label *unknown* also backgrounds the possibility of deliberately leaving out the information about the author, which is often not the case. For instance, consider the impossibility of attributing proverbs to a specific speaker or time when such a folklore knowledge was initially used as a simple, witty remark, cf. (7).

The Reported Speaker-parameter is necessary for the following reason. I expect each type of report to reflect idiosyncrasies regarding the participants, which would affect their marking via person indexes. It can be assumed that Reported Speaker and Reporter coincide in self-quotations. Hence, they shall be marked identically, but as is demonstrated in Section 5, some non-Finno-Ugric languages may still employ

¹⁷ Aikhenvald (2004: 10) acknowledges the possibility of using lexical means to express evidentiality and even hypothesizes that lexical strategies are probably universal in the world’s languages compared to grammatical (arguably often unfamiliar to European languages). However, the focus of her work remains on grammatical expression of evidentiality. In contrast, this study disregards grammatical expression of reported evidentiality and focuses on lexical strategies.

different pronouns for Reported Speaker and Reporter in self-quotations. RT in self-quotations, in turn, shall also subsume Reported Addressee due to egocentricity of thoughts. In quotations, Reported Speaker shall always be distinct from Reporter, but Reporter may fulfill other roles, i.e. Reported Addressee, Interlocutor, or Other. Note that Reporter can only occupy the role of Reported Addressee when RS is presented; in RT, only Reported Speaker can be Reported Addressee due to egocentricity of thoughts. In quotation with an unknown source, Reported Speaker, Reported Addressee, and Interlocutor(s) shall remain covert due to the basic semantic characteristics of this report type (see above). Therefore, neither these participants nor their perspectives shall be reflected in this type of report. Table 1 summarizes these idiosyncrasies and their expected effect on participants. Their realization will be illustrated with the corpus data and further discussed in Section 4.

Types of report	Participants	
	RS	RT
Self-quotations:	Rep. Sp. [= Reporter] Rep. Addr. Interlocutor Other	Rep. Sp. [= Rep. Addr. = Reporter] Interlocutor Other
Quotation:	Rep. Sp. [≠ Reporter] Rep. Addr. (= Reporter) Interlocutor (= Reporter) Other (= Reporter)	Rep. Sp. [= Rep. Addr. ≠ Reporter] Interlocutor (= Reporter) Other (= Reporter)
Quotation with an unknown source:		*Rep. Sp., Addr., Interlocutor
	Other (= Reporter)	Other (= Reporter)

Table 1: Types of report and expected distribution of participants [square brackets stand for semantic and round for pragmatic features].

3. Method and data

To investigate how person alignment is arranged in quotations, self-quotations, and quotations with an unknown source, I have compiled a database of reported speech constructions in six Finno-Ugric languages: Hungarian, Estonian, Finnish, Erzya, Udmurt, and Komi. I used data from social network sites (SNS). Considering the principle “Yesterday’s discourse is tomorrow’s grammar” (see e.g. Du Bois 1985), SNS provide relevant material. It reflects the language use in dynamic synchrony and

contains features of colloquial speech and standard writing that often harmonically combine within one text (Tagliamonte & Denis 2008; Helasvuo et al. 2014). Even though such data contain some original features peculiar to a written modality in general and specific to online communications (e.g. emoticons, unstandardized shortenings, combinations of orthographic symbols), to a certain degree, language use on SNS is a written approximation of spoken language.

The data are derived from social media corpora available in open access online (see Table 2). The material used is only a data sample, meaning that the data represent a limited number of examples selected randomly from the corpora but queried with specific tactics in mind (see below). The number of strategies queried differs across languages, as the size of social media corpora does. Therefore, the number of collected examples varies from 400 to 1000 per language.

Language	Corpus	# exx.
Erzya	Erzya social media corpus (ESmC): 968k words	423
	Erzya Corpora, blogs (EC_blogs): 138k words	
Estonian	etTenTen19: 185 mil. words	620
Finnish	Internet communications corpus (IKA): 6,95 bil. words	1078
Hungarian	Hungarian National Corpus (MNSz), Personal subcorpus: 18,6 mil. words	570
Komi-Zyrian	Komi-Zyrian social media corpus (KoZSmC): 1,85 mil. words	466
Udmurt	Udmurt social media corpus (UdSmC): 2,66 mil. words	644
	Udmurt Corpora, blogs (UdC_blogs): 488k words	

Table 2: The corpora and amount of examples.

For the database, I used the typology of quotative constructions in these languages (Teptiuk 2019, 2020) to query quotations, self-quotations, and quotations with an unknown source. Table 3 presents glosses of the constructions used to introduce these types of report.¹⁸ For each type of gloss in Table 3, at least 100 examples were extracted from the corpora, when available. The strategies corresponding to the glosses and used as a query are overviewed at the beginning of Section 4 and are listed in Appendix. I also briefly address limitations concerning the tense forms in the query in Section 4.1. Accidental repetitions and random collocations used in a

¹⁸ Slash (/) stands for a disjunction of values ('or') here and elsewhere in the tables.

function other than report-introducing were excluded. That said, the number of examples for each strategy differs across languages and within one language.

	Self-quotations	Quotations ¹⁹	Quotations with unknown source
Speech:	'I said'	'(s)he/they said'	'they _{UNKNOWN} said'
Thought:	'I thought'	'(s)he/they thought'	'they _{UNKNOWN} thought'

Table 3: Types of RS and RT and constructions introducing them.

The collected examples were compiled in MS Excel database. Different reports were placed in separate sheets, additionally distinguishing RS from RT. This division resulted in 6 different sheets. The examples were manually annotated for different types of categories. As mentioned in Section 1, only person indexing is discussed in this paper; however, the database can be used in the future for other types of shifters. Attention was paid to the correspondence of person indexes to the reported (E_N) or current speech event (E_S), for which abbreviated labels 'en' and 'es' were used. The realization of this category was specified in a separate column, e.g. personal deixis: 'en', '2sg', which allowed further investigation on how such correspondence is encoded linguistically.

This is a pilot study with an overall goal to illustrate differences between types of report. The empirical data are demonstrated here to support the qualitative description and prepare a theoretical basis for further quantitative investigation. This means that any quantitative representation of the data will be postponed for the future, and some tendencies outlined here would need to be further tested with statistical methods.

The data in the paper are presented without any corrections of spelling or punctuation. Examples from Eastern Finno-Ugric languages are transcribed and illustrated without original in Cyrillic to save space. Occasional code-switches into Russian in these languages are presented in transliteration with non-italics in the example line. In the further presentation of data, where possible I will recycle the colors employed in Figure 1 and use blue color to designate P_N and red color for P_S in

¹⁹ For now, 2SG and 2PL forms are disregarded, representing a specific type of quotation where speech and thought are attributed to Current Addressee. This characteristic sets it apart from quotations attributed to Other (or more rarely Interlocutor) in E_S and would require a study of its own, where reports attributed to Current Addressee could be further explored.

glosses and occasionally in translation lines. Ambiguous cases and other highlights will be marked merely in bold. To save space, some more trivial examples will be illustrated in the text in English. These examples, however, are derived from free translations of actual examples in the database unless indicated otherwise.

4. Person alignment in different types of report according to Reported Speaker

Before discussing person alignment patterns in Section 4.2, I will briefly introduce the strategies used to query different types of report in Section 4.1. A list of these strategies can be found in Appendix.

4.1. Quotative strategies in Finno-Ugric languages and their use with different types of report

Person indexing in Finno-Ugric languages considered in this study is flagged via verbal personal endings, free pronouns, and possessive suffixes (except Estonian). Personal endings and free pronouns shall be coreferential, but the latter are not obligatory and can often be omitted. This condition mainly stems from the fact that the highest-ranking argument of the clause is usually obligatorily marked on the verb (e.g. Fin. *syö-n joka päivä* ‘eat.PRS-1SG every day’), excluding non-canonical realizations (e.g. experiencer, possessor: Fin. *minu-lla on pallo* ‘1SG-ADE be:PRS.3SG ball’ pro ‘I have a ball’). Possessive suffixes typically appear on nominal categories, nonfinite verb forms and some adpositions. In contrast to other languages with possessive suffixes, colloquial Finnish is quite relaxed about their use and more often highlights the possessor with free pronouns in the genitive case (8a) instead of using the possessive suffix with or without the pronoun in the genitive case (8b).

(8) Finnish (Uralic; IKA)

- a. ...*mun äiti ei tiedä siitä*
1SG:GEN mother NEG:3SG know.PRS.CN DEM.DIST:ELA
vieläkään.
still
‘...my mom still doesn’t know about it.’

- b. ... *niin* (*minun*) *äiti-ni* *teki* *kun* *olin*
 so 1SG:GEN mother-1SG do:PST.3SG when be:PST:1SG
pieni.
 little
 ‘...my mom did so when I was little.’

Each language investigated here has in its inventory at least one generic speech verb ‘say’ and mental verb ‘think’ (see Appendix). These verbs used in the past tense can introduce speech and thought attributed to different speakers. Finno-Ugric languages also exhibit cases of historical present tense. The majority also possess more than one past tense. For this study, only basic past tense forms were investigated unless the form is conventionally used in the present tense (e.g. Hung. *mondok* ‘I say’ pro ‘I said/thought’, cf. Teptiuk 2021a). Differences between tenses in quotative constructions (see e.g. Sakita 2002 for English) are beyond the scope of this study and shall be confronted in the future.

Where possible, I gave preference to a more colloquial variant, e.g. the contracted variant *ütsin* in colloquial Estonian instead of standard *ütlesin* (pro ‘I said’), or the contracted variant *aszonta* in Hungarian instead of standard *azt mondta* (pro ‘(s)he said’). This choice was mainly motivated by descriptive goals since only a few studies (if any) focus on these colloquial variants. By investigating these strategies, I also attempted to see if the contraction and change in the marker’s form impact its use.

For self-quotations, 1SG forms were checked; for quotations, 3SG forms were preferred over other possible manifestations (see fn. 19); for quotations with an unknown source, 3PL forms were investigated. In some cases, 3PL Reported Speakers can be identified from the context; in others, Reported Addressee can be identified. When either of these two conditions was realized, such types of report were labeled as quotations. Note that although quotations and quotations with an unknown source can be introduced formally identically, the differences regarding their participants are expected to remain. This will be one of the topics picked up in Section 4.2.

The distribution of speech and mental verbs across different report types can be uneven. For instance, Komi possesses three different mental verbs that can be roughly translated as ‘think’: *ćajtny*, *mövpavny*, and *dumajtny* (< Russian *dumat* ‘think’). I have checked all these verbs, but as is shown in Appendix, only *ćajtny* is used across all report types, while *dumajtny* appears only in self-quotations and *mövpavny* in

quotations only. The reason behind such uneven distribution of verbs is unclear for now and would require a separate investigation in the future.

Besides strategies with speech and mental verbs, Finnish and Estonian possess ‘new quotatives’ (see Buchstaller & Van Alphen 2012), consisting of the equational verb ‘be’ and etymologically non-reportative elements, e.g. Fin. *se oli niinku* ‘she was like’, Est. *ma olin mingi et* ‘I was like’ (lit. ‘I was something that’) (see Appendix; for more details on the strategies see e.g. Teptiuk 2019: Ch. 4). By default, these strategies can be used with both RS and RT, although they are attracted more to RT in self-quotations than other report types (cf. Teptiuk forth.). The current corpus investigation shows that these strategies are not used for quotations with an unknown source, and in Estonian are limited to quotations of speech while in self-quoting contexts can also introduce thought.

In addition, all languages have non-clausal units introducing RS (and more rarely also RT). For instance, self-quotative particles in Permic languages (Komi *miša* and Udmurt *pöj*) were included for the glosses ‘I said’ and ‘I thought’ in Table 3. These particles are restricted in their use to contexts where Reporter introduces her own speech or thought (see Appendix; for more details on the particles, see Teptiuk 2021a). The same holds for quotative particles introducing speech and thought belonging to both known and unknown speakers. Thus, besides lexical evidential constructions with the verb ‘say’, grammaticalized quotative/reported evidential particles available in each language were also used as a query. There are several such particles in some languages, e.g. Est. *kuuldavasti*, *väidetavalt*, both ‘allegedly’. In addition to the use of autochthonous particles, Komi, Udmurt and Erzya borrow quotative particles *mol* and *tipa* from the contact Russian language (see e.g. overviews in Teptiuk 2020, 2021b). However, the number of the borrowed particles in the corpora is limited with usually less than 100 examples in the entire corpus. They are also limited to colloquial speech and are used by the speakers who are more relaxed about borrowing functional and lexical elements from Russian and eventual code-mixing with Russian. A similar situation is observed in the distribution of new quotative constructions with the equational verb ‘be’ in Finnish and Estonian, where some strategies are more numerous than others and are typical for colloquial speech only.

To limit the scope of this investigation, I have excluded some strategies and did not use them as a query for different types of report for now. One type not included in the current investigation is turn-taking strategies highlighting the reported speaker

(e.g. Hungarian *erre ő* ‘upon this (s)he’, see Teptiuk 2019: 211-218). Another such strategy is the quotative construction consisting of non-reportative verbs, e.g. Hungarian *azzal jött/em* ‘I/she said’, lit. ‘I/she came with that’.

Search results could be extremely numerous if only the verbs are queried. To make them more effective, I added complementizers to some strategies. Unlike many European languages, Finno-Ugric are less sensitive in perspectivization when complementizers are present in Matrix. Strategies with complementizers can introduce both Reporter’s and Reported Speaker’s perspectives in colloquial speech. Thus, adding one to Matrix would not necessarily lead to over-representation one perspective over the other in the database.²⁰ As for the query, adding the complementizer increases the chances that the quotative construction would be followed by an independent sentential unit forming the report. Thus, accidental collocations²¹ were excluded to a certain extent by the query itself. Already grammaticalized quotative particles were queried on their own, although cases, where they occur as a part of more complex construction with speech or mental verbs, were also considered.

Although the queries were designed to search for different types of report, some types were more numerous than others (see Table 4). The most numerous in the database is the quotation of speech, which included besides quotations attributed to specified 3SG speakers also examples with 3PL Reported Speakers. The least numerous is the quotation of thoughts attributed to an unknown source. Interestingly, my database also contains more examples of RT in self-quotations than in quotations. The same holds for the difference between RT and RS in self-quotations. Even though it is too early to make any robust generalizations, the data hint that speakers report their own thoughts more than the thoughts of others, and the least so when the other is unknown. Although the data for the latter are limited, they still reflect some tendencies across languages (for more details, see Section 4.2.3).

²⁰ As an anonymous reviewer has pointed out, adding a complementizer could still introduce a non-categorical bias or a preference regarding the perspective. I find this point valid and acknowledge that a better tactic shall be developed to limit the number of examples in the corpora. However, the focus of this study is not on the quantitative representation of different perspectives in RST. It merely explores the person alignment in highlighting different perspectives. Therefore, I find the query tactics implemented here tolerable for collecting a more significant number of valid examples that still demonstrate different perspectives, regardless of the presence of a complementizer.

²¹ For instance, Fin. *sanoin vaan mielipiteeni* ‘I told my opinion’, *mä olin niinku se jätka* ‘I was like this dude’ (IKA).

	RS	RT
Self-quotations	684	727
Quotations	1598	387
Quotations with an unknown source	378	32

Table 4: Types of report in the database.

4.2. Person alignment in Finno-Ugric languages

Person indexing in RS and RT exhibits the following systemic tendency. The behavior of person indexes can be summarized in the view of their correspondence with E_N , E_S or lead to ambiguity. The latter means that the reference does not lead to interpretation favoring only one perspective. This characteristic mainly stems from the interaction between pragmatic conditions and the reference to concrete participant roles. The reference to concrete participants depends on the idiosyncrasies theorized for different report types in Section 2.

The possibility of mixing perspectives within one report, as in (9), has been previously discussed for Udmurt in Winkler (2011: 170). However, in my database, examples like (9) have occurred neither in Udmurt nor in other languages. The database's only cases with mixed perspectives contain multipart RST, as in (10). In such RST, each perspective is demonstrated separately and does not lead to the mixture of forms in one clause, as in (9).

(9) Udmurt (Uralic; Winkler 2011: 170; glossing and translation are modified)

Vladimir syče kurašky-sa as-s-e uli kari-ško šuyasa
 PN such beg-CV self-3SG-ACC under do-PRS.1SG COMP
malpa-m.

think-PRF.3SG

‘Vladimir_i apparently thought that he_i humbles himself_i (lit. I humble himself) by begging in such a manner.’

(10) Erzya (Uralic; ESmC)

Ťevem, keľa, velese lamot, eřavi
 work:1SG QUOT village:INE many:PL have.to:PRS.3SG
kardajse-piřese ľezdams avanstenze-ťeťanstenze...
 yard:INE-garden:INE help:INF mother:DAT.3SG-father:DAT.3SG

‘He says, **I** have a lot of work, **he** needs to help **his** parents with the yard and the garden.’

Even though technically the parts in (10) belong to one stretch of the report, when approached analytically, they shall not be considered cases of mixed perspective in the strict sense, as examples like (9) would typically be. Thus, in the further presentation of data, I will not pay special attention to such cases, and they will be discussed together with other cases where person indexes demonstrate only one perspective in a stretch of RST.

This subsection is split into three smaller subsections dedicated to each report type. In what follows, I review how these idiosyncrasies reflect on the indexing of participants and what pragmatic conditions influence the choice of one perspective over another or lead to the ambiguity.

4.2.1. Person alignment in self-quotations

In this section, I first turn to the discussion of self-quotations of speech and illustrate the peculiarities of person alignment therein with examples. Second, I discuss self-quotations of thought and illustrate the differences between the two types of report.

1SG forms in self-quotation of speech reflect ambiguity in Finno-Ugric languages. These forms simultaneously refer to Reporter as P_S and Reported Speaker as P_N . Thus, it is impossible to point out if Reporter presents the report from the standpoint of E_S or E_N solely from the linguistic encoding of this participant, cf. (11).

(11) Estonian (Uralic; etTenTen19)

... küsis une kohta ja ütsin, et
 ask:PST.3SG sleep:GEN about and say:PST:1SG COMP
 vahepeal kõnnin unes a seda lapsest
 sometimes walk:PRS.1SG sleep:INE but DEM:PRTV child:ELA
 saati.

from

‘...she asked about sleep, and I said that I sometimes walk in sleep, but (I do) it starting from childhood.’

At first glance, this observation might seem a bit trivial. However, theoretically one could expect to find a language reflecting the differentiation between two sources of consciousness, flagged with personal markers. As was pointed out by Güldemann (2008: 7), “[e]ven in self-quoting, [...] two centers of consciousness differing from each other at least on the time dimension must be recognized.” However, as becomes apparent from this investigation, such differentiation appears as a contextual implicature and, at least in Finno-Ugric languages, is not marked via personal markers.

In contrast, 1PL forms in my data can be considered primarily indicative of Reported Speaker’s perspective since they typically subsume Reported Speaker and Reported Addressee(s), as in (12a). However, certain ambiguity is observed when the 1PL form allow an inclusive generic interpretation. When such is pragmatically possible, 1PL forms may subsume Reported or Current Addressee or even simultaneously refer to both, and none of these interpretations excludes the others, as in (12b).

(12)

a. Estonian (Uralic; etTenTen19)

süis ma ütšin, ee Liigume [sic!] äki?
 then 1SG say:PST:1SG INTERJ move:NPST.1PL PTCL
 ‘then I said, ee should **we** move?’

b. Erzya (Uralic; ESmC)

(Vesť žardo-buťi moň kevksťimiž, meks, kelá ton šormadat Pazońť langa?)

Meřiň: mińek od šormadićanok uliť.
 say:PST:1SG 1PL.GEN new writer:1PL be:PRS.3PL

(Kijak ejsteděst šormadi Pazdońť?...)

‘(Once I was asked: why do you write about God?) I said: **we** have new/young writers. (Does anyone of them write about God?...)’

2SG and 2PL forms typically refer to addressees in E_N and E_S, as in (13).

(13) English (Indo-European)

a. *I said: “Don, **you** should quit smoking.”*

b. *I said to Bill earlier today that **you** should quit smoking.*

The ambiguity in using these forms can also be observed when Reported Addressee is also Current Addressee or vice versa, as in (14a). Even though the cases when 2SG and 2PL refer to Current Addressees are few in my database, interestingly, all of them exhibit such ambiguity. Thus, examples like (13b) where Current Addressee is absent in E_N and thus fulfills the role of Reported Other are lacking in my database. I will return to this type of reference once again below. In addition to Addressees, 2SG forms can be used for generic reference, as in (14b). Similarly, they lead to ambiguity since they refer to P_N and P_S equally.

(14)

a. Finnish (Uralic; IKA)

(sää sanoit et sää oot siellä.. ja pysyt etkö sanonuki)

... mut mähän sanoin et mä voin lentää
 but 1SG:PTCL say:PST:1SG COMP 1SG can:PRS.1SG fly:INF
 sun luo...

2SG:GEN to(wards)

‘(you said that you are there... and you will stay, didn’t you?) ...but I said that I can fly to **you**...’

b. Komi (Uralic; KoZSmC)

*Me na kyvlyšli, miša, on kö
 1SG PCTL respond:PST.1SG QUOT.SELF NEG.PRS:2SG COND
 udžav, olan prösta – byd lun kažitcö kužön
 work.CN live:PRS.2SG empty:ADV whole day seem:PRS.3SG long:INSTR
 da ydžydön.*

and hard:INSTR

‘I even replied, if **you** don’t work, **you** live emptily – the whole day seems long and hard.’

3SG and 3PL forms can also lead to similar ambiguity when used for generic reference (15a). In addition, ambiguity is also observed when 3SG and 3PL forms refer to Reported Other/s who are also absent in E_S (15b).

(15)

a. Finnish (Uralic; IKA)

*En sanonut et pitää olla tiedossa,
 NEG:1SG say:PST.CN COMP have.to:PRS.3SG be:INF knowledge:INE*

sanoin et MONET ketkä menee
say:PST:1SG COMP many:PL who:PL go:PRS.3SG
kihloihin niin niillä on jo
engagement:PL:ILL SO 3PL:ADE be:PRS.3SG already
häpäivä tiedossa..
wedding:day knowledge:INE

‘I didn’t say that it should be known, I said that many who get engaged, **they** know already the day of the wedding...’

b. Erzya (Uralic; ESmC)

Šekskak mon meřiř: sonze ojmeze kavtov javś.
therefore 1SG say:PST:1SG 3SG.GEN soul:3SG two:LAT divide:PST.3SG

‘Therefore, I said: **his** soul has divided into two [parts].’

Even though circumstantially Reported Other and Current Other often coincide, this should not always be the case. There are examples where these two participants fulfill different roles in E_N and E_S , respectively. For instance, there are situations where Current Other is Reported Addressee in E_N (17). Although an example of the opposite where Reported Other is Current Addressee is lacking in the database, it appears to be theoretically possible, cf. (13b). Thus, the lack of such examples in my data might be merely accidental.

The ambiguity in perspective also appears among the honorific uses of 3SG and 3PL in Hungarian when the pronoun is not explicitly expressed.²² Consider (16a) where the pronoun drop leads to two possible interpretations of the 3SG imperative form in the report, which can equally refer to Reported Addressee from the perspective of E_N or E_S . Note that the pronoun drop may cause problems in interpretation in similar contexts beyond RST. (16b) taken from the translation of the novel *Good men* (Hung. *Jó emberek*) by Arturo Pérez-Reverte demonstrates such ambiguity in a regular conversation between two characters.

²² The Hungarian honorific pronouns *Őn/Önök* are formally different from the regular pronouns *ő/ők*, but both honorific and non-honorific pronouns use 3SG/3PL morphology elsewhere.

(16) Hungarian (Uralic)

a. (MNSz)

Aszontam neki, hogy nézzen rám.
 DEF.say:PST:1SG DAT:3SG COMP look:IMP.HON/3SG DAT:1SG

‘I said to her: look at me.’ or ‘I said to her to look at me.’

b. (Pérez-Reverte 2017: 390; glossing and translation added)

– *Az erénye... – kezd bele, de aztán elakad.*
 DEF honor:HON/3SG start.PRS.3SG PRE but then freeze.PRS.3SG

– *Mi van az erényemmel?*

– ***Nem az önéről beszélek, uram. Hanem a lányoméről.***

‘– **Her/Your** honor... – he starts but then freezes.

– What’s with my honor?

– **I don’t speak about Your honor**, my lord. **But about my daughter’s.**’

In other situations, 3SG and 3PL forms are used as shifters and refer to Reported Addressee absent in E_s , similarly to the second interpretation in (16a).

(17) Udmurt (Uralic; UdSmC)

Vaşaly šuiško, soly, pöj, instagram

PN:DAT say:PRS.1SG 3SG:DAT QUOT.SELF PN

íléfonaz puktono.

phone:ILL.3SG install:PTCP.NEC

‘I said to Vasya that Instagram shall be installed on **his** telephone.’

Person alignment in self-quotations of speech is summarized in Table 5. As in Table 1, I use square brackets to mark semantic and round brackets for pragmatic relations between the participants.

Now let us turn to self-quotations of thought. This type of report exhibits different behavior of person indexing when compared to RS. First, the 1SG forms subsume not only Reported Speaker and Reporter but also Reported Addressee. This is the most apparent in the situations when unverbilized egocentric speech is introduced with a speech verb and a reflexive pronoun indicating Reported Addressee in M, as in (18).

Person indexing	Perspective	Participant role
1SG	ambiguous	Reported Speaker [= Reporter]
1PL	E _N	Reported Speaker [= Reporter] [+ Reported Addressee]
	ambiguous	Reported Speaker [= Reporter] [+ Reported/Current Addressee]
2SG	ambiguous	Generic
2SG/PL	E _N	Reported Addressee/s
	E _S	Current Addressee/s (= Reported Other/s)
	ambiguous	Reported Addressee/s [= Current Addressee/s]
3SG/PL	E _S	Reported Addressee/s [= Current Other/s]
	ambiguous	Reported Other/s [= Current Other/s]
		Generic
3SG/PL.HON	E _N	Reported Addressee/s
3SG.(+/-HON)	ambiguous	Reported Addressee [= Current Other]

Table 5: Person alignment and their relation to different perspectives in self-quotation of speech.

(18) Udmurt (Uralic; UdSmC)

Mon soleś kinoze ućkysa vdoxnovilsja,
 1SG 3SG:ABL movie:ACC.3SG watch:CV inspire:PST.M:REFL
kužym bašti no aslym šui: “ved’
 power get:PST.1SG and self:DAT:1SG say:PST.1SG PTCL.ENIM
mon no bygato ta užyn asleštym
 1SG and be.able:FUT.1SG DEM work:INSTR own:ABL.1SG
bygatonlykme vožmatyny.”
 ability:ACC.1SG show:INF

‘After watching his movie, I got inspired, gathered my forces and **said to myself**:
 “I can show my abilities with this work after all.”’

Even though the egocentricity of thoughts/unverbalized speech might not always be overtly marked in M, the situation is not different therein, and the speaker’s thoughts shall be viewed as originally egocentric (cf. Vygotsky 1986; Teptiuk forth.). Also, note that only Reported Addressee is identical with Reported Speaker and Reporter, but Current Addressee/s differ from these participants. Thus, when 2SG and 2PL forms are used as shifters, they refer to Current Addressee/s, e.g. *I thought, you were Erzya*. On other occasions, they refer to Reported Interlocutor/s, which is discussed below.

1PL forms subsume Reported Interlocutor or Current Addressee in addition to Reporter and her other roles in E_N . In a handful of examples in my database from all languages except Estonian, it was impossible to tease apart if the form subsumes P_N or P_S .²³ On the one hand, 1PL forms always involve Reporter as a part of self-reference. Thus, Reporter may refer to P_N and P_S with the same form without explicitly distinguishing them. On the other hand, RT is always egocentric and does not require other participants to be immediately present in E_N . Reporter can adapt her thoughts to E_S and involve P_S where necessary. Furthermore, some context hint at an immediate representation of thoughts in E_S . Thus, the difference between E_N and E_S can be quite insignificant. For instance, consider (19), where Reporter apparently almost instantly quotes her thoughts after they had occurred to her. The passive form often used in Finnish instead of 1PL may be interpreted as involving (i) Current Addressee and consequently demonstrating Reporter's perspective (we = 'I + you' in E_S) or (ii) Reported Other and consequently demonstrating Reported Speaker's perspective (we = 'I + s/he' in E_N).

(19) Finnish (Uralic; IKA)

(Onx sul e-mailii?)

Mä vaan ajattelin et voitais keskustella

1SG just think:PST:1SG COMP can:PASS:COND talk:INF

tästä samasta aiheesta...

DEM.PROX:ELA same:ELA topic:ELA

'(Do you have an e-mail?) I just thought that we could talk about the same topic...'

Note that among Finno-Ugric languages investigated here, only Udmurt marks inclusivity or exclusivity of addressees/interlocutors by using either reflexive pronoun

²³ Examples of 1PL lack among Estonian self-quotations of thought in the database. However, few examples in etTenTen19, searched with the query *mõtlesin et me* 'I thought that we', show that they lead to the same ambiguity. Consider the following example where 1PL refers to Reported Interlocutor who is Current Other: *Einar tegi mulle plaanikuga sellist lendu et mul siinamaani irvitus näol ma juba vahepeal mõtlesin et me lendame yakiga mitte plaaneriga* 'Einar organized such a flight with the aeroplane that I have the grimace on the face till now, I already thought that we (i.e. I + P_N [Einar]) fly with Yak and not with aeroplane.' One more case among queried shows the same, while one more indicates the ambiguity described for the second case in Table 6: 1PL refers to Reported Other who is Current Addressee.

aśmeos ‘we ourselves (+ addressee)’ or *mi* ‘1PL (-addressee)’ (cf. Winkler 2011: 69, 71; also see Norvik et al. 2022). However, in Udmurt, such forms do not occur in RT, even outside the new media genre. As for RS, they are found in the few examples in Udmurt Corpus and Udmurt social media corpus in self-quotations but are more typical for quotations. (20) illustrates the use of the inclusive form that subsumes Reported Addressee in self-quotations of speech. The same holds for other examples in (self-)quotations. It is possible to assume that on the condition of being used in (self-)quotations of thought, such a feature could have possibly hinted at whether the form involves Current Addressee or not. However, for now, this condition is only theoretically possible and is yet to be empirically proven.

(20) Udmurt (Uralic; UdC)

Šui soosly: *aśmeos* ke öm,
 say:PST.1SG 3PL:DAT ourselves PTCL.COND NEG.PST:**1PL**
kin udmurt kul'turajez konkursyn voźmatoz.
 who Udmurt culture:ACC competition:INE demonstrate:FUT:3SG
 ‘I_{SP} said to them_{ADDR}: if **we**_(SP + ADDR) didn’t (do it), who would represent Udmurt
 culture on the competition.’

Among the similarities between RS and RT, we can observe that 2SG and 3SG forms can be used to mark generic referents. Similarly, in the environment when Reported Other/s are also absent in E_s, it is impossible to distinguish if 3SG and 3PL forms refer to Reported Other(s) or Current Other(s), or if they simply overlap. Interestingly, even though RT is always egocentric and thus could be expected to use 2SG and 2PL only for Current Addressee(s), there are cases where these forms can also refer to Reported Interlocutor(s). Such cases represent what can be labeled as ‘unverbalized speech’. Unverbalized speech is formally identical to RS but pragmatically denotes RT and can be considered an intermediate category reflecting properties of both types of report. (21) from Udmurt demonstrates such type of report.

(21) Udmurt (Uralic; UdC_blogs)

(*Łukaškem dyšetišjos šory učki, učki no, kylziški, kylziški no,*)
malpaj, eee, nylaška, myn bert. Myn
 think:PST.1SG INTERJ girl:DIM go.IMP.**2SG** back go.IMP.**2SG**

aj dyšetsky no dyšetsky na.

PTCL study.IMP.2SG and study.IMP.2SG still

‘(I looked at the students gathered, looked and listened and) I thought, eee, girl, go back. Go and study more.’

Note that Reported Speaker still conceptually subsumes the role of Reported Addressee, and the 2SG forms in (21) are used to refer to Reported Interlocutor.

More canonically, Reported Interlocutors in RT are referred to with 3SG and 3PL forms that indicate that thinking happens outside of their reach, and other P_N cannot access the Reported Speaker’s thoughts. Such a reference to P_N indicates no referential shift and can be contrasted with the 2SG and 2PL forms used in RS to refer to Reported Addressees and signal the same lack of referential shift. This leads to the conclusion that certain person indexes behave oppositely in RS and RT among self-quotations. Compare the two examples in (22) where pronouns correspond to E_N.

(22) Estonian (Uralic; etTenTen19)

a. RT

(...ütles üks tüdruk, et ta “leidis hullult mugava voodi kuhu minna” ja kutsus mind katsuma-testima seda. Olin täis ja)

mõtsin mis sitta ta ajab, mis voodi,

think:PST:1SG what shit:PRTV 3SG drive:PRS.3SG what bed

mis katsumine ja läksin pitse lammutama.

what try.out:AN and go:PST:1SG shot.glass:PL.PRTV take.down:INF

‘(...one girl_i said that she_i “found crazily comfortable bed where to go” and invited me to test it. I was smashed and) I thought: what the hell is she_i talking about, what bed, what testing and went to take shots.’

b. RS

Siis tuli ja küsis, ja ma ütisin et

then come:PST.3SG and ask:PST.3SG and 1SG say:PST:1SG COMP

sinu asi vä?

2SG:GEN thing Q

‘Then she came and asked, and I said: is it your business?’

Now notice how they shift to match E_s in (23), reflecting the opposite use of pronouns as above.

(23) Estonian (Uralic; etTenTen19)

a. RT

Mõtsin, et pommitan sind ka selle
think:PST:1SG COMP bomb:PRS.1SG **2SG:PRTV** also DEM:GEN
küsimusega.
question:COM

‘I thought that I would bomb **you** too with this question.’

b. RS

...jah ma ütisin talle, et ma armastan teda...
yes 1SG say:PST:1SG 3SG:ALL COMP 1SG love:PRS.1SG **3SG:PRTV**

‘...yes, I said to him that I love **him**...’

Of course, the opposite use of personal markers mainly stems from the change of participant roles with the change of perspective. However, it is interesting that opposite forms signal (the lack of) the perspective shift in RS and RT. Such differences in person alignment can distinguish RT from RS where Matrix does not point explicitly at one type of report, as in (24). Obviously, when attempting such a distinction, one must consider the surrounding context and pragmatics of the situation in addition to the person indexing.

(24) Finnish (Uralic; IKA)

a. RT

(...Toni juoksi meidän perään ja alkoi kyseleä et mihkä ollaan menos.)

Silloin olin ihan et “Hyi kuka toiki on?!”
then be:PST:1SG totally COMP INTERJ who DEM.PROX:ADD be:PRS.**3SG**

(no sit vastattiin että kävelyllähän me...)

‘(...Toni ran after us and started asking where we were going.) Then **I was like** “Oh, who’s that?!” (but then answered that we [went] for a walk...)’

b. RS

(*Ja se niinku höpis jostain rehtorisista [sic!] kysymyksistä*)

ja mää olin ihan et “näytänks mää susta
 and 1SG be:PST:1SG totally COMP appear:PRS.1SG:Q 1SG 2SG:ELA
koululaiselta”
 school.kid:ABL

(*ja sit se ei vastannu.*)

‘(And he kinda blabbered something about rhetorical questions) and **I was like** “according to **you**, do I look like a school kid?” (and then he didn’t answer.)’

The forms used in self-quotations of thought are summarized in Table 6.

Person indexing	Perspective	Participant role
1SG	ambiguous	Reported Speaker [= Rep. Addressee = Reporter]
1PL	ambiguous	Rep. Speaker [= Rep. Addr. = Reporter] [+ Rep. Interloc. = Cur. Other] Rep. Speaker [= Rep. Addr. = Reporter] [+ Rep. Other = Cur. Addr.]
2SG	ambiguous	Generic
2SG/PL	E_N	Reported Interlocutor/s
	E_S	Current Addressee/s (= Reported Other/s)
3SG/PL	E_N	Reported Interlocutor/s
	ambiguous	Reported Other/s [= Current Other/s] Generic

Table 6: Person alignment and their relation to different perspectives in self-quotations of thought.

4.2.2. Person alignment in quotations

Person alignment in quotations shows the expected difference compared to self-quotations. Even though some roles are realized differently in quotations, there are also some similarities with self-quotations in terms of reference to the participants in E_N and E_S , and ambiguity in perspectivization.

The 1SG and 1PL forms do not subsume Reporter in E_N ; however, they do refer to Reporter when these forms shift and refer to P_S , e.g. *They said that **I** live in a dream world*. When speech is attributed to multiple speakers, the 1PL forms mark Reported Speakers, e.g.: *They said, **we** understand only some words*. In contrast, when it is

attributed to one speaker, the 1PL form can also subsume Reported Addressee(s), Interlocutor(s), or Others, as in (25).

(25) English (Indo-European; invented)

- a. *She said: let's go.*
- b. *She said: thanks for advice, but Gary and I, we should go now.*
- c. *She_i said: I will call my husband_z and tell him that we_{i+z} should go there tomorrow.*

When 1SG and 1PL forms are used as shifters, Reporter often fulfills the role of Reported Addressee in E_N. The same holds for the 2SG and 2PL forms canonically used to mark Reported Addressees in E_N, as in (26).

(26) Finnish (Uralic; IKA)

(mutta hermostuin sitten kun pentuja ei voinut mennä kattomaan,)

ja nämä ihmiset oli tyliin ota tai

and DEM.PROX.PL person:PL be:PST.3SG like take.IMP.2SG or

jätä älä jahkaile...

take.IMP.2SG NEG.IMP.2SG stall.PRS.CN

‘(but I got nervous when it was not possible to go and check puppies,) and these people were like take [it] or leave [it], don’t stall...’

Based on these observations, an interesting tendency can be drawn from the database, showing that Reporters often present speech originally addressed to them. Even though the statistical significance of this tendency shall be confronted in a separate study, to my knowledge, a similar tendency has not been previously discussed in the studies focusing on RST. It could be an interesting topic to explore further in the future if it holds for this and other colloquial genres.

1PL used as shifters typically subsume Reporter and another participant present in E_N. The two participants usually fulfill the role of Reported Addressees, but this condition appears pragmatic rather than semantic. For instance, the Reported Addressee condition is fulfilled in (27a), while (27b) demonstrates the situation when the 1PL form involves Reporter but the speech is unlikely to be addressed directly at Reporter. Instead, the 1PL form subsumes Reporter and Current Addressee/s, marked with the addressee-inclusive form *ašme* ‘our own’. Note that unlike in (20) where the

lexical cognate *aśmeos* ‘we ourselves’ is used to subsume Reported Speaker and Reported Addressees, *aśme* in (27b) is used as a shifter. Thus, contextual variation can be observed in RS when the addressee-inclusive form is used, showing that it can pragmatically subsume either Reported or Current Addressee(s).

(27)

a. Finnish (Uralic; IKA)

*Poliisit saapuivat paikalle ja sanoivat et jos
 police:PL arrive:PST:3PL place:ALL and say:PST:3PL COMP if
 meil ois ollu vyöt ni oltais kuoltu
 1PL:ADE be:COND.3SG be:PP belt:PL so be:PASS:COND die:PASS.PP
 kaikki!
 all*

‘The police arrived and said that if **we** had had belts on, **we** all would have died!’

b. Udmurt (Uralic; UdSmC)

*Juneskoys' ekspertjos šuizy, aśme deputatjos dory
 PN:ELA expert:PL say:PST.3PL our deputy:PL to:ILL
 kule važiškyny no soosen veraškyny.
 have.to:PRS.3SG turn.to:INF and 3PL:INSTR speak:INF*

‘The UNESCO experts said that one has to turn to **our** deputies and speak with them.’

Furthermore, in specific contexts, 1PL forms may subsume both Reported Speaker and Reporter. Such subsumption of P_N and P_S leads to ambiguity in perspective (28). Note that (28) also reflects the tendency where reported information was initially addressed to Reporter. Nonetheless, the difference between the two perspectives is not highlighted in the person marking of participants.

(28) Udmurt (Uralic; UdSmC)

*Lyktysa todi, čto Buranovoje, pe, myniškom.
 come:CV know:PST.1SG COMP PN:ILL QUOT go:PRS:1PL*

‘When we were arriving, I got to know that **we** are going to Buranovo, as they said.’

The 2SG forms can be used for generic reference, similarly to self-quotations. More typically, they mark Current or Reported Addressee, depending on whether they shift or not, e.g. *She asked who are you?* / *who you were* (invented). The same holds for 2PL forms, referring to multiple addressees, Reported or Current. As for 3SG and 3PL forms, they are used as shifters and refer to multiple Reported Speakers when speech is attributed to such, e.g. *They_i said they_i would come* (invented). Alternatively, when it is attributed to a single speaker, 3PL forms may subsume Reported Speaker and Reported Other, as in (29).

(29) Estonian (Uralic; etTenTen19)

...sugulane töötab toyotas ja ütles, et
 relative work:PRS.3SG PN:INE and say:PST.3SG COMP
 neil neli kasutatud volvot ka sees kasutatud
3PL:ADE four used PN:PRTV also inside used
 autodes.
 car:PL:INE

‘...[my] relative works in Toyota and said that **they** have 4 used Volvos among the used cars.’

3SG and 3PL forms can also be used for generic reference (e.g. *he said, one has to paint it fast*) or refer to Reported Other(s) (e.g. *The commentator said (that) she’s Erzya*). When such a reference happens, they do not pertain to one perspective and can be considered ambiguous. The honorific forms in Hungarian exhibit the same tendencies as in self-quotations when free pronouns are elliptic (see e.g. ex. (16)). Otherwise, they refer to Reported Addressees in E_N, similarly to non-honorific 2SG and 2PL forms, or honorific 2PL forms in other languages. Table 7 provides a summary of person alignment in quotation of speech.

Person indexing	Perspective	Participant role
1SG	E _N	Reported Speaker
	E _S	Reporter (= Reported Addressee)
1PL	E _N	Rep. Speakers
		Rep. Speaker + Rep. Other/Interlocutor/Addressee (= Reporter)
	E _S	Reporter + Rep. Interlocutor (= Rep. Addressees)
		Reporter + Current Addressee/s
ambiguous		Generic

Reported Speaker + Reporter		
2SG	E _N	Reported Addressee (= Reporter)
	E _S	Current Addressee (= Reported Other)
	ambiguous	Generic
2PL	E _N	Reported Addressees (+ Reporter)
	E _S	Current Addressees (= Reported Others)
3SG	E _S	Reported Speaker [= Current Other]
	ambiguous	Reported Other [= Current Other]
		Generic
3PL	E _S	Reported Speakers [= Current Others]
		Reported Speaker + Rep. Other
	ambiguous	Reported Others [= Current Others]
		Generic
3SG/PL.HON	E _N	Reported Addressee/s
3SG/PL(. + /-HON)	ambiguous	Reported Addressee/s [= Current Other/s]

Table 7: Person alignment and their relation to different perspectives in quotation of speech.

Now let us turn to person alignment in quotations of thought. In general, one may notice quite a few similarities with person alignment between RS and RT. Since we are dealing with RT, obviously, we find some differences in the distribution of roles. For instance, Reported Speaker automatically becomes Reported Addressee of her own thoughts, even though practically such thoughts are always produced by Reporter and are only attributed to Reported Speaker (cf. Teptiuk forth.).

When 1SG and 1PL pronouns shift they refer to Reporter, and Reporter plus Current Addressee, respectively. Interestingly, 1PL forms referring to P_N do not subsume Reporter in my data, although it could have been practically possible if Reporter had carried the role of Reported Interlocutor in E_N. Occasionally, 1PL is used to refer to both Reported Speaker and Reporter. Such reference leads to ambiguity in perspective, and it is impossible to distinguish if Reporter presents E_N from Reported Speaker's or her own perspective. Consider (30), where the passive form is again used instead of 1PL in colloquial Finnish and subsumes both Reported Speaker and Reporter.

(30) Finnish (Uralic; IKA)

(Tavallaan olin jo jättämässä mun nykyisen ja uus jätkä "odotti" mua, oli jo odottanut yli vuoden ja totta kai oli nyt hiton onnellinen)

kun ajatteli et viimein oltais yhessä...
 when think:PST.3SG COMP finally be:PASS:COND together
 ‘(In a way I_i was already about to leave my current [boyfriend] and a new guy_z
 “waited” for me, he_z was already waiting for more than a year and of course he_z
 was damn happy) since he_z thought that finally we_{i+z} could be together...’

Similarly to self-quotations, there is a small number of RT-constructions in quotations consisting of unverballed speech. They can be distinguished from other cases of RT by the 2SG forms referring to Reported Interlocutor, as in (31).

(31) Estonian (Uralic; etTenTen19)

(– *Kurat, sul ju põrand verd täis!*)

“Ahhh, mine persse, tead!” – mõtles Ene omaette...

INTERJ go.IMP.2SG ass:ILL know:PRS.2SG think:PST.3SG PN in.privacy

‘(– Goddammit, your floor is full of blood!) “Ahhh, **go to hell**, you know!” – silently thought Ene...’

In other cases, 2SG forms are used as shifters and refer to Current Addressee (e.g. *she probably thought that **you** are too young*) or reflect ambiguity when used for the generic reference (e.g. ***you** won’t step into the same river twice, the guy thought*).

Interestingly, 2PL forms are few in my database for quotations of thought and are used only as shifters subsuming Reported Speaker and Current Addressee, e.g. *or he_i thought (that) what if someone among **your** common friends had seen him_i there*. Furthermore, one formal reference missing in my database for RT in quotations is the use of honorific 3SG and 3PL forms in Hungarian. The same is true for self-quotations of thought (see Table 6). In general, it is not that surprising. In practice, such reference can be expected only in situations when unverballed speech is presented. However, since thoughts remain silent to the interlocutors, the use of honorifics in such cases can be considered redundant. Furthermore, there might be a reason why such speech remained unverballed. In my database, this subtype of RT often conveys a negative attitude that Reported Speaker expresses towards Reported Interlocutor, as e.g. in (31). Such an attitude conflicts with the meaning honorific forms tend to convey, e.g. respect, courtesy, esteem. Therefore, it goes without surprise that what remains

unverbalized hardly ever contains an honorific attitude towards Reported Interlocutor and consequently explains the lack of honorific forms.²⁴

3SG and 3PL forms used as shifters correspond to their use in quotations of speech and refer to Reported Speaker, e.g. *she_i thought that **she_i** would wait a bit*. Also, cases of ambiguity can be observed when 3SG and 3PL refer to Reported Others (e.g. *he thought that **these cops** won't catch (him on) a new Mercedes*), or they are used for a generic reference (*the rascal thought that **no-one of landlords** pays taxes as a rule*). As noted for quotations of speech, 3PL forms can subsume several Reported Speakers (e.g. *they_i thought that **they_i** could get a dog themselves*) or refer to Reported Speaker and Reported Other (e.g. *he_i thought that **they_{i+z}** had once again established themselves_{i+z} permanently in power*). In addition to these uses, 3SG forms may also refer to Reported Interlocutor. Such use was already described for self-quotations of thought (cf. exx. (22a), (24a)). (32) demonstrates a similar use in quotations.

(32) Finnish (Uralic; IKA)

(– *vitsit sä oot kyllä laihtunu! se sano ja katto ku ihaillen.*)

*varmaan ajatteli et kerranki **se** **läski***

surely think:PST.3SG COMP for.once DEM.DIST fatso

on saanu jotain aikaseks.

be:PRS.3SG get:PP something.PRTV early:TRANSL

‘(– are you kidding, you have slimmed very much! he said and looked as if surprised.) surely he thought (that) finally **this fatso has** accomplished something.’

Note that E_N in (32) demonstrates the dialogue between Reported Speaker and Reporter, the latter of which fulfills the role of Reported Interlocutor in E_N. Thus, in addition to Reported Interlocutor, the 3SG forms in quotations can be used by Reporter for self-reference, creating an extra distancing and impartial effect on the presentation of someone else's thought (or speech). A hypothetical assumption about other people's thoughts, especially portraying them in a negative light as in (32), may put extra pressure on Reporter. Hence, an additional requirement may arise to create extra

²⁴ Counterexamples to this tendency can sometimes occur. See e.g. ex. (15b) from Komi in Teptiuk (2021a: 223-224), containing an honorific reference in the report of unverbalized speech: ‘My tongue was very itching to answer that **You** yourself appointed and we announced. But I didn't manage (to answer).’

distancing on such occasions, leading to such an atypical self-reference. These and other types of person alignment discussed above for quotations of thought are summarized in Table 8.

Person indexing	Perspective	Participant role
1SG	E_N	Reported Speaker [= Reported Addressee]
	E_S	Reporter
1PL	E_N	Rep. Speakers Rep. Speaker + Rep. Other/Interlocutor (\neq Reporter)
	E_S	Reporter + Current Interlocutor
	ambiguous	Reported Speaker + Reporter
2SG	E_N	Reported Interlocutor
	E_S	Current Addressee (= Reported Interlocutor/Other)
	ambiguous	Generic
2PL	E_S	Current Addressee (= Reported Other) + Reported Speaker
3SG	E_N	Reported Interlocutor (= Reporter)
	E_S	Reported Speaker [= Current Other]
	ambiguous	Reported Other [= Current Other] Generic
3PL	E_S	Reported Speakers [= Current Others] Reported Speaker [= Current Other] + Rep. Other
	ambiguous	Reported Others [= Current Others] Generic

Table 8: Person alignment and their relation to different perspectives in quotation of thought.

4.2.3. Person alignment in quotations with an unknown source

Compared to quotations and self-quotations, quotations with an unknown source exhibit noticeable systemic differences in person alignment. Although one may still find similar relation between the person marking and participant roles as in quotations and self-quotations, some types of perspective are not reflected in quotations with an unknown source.

None of the forms refers to the participants in E_N . As was expected in Section 2, this happens because the whole situation behind the report and its original participants remain unspecified by Reporter if not totally unknown. As follows, Reporter cannot tie the report to the perspective of P_N . In contrast, it is possible to

adapt the report to E_s, and therefore 1SG and 1PL, as well as 2SG and 2PL forms can be used as shifters. (33) and (34) demonstrate such uses.

(33) Udmurt (Uralic; UdC_blogs)

A, valamon, šuízy val, odig nylkyšno mi
 INTERJ understand:PTCP say:PST.3PL PST.AUX one woman **1PL**
pölyn šékyten šuysa.
 among pregnant COMP
 ‘Ah, understood, it was said that one woman among **us** is pregnant.’

(34) Finnish (Uralic; IKA)

Te ette kuulemma osaa käyttää kytä...
2PL NEG:**2PL** QUOT know:PRS.CN use:INF clutch:PRTV
 ‘Allegedly, **you** don’t know how to use the clutch...’

Similarly to quotations and self-quotations, 1PL, 2SG, 3SG, and 3PL forms are also used for generic reference. Technically, they exhibit ambiguity in terms of perspective, even though there is no possibility to tie them to any P_N, e.g. *it is said that **one shouldn’t plant potatoes in the same spot for a couple of years.*** The system of person marking in quotations of speech with an unknown source is summarized in Table 9.

Person indexing	Perspective	Participant role
1SG	E _s	Reporter
1PL	E _s	Reporter + Reported Other Reporter + Current Addressee(s)
	ambiguous	Generic
2SG	E _s	Current Addressee [= Reported Other]
	ambiguous	Generic
2PL	E _s	Current Addressees [= Reported Others]
3SG/PL	ambiguous	Reported Other/s [= Current Other/s] Generic

Table 9: Person alignment and their relation to different perspectives in quotations of speech with an unknown source.

Now let's turn to RT in quotations with an unknown source. As briefly mentioned at the beginning of this section, such cases are few in the database. Therefore, I can discuss only some participant roles and their linguistic coding. All possibilities summarized below in Table 10 correspond to the options observed for RS in quotations with an unknown source. Interestingly, my database does not contain cases where quotations of thought with an unknown source involve or relate to Current Addressee/s, e.g. *some/people thought that you were a snob* (invented). This lack of data might be merely accidental; another possibility suggests a cross-linguistic tendency where Reporters tend to report thoughts about themselves, as in (35), but, for some reason, prefer not to do so regarding P_S if thoughts are not attributed to a concrete source. However, a further investigation of this tendency is beyond the scope of the current study.

(35) Hungarian (Uralic; MNSz)

Tudom, sokan gondolták, hogy én a
 know:PRS.1SG.DEF many think:PST:3PL.DEF COMP 1SG DEF
MIÉP fizetett alkalmazottja vagyok...
 PN pay:PP employ:PP:3SG be.PRS:1SG

'I know that a great many people thought that I am a paid employee of MIÉP²⁵...'

The lack of reference to P_N is explained with the same considerations as for quotations of speech with an unknown source. Namely, the situation when the report has occurred originally, and its original author and addressees are unknown (see Section 2). Consequently, the reference to P_N remains irrelevant if not impossible when this report is produced. Table 10 summarizes the observed possibilities for person alignment in quotations of thought with an unknown source.

Person indexing	Perspective	Participant role
1SG	E _S	Reporter
1PL	ambiguous	Generic
3SG/PL	ambiguous	Reported Other/s [= Current Other/s] Generic

Table 10: Person alignment and their relation to different perspectives in quotations of thought with an unknown source.

²⁵ *Magyar Igazság és Élet Pártja* 'The Hungarian Justice and Life Party' or *MIÉP* was a right-wing political party in Hungary.

5. Summary and discussion

This section summarizes results for Finno-Ugric languages discussed here and put them into a broader context without keeping strict boundaries between summary and discussion. I will focus on the effect of the reference to participants on perspectivization in discourse reporting.

Based on the empirical data illustrated in Section 4.2, one can conclude that the situation appears to be more complex than theorized in Section 2. Among Finno-Ugric languages, any (re)presentation of speech and thought is indeed connected to the two perspectives, i.e. Reported Speaker's and Reporter's. However, we also find many ambiguous cases in addition to the two perspectives. Even though these results are drawn based on only a handful of Finno-Ugric languages, they could be extended to other languages keeping the two perspectives canonically distinct in RST and not exhibiting specific cultural restrictions on reports of speech and thought.

The Reported Speaker's perspective brings forth P_N . Such perspectivization is available only for quotations and self-quotations and excludes the third type of report, i.e. quotations where the source (i.e. Reported Speaker) remains at best unspecified if not completely unknown. What characterizes such type of perspective is that although P_N may fulfill some role in E_S , Reporter chooses to neglect them. This also holds for the cases when such a participant is Reporter. Thus, Reporter becomes impartial in speech and thought (re)presentation. This is also true for the marking of other participants in self-quotations. Thus, when other participants are displayed from the perspective of E_N in self-quotations, Reporter basically neglects her role in E_S and reports it from the Reported Speaker's perspective, i.e. the earlier/different self.

The Reporter's impartiality in self-quotations can indicate the differentiation of two sources of consciousness in this type of report, briefly touched upon by Gldemann (2008: 7). Namely, Gldemann (2008: 7) indicates that even in self-quotations, one can still differentiate two speakers/cognizants at least on the time dimension, i.e. 'I-now' (Reporter) and 'I-then' (Reported Speaker). Although such differentiation has not been found in the marking of Reporter/Reported Speaker in self-quotations throughout this investigation (see below), self-quotations can still indicate differences in two perspectives by highlighting the roles that other participants occupy in E_N .

In contrast, the Reporter's perspective is highlighted when she chooses to neglect the perspective of Reported Speaker. This type can be observed not only in quotations and self-quotations, but also in quotations with an unknown source. As a matter of

fact, only this type of perspective remains for quotations with an unknown source; otherwise, the perspective is ambiguous or rather unassigned.

Different reasons stand behind the selection of Reporter's perspective over the Reported Speaker's. In some cases, the Reporter's perspective is motivated by the absence of P_N in E_S . For instance, when this condition holds, the Reporter's perspective can be observed in the marking of Reported Speaker in quotations or Reported Addressee in self-quotations. In other cases, it is motivated by the information in the report. An important characteristic leading to the choice of this perspective over the possibility of staying impartial is the report containing information about Reporter. This is crucial for quotations with an unknown source, allowing only this type of perspectivization. As is mentioned above, in quotations and self-quotations Reporter may also choose to be impartial. Impartiality in the presentation of the report also interacts with other characteristics discussed for so-called 'direct' modes of presentation, e.g. vividness, dramatization, involvement (cf. Wierzbicka 1974; Li 1986). Therefore, the report containing information about Reporter is not always presented from the Reporter's perspective.

However, when the report contains information about Current Addressee, Reporter always selects her own perspective. On the one hand, this selection is motivated by discourse conditions in E_S : Reporter delivers her own or someone else's speech or thought to no one else but Current Addressee. On the other hand, this might also be influenced by Current Addressee's role in E_N . In all instances among quotations of speech, Current Addressee is absent in E_N , and only in a few examples of quotations of thought, she is Reported Interlocutor in E_N . Regarding the latter, even the direct involvement of Current Addressee in E_N does not seem to override the requirement for Reporter to stick to her own perspective instead of turning to the Reported Speaker's. Thus, Current Addressee's role in E_N might have little involvement in the choice of perspective after all. Nonetheless, the frequent absence of Current Addressee in E_N is interesting and somewhat characterizes reports about this participant.

Among ambiguous cases, four different scenarios can be pointed out. The first scenario is characterized by the presence of the participant with distinct roles in E_N and E_S . These roles appear to be technically the same but realized in different events. This characteristic can directly affect the core participants like Reporter and Reported Speaker. For instance, this realization can be observed on the semantic level in self-quotations since Reporter is also Reported Speaker. I have not found any differences in marking these participants in self-quotations, which would highlight differences

between the two sources of consciousness briefly mentioned above. Outside the core participants, the first scenario is observed when Reported Addressee is also Current Addressee. Since these roles are identically marked in E_N and E_S , it is impossible to distinguish the two perspectives that blend in one report, at least on the level of person marking. Finally, the same ambiguity arises when Other/s in E_N is/are equally absent in E_S . No matter who Reporter or Reported Speaker is, the absence of such participants in both events leads to identical person marking across different report types.

In contrast to the first scenario, the second scenario involves two distinct participants, P_N and P_S , referred to with the same forms. Only one such case was observed in quotations when 1PL forms refer to Reporter and Reported Speaker. Since two core participants are subsumed under one form, such personal reference blends the two perspectives in one report. However, it does not mean that other shifters cannot interact with perspectivization. Thus, other markers may still be used to highlight one perspective over another.

The third scenario involves the use of person marking for a generic reference. Dahl (2000) has observed that the Swedish generic pronoun *man* behaves similarly to other egophoric pronouns used to encode speech act participants, i.e. P_S in RST-constructions. The use of generic reference in different report types shows that they may involve P_S and P_N equally. Outside RST, generic 2SG is quite typical for Estonian (Lindström et al. 2020, 2022) and Finnish (Suomalainen 2020), 3SG marked on the verb with null arguments are occurring in Finnish (Kaiser 2015), Estonian (Lindström et al. 2022) and Hungarian, the latter of which also uses 3PL (Dalmi 2022). Similarly to Hungarian, Eastern Finno-Ugric languages seem to use 3SG and 3PL forms (Gulyás 2019) in addition to 2SG typical for Russian (Leinonen 1983), dominant in the region. In addition to those markers, I have observed a certain amount of 1PL forms used for a generic reference in my data. For instance, 1PL forms may be contextually vague in self-quotations of speech and include besides Reported Speaker [= Reporter] also Reported or Current Addressees, since the form subsumes a more generic group of people like ethnos or humanity in general. The question of which interpretation is more accurate remains open since reported information with generic reference equally applies to P_N and P_S . The same holds for 3SG forms creating generic meaning in necessity constructions (e.g. ‘one cannot do so’) or 3PL arguments referring to a non-specific inclusive group of people (e.g. ‘the Erzya people’, ‘everyone’).

The fourth scenario is quite close to the 1PL examples discussed in the previous paragraph since it reflects a contextually caused ambiguity when formal reference appears to be somewhat incomplete to achieve a specific interpretation. One such case involves honorific forms in Hungarian when Reported Addressee is absent in E_S , and the pronoun is elided from the report. Since formally such interpretation allows viewing the report as containing shifters or lacking them, the assignment of one perspective is problematic. Another such case is when 1PL forms are used in self-quotations of thought. Besides Reporter who automatically fulfills the role of Reported Speaker/Addressee, such forms may subsume two types of participants. Notably, these participants fulfill different roles in E_N and E_S . The first type involves Reported Interlocutor absent in E_S ; the second type – Current Addressee absent in E_N . Even though these participants fulfill distinct roles in two events, it is impossible to say if Reporter refers to P_N or P_S when she uses 1PL forms.

The lack of differentiation in marking Reporter and Reported Speaker in self-quotations discussed for the first scenario is not restricted to Finno-Ugric languages investigated here and reflects a cross-linguistic tendency. I have conducted a short investigation of the available literature on discourse reporting and raised a query in the LINGTYP mailing list (21.01.2022). According to what I have managed to find out, we do not (yet) know about languages that would pinpoint the following difference in self-quotations: ‘I-now’ as Reporter vs. ‘I-then’ as Reported Speaker, unless sociopragmatic conditions require it. Such a sociopragmatic requirement can be seen in (36) from Pontianak Malay, an example kindly provided to me by David Gil via the list (21.01.2022):

(36) Pontianak Malay (Austronesian; MPIEVA Jakarta Field Station Corpus)²⁶

aku tanya? kalɔʔpake? bəs saya maɔ? sɛwə mɔbil kamu tu.
 1SG ask TOP use bus 1SG want rent car 2SG DEM.DIST
 ‘I (*aku*) asked whether, instead of taking a bus, I (*saya*) could rent your car.’

According to David Gil’s comment, “the speaker uses *aku* in the main clause when talking to a friend, but *saya* in the embedded clause, in which the reported situation is a more impersonal one involving a commercial transaction” (Gil p.c.). David Gil’s explanation, however, states that “the relevant factor governing the choice of 1SG pronoun is not reported speech per se, but rather the different politeness conditions

²⁶ Glossing and translation by David Gil, emphasis added.

associated respectively with the main and embedded clauses [here: Matrix and Report, DT].” Quite a few (South)east Asian languages may exhibit similar cases. However, those cases would not be restricted to discourse reporting and would largely depend on similar sociopragmatic conditions.²⁷

Theoretically, one could also think of a language that oppositely to the scenario ‘I-now’ vs. ‘I-then’ would systematically encode co-referentiality between Reporter and Reported Speaker in self-quotations by using dedicated logophoric markers. Many studies focusing on the phenomenon of logophoricity specify that it almost never concerns first-person referents; less rarely but still it can be applied to the second person (see e.g. Comrie & Hyman 1981; Roncador 1992; Nikitina 2012b). After browsing the literature, I have found only two African languages using dedicated logophoric forms in self-quotation: Ngbaka-Ma’bo (Ubangi; CAR, DRC) and Gokana (Ogoni; Nigeria). As discussed in Roncador (1988: 166), Ngbaka-Ma’bo remains an open case since the data in Thomas (1963), illustrating such use, allows various interpretations. For Gokana, Comrie & Hyman (1981: 23) specify that the logophoric suffix on the main verb in self-quotations is “superfluous and dispreferred”, since “it is *not possible* to get the two first-person singular pronouns to be non-coreferential” (Comrie & Hyman 1981: 23, emphasis added).

A case where the logophoric pronoun is used in self-quotations is also illustrated in Nikitina (2020: 90) for another African language Wan (Mande; Ivory Coast). However, it would rather qualify as an example where (socio)pragmatic factors affect the appearance of logophoric pronoun in self-quotation. In that example, the reporter uses 2SG pronoun for self-reference, which creates extra distancing. Hence, the logophoric pronoun is used to signal co-referentiality between the 2SG argument in Matrix and the participant in Report.

Different possibilities for logophoric modes of organizing discourse reporting in Finno-Ugric languages are yet to be extensively studied, especially for self-quotations. Nevertheless, I have noticed an interesting tendency while compiling a database. All languages investigated here to a different degree may signal such co-referentiality in

²⁷ In response to my query, I was also suggested to check different strategies of gender indexing. Although social roles and sociopragmatic conditions do influence gender indexing in some languages (cf. Rose 2013), such languages do not use two 1st person pronouns that would differ depending on the gender of the addressee. Although it remains a theoretical possibility, an updated database of genderlects so far shows a lack of a similar system (Rose p.c.). Otherwise, a language containing such a system could show the use of two different pronouns in self-quotations, similarly to (36).

quotations by eliding free pronouns from the report, e.g. '(s)he_i said, Ø_i comes'. However, such 'logophoric strategies' are not used systematically. Furthermore, all languages also exhibit opposite cases where co-referentiality is marked by the presence of pronouns, e.g. '(s)he_i said, (s)he_i comes'. Thus, some other factors may affect the presence or absence of free pronouns co-referential with the main argument in Matrix. A similar situation is observed in self-quotations. Thus, until there is evidence to the contrary, Reported Speaker and Reporter two participants acquire identical marking in self-quotation cross-linguistically, at least as far as person indexing of these two participants is concerned.²⁸

Although there are several ambiguous scenarios in Finno-Ugric languages, they do not seem to create any discursive problems. Even if various interpretations are possible, they do not seem to be crucial for the success of the narrative or communicative act involving different types of reports. The distribution of participants may also appear to be mostly theoretically complex but would not create practical complications for the successful construal of their formal marking in the context. However, this issue shall be separately studied using experimental methods.

There is also a system behind the complexity. The reasons behind the selection of different perspectives in RST are numerous, but they have a limited number of realizations. Person alignment interacts with the distribution of participants and leads to assigning different roles to forms used in them. When scrutinized, idiosyncrasies observed among different report types stem from the semantics of these types and the limited set of possible participants.

²⁸ Note that this remark is mainly relevant for pronominal marking. Different verbal indexing in self-quotations and clauses introducing them was mainly beyond this short investigation's scope and would require a separate study. Another point requiring a separate check is the use of other reference types for such a differentiation. For instance, in some East Asian languages, it is common to refer to speech act participants by the person's social role. Beyond this region, proper names can also be used. I owe the knowledge about the former to Pavel Ozerov (p.c.), and I am grateful to him for the latter remark about proper names. Also see Haiman (1995) discussing English where third-person construals are sometimes used in viewing the first person. Such cases shall be further investigated if they are used to differentiate two sources of consciousness and if some language has already established a system based on these distinctions in self-quotations.

6. Conclusion

This paper dealt with the person alignment in different types of reports. Data from internet communications of six Finno-Ugric languages were used to illustrate person alignment in six report types. In contrast to previous studies mainly focusing on representations of speech belonging to a speaker different from Reporter, I have included three types of speech and thought reports according to Reported Speaker: self-quotations, quotations, and quotations with an unknown source. I have shown that person alignment largely depends on the distribution of participant roles, and the report types behave differently in marking core participants. Among other participant types, they often show similarities, but again largely depend on the distribution of participant roles in the narrated and current speech event. The overlap of participants in E_N and E_S , as well as common reference to P_N and P_S with one form often cause ambiguity in perspectivization, which is otherwise connected to the perspective of Reporter as P_S and Reported Speaker as P_N .

Some results and ideas provided here need to be further explored, as well as the statistical significance of some observations need to be confirmed in the future. Although theoretical implications made here are meant to be extended to other languages beyond the six Finno-Ugric, the phenomenon of discourse reporting shall be further studied in individual languages to confirm or disclaim their universality. However, I hope I managed to convince the reader in need to further investigate discourse reporting in its complexity and look at the phenomenon of discourse reporting beyond a mere (re)presentation of speech attributed to other speakers.

Acknowledgements

This work was supported by the Estonian Research Council grant PRG1290. I am greatly thankful to the Kone Foundation for supporting my stay at the Helsinki Collegium for Advanced Studies in 2021, during which many ideas outlined in this paper were formulated, and the compilation of the database of reported speech and thought constructions was initiated. I am grateful to the colleagues at the Collegium and the University of Helsinki for the possibility of discussing some earlier ideas that lead to this publication. I want to thank all responders to my query in the LINGTYP mailing list and Marili Tomingas for taking the time to read and comment on the

original draft. I am enormously obliged to the two anonymous reviewers for their fruitful comments and suggestions. The responsibility for all remaining shortcomings is entirely mine.

Abbreviations

1 = 1 st person	ENIM = enimitive	PP = past participle
2 = 2 nd person	E _{NS} = narrated speech event	PRE = preverb
3 = 3 rd person	E _S = speech event	PRF = perfect
ABL = ablative	F = feminine	PROX = proximate
ACC = accusative	FUT = future	PRS = present
ADD = additive	GEN = genitive	PRTV = partitive
ADE = adessive	HON = honorific	P _S = participant in speech event
ADV = adverbial	ILL = illative	PST = past
ALL = allative	IMP = imperative	PTCL = particle
AN = action noun	INE = inessive	PTCP = participle
ANAPH = anaphoric	INF = infinitive	Q = question particle
ASS = assertion	INSTR = instrumental	QUOT = quotative particle
AUX = auxiliary verb	INTERJ = interjection	QUOT.SELF = self-quotative particle
CN = connegative	ITRV = iterative aspect	REFL = reflexive
COMP = complementizer	LAT = lative	RPRT = reportative
COND = conditional	M = masculine	RS = reported speech
CV = converb	NEC = necessitative	RST = reported speech and thought
DAT = dative	NEG = negative	RT = reported thought
DEF = definite	NPST = non-past	S = subject prefix
DEM = demonstrative	O = object prefix	SNS = social network sites
DIM = diminutive	PASS = passive	SG = singular
DIST = distal	PL = plural	TERM = terminative
ELA = elative	P _N = participant in the narrated event	TOP = topic
E _N = narrated event	PN = proper noun	TRANSL = translative

References

- Aikhenvald, Alexandra Y. 2004. *Evidentiality*. Oxford: Oxford University Press.
<https://doi.org/10.1086/509492>
- Aikhenvald, Alexandra Y. 2008. Semi-direct speech: Manambu and beyond. *Language Sciences* 30(4). 383-422. <https://doi.org/10.1016/j.langsci.2007.07.009>
- Aikhenvald, Alexandra Y. 2011. Semi-direct speech in typological perspective. In Alexandra Y. Aikhenvald & Robert M. W. Dixon (eds.), *Language at large: Essays on*

- syntax and semantics*, 327-369. Leiden: Brill.
<https://doi.org/10.1163/ej.9789004206076.i-606.74>
- Bakhtin, Mikhail. 1981. *The dialogic imagination*. Austin / London: University of Texas Press. <https://doi.org/10.2307/1772435>
- Besnier, Niko. 1993. Reported speech and affect on Nukulaelae Atoll. In Jane H. Hill & Judith T. Irvine (eds.), *Responsibility and evidence in oral discourse*, 161-181. Cambridge: Cambridge University Press.
- Buchstaller, Isabelle & Ingrid van Alphen. 2012. Introductory remarks on new and old quotatives. In Isabelle Buchstaller & Ingrid van Alphen (eds.), *Quotatives: Cross-linguistic and cross-disciplinary perspectives*, xii-xxx. Amsterdam: John Benjamins. <https://doi.org/10.1075/celcr.15.02pre>
- Clark, Herbert H. & Richard J. Gerrig. 1990. Quotations as demonstrations. *Language* 66(4). 764-805. <https://doi.org/10.2307/414729>
- Clark, Herbert H. 2016. Depicting as a method of communication. *Psychological Review* 123(3). 324-347. <https://doi.org/10.1037/rev0000026>
- Comrie, Bernard & Larry M. Hyman. 1981. Logophoric reference in Gokana. *Journal of African Languages and Linguistics* 3(1). 19-37. <https://doi.org/10.1515/jall.1981.3.1.19>
- Dahl, Östen. 2000. Egophoricity in discourse and syntax. *Functions of Language* 7(1). 33-77. <https://doi.org/10.1075/fof.7.1.03dah>
- Dalmi, Gréte. 2022. Who on earth is pro? – Licensing null arguments in Hungarian matrix and dependent clauses. In Gréte Dalmi, Egor Tsedryk & Piotr Cegłowski (eds.), *Null subjects in Slavic and Finno-Ugric. Licensing structure and typology*, 253-280. Berlin / Boston: Mouton De Gruyter. <https://doi.org/10.1515/9781501513848-009>
- Du Bois, John W. 1985. Competing motivations. In John Haiman (ed.), *Iconicity in Syntax*, 343-365. Amsterdam / Philadelphia: John Benjamins. <https://doi.org/10.1075/tsl.6.17dub>
- Evans, Nicholas. 2013. Some problems in the typology of quotation: A canonical approach. In Dunstan Brown, Marina Chumakina & Greville G. Corbett (eds.), *Canonical morphology and syntax*, 66-98. Oxford: Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199604326.003.0004>
- Goffman, Erving. 1979. Footing. *Semiotica* 25(1-2). 1-30.
- Goffman, Erving. 1981. *Forms of talk*. Philadelphia: University of Pennsylvania Press.

- Gulyás, Nikolett F. 2019. Impersonals in Finno-Ugric. Paper presented at 3rd SOUL (Syntax of Uralic Languages) (keynote talk). University of Tartu, June 19.
- Güldemann, Tom. 2001. *Quotative constructions in African languages: A synchronic and diachronic survey*. Habilitation Thesis: Institute für Afrikanistik, Universität Leipzig.
- Güldemann, Tom. 2008. *Quotative indexes in African languages: A synchronic and diachronic survey*. Berlin: Mouton de Gruyter.
<https://doi.org/10.1515/9783110211450>
- Haiman, John. 1995. Grammatical signs of the divided self: A study of language and culture. In Werner Abraham, Talmy Givón & Sandra A. Thompson (eds.), *Discourse grammar and typology: Papers in honor of John W. M. Verhaar*, 213-234. Amsterdam: John Benjamins. <https://doi.org/10.1075/slcs.27.17hai>
- Helasvuo, Marja-Liisa, Marjut Johansson & Sanna-Kaisa Tanskanen. 2014. Johdatus digitaalisen vuorovaikutukseen [Introduction to the digital interaction]. In Marja-Liisa Helasvuo, Marjut Johansson & Sanna-Kaisa Tanskanen (eds.), *Kieli verkossa: Näkökulmia digitaaliseen vuorovaikutukseen*, 9-29. Helsinki: Suomalaisen Kirjallisuuden Seura.
- Holvoet, Axel. 2018. Epistemic modality, evidentiality, quotativity and echoic use. In Zlatka Guentchéva (ed.), *Epistemic modalities and evidentiality in cross-linguistic perspective*, 242-259. Berlin / Boston: Walter de Gruyter.
<https://doi.org/10.1515/9783110572261-011>
- Jakobson, Roman. 1971 [1957]. Shifters, verbal categories and the Russian verb. In *Selected writings, vol. II, word and language*, 130-148. The Hague / Paris: Mouton.
<https://doi.org/10.1515/9783110873269.130>
- Jakobson, Roman. 1990. Langue and parole: Code and message. In Linda R. Waugh & Monique Monville-Burston (eds.), *Roman Jakobson, 1896-1982. On language*, 80-110. Cambridge / London: Harvard University Press.
- Kaiser, Elsi. 2015. Impersonal and generic reference: A cross-linguistic look at Finnish and English narratives. *ESUKA – JEFUL* 6(2). 9-42.
<http://dx.doi.org/10.12697/jeful.2015.6.2.01>
- Knyazev, Mikhail. 2022. SAY-complementizers and indexical shift in Poshkart Chuvash: With emphasis on communicative reception reports. *Studies in Language* 46(2). 402-452. <https://doi.org/10.1075/sl.19078.kny>
- Leinonen, Marja. 1983. Generic zero subjects in Finnish and Russian. *Scando-Slavica* 29(1). 143-161. <https://doi.org/10.1080/00806768308600841>

- Lindström, Liina, Nicole Nau, Birutė Spraunienė & Asta Laugalienė 2020. Impersonal constructions with personal reference. Referents of deleted actors in Baltic and Estonian. *Baltic Linguistics* 11. 129-213. <https://doi.org/10.32798/bl.700>
- Lindström, Liina, Maarja-Liisa Pilvik & Helen Plado. 2022. The use of 2nd person singular forms in Seto. Paper presented at *Minor Finnic Languages 1: Historical and Current Perspectives*. University of Uppsala, June 3.
- Li, Charles N. 1986. Direct and indirect speech: A functional study. In Florian Coulmas (ed.), *Direct and indirect speech*, 29-45. Berlin: Mouton de Gruyter. <https://doi.org/10.1515/9783110871968.29>
- Michael, Lev. 2015. The cultural bases of linguistic form: The development of Nanti quotative evidential. In Rik De Busser & Randy J. LaPolla (eds.), *Language structure and environment: Social, cultural, and natural factors*, 99-133. Amsterdam / Philadelphia: John Benjamins. <https://doi.org/10.1075/clsc.6.05mic>
- Nikitina, Tatiana. 2012a. Personal deixis and reported discourse. Towards a typology of person alignment. *Linguistic Typology* 16(2). 233-263. <https://doi.org/10.1515/lity-2012-0008>
- Nikitina, Tatiana. 2012b. Logophoric discourse and first person reporting in Wan (West Africa). *Anthropological Linguistics* 54(3). 280-301. <https://doi.org/10.1353/anl.2012.0013>
- Nikitina, Tatiana. 2020. Logophoricity and shifts of perspective: New facts and a new account. *Functions of Language* 27(1). 78-99. <https://doi.org/10.1075/fol.20001.nik>
- Nikitina, Tatiana & Anna Bugaeva. 2021. Logophoric speech is not indirect: Towards a syntactic approach to reported speech constructions. *Linguistics* 59(3). 609-633. <https://doi.org/10.1515/ling-2021-0067>
- Norvik, Miina, Yingqi Jing, Michael Dunn, Robert Forkel, Terhi Honkola, Gerson Klumpp, Richard Kowalik, Helle Metslang, Karl Pajusalu, Minerva Piha, Eva Saar, Sirkka Saarinen & Outi Vesakoski. 2022. Uralic typology in the light of new comprehensive data sets. *Journal of Uralic Linguistics* 1(1). 4-42. <https://doi.org/10.1075/jul.00002.nor>
- Pascal, Roy. 1977. *The dual voice: Free indirect speech and its functioning in the nineteenth-century European novel*. Manchester: Manchester University Press. <https://doi.org/10.2307/3726924>
- Pérez-Reverte, Arturo. 2017. *Jó emberek [Good people]*. Budapest: Libri Könyvkiadó Kft.

- Roncador, Manfred von. 1988. *Zwischen direkter und indirekter Rede. Nichtwörtliche direkte Rede, erlebte Rede, logophorische Konstruktionen und Verwandtes [Between direct and indirect speech. Nonverbal direct speech, free indirect speech, logophoric constructions and related phenomena]*. Tübingen: Niemeyer.
<https://doi.org/10.1515/9783111678764>
- Roncador, Manfred von. 1992. Types of logophoric marking in African languages. *Journal of African Languages and Linguistics* 13(2). 163-182.
<https://doi.org/10.1515/jall.1992.13.2.163>
- Rose, Françoise. 2013. Le genre du locuteur et de l'allocutaire dans les systems pronominaux: Genre grammatical et indexicalité du genre [The gender of the speaker and the addressee in pronominal systems: grammatical gender and indexicality of gender]. *Bulletin de la Société de Linguistique de Paris* 108(1). 381-417.
- Sakita, Tomoko I. 2002. *Reporting discourse, tense, and cognition*. Oxford: Elsevier.
<https://doi.org/10.1163/9789004487215>
- Spronck, Stef. 2012. Minds divided: Speaker attitudes in quotatives. In Isabelle Buchstaller & Ingrid van Alphen (eds.), *Quotatives: Cross-linguistic and cross-disciplinary perspectives*, 71-117. Amsterdam: John Benjamins.
<https://doi.org/10.1075/celcr.15.07spr>
- Spronck, Stef. 2015. *Reported speech in Ungarinyin: grammar and social cognition in a language of the Kimberley region, Western Australia*. PhD dissertation: Australian National University.
- Spronck, Stef & Tatiana Nikitina. 2019. Reported speech forms a dedicated syntactic domain. *Linguistic typology* 23(1). 119-159. <https://doi.org/10.1515/lingty-2019-0005>
- Suomalainen, Karita 2020. *Kuka sinä on?: Tutkimus yksikön 2. persoonan käytöstä ja käytön variaatiosta suomenkielisissä arkikeskusteluissa [Who is you?: The study on the use of 2SG person and on the variation of use in Finnish everyday conversations]*. PhD dissertation: University of Turku.
- Tagliamonte, Sali A. & Derek Denis. 2008. Linguistic ruin? LOL! Instant messaging and teen language. *American Speech* 83(1). 3-34.
<https://doi.org/10.1215/00031283-2008-001>
- Teptiuk, Denys. 2019. *Quotative indexes in Finno-Ugric (Komi, Udmurt, Hungarian, Finnish and Estonian)*. PhD dissertation: University of Tartu.

- Teptiuk, Denys, 2020. Quotative indexes in Erzya: a typological overview. *Finnisch-Ugrische Mitteilungen* 44. 47-79.
- Teptiuk, Denys, 2021a. Self-quotative markers in Permic and Hungarian. *Linguistica Uralica* 57(3). 213-232. <https://doi.org/10.3176/lu.2021.3.04>
- Teptiuk, Denys, 2021b. Quotative indexes in Permic: Between the original strategies and Russian. In Diana Forker & Lenore Grenoble (eds.), *Language contact in the territory of the former Soviet Union*, 217-259. Amsterdam: John Benjamins. <https://doi.org/10.1075/impact.50.08tep>
- Teptiuk, Denys. Forthcoming. Self-quotations of speech and thought, and how to distinguish them. In Daniela Casartelli, Silvio Cruschina, Pekka Posio & Stef Spronck (eds.), *Grammar of thinking*. Berlin: Mouton de Gruyter.
- Thomas, Jacqueline M.C. 1963. *Le parler Ngbaka de Bokanga. Phonologie, morphologie, syntaxe* [*The Ngbaka dialect of Bokanga. Phonology, morphology, syntax*]. The Hague / Paris: Mouton.
- Vandelanotte, Lieven. 2021. Clearer contours: The stylization of free indirect speech in nineteenth century fiction. In Peter J. Grund & Terry Walker (eds.), *Speech representation in the history of English*, 131-155. Oxford: Oxford University Press. <https://doi.org/10.1093/oso/9780190918064.003.0006>
- Voloshinov, Valentin N. 1973 [1931]. *Marxism and the philosophy of language*. New York / London: Seminar Press.
- Vygotsky, Lev. 1986 [1934]. *Thought and language: Revised edition*. Cambridge, MA: MIT Press.
- Wierzbicka, Anna. 1974. The semantics of direct and indirect discourse. *Research on Language & Social Interaction* 7(3-4). 267-307. <https://doi.org/10.1080/08351817409370375>
- Winkler, Ekkehard. 2011. *Udmurtische Grammatik* [The Udmurt Grammar]. Wiesbaden: Harrassowitz.

Corpora

Erzya Corpora, blogs (EC_blogs)

http://erzya.web-corpora.net/erzya_corpus/search

Erzya Corpora, Social Media Corpus (ESmC)

http://erzya.web-corpora.net/erzya_social_media/search

English Web 2020 (enTenTen20)

https://app.sketchengine.eu/#dashboard?corpname=preloaded%2Fententen20_t31_1&corp_info=1

English Web 2019 (enTenTen19)

https://app.sketchengine.eu/#dashboard?corpname=preloaded%2Fententen19_fi12

Estonian Web 2019 (etTenTen19)

https://app.sketchengine.eu/#dashboard?corpname=preloaded%2Fententen19_fi12

Internet-keskusteluaineistoja (IKA)

https://korp.csc.fi/korp-old/#?stats_reduce=word&cqp=%5B%5D&corpus=s24_2001,s24_2002,s24_2003,s24_2004,s24_2005,s24_2006,s24_2007,s24_2008,s24_2009,s24_2010,s24_2011,s24_2012,s24_2013,s24_2014,s24_2015,s24_2016,s24_2017,s24_2018,s24_2019,s24_2020,s24_001,s24_002,s24_003,s24_004,s24_005,s24_006,s24_007,s24_008,s24_009,s24_010,s24,ylilauta

Komi-Zyrian Corpora, Social Media Corpus (KoZSmC)

http://komi-zyrian.web-corpora.net/index_en.html

Magyar Nemzeti Szövegtár, személyes alkkorpusz (MNSz)

<http://mnsz.nytud.hu>

Udmurt Corpora (UdC)

http://udmurt.web-corpora.net/udmurt_corpus/search

Udmurt Corpora, blogs (UdC_blogs)

http://udmurt.web-corpora.net/udmurt_corpus/search

Udmurt Corpora, Social Media Corpus (UdSmC)

http://udmurt.web-corpora.net/index_en.html

Appendix

Below I present the list of quotative strategies used to query different types of reports, arranged language-wise. Next to the strategy, separated by semi-colons I present the gloss and approximate translation to show the range of use for a concrete strategy in the database. Even though the strategies are ordered with the following principle: ‘speech verbs > mental verbs > other clausal units > non-clausal units’, the letters

next to strategies are not indicative and shall not be used to compare the strategies between the languages.

1. Erzya:
<p>1a. <i>mefiń</i>; ‘say:PST:1SG’; ‘I said’, ‘I thought’; 1b. <i>mefś</i>; ‘say:PST.3SG’; ‘(s)he said’, ‘(s)he thought’; 1c. <i>mefśt</i> ‘say:PST.3PL’; ‘they said’, ‘they_{UNKNOWN} said’; 1d. <i>arśiń</i>; ‘think:PST:1SG’; ‘I thought’; 1e. <i>arśeś</i> ‘think:PST.3SG’; ‘(s)he thought’; 1f. <i>arśeśt</i> ‘think:PST.3PL’; ‘they thought’; ‘they_{UNKNOWN} thought’; 1g. <i>kéla</i>; ‘QUOT’; ‘I said’, ‘I thought’, ‘(s)he said’, ‘(s)he thought’, ‘they_{UNKNOWN} said’, ‘they_{UNKNOWN} thought’; 1h. <i>mol</i>; ‘QUOT’; ‘(s)he said’, ‘they_{UNKNOWN} said’.</p>
2. Estonian:
<p>2a. <i>ütsin</i>; gloss: ‘say:PST:1SG’; approximate translation: ‘I said’, ‘I thought’; 2b. <i>ütles</i>; ‘say:PST.3SG’; ‘(s)he said’; 2c. <i>üt(le)sid</i>; ‘say:PST.3PL’; ‘they said’, ‘they_{UNKNOWN} said’; 2d. <i>mõtsin</i>; ‘think:PST:1SG’; ‘I thought’; 2e. <i>mõtles</i>; ‘think:PST.3SG’; ‘(s)he thought’; 2f. <i>mõt(le)sid</i>; ‘think:PST.3PL’; ‘they thought’, ‘they_{UNKNOWN} thought’; 2g. <i>olin nagu</i>; ‘be:PST:1SG like’; ‘I said’, ‘I thought’; 2h. <i>olin mingi</i>; ‘be:PST:1SG something’; ‘I said’, ‘I thought’; 2i. <i>mul oli (nü) et</i>; ‘1SG:ADE be:PST.3SG (so) COMP’; ‘I said’, ‘I thought’; 2j. <i>mul oli nagu (et)</i> ‘1SG:ADE be:PST.3SG like (COMP)’; ‘I thought’; 2k. <i>oli (nü) et</i>; ‘be:PST.3SG (so) COMP’; ‘(s)he said’; 2l. <i>olid et</i>; ‘be:PST.3PL COMP’; ‘they said’; 2l. <i>oli nagu (et)</i> ‘be:PST.3SG like (COMP)’; ‘(s)he said’; 2m. <i>oli lihtsalt</i> ‘be:PST.3SG simply’; ‘(s)he said’; 2n. <i>väidetavalt</i>; ‘allegedly’; ‘they_{UNKNOWN} said’; 2o. <i>kuuldavasti</i>; ‘allegedly’; ‘they_{UNKNOWN} said’; 2p. <i>kuulukse</i>; ‘QUOT’; ‘they_{UNKNOWN} said’.</p>
3. Finnish:
<p>3a. <i>sanoin et</i>; gloss: ‘say:PST:1SG COMP’; approximate translation: ‘I said’; 3b. <i>sanoi et</i>; ‘say:PST.3SG COMP’; ‘(s)he said’; 3c. <i>sanoivat et</i>; ‘say:PST:3PL COMP’; ‘they said’; 3d. <i>ajattelin et</i>; ‘think:PST:1SG COMP’; ‘I thought’; 3e. <i>ajatteli et</i>; ‘think:PST.3SG COMP’; ‘(s)he thought’; 3f. <i>ajattelivat et</i>; ‘think:PST:3PL COMP’; ‘they think’, ‘they_{UNKNOWN} thought’; 3g. <i>olin (ihan) et</i>; ‘be:PST:1SG completely COMP’; ‘I said’, ‘I thought’; 3h. <i>olin tyyliin (et)</i>; ‘be:PST:1SG like (COMP)’; ‘I said’; 3i. <i>olin niinku (et)</i>; ‘be:PST:1SG like (COMP)’; ‘I thought’;</p>

<p>3j. <i>olin silleen et</i>; ‘be:PST:1SG so COMP’; ‘I said’, ‘I thought’; 3k. <i>olin vaan et</i>; ‘be:PST:1SG just COMP’; ‘I said’, ‘I thought’; 3l. <i>oli et</i>; ‘be:PST:3SG COMP’; ‘(s)he said’; 3m. <i>oli ihan et</i>; ‘be:PST:3SG completely COMP’; ‘(s)he said’, ‘(s)he thought’; 3n. <i>oli niinku et</i>; ‘be:PST:3SG like COMP’; ‘(s)he said’; 3o. <i>oli tyliin</i>; ‘be:PST:3SG like’; ‘(s)he said’; 3p. <i>oli silleen et</i>; ‘be:PST:3SG so COMP’; ‘(s)he said’, ‘(s)he thought’; 3q. <i>oli vaan et</i>; ‘be:PST:3SG just COMP’; ‘(s)he said’, ‘(s)he thought’; 3r. <i>kuulemma</i>; ‘QUOT’; ‘they_{UNKNOWN} said’.</p>
<p>4. Hungarian:</p> <p>4a. <i>mondok / mondom</i>; gloss: ‘say:PRS.1SG/.DEF’; approximate translation: ‘I said’, ‘I thought’; 4b. <i>aszontam</i>; ‘DEF.say:PST:1SG’; ‘I said’, ‘I thought’; 4c. <i>aszongya</i>; ‘DEF.say:PRS.3SG.DEF’; ‘(s)he says’; 4d. <i>aszonta</i>; ‘DEF.say:PST:3SG.DEF’; ‘(s)he said’; 4e. <i>aszongyák</i>; ‘DEF.say:PRS:3PL.DEF’; ‘they said’, ‘they_{UNKNOWN} said’; 4f. <i>aszonták</i>; ‘DEF.say:PST:3PL.DEF’; ‘they said’, ‘they_{UNKNOWN} said’; 4g. <i>gondoltam</i>; ‘think:PST:1SG’; ‘I thought’; 4h. <i>gondolta</i>; ‘think:PST:3SG.DEF’; ‘(s)he thought’; 4i. <i>gondolták</i>; ‘think:PST:3PL.DEF’; ‘they thought’; ‘they_{UNKNOWN} thought’; 4j. <i>állítólag</i>; ‘allegedly’; ‘they_{UNKNOWN} said’, ‘(s)he said’.</p>
<p>5. Komi:</p> <p>5a. <i>šui</i>; gloss: ‘say:PST.1SG’; approximate translation: ‘I said’; 5b. <i>šuis</i>; ‘say:PST.3SG’; ‘(s)he said’; 5c. <i>šuisny</i>; ‘say:PST.3PL’; ‘they said’, ‘they_{UNKNOWN} said’; 5d. <i>ćajti</i>; ‘think:PST.1SG’; ‘I thought’; 5e. <i>dumajti</i>; ‘think:PST.1SG’; ‘I thought’; 5f. <i>ćajtis</i>; ‘think:PST.3SG’; ‘(s)he thought’; 5g. <i>mövpalis</i>; ‘think:PST.3SG’; ‘(s)he thought’; 5h. <i>ćajtisny</i>; ‘think:PST.3PL’; ‘they thought’, ‘they_{UNKNOWN} thought’; 5i. <i>miša</i>; ‘QUOT.SELF’; ‘I said’, ‘I thought’; 5j. <i>pö</i>; ‘QUOT’; ‘(s)he said’, ‘(s)he thought’, ‘they said’, ‘they_{UNKNOWN} said’; 5k. <i>mol</i>; ‘QUOT’; ‘(s)he said’; 5l. <i>úpa</i>; ‘like’; ‘I said’, ‘(s)he said’, ‘they said’, ‘they_{UNKNOWN} said’.</p>
<p>6. Udmurt:</p> <p>6a. <i>šui</i>; gloss: ‘say:PST.1SG’; approximate translation: ‘I said’, ‘I thought’; 6b. <i>šuiž</i>; ‘say:PST.3SG’; ‘(s)he said’; 6c. <i>šuižy</i>; ‘say:PST.3PL’; ‘they said’, ‘they_{UNKNOWN} said’; 6d. <i>malpaj</i>; ‘think:PST.1SG’; ‘I thought’; 6e. <i>malpaz</i>; ‘think:PST.3SG’; ‘(s)he thought’; 6f. <i>malpazy</i>; ‘think:PST.3PL’; ‘they thought’, ‘they_{UNKNOWN} thought’; 6g. <i>pöj</i>; ‘QUOT.SELF’; ‘I said’, ‘I thought’; 6h. <i>pe</i>; ‘QUOT’; ‘(s)he said’, ‘they said’, ‘they_{UNKNOWN} said’;</p>

- 6i. *mol*; 'QUOT'; 'I said', 'I thought', '(s)he said', 'they said';
6j. *tipa*; 'like'; 'I said', '(s)he said', 'they said'.

Table 11: Quotative strategies used as a query of different report types.

CONTACT

denys.teptiuk@ut.ee

Predicting grammatical gender in Nakh languages: Three methods compared

JESSE WICHERS SCHREUR¹, MARC ALLASSONNIÈRE-TANG², KATE BELLAMY³,
NEIGE ROCHANT⁴

¹LEIDEN UNIVERSITY/GOETHE UNIVERSITY FRANKFURT/EPHE PARIS, ²CNRS/MNHN/UNIVERSITY
PARIS CITY (EA UMR 7206), ³LEIDEN UNIVERSITY, ⁴SORBONNE NOUVELLE/LACITO UMR
7107/LLACAN UMR 8135

Submitted: 11/03/2022 Revised version: 12/10/2022

Accepted: 20/11/2022 Published: 22/12/2022

Abstract

The Nakh languages Chechen and Tsova-Tush each have a five-valued gender system: masculine, feminine, and three “neuter” genders named for their singular agreement forms: B, D and J. Gender assignment in languages is generally analysed as being dependent on both form and semantics (e.g. Corbett 1991), with semantics typically prevailing over form (e.g. Bellamy & Wichers Schreur 2022, Allasonnière-Tang et al. 2021). Most previous studies have considered only binary or tripartite gender systems possessing one masculine, one feminine, and one neuter value. The five-valued system of Nakh thus represents a more complex and insightful case study for analysing gender assignment. In this paper we build on the existing qualitative linguistic analyses of gender assignment in Tsova-Tush (Wichers Schreur 2021) and apply three machine-learning methods to investigate the weight of form and semantics in predicting grammatical gender in Chechen and Tsova-Tush. Our main aim is thus to show how three different computational classifier methods perform on a novel set of non-Indo-European data. The results show that while both form and semantics are helpful for predicting grammatical gender in Nakh, semantics is dominant, which supports findings from existing literature (Allasonnière-Tang et al. 2021), as well as confirming the utility of these computational methods. However, the results also show that the coded semantic information could be further fine-grained to improve the accuracy of the predictions (see also Plaster et al. 2013). In addition, we discuss the implications of the output for our understanding of language-internal and family-internal processes of language change, including how loanwords are integrated from Russian, a three-gender language.

Keywords: grammatical gender; Nakh languages; computational classifiers; gender assignment.

1. Introduction

Humans intuitively categorise the world around them (Senft 2000). This categorisation emerges in language, such as through the presence of grammatical gender as a means of identifying and tracking referents in discourse (Contini-Morava & Kilarski 2013). Perhaps the most familiar gender systems are the binary masculine-feminine ones found in, for example, French, Spanish and Italian, or the tripartite masculine-feminine-neuter system of German. Yet gender is a pervasive grammatical category, being present in around half of the world's sampled languages (Corbett 2013). In a grammatical gender system, all nouns are assigned a gender, which is then formally reflected in agreement markers (their *exponence*) on other associated elements in the clause (Hockett 1958; Corbett 1991). These elements can include, but are not limited to, articles, adjectives or verbs, depending on the language in question.

Gender is assigned to nouns on the basis of semantic and/or formal (phonological and morphological) properties, with semantic features usually taking precedence over form (Corbett & Fraser 2000; Bellamy & Wichers Schreur 2022, Allasonnière-Tang et al. 2021). In some systems this assignment is completely transparent, whereas in others it is more opaque and, particularly for the L2 learner, can be difficult to systematise (e.g. Sokolik & Smith 1992 regarding French). Yet children acquire gendered languages of varying levels of transparency with equal ease, as they do languages possessing differing levels of complexity in other domains (cf. Karmiloff-Smith 1979). This begs the question, therefore, as to what the specific principles are that underpin these complex assignment systems. In this paper, we will use three computational methods to test which principles, or factors, most adequately predict the gender of nouns in Chechen and Tsova-Tush, two Nakh languages of the East Caucasian family that each possess a five-value gender system.

The present paper builds on a relatively small but expanding set of quantitative and qualitative studies concerning gender prediction in East Caucasian and other languages. Using Russian nominals as a case study, Corbett and Fraser (1993) introduced Network Morphology, a framework for analysing inflectional morphology. They followed this up with a more detailed analysis of the interrelation of meaning, gender, declension class and phonology in Russian, augmenting the Network Morphology approach with a lexical knowledge representation language known as DATR (Fraser & Corbett 1995, building on Corbett 1982; see also Evans & Gazdar 1989a 1989b 1996). The same approach has also been used to model the gender systems of Arapesh (Northern Papua; Fraser & Corbett 1997), Polish (West Slavic;

Brown 1998), and Mayali (northern Australia; Evans et al. 2002). While these studies provide intricate and informative representations of the gender systems they aim to model, they are not predictive in nature since the gender value constitutes one of the noun's attributes in their notation. As such, they are not testing previous assumptions in the same way that, later, predominantly tree-based approaches do.

Computational modelling of gender systems dates back to the early 1990s, notably Sokolik and Smith (1992) on French. In this study, the authors used a connectionist or 'parallel distributed processing' model to test "whether the information inherent within the structure of individual French nouns is sufficient to allow gender to be correctly assigned without reliance upon other types of information" (Sokolik & Smith 1992: 41). They selected 600 nouns (300 masculine and 300 feminine) from introductory French language textbooks, with 450 serving as the training set and the remaining 150 as the test set, to identify whether the model could assign the correct gender to nouns it had not yet encountered. In short, the model was able to do so with a high level of accuracy (over 76%), having 'learnt' that certain orthographic features correlate with either masculine or feminine gender. This suggests that L2 learners could also learn gender in a similar way, without having recourse to the often ambiguous cues appearing alongside a noun whose gender is not yet known, such as the singular prevocalic definite article *l'* and the plural definite article *les*.

Also in relation to French, and also using a connectionist model, Polinsky and van Everbroeck (2003) investigated the reanalysis of the Latin gender system as it transitioned to Old French. The simulations of the frequency-based neural network model aimed to uncover which factors are sufficient to lead to language change, in this case the evolution of the gender system in Old French. The authors built a training corpus from the 500 most commonly occurring nouns (excluding proper nouns and clear Greek borrowings) in the fifth-century Vulgate, in which the token frequency of each of the six possible case and number forms for each noun was calculated (for more details on building the corpus, see Polinsky & van Everbroeck 2003: 369-370). Using a feed-forward network with one hidden layer (see Section 3 for details of the neural network architecture used in the present study), the authors found that the model could adequately - around 60% at the end of the ninth generation - learn the gender of nouns in Late Latin, with a strong reliance of formal cues (i.e. case endings). Its decreasing performance over generations mirrored changes to the gender system largely instigated by phonological changes from Late Latin to Old French. Moreover, the model was also able to reflect the proposed influence of Gaulish (three genders) on gender assignment in Latin.

Bateman and Polinsky (2010) used decision trees to reconsider the gender system of Romanian, allowing them to posit a two-value rather than the traditional three-value analysis. They use the C4.5 Decision Tree Algorithm designed by Quinlan (1993 1996) to establish the rules of plural formation for Romanian nouns. Notably, the first cut relates to a semantic feature, namely whether the noun possesses masculine semantic features. The following cuts all pertain to formal features, that is, the final segment of the noun, its number of syllables, and whether the root contains a diphthong. A semantics-first assignment principle has already been observed in previous studies outlined in this section.

Indeed, Plaster, Polinsky and Harizanov (2012) applied a similar approach to what they call noun classification (and we call a gender system) in Tsez (Dido), an endangered East Caucasian language (Tsezic branch) that possesses a five-value gender system, like Tsova-Tush and various other languages of the family. Their goal was to “identify a set of formal and semantic features of the sort to which young children acquiring language are known to be sensitive and to produce a decision tree containing these features that will predict the classification of nouns in Tsez” (Plaster et al. 2012: 7). Over 3,500 Tsez nouns (including loanwords), collected from dictionaries (Khalilov 1999; Rajabov, undated), were coded for formal (at least seven) and semantic (at least nine) features and then tested and run through the decision tree module of the “Orange” data mining tool (see Demsar et al. 2004). Their results demonstrate that the semantic features of a noun were most predictive of its gender, with such features overriding formal ones, as has been observed in other mixed assignment type systems (e.g. Corbett 1991). The decision tree model produced is able to predict around 70% of nouns in Tsez, with assignment to the remaining 30% possibly complicated by their status as loanwords or dialectal variants (Plaster et al. 2013: 11). Smaller semantic fields than those expressing core features such as animacy or biological sex, such as [berry] and [stone], are also highlighted as being predictive in the model. It is noteworthy, however, that the preference for semantic features in the computational model stands at odds somewhat with results from studies on Tsez language acquisition. Gagliardi and Lidz (2014) found that the Tsez-speaking children in their study had a preference for using phonological rather than semantic information for classifying nonce words, despite there being a statistical asymmetry that prefers the opposite. They suggest that this phonological bias relates to the higher value placed on phonological information in the intake: such information is available to children long before they know what a word means, and is more reliable than a semantic form (Gagliardi & Lidz 2014: 81).

The most recent and the most methodologically relevant previous study is Allasonnière-Tang, Brown and Fedden (2021), which applies three different computational classifier methods - simple decision trees, random forest trees and neural networks - to test the claim that Mian, an Ok language of Trans New Guinea, assigns gender predominantly on the basis of a noun's semantics. The accuracy in predicting the gender of the test nouns (30% of the nouns in the Mian dictionary used) was barely better than the majority baseline when only formal (i.e. phonological) features were fed to the computational classifiers. However, this improved considerably when only semantic features were fed in, and a little better again when both sets of features were used. The tree-based methods were the most accurate, with the random forest model only improving accuracy over a single tree by 1%. Moreover, the random forest method supports previous descriptive linguistic research on Mian (Fedden 2011), in that the top-ranked variables for predicting gender in the language are all semantic. Indeed, formal features only play a fine-grained discriminatory role in one semantic class, namely birds. These results support existing qualitative analyses of the Mian gender system as being semantics-dominant, as well as providing support for the methods used. The authors also highlight the importance of investigating and testing gender systems beyond those found in Indo-European languages, in order to contribute to the investigation of nominal classification more broadly. The present paper also represents a response to this call.

As such, the main aim of this paper is to show how three different computational classifier methods will perform on a novel set of non-Indo-European data. Some qualitative work on gender assignment in Nakh languages has been undertaken with limited results (see Section 2), therefore more fine-grained levels of description are needed. Although this paper contributes to finding semantic correlates to gender classes, our principal goal is a methodological assessment of the computational classifiers.

2. Background

2.1. Introduction to Chechen and Tsova-Tush

The Nakh languages form an outlying branch of the East Caucasian family (also known as Northeast Caucasian or Nakh-Daghestanian; Nichols 2003), and includes only three languages: Chechen (ISO 639-3 che), Ingush (ISO 639-3 inh) and Tsova-Tush (ISO 639-3 bbl). Chechen and Ingush are more closely related to each other than

to Tsova-Tush, so in this paper we focus on one language from each sub-group within Nakh: Chechen and Tsova-Tush.

Chechen speakers mostly live in their ancestral homeland, Chechnya, located on the northern slopes of the Greater Caucasus Mountains and now a semi-autonomous republic of the Russian Federation. As of the 2010 Census, Chechens constitute 95.3% of the Chechen republic's population, and Standard Chechen is one of the two official languages of the republic alongside Russian. Some Chechen speakers can also be found in the Pankisi Gorge of neighbouring Georgia and in several villages in the Daghestanian lowlands.

Tsova-Tush (also known as Bats or Batsbi) is spoken in the village of Zemo Alvani in eastern Georgia by approximately 500 people who are all fluent in Georgian. These speakers have stopped transmitting Tsova-Tush to the next generation, which is why the language is considered severely endangered (Wurm et al. 2001). The Tsova-Tush ethnically self-identify as Georgian, and their language has been under the influence of Georgian for over four centuries (Desheriev 1953). Where Tsova-Tush shows little to no dialectal variation, Chechen has several distinct regional dialects. During the Soviet period, a written standard language was created for Chechen, whereas Tsova-Tush remains largely unwritten to this day.

Both Chechen and Tsova-Tush possess five gender classes, and agreement is marked by the same four consonantal prefixes. Agreement targets include a third of all underived verbs and a small number of adverbs, which agree with the nominative argument of the clause, as in Examples (1) and (2), as well as approximately ten underived adjectives and the numeral ‘four’, which agree with the head noun they are modifying (see Example (1)).

(1) Tsova-Tush (Kadagidze 2009: 44)

d-aqqoⁿ xi d-ujt'-ǔ
 D-big water D-go-PRS
 ‘The big river is flowing.’

(2) Chechen (Nichols 1994: 37)

ča jiett āra b-ēl-ir
 one cow(B) out B-go-AOR
 ‘One cow went out.’

The way in which the agreement markers are distributed through the gender classes varies between Chechen and Tsova-Tush. Firstly, as shown in Table 1, Chechen has a unified human plural marker *b-*, whereas Tsova-Tush differentiates between masculine plural *b-* and feminine plural *d-*.

Gender class	Semantics	Markers (sg/pl)	
		Tsova-Tush	Chechen
M	male rationals	v- / b-	v- / b-
F	female rationals	j- / d-	j- / b-
B	animals, inanimates	b- / d-	b- / d-
J	animals, inanimates	j- / j-	j- / j-
D	animals, inanimates	d- / d-	d- / d-
Bb	animals, inanimates	b- / b- (6 items)	b- / b- (22 items)
Bj	body parts, 'step', 'kick'	b- / j- (17 items)	-
Dj	body parts	d- / j- (6 items)	-

Table 1: Gender classes and their agreement markers in Tsova-Tush and Chechen.

Secondly, both Tsova-Tush and Chechen have several lexical items that show an agreement pattern that is different to the five main genders. In addition to several abstract nouns and nouns denoting materials that do not form a plural, both languages have words that require the agreement marker *b-* in both singular and plural contexts. Additionally, most Tsova-Tush nouns denoting body parts take *j-* in the plural and *d-* or *b-* in the singular. It has been argued that these groupings constitute additional gender classes, such that Tsova-Tush and Chechen have eight and six genders, respectively. Alternatively, these items can be viewed as exceptions or anomalies, and belong to so-called 'inquate' genders, in that the number of items in these gender classes is too low to form a quorum (Corbett 1991: 170–175). Regardless of the analysis, these items will not be included in the present study.

2.2. Gender assignment in Chechen and Tsova-Tush

While the two human classes M and F mostly contain male and female rationals respectively, assignment to the three non-human classes, which can be called 'neuter', is far from fully predictable. On the one hand, it seems that semantics, morphology and phonology play a role in the assignment of the gender of non-rationals to one of

the different neuter classes in both Tsova-Tush and Chechen. On the other hand, this role is very marginal, as the tendencies described below can only predict the gender of a small portion of non-rationals:

- Semantics: both Chechen and Tsova-Tush feature some semantically-based clusterings of genders, which allow the gender of only 15% of Tsova-Tush non-rationals to be predicted (Nichols 2007; Wichers Schreur 2021).
- Morphology: in Tsova-Tush, all verbal nouns (e.g. *st'exar* 'waiting' from *st'ex-* 'to wait') are assigned to the D class, and most de-adjectival abstract nouns (e.g. *must'ol* 'acidity' from *must'in* 'sour') are assigned the J class (Wichers Schreur 2021). Nichols (2007) also showed that a number of derived abstract nouns in Chechen are assigned D or J depending on the degree of abstractness (e.g. *goomalla* (D) 'crookedness, bend', (J) 'enmity, hostility' from *goma* 'bent, crooked'). Inflectional morphology is not related to any gender assignment tendency.
- Phonology: Wichers Schreur (2021) showed that 60-65% of all Tsova-Tush nouns starting in *b-*, *d-* and *j-* belong to B, D and J gender classes respectively. This phenomenon is assumed to be a consequence of two processes: autogender and alliterative concord. The former refers to the phenomenon of fossilised gender markers on nouns themselves (cf. Nichols 2011: 147), while the latter refers to a phenomenon where the gender assignment of a noun is in fact influenced by its phonology (Corbett 1991: 117).

In cases where “neutral gender agreement” is required (see Corbett 1991: 205), Tsova-Tush agreement targets default to the D gender (Bellamy & Wichers Schreur 2022). Such an agreement pattern occurs in cases where the speaker leaves the gender of a human referent unspecified, when a word agrees with two or more human nouns that have different genders, or when agreement is with a clause. Additionally, a set of nouns denoting humans whose social gender is unspecified (e.g. 'friend', 'godparent', 'Christian') also triggers D agreement. However, these nouns are annotated as MF in the source dictionaries, a label which we have taken over. Neutral gender agreement is assumed to be highly similar in Chechen, as it is also found in Ingush (Nichols 2011: 435).

Gender assignment of loanwords has not been described in detail for Chechen, but Nichols (2007) notes that new loanwords are often given the gender of a near

synonym or an immediate generic. However, a cursory query of the Chechen data (see below), shows us that all recent Russian loans have gender J. As regards Tsova-Tush loanwords, Wichers Schreur (2021) shows that they follow the same set of semantic and phonological tendencies as native nouns.

In sum, although a number of semantic, morphological and phonological tendencies have been observed, gender assignment for the vast majority of nouns in Chechen and Tsova-Tush remains poorly understood. This gap in our understanding holds for both qualitative and quantitative approaches, a void we aim to begin filling with the present study.

3. Data and Methods

3.1. Data

Two lexical databases were used in the present study. For Chechen we used the database developed by Chechen language specialist Erwin Komen, consisting of the combined dictionaries of Matsiev (1961) and Jamalkhanov and Aliroev (1991). The database contains 4339 nouns with gender annotation and Russian translation. After removing nouns that only occur in singular or plural, proper nouns, and nouns with ambiguous or inquate genders, we arrive at 2673 items. We used these items as our dataset and annotated each item with a value “yes/no borrowed”.¹ 298 items were classified as borrowings. The distribution of the gender categories found in the Chechen data is visualised in Figure 1a.

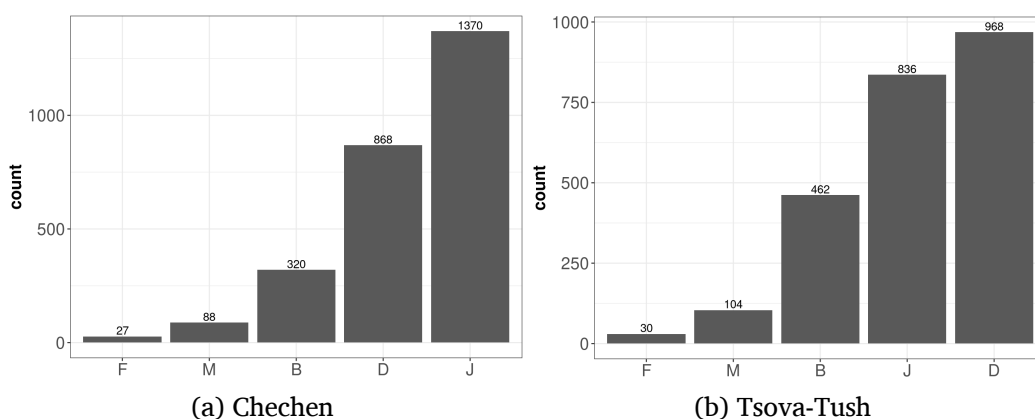


Figure 1: The distribution of gender categories in different languages of the data.

¹ Where “yes” means ‘obvious borrowing’ (i.e. the form differing in one or two segments from its Russian translation), and “no” ‘not known to be a borrowing’.

For Tsova-Tush, we used the Kadagidze and Kadagidze (1984) dictionary, containing 2775 nouns with gender annotation and Georgian and Russian translation. After removing nouns that only occur in singular or plural, proper nouns, and nouns with ambiguous or inquate genders, we arrive at 2400 items. The distribution of the gender categories found in the Tsova-Tush data is visualised in Figure 1b. We translated all nouns to English and annotated each item with a value “yes/no borrowed”.² Within all the items, 1365 are marked as borrowed from Georgian, whereas 72 additional items are marked as borrowed from other languages (mostly Russian).

For both datasets, several annotations were added that classify items in semantic categories. Firstly, the following broad semantic categories were added: Male, Female, Human (i.e. gender-unspecified human), Animal, Inanimate. For example, in the Tsova-Tush dataset, we marked 2081 items as Inanimate, 175 items as Animal, 104 items as Male, 13 as Human, and 27 as Female. Secondly, English translations were matched against their corresponding Concept Sets in the online Concepticon (List et al. 2016). This database was developed to serve as a reference concept list to aid in studies on semantic change, cross-linguistic polysemies, and semantic associations. Both the abstract concept (Concept Set in Concepticon’s terminology), as well as the semantic category of this concept (e.g. ‘agriculture and vegetation’ or ‘the body’) were added as annotations to items in the Tsova-Tush and Chechen databases where the English translation matched a Concepticon Concept Set. This was possible for 57% (1363/2400) of Tsova-Tush items and 44% (1164/2673) of Chechen items. Thirdly, a more fine-grained semantic annotation was added for 69% (1655/2400) of Tsova-Tush items and 1490/2673 (56%) for Chechen. These annotations consisted of semantic domains we chose, to represent the common practice of linguists inventing semantic domains based on intuitive and language-internal principles. As an example, within the Tsova-Tush data, 228 items are annotated as ‘Abstract’, 87 items are marked as ‘Person’, 137 as ‘Implement’, and 96 as ‘Food’. Annotating our data with three independent semantic layers allows us to test which is preferred by the different computational methods described below. It is necessary to distinguish between synonymous labels of different annotation layers, i.e. between the label [person] in the fine-grained semantic domain layer (indicating nouns that refer to professions and other identities that humans have, other than kinship relations and ethnicities),

² Where “yes” means ‘obvious borrowing’ (i.e. the form differing in one or two segments from its Georgian or Russian translation), and “no” ‘not known to be a borrowing’.

Concepticon’s label [person/thing] (as opposed to [action/process]), and the label [human] in the layer in the broad semantics layer (indicating any human that is not explicitly female or male).

In terms of form, we included the information of the first three and the last three phonemes of each noun. This choice was motivated by the fact that nominal features such as gender are generally found at the start and/or the end of nouns (Dryer 2013; Basirat et al. 2021). Information regarding word length was also included. As an example, the first three phonemes of Tsova-Tush *haer* ‘air’ are /h/, /a/, and /e/, and its word length is counted as 4. An example of the raw data used for Tsova-Tush is provided in Table 2. The full raw data is provided in the supplementary materials.

Noun	haer	mar
Gloss	air	husband
Gender	J	M
Concepticon_category	Person/thing	Person/thing
Concepticon_field	Physical world	Kinship
Semantic_broad	Inanimate	Male
Semantic_domain	Natural	Kinship
Borrowed_Arab	0	0
Borrowed_GE	1	0
Borrowed_Turk	0	0
Borrowed_Russian	0	0
Word length	4	3
Last first phoneme	r	r
Last second phoneme	e	a
Last third phoneme	a	m
First phoneme	h	m
Second phoneme	a	a
Third phoneme	e	r

Table 2: A sample of the raw Tsova-Tush data used in this paper.

3.1. Method

We used three computational classifiers to evaluate the predictive power of form and semantics on the grammatical gender of nouns in Chechen and Tsova-Tush. The first two classifiers apply the method of binary recursive partitioning (Breiman et al. 1984). First, one of the classifiers generates a decision tree by recursively partitioning the data in a binary way to create homogeneous groups. The output is represented as a decision tree that allows us to visualise the hierarchical interaction of the semantic and formal variables when it comes to predicting the grammatical gender of nouns. As an example, if both a formal and a semantic feature are helpful for predicting the grammatical gender of nouns, the decision tree will display which variable should be considered first during the decision-making process. Such a tree could be subject to overfit when it comes to multiple replications, therefore the second classifier generates a forest of trees instead of a single tree, hence its name: *random forests*.

The random forests computational classifier creates a forest of 300 decision trees that are considered as a whole to evaluate the relevance of formal and semantic variables for predicting the grammatical gender of nouns. This classifier uses the same algorithm as a single decision tree, but it creates a forest of trees instead of only one tree. For each tree in the sample of 300 trees, the classifier takes a bootstrap sample of the data along with a subset of the variables of the data. In other words, the classifier extracts a random subset of rows and columns in the data to generate each of the 300 trees. A statistical test carried out for each sample shows if an interaction between some variables is consistently observed across all the samples. This process of random sampling is one of the main strengths of random forests, as it allows for the analysis of different data sizes (ranging from small to large), as well as the consideration of potential auto-correlation between variables (Tagliamonte & Baayen 2012). Another main strength of decision tree-based classifiers such as random forests, is that they allow for a relatively transparent understanding of the interaction between the variables. For example, such classifiers have been used to investigate linguistic universals (One-Soon & Tang 2020), automatic identification of sounds based on phonetic features (Ulrich et al. 2021), tone paradigms (Lemus-Serrano et al. 2021), and also the gender affiliation of nouns (Allasonnière-Tang et al. 2021). More specifically, random forests provide information on the relative importance of the predictor variables. The larger the importance of a variable, the more predictive it is. For instance, if the accuracy of the classifier drops the most when it does not take

into account a specific feature, this feature can be considered to have the highest ranking within all the variables.

However, the transparency of decision-tree based classifiers can also be their weakness, as they might fail to capture extremely complex interactions between the variables. Therefore, the third classifier we consider in our study has a neural network architecture (Haykin 1998; Parks et al. 1998). *Neural network* is a non-linear discriminative classifier that identifies boundaries between the data points with regard to their predicted variable. While it is much more complex to extract the interactions between the variables using a neural network, we can use such a classifier to assess whether decision-tree based classifiers missed some non-linear information encoded in the data. For instance, if decision-tree based classifiers and neural networks reach a similar level of accuracy for predicting the grammatical gender of nouns, we can assume that both classifiers are capturing most of the information encoded in the data and that additional information is needed to further improve the performance of the classifiers.

In this study, we use a simple feed-forward neural network that consists of an input layer, a hidden layer, and an output layer. Each layer has a specific number of neurons that are connected to each other. The input layer has one neuron for each predictor (i.e., each variable). The number of hidden layers and their quantity of neurons is flexible. The size of the output layer is equal to the number of categories to predict. As an example, when predicting gender in Tsova-Tush, the output layer has five neurons, which represent the gender categories M, F, B, D and J. In our experiments, we set the size of the hidden layer to ten. More experiments could be conducted to finetune the number and size of hidden layers. For example, we could have an architecture with hundreds of neurons and a dozen hidden layers. Nevertheless, since we are only interested in the relative performance of formal and semantic features, we do not perform such experiments in this study.

Some parameters are shared across all of the classifiers. First, each classifier is trained with 70% of the data. Then, the trained classifier is evaluated when predicting the other 30% of the data that it did not encounter during the training. That is to say, the training set and the test set do not overlap. If a noun is used in the training set, it will not be used in the test. To avoid coincidental biases from the random sampling between the training and test sets (which can never be precisely 70% and 30% respectively), this sampling and evaluation process is repeated ten times for each classifier. If the results are similar across the ten replications, it shows that the results

are robust. We also tested the experiments based on 100 and 1000 replications, which gave similar results. For sake of simplicity and to avoid over-computation, we thus report the results based on ten replications. The code in the supplementary materials allows for additional tuning of the parameters, such as conducting more iterations and only selecting specific gender categories for making predictions.

The performance of the classifiers is evaluated using accuracy, precision, and recall. First, the accuracy shows how good the classifier is at predicting grammatical gender on the data. The accuracy is equal to the ratio of all the correctly retrieved tokens within the entire data. For example, if the classifier correctly predicted the gender of 70 out of 100 nouns, the accuracy of the classifier is 70%. This accuracy value needs to be compared to a baseline in order to be interpreted as good or bad. In this study, we compare the accuracy to a random baseline and a majority baseline. The random baseline constitutes what the model would obtain by making totally random guesses. Taking a binary classification as an example, we could imagine a model that has to guess if a coin toss will be heads or tails. The random baseline is calculated by adding the probability of heads in the data multiplied by the probability that you guess heads, with the probability of tails in the data multiplied by the probability that you guess tails. Assuming that the coin is fair and that there is a 50-50 chance to get heads or tails, the random baseline equals to $0.5 * 0.5 + 0.5 * 0.5 = 0.5$. Taking Chechen as an example, the random baseline is equal to the sum of the square of the proportion of each gender category in the data, i.e., $(320/2673) * (320/2673) + (868/2673) * (868/2673) + (27/2673) * (27/2673) + (1370/2673) * (1370/2673) + (88/2673) * (88/2673) = 38.3\%$. This accuracy is what the model would guess by making a random guess for each of the nouns in the data. That is to say, for each noun, the model has a probability to guess that the noun belongs to a gender category based on the proportion of nouns belonging to that gender category. Conducting a random shuffling of gender labels of the nouns would also get this level of accuracy. If the accuracy of a classifier is above this baseline, it performs better than chance.

We also consider the majority baseline, which is generally higher (and thus harder) than the random baseline. The majority baseline represents what the model would obtain by making an informed guess and affiliating each noun in the test set to the largest gender category in the data. For example, in Chechen, the J gender category is equal to 51.3% (1370/2673) of the data. The classifier could thus reach an accuracy of 51.3% just by guessing that all nouns belong to the J gender category. Therefore,

the accuracy of the classifiers trained with the information of forms and/or semantics should at least exceed the accuracy of 51.3% to be considered as having good discriminatory power.

Second, precision and recall are used to assess the performance of the classifier on each of the gender categories in the target language. Precision assesses how many tokens are correct of all the tokens assigned to a gender category by the classifier, while recall evaluates how many tokens belonging to a gender category are correctly retrieved by the classifier. These measures are used in a similar way as Suppliance in Obligatory Context (SOC) and Target-Like Usage (TLU) in studies of language acquisition (Ting 2010). Using values such as precision and recall allows us to have a more precise understanding of which gender categories are more difficult to predict for the classifiers, as opposed to the accuracy, which only provides an overview of the performance of the classifiers.

4. Results

In the following subsections, we present the quantitative analyses for both Chechen and Tsova-Tush. For each language, three computational classifiers are trained and tested ten times: single decision tree, random forests, and neural networks. Each classifier is fed with three types of input data. First, the classifier is trained with information on form to predict gender in a given language. Second, the classifier is trained with semantic information to predict gender. Finally, the classifier is trained with both formal and semantic information.

4.1. Chechen

The accuracy of each combination of parameters is first compared to visualise the effect of form and semantics for predicting the gender of nouns. In Figure 2, each point represents the accuracy of a parameter across its ten iterations. First, we observe that the accuracy of a parameter does not vary much visually across the ten iterations, showing that the classifier is likely not influenced by random splits between the training and test sets. Second, we observe that the combination of formal and semantic information generally results in the highest accuracy for each classifier (random forests, single tree, neural networks). The performance of neural networks is not higher than decision trees, showing that the output of decision trees is capturing

as much information as can be captured in the variables of the data. Third, the accuracy of the classifiers is far above the random baseline, demonstrating that the interaction of the variables captured by decision trees can be further analysed.

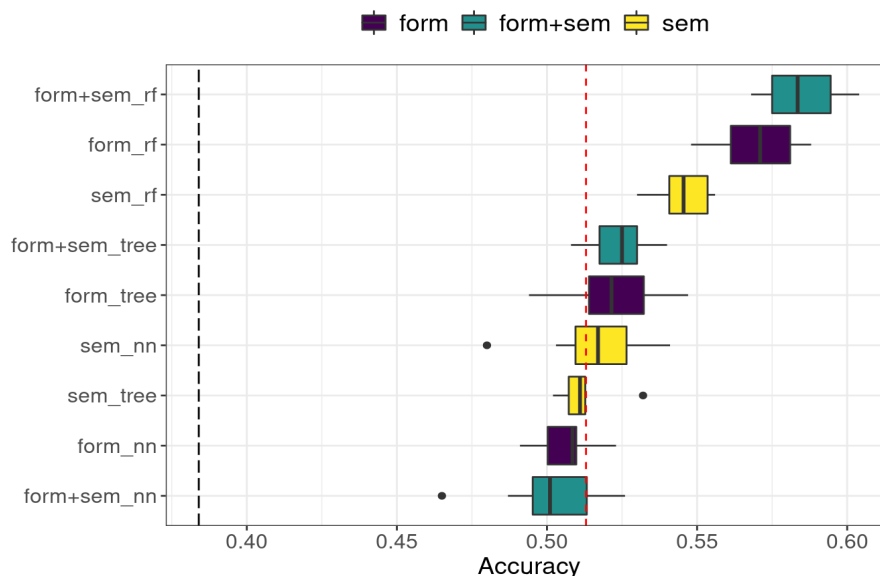


Figure 2: The accuracy of different parameters for predicting the gender of nouns in Chechen. The abbreviations are interpreted as follows: sem = semantics, rf = random forests, tree = single decision tree, nn = neural networks. The red dashed line indicates the majority baseline. The black dashed line indicates the random baseline.

Figure 2 only shows the overall accuracy of the model. However, it is also important to analyse how the classifiers perform on each gender category found in Chechen. For example, a classifier could have a high accuracy but only be good at predicting two gender categories, even when a language possesses more categories than that. Therefore, in Table 3 we visualise the precision and recall of each classifier on each gender category found in Chechen. We can observe that the classifiers have difficulties identifying items from the categories B and F. This is expected from a quantitative point of view, as these categories have a smaller number of tokens in the data. The detailed confusion matrices of the classifiers are provided in the supplementary materials.

Since the performance of combining formal and semantic information results in the highest accuracy for both single tree and random forests classifiers, we only display the output of combining formal and semantic information in this paper. However, the detailed output of each parameter with each classifier is provided in the supplementary materials.

Classifier	Setting	Mean Acc	Pr/Rc [B]	Pr/Rc [D]	Pr/Rc [F]	Pr/Rc [J]	Pr/Rc [M]
Tree	form	52.2 (51.2- 53.3)	0/0	0/14.7	0/0	53.1/92.5	0/0
Tree	sem	51.2 (50.6- 51.7)	0/0	0/5.3	0/0	52.0/96.5	0/8.6
Tree	form + sem	52.4 (51.6- 53.1)	0/0	0/7.2	0/0	52.3/97.1	0/12.2
RF	form	56.9 (56.0- 57.9)	46.5/1 3.6	50.5/4 5.4	0/0	60.5/78.7	61.7/4.7
RF	sem	54.5 (53.9- 55.2)	0/1.0	46.8/2 2.9	0/23.8	56.4/87.8	54.7/56.7
RF	form + sem	58.5 (57.6- 59.4)	49.8/9. 3	52.8/4 4.9	73.8/36. 2	61.0/80.8	64.0/33.9
NN	form	50.6 (49.9- 51.3)	26.1/2 0.9	44.3/4 0.9	0/0	58.9/68.0	0/0
NN	sem	51.7 (50.4- 52.9)	20.8/9. 3	43.3/3 5.7	0/17.3	58.5/72.8	49.0/44.7
NN	form + sem	50.1 (48.9- 51.4)	23.5/2 1.0	45.2/4 6.3	0/0	61.2/62.4	20.8/18.3

Table 3: The performance of the classifiers across ten replications ranked according to their mean accuracy when predicting gender in Chechen. The numbers in parentheses indicate the upper and lower confidence intervals of the accuracy. The abbreviations are interpreted as follows: Acc = accuracy; Pr = precision; Rc = recall. The values in bold indicate the parameters with the highest accuracy for each classifier.

In Figure 3, we first present a single decision tree that is generated when feeding both formal and semantic information to the classifier. Since the accuracy of the ten iterations does not vary much, we consider the last tree as an example. However, it is important to point out that this tree is displayed as a visualisation of how the classifier works rather than an absolute truth of gender assignment in the language under discussion. The tree can be read in the following way: the buckets at the bottom of the tree indicate with different colours the predicted gender. For example, the nouns predicted as M by the classifier are coloured in green. The numbers in each bucket indicate the number of predictions and how many of them are correct. As an example, in node 3 (bottom right) 206 nouns are predicted as J by the classifier, and of those 206 nouns, 173 do indeed belong to J gender. Each prediction is read starting from the top of the tree, descending until a bucket is reached. As another example, if the word is not a Russian loan (node 1 to node 2), but the semantic field is kinship (node 2 to node 5), the predicted gender is M. Following the path, 32 nouns are predicted as M, and 16 of them indeed belong to the M gender, which results in an accuracy of $16/32 = 50\%$. This accuracy is quite low, although it is important to emphasise that this accuracy should be compared with either the random baseline (38.4%) or the proportion of the predicted category. As an example, node 5 relates to gender M, which has a proportion of 3.3% in the entire data. The same reading can be applied for the other branches of the tree.

Semantic features are found in the tree: the tree considers the information of ‘Kinship’ (node 2) and ‘Person’ (node 8). As an example, if a noun is coded as possessing kinship semantics, it is more likely to be the M category. Formal features are found lower in the tree. For example, if a noun ends with /e/ (node 4), it is more likely to be from the J category. As another example, if a noun starts with /d/ (node 16), it is more likely to be from the D category (recall Section 2). Since different trees generated by the classifier could result in the use of different variables, we consider the forest of trees generated by the random forests classifier and extract the importance of variables based on three metrics.

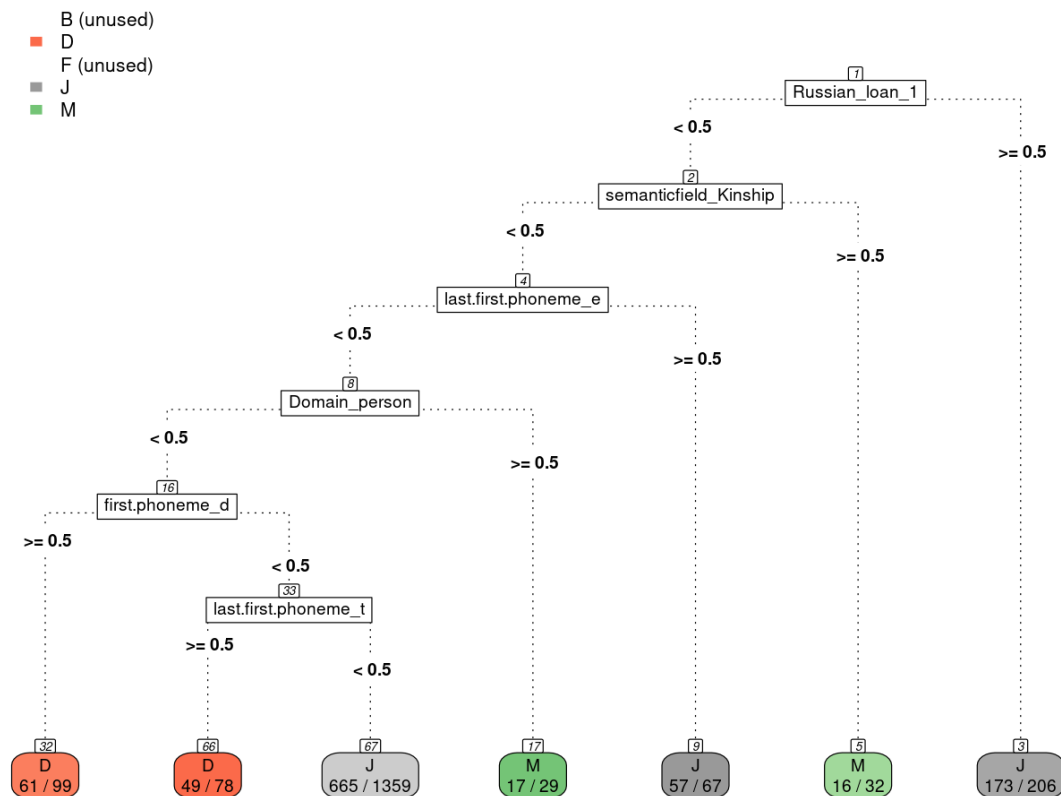


Figure 3: One of the ten decision trees generated for predicting gender in Chechen based on formal and semantic information. Labels marked as unused refer to categories that the models did not predict based on this sample tree.

First, we consider the minimal depth of a variable in a decision tree. For example, in Figure 3, the variable ‘Russian loan’ is at the top of the tree, which means that it has a depth of 0. As another example, the semantic category Kinship appears immediately after the root of the decision tree, which gives it a depth of 1. The closer a variable is found to the root, the larger group it creates within the data, which gives it higher importance. Second, we consider the decrease of accuracy for the classifier when a variable is removed. For example, if the tree has an accuracy of 60% with all variables considered, and the accuracy drops by 30% if we remove the semantic category of Person, it means that the semantic category Person is highly relevant for the classification task. Third, we consider the decrease of purity in predictions (i.e., the Gini coefficient). As an example for calculating the Gini coefficient, if the predictions of a node are all correct, it results in a high purity of the predictions and therefore a high Gini coefficient. In contrast, the predictions of node 5 have a lower purity/Gini coefficient, as they are correct at 50% (16/32). The top ten most

important variables according to these three metrics are shown in Table 4. The full ranking of the variables is provided in the supplementary materials.

Ranking	Minimap depth	Mean accuracy	decrease	Mean coefficient	decrease	Gini
1	Word length	Person		Word length		
2	Person.thing	Russian.loan_1		Russian.loan_1		
3	Russian.loan_0	Kinship		Person		
4	Russian.loan_1	First.phoneme_d		First.phoneme_d		
5	Person	Last.second.phoneme_u		Last.first.phoneme_i		
6	First.phoneme_d	Last.first.phoneme_e		Person.thing		
7	Last.first.phoneme_a	Ethnicity		Russian.loan_0		
8	Second.phoneme_a	Word length		Last.first.phoneme_a		
9	Third.phoneme_r	Last.first.phoneme_i		Third.phoneme_r		
10	Inanimate	Wild.plant		Kinship		

Table 4: The importance of variables in random forests according to different metrics for predicting gender in Chechen. Only the top ten variables for each metric are listed. The variables highlighted in grey are the variables that are found in the top ten of all three metrics.

Four variables are found consistently across the three rankings. First, borrowing from Russian has an effect on gender assignment. Second, the variable ‘Person’ shows that the human/inanimate distinction is relevant for grammatical gender assignment. Finally, in terms of form, we also see that word length is relevant, combined with the use of /d/ as the first phoneme. Additional linguistic interpretation of these variables is provided in Section 5.

4.2. Tsova-Tush

The same analysis was conducted for the Tsova-Tush data. First, we compare the accuracy of each combination of parameters. see Figure 4. As found with Chechen, we see that the accuracy does not vary much across the ten iterations, which indicates that the output of the models is stable. We can observe that, as in Chechen, combining

formal and semantic information results in the highest accuracy for each of the models. As an example, the accuracy of random forests is at its highest when combining formal and semantic information. The accuracy of all models is also above the majority baseline.

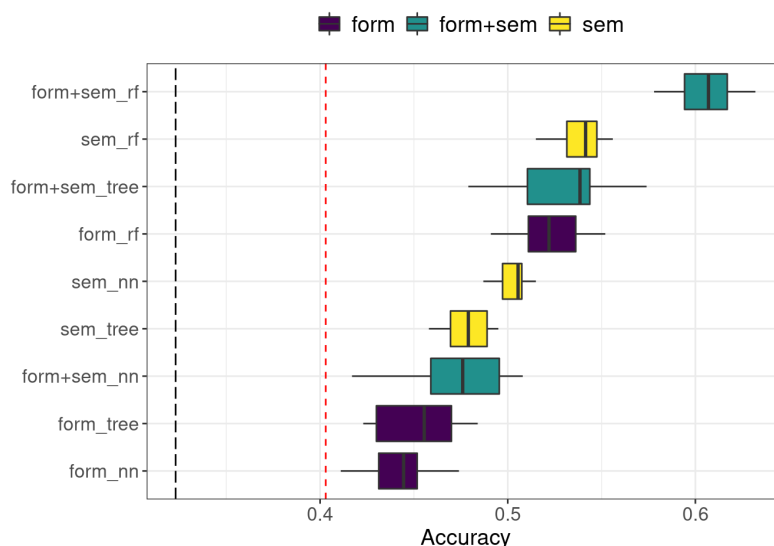


Figure 4: The accuracy of different parameters for predicting the gender of nouns in Tsova-Tush. The abbreviations are interpreted as follows: sem = semantics, rf = random forests, tree = single decision tree, nn = neural networks. The red dashed line indicates the majority baseline. The black dashed line indicates the random baseline.

The detailed performance of each classifier and each combination of parameters is shown in Table 5. In general, we see that the classifiers perform well on the M and F categories, but have more difficulties identifying the nouns from the B, D, and J categories. For example, the precision and recall of the gender category J are generally low, since the model did not correctly identify the majority of nouns belonging to this category. The detailed confusion matrices of the classifiers are provided in the supplementary materials.

As with Chechen, in the following analyses we only consider the models that combine formal and semantic features, since they have the highest accuracy. Nevertheless, the details regarding other combinations of parameters are included in the supplementary materials. In Figure 5, we visualise one of the ten decision trees generated based on formal and semantic information for predicting the gender of nouns in Tsova-Tush. The decision tree uses both formal and semantic variables. In terms of formal features, the tree considers the first phonemes /b/ and /n/, along

with the penultimate phoneme /a/, and word length (the cut-off point being five phonemes). With regard to semantic features, the tree uses the broad semantic categories of Male, Female, and Inanimate. It also considers the more specific category of ‘Wild plants’ and ‘The physical word’.

Classifier	Setting	Mean Acc	Pr/Rc [B]	Pr/Rc [D]	Pr/Rc [F]	Pr/Rc [J]	Pr/Rc [M]
Tree	form	45.2 (43.6-46.8)	55.7/15 .9	46.4/66 .8	0/0	41.8/43 .1	0/0
Tree	sem	47.9 (46.9-48.7)	0/0.6	50.4/59 .4	0/65.5	42.0/54 .1	100/100
Tree	form + sem	52.8 (50.7-54.9)	56.9/15 .9	52.9/63 .7	0/67.3	46.5/54 .0	100/100
RF	form	52.2 (50.9-53.4)	52.5/29 .1	52.1/70 .4	0/0	52.0/50 .6	85.7/13. 2
RF	sem	53.9 (52.9-54.7)	44.1/11 .7	54.0/64 .7	90.8/84 .9	49.2/57 .9	98.4/10 0
RF	form + sem	60.7 (59.4-61.9)	62.6/27 .1	58.5/71 .8	93.0/76 .7	57.6/60 .6	99.8/10 0
NN	form	44.4 (42.9-45.8)	33.6/25 .7	48.6/56 .6	0/1.7	45.0/47 .3	0/3.3
NN	sem	50.3 (49.7-50.9)	34.2/15 .0	53.8/53 .6	78.8/76 .7	46.0/59 .8	94.0/95. 3
NN	form + sem	47.4 (45.2-49.5)	33.0/33 .7	51.9/57 .8	0/24.8	49.2/42 .8	0/57.5

Table 5: The performance of the classifiers across ten replications ranked according to their mean accuracy when predicting gender in Tsova-Tush. The numbers in parentheses indicate the upper and lower confidence intervals of the accuracy. The abbreviations are interpreted as follows: Acc = accuracy; Pr = precision; Rc = recall. The values in bold indicate the parameters with the highest accuracy for each classifier.

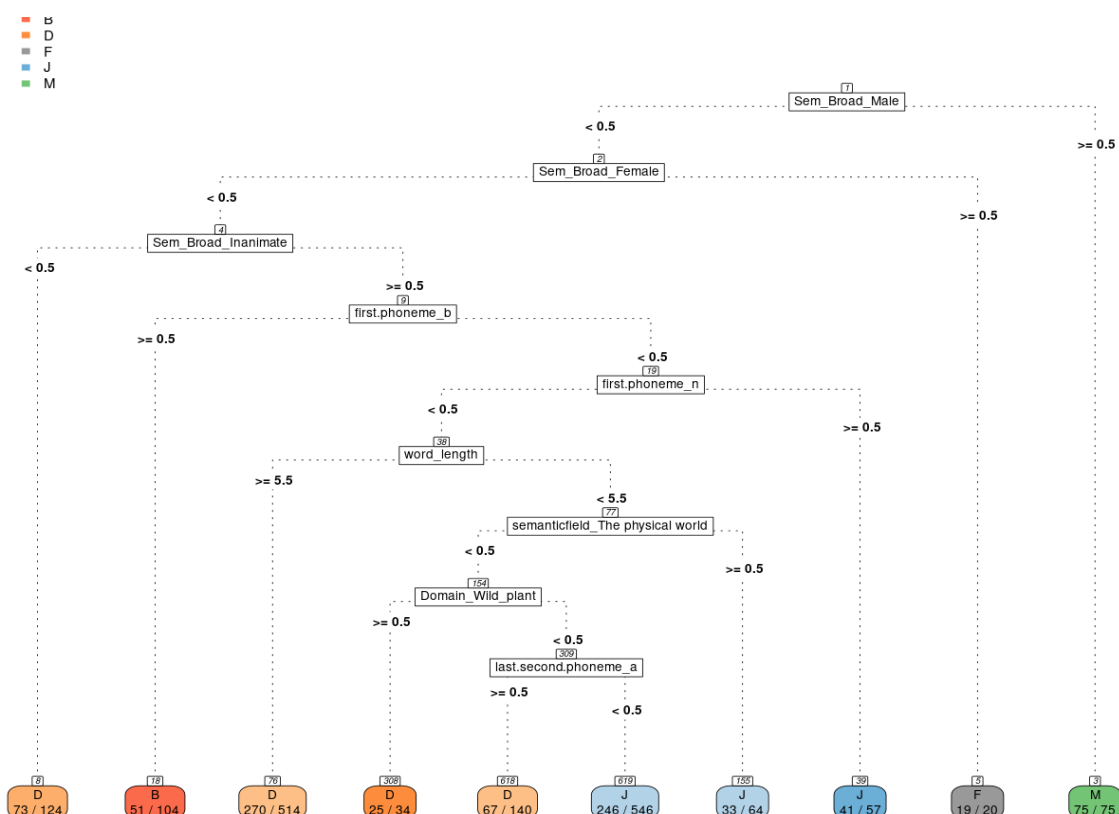


Figure 5: One of the ten decision trees generated for predicting gender in Tsova-Tush based on formal and semantic information.

We then consider the output from the random forests classifier to consider the overall importance of the variables when generating a forest of trees. We take into account the three metrics minimal depth, mean decrease of accuracy, and Gini coefficient. The top ten most important variables according to these three metrics are listed in Table 6. The full ranking of the variables is provided in the supplementary materials.

We can observe that the recurrent variables are word length, the first phoneme /b/, the penultimate phoneme /a/, along with the semantic information of male and inanimate. The information of Person also seems to be relevant, as it is encoded by different variables that are found in the rankings. For example, the metric of mean decrease of Gini coefficient has the Concepticon category ‘Person.thing’ in its top ten, while the rankings of accuracy and purity consider the dictionary information regarding person. Additional linguistic interpretation of these variables is provided in Section 5.

Minimap depth	Mean decrease in accuracy	Mean decrease in Gini coefficient
Word_length	Male	Male
Borrowed_GE	Inanimate	Word_length
Inanimate	First.phoneme_b	Inanimate
Person.thing	Word_length	Person
Last.second.phoneme_a	Female	Borrowed_GE
First.phoneme_b	Abstract	First.phoneme_b
Second.phoneme_a	Animal	Last.second.phoneme_a
Male	First.phoneme_d	Person.thing
Last.first.phoneme_r	Person	Female
First.phoneme_d	Last.second.phoneme_a	Animal

Table 6: The importance of variables in random forests according to different metrics for predicting gender in Tsova-Tush. Only the top ten variables for each metric are listed. The variables highlighted in grey are the variables that are found in the top ten of all three metrics.

5. Discussion

5.1. Methodological and technical aspects

In terms of accuracy, we observe that the classifiers generally perform above the random and the majority baseline, which indicates that interpreting the interaction of the variables detected by the classifiers is worthwhile. However, the accuracy stabilises at around 60%, which shows that the classifiers are still far from being able to perfectly predict the gender of nouns. While similar results are found in the literature, widening and deepening the variables is likely to improve the performance of the classifiers. On the one hand, additional variables related to form and semantics could provide additional information to the classifiers. For example, it is possible that adding frequency information could help the classifiers to find different rules for frequent and less-frequent words. On the other hand, the semantic information used in our experiments could be further refined. As an example, the semantic domains

could be annotated as individual variables rather than being merged as a single variable, as a noun may belong to different semantic categories simultaneously. Furthermore, additional tuning of the classifiers is also likely to improve their accuracy. In the current experiments, we did not tune the classifiers to find the parameters that fit best the gender classification task. Running additional tests to identify those parameters will also improve the accuracy of the classifiers. As an example, preliminary tests show that allowing the tree to split the data into smaller buckets (e.g., buckets of two units) and/or allowing the tree to go much deeper (i.e., have more branches) does not alter the general results. Nevertheless, additional testing could help to identify the settings that result in the highest possible accuracy.

5.2. Language-internal and family-internal aspects

For Tsova-Tush, the results in Figure 5 and Table 6 corroborate findings by Bellamy and Wichers Schreur (2022) and Wichers Schreur (2021): Firstly, in our classifier experiments, the variable “Male” has the most predictive power when all metrics are considered (Table 5), pointing to 100% accuracy in Table 6. It is known from corpus research (Wichers Schreur 2021) that the variable “Female” holds the same predictive power. That this is not shown in our outcomes is likely due to the low number of items in this category. Semantically, we find few other variables that serve as cues for gender assignment. Only the label “wild plant” points to gender D, which was not observed in previous literature, and the broad label “Inanimate” (after subtracting phonological clues and “wild plants”) points to gender J. In terms of phonology, we see that nouns starting in *b-* correlate with gender B, as expected from previous qualitative and descriptive studies. Also as in Tsova-Tush, in terms of phonology, we find words starting in *d-* to correlate with gender D.

The variable “word length” (i.e., possessing more than 5 phonemes) is an unexpected predictor, especially since it scores highly in all three metrics presented in Table 5. Word length corresponds highly with gender D, which we explain by the high percentage of abstract nouns formed by suffixes (increasing the word length), many of which trigger gender D (see Wichers Schreur 2021). Conversely, one suffix forming abstract nouns that triggers J gender is the suffix *-ol*, which was unexpectedly not used as variable by any of the decision trees. As expected from previous literature, loanword status is not used as a variable, as it does not correlate with any particular gender.

In our Chechen results, “word length” is similarly an important variable, although in this language it is correlated with gender J. It remains unclear whether there is a larger number of suffixes deriving J gender nouns in Chechen (as opposed to Tsova-Tush), or whether there are simply more attestations of such suffixes in our dataset; further research is thus necessary in this regard. Additionally, the majority of Russian loanwords (coded as such, see Section 3.1) in the dataset are assigned gender J. As indicated in Figure 1a, J is the largest gender category in Chechen, comprising just over half (51.3%) of the nouns included in the present study. It is hard to detect any formal (i.e. phonological) or semantic patterns in this group of loanwords, therefore it is tempting to suggest that they are being assigned to the unmarked gender (also referred to as a ‘default assignment strategy’; Bellamy & Parafita Couto 2022; Corbett 1991: 77; Poplack et al. 1982: 21-23). That said, in reference to German loanwords in Russian, Corbett (1991) argued that the morpho-phonological form of the former dictated their assignment to the appropriate declension classes of the latter. Most of the exceptions to the J gender assignment rule are animate nouns, which can more transparently be assigned a gender in the recipient language, especially since Chechen possesses clear M and F categories. Nonetheless, further investigation is needed to verify whether the inanimate exceptions, which represent 22% of Russian borrowings in this dataset, could be explained by semantic analogy or other factors (ibid.).

Both the Chechen and Tsova-Tush datasets contain clusters of items correlating strongly with one particular gender, which are not used by the decision tree algorithm and which score low on all importance metrics (Tables 4 and 6). Examples of this include abstract nouns in *-lo* (J) in Tsova-Tush, and nouns annotated as “Male” (M) in Chechen, and nouns annotated as “Female” (correlating with F) in both languages. The fact that these variables are not used can be explained by (i) the number of items in a category being too low; (ii) this cluster being split up into other categories that are deemed more important by the algorithm; or (iii) there was no need to explicitly separate this cluster in the decision tree, as it is subsumed under larger categories, most importantly “inanimate” and “word length”.

These results tie in well with hypotheses formulated elsewhere regarding gender assignment in East Caucasian. Particularly, they corroborate the observation that broad semantic categories such as animacy, humanness and abstractness show a high correlation with certain gender classes (Carling et al. 2021). On the other hand, whereas more fine-grained categories such as “animal”, and “metal” show a general tendency of correlating with certain gender classes in East Caucasian (B for animals,

D for metals), no such tendency has been found in our experiments. Animals in Nakh are divided between the three neuter classes, and metals do correlate with gender D, but form too small a category to be picked up by the models used here.

5.3. Broader linguistic implications

Other studies, both qualitative and quantitative, have shown that semantics plays a more important role in gender assignment than (phonological-morphological) form (see overview in Section 1), despite many systems relying on both features to different degrees (e.g. Corbett 1991). Indeed, we found that the form alone of a noun is always less predictive than purely its semantics, but that form and semantics together are always more informative, in all models. The Nakh languages therefore sit towards the semantic end of the spectrum of possible gender assignment types, albeit not as categorical as, for example, Mian (Allasonnière-Tang et al. 2021). An advantage of the computational methods presented here is that they can test and validate, or not, descriptive analyses of gender assignment systems. Moreover, they can also identify the relative weights of semantics and form in assignment, as well as the relative importance of the various semantic domains involved.

However, the prevalence of semantics-based assignment in the Nakh systems, amongst many others (such as Mian), seems difficult to reconcile with the observation that children pay more attention to phonetic cues when acquiring gender systems, especially those found noun-externally, namely the agreement markers that occur on other elements in the clause (see Gagliardi & Lidz 2014 for such evidence from Tsez, another Daghestanian language). This reliance on phonetic input is to be expected, since in the earliest stages of language acquisition, children do not have access to the meanings of all the input they receive. There appears, therefore, to be a discrepancy between the earliest stages of assignment and later processes, including the integration of loanwords, nonce words or neologisms into the language. Such a discrepancy begs the question as to what happens with regard to the processing and storage of gender assignment information between (early) childhood and adulthood. How can different cues gain greater prominence as age and language develop? More comparative studies of children of different ages and adults speaking the same language are required to investigate this issue further.

That said, certain semantic cues are likely to be prominent in gender assignment from early childhood, notably animacy. We have seen how Nakh languages have clear

masculine and feminine animate categories (labelled M and F respectively), and humanness is almost exceptionlessly diagnostic for gender assignment in many Indo-European languages. Moreover, there is evidence from code-switched speech that animacy is also the most important feature of a noun from one language inserted into a phrase or clause of another. Cruz (2021) highlights how English nouns representing feminine animates are assigned feminine gender when inserted into Spanish, despite there otherwise being a preference for a default masculine assignment strategy in inanimates (cf. Balam 2016 who finds feminine animates also assigned masculine gender in code-switching mode).

Finally, it should be stressed here that computational models of gender assignment do not necessarily represent mental classification, and an interdisciplinary approach to investigating gender assignment still remains to be undertaken.

6. Conclusion

The main aim of this paper was to show how three different computational classifier methods perform on a novel set of non-Indo-European data. We applied three machine-learning methods to investigate the relative weight of (phonological) form and semantics in predicting grammatical gender in the Nakh languages Chechen and Tsova-Tush. The results showed that the combination of form and semantics gives the best results for both languages, and that semantics is dominant in Tsova-Tush, which supports findings from existing literature. However, the results also suggest that making the coded semantic information more fine-grained could improve the accuracy of the gender predictions.

Our results confirm observations about the East Caucasian family, which relies heavily on humanness and abstractness as classifiers for gender assignment. Additionally, the first segment of a phonological form of a given noun being /b/ or /d/ is once again found to correlate highly with the corresponding genders B and D, respectively (see e.g. Nichols (1989; 2011: 147) for the concept of autogender, and its possible historical explanations).

As the many descriptive analyses demonstrate, gender assignment is language-specific, especially with respect to the specific semantic domains that emerge as important. Nonetheless, we have presented results here supporting the claim that certain common trends can be identified, notably the greater primacy of semantics as a categorisation principle, and within this, the importance of animacy. Whether this

principle holds across all gendered languages (that is, whether it can be considered ‘universal’) requires further empirical testing, from both qualitative and quantitative perspectives.

Acknowledgements

The authors would like to thank two anonymous reviewers and Francesca Di Garbo for their comments on earlier versions of this paper. The contributions by JWS were in part thanks to funding of the Linked Open Dictionaries project, an Early Career Research Group funded by the German Ministry for Education and Research. KB acknowledges funding from a Postdoctoral Mandate at KU Leuven, and the Dutch Research Council (NWO) Veni grant number VI.Veni.211C.077. MAT is thankful for the support from the French National Research Agency (ANR-20-CE27-0021). NR acknowledges funding from a doctoral contract granted by École Normale Supérieure - PSL, hosted at Sorbonne Nouvelle University (Paris).

Abbreviations

AOR = aorist

B = B gender

D = D gender

PRS = present

References

- Allasonnière-Tang, Marc & Dunstan Brown & Sebastian Fedden. 2021. Testing semantic dominance in Mian gender: Three machine learning models. *Oceanic Linguistics* 60(2). 302–334. <https://doi.org/10.1353/ol.2021.0018>
- Balam, Osmer. 2016. Semantic categories and gender assignment in Contact Spanish: Type of code-switching and its relevance to linguistic outcomes. *Journal of Language Contact* 9(3). 405–435. <https://doi.org/10.1163/19552629-00903001>
- Basirat, Ali & Marc Allasonnière-Tang & Aleksandrs Berdicevskis. 2021. An empirical study on the contribution of formal and semantic features to the grammatical gender of nouns. *Linguistics Vanguard* 7(1). 20200048. <https://doi.org/10.1515/lingvan-2020-0048>
- Bellamy, Kate & M. Carmen Parafita Couto. 2022. Gender assignment in mixed noun phrases: State of the art. In Dalila Ayoun (ed.), *The acquisition of gender: Crosslinguistic perspectives*. 14–48. Amsterdam / Philadelphia: John Benjamins.

- Bellamy, Kate & Jesse Wichers Schreur. 2022. When semantics and phonology collide: Gender assignment in mixed Tsova-Tush-Georgian nominal constructions. *The International Journal of Bilingualism* 26(3). 257–285.
<https://doi.org/10.1177/13670069211039559>
- Breiman, Leo & Jerome H. Friedman & Richard A. Olshen & Charles J. Stone. 1984. *Classification and regression trees*. Boca Raton: Routledge.
- Brown, Dunstan. 1998. Defining ‘sub-gender’: Virile and devirilized nouns in Polish. *Lingua* 104(3-4). 187–233.
- Carling, Gerd & Kate Bellamy & Jesse Wichers Schreur. 2021. *Gender stability in Nakh-Daghestanian*, paper presented at Languages, Dialects and Isoglosses of Anatolia, the Caucasus and Iran, March 2021, Paris.
- Contini-Morava, Ellen & Marcin Kilarski. 2013. Functions of nominal classification. *Language Sciences* 40. 263–299. <https://doi.org/10.1016/j.langsci.2013.03.002>
- Corbett, Greville. 1982. Gender in Russian: An account of gender specification and its relationship to declension. *Russian Linguistics* 6(2). 197–232.
- Corbett, Greville. 1991. *Gender*. Cambridge: Cambridge University Press.
- Corbett, Greville. 2013. Systems of gender assignment. In Matthew S. Dryer & Martin Haspelmath (eds.), *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at <http://wals.info/chapter/32>, Accessed on 2022-02-02)
- Corbett, Greville G. & Norman M. Fraser. 1993. Network Morphology: A DATR account of Russian nominal inflection. *Journal of Linguistics* 29(1). 113–142.
<https://doi.org/10.1017/S0022226700000074>
- Corbett, Greville & Norman M. Fraser. 2000. Default genders. In Barbara Unterbeck & Matti Rissanen (eds.), *Gender in grammar and cognition I: Approaches to gender*. 55–98. Berlin: Mouton de Gruyter.
- Cruz, Abel. 2021. A syntactic approach to gender assignment in Spanish–English bilingual speech. *Glossa: a journal of general linguistics* 6(1). 1–40.
<https://doi.org/10.16995/glossa.5878>
- Desheriev, Y. D. [Дешериев]. 1953. *Vacbijskij jazyk: fonetika, morfologija, sintaksis, leksika* [The Tsova-Tush language: phonetics, morphology, syntax, lexicon]. Moscow: Izdatel'stvo AN SSSR.
- Demsar, Janez & Blaz Zupan & Gregor Leban & Tomaž Curk. 2004. Orange: From experimental machine learning to interactive data mining, white paper. *European Conference of Machine Learning: 2004; Pisa, Italy* 3202. 537–539.

- Dryer, Matthew S. 2013. Prefixing vs. suffixing in inflectional morphology. In Matthew S. Dryer & Martin Haspelmath (eds.), *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at <http://wals.info/chapter/33>, Accessed on 2022-02-02)
- Evans, Nicholas & Dunstan Brown & Greville Corbett. 2002. The semantics of gender in Mayali: Partially parallel systems and formal implementation. *Language* 78(1). 109–153.
- Evans, Roger & Gerald Gazdar. 1989a. Inference in DATR. *Proceedings of the fourth conference of the European Chapter of the Association for Computational Linguistics, Manchester, England*. 66–71.
- Evans, Roger & Gerald Gazdar. 1989b. The semantics of DATR. In A. G. Cohn (ed.), *Proceedings of the seventh conference of the Society for the Study of Artificial Intelligence and Simulation of Behaviour*, 79–87. London: Pitman/Morgan Kaufmann.
- Evans, Roger & Gerald Gazdar. 1996. DATR: A language for lexical knowledge representation. *Computational Linguistics* 22(2). 167–216.
- Fedden, Sebastian. 2011. *A Grammar of Mian*. Berlin / Boston: Mouton de Gruyter.
- Fraser, Norman M. & Greville G. Corbett. 1995. Gender, animacy, and declensional class assignment: A unified account for Russian. In Geert Booij & Jaap van Marle (eds.), *Yearbook of Morphology 1994*, 123–150. Amsterdam: Kluwer Academic Publishers.
- Fraser, Norman M. & Greville G. Corbett. 1997. Defaults in Arapesh. *Lingua* 103(1). 25–57.
- Gagliardi, Annie & Jeffrey Lidz. 2014. Statistical insensitivity in the acquisition of Tsez noun classes. *Language* 90(1). 58–89. <https://doi.org/10.1353/lan.2014.0013>
- Haykin, S. 1998. *Neural networks: A comprehensive foundation*. Prentice-Hall: Englewood Cliffs.
- Her, One-Soon & Marc Tang. 2020. A statistical explanation of the distribution of sortal classifiers in languages of the world via computational classifiers. *Journal of Quantitative Linguistics* 27(2). 93–113. <https://doi.org/10.1080/09296174.2018.1523777>
- Hockett, Charles F. 1958. *A course in modern linguistics*. New York: MacMillan.
- Jamalkhanov, Z. D. [Джамалханов] & Aliroev, I. Y. [Алироев]. 1991. *Slovar' pravopisanija literaturnogo čečenskogo jazyka* [Orthographical dictionary of literary Chechen]. Grozny: Kniga.

- Kadagidze, E. [ქადაგიძე]. 2009. *C'ova-tušuri t'ekst'ebi* [Tsova-Tush texts]. Tbilisi: TSU gamomcemloba.
- Kadagidze, D. & N. Kadagidze. [ქადაგიძე]. 1984. *C'ova-tušur-kartul-rusul leksik'oni* [Tsova-Tush-Georgian-Russian dictionary]. Tbilisi: Mecniereba.
- Karmiloff-Smith, Annette. 1979. *A functional approach to child language*. New York / London: Cambridge University Press.
- Khalilov, M. S. [Халилов]. 1999. *Cezsko-russkij slovar'* [Tsez-Russian dictionary]. Moscow: Academia
- Lemus-Serrano, Magdalena & Marc Allasonnière-Tang & Dan Dediú. 2021. What conditions tone paradigms in Yukuna: Phonological and machine learning approaches. *Glossa: a journal of general linguistics* 6(1). 60. <https://doi.org/10.5334/gjgl.1276>
- List, Johann-Mattis & Michael Cysouw & Robert Forkel. 2016. *Concepticon: A resource for the linking of concept lists*. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). 2393–2400.
- Matsiev, A. G. [Мациев]. 1961. *Slovar' čečenskogo jazyka* [Chechen dictionary]. Moscow: Gosudarstvennoe izdatel'stvo inostrannyx i nacional'nyx sloverej.
- Nichols, Johanna. 1989. The Nakh evidence for the history of gender in Nakh-Daghestanian. In Howard I. Aronson (ed.), *The non-Slavic languages of the USSR: linguistic studies*, 158–175. Chicago: Chicago Linguistic Society, University of Chicago.
- Nichols, Johanna. 1994. Chechen. In Riks Smeets (ed.) *The North East Caucasian languages, part 2*, 1–78. Delmar: Caravan.
- Nichols, Johanna. 2003. The Nakh-Daghestanian consonant correspondences. In Dee Ann Holisky & Kevin Tuite (eds.), *Current trends in Caucasian, East European and Inner Asian linguistics: Papers in honor of Howard I. Aronson*, 207–264. Amsterdam: John Benjamins.
- Nichols, Johanna. 2007. Chechen morphology with notes on Ingush. In Alan S. Kaye (ed.), *Morphologies of Africa and Asia*, 1188–1207. State College: Penn State University Press. <https://doi.org/10.1515/9781575065663-044>
- Nichols, Johanna. 2011. *Ingush grammar*. Berkeley, Los Angeles: University of California Press.
- Parks, Randolph & Daniel S. Levine & Debra L. Long. (eds.). 1998. *Fundamentals of neural network modeling: Neuropsychology and cognitive neuroscience*. Boston: MIT Press.

- Plaster, Keith & Maria Polinsky & Boris Harizanov. 2013. Noun classes grow on trees: Noun classification in the North-East Caucasus. In Balthazar Bickel, Lore A. Grenoble, David A. Peterson & Alan Timberlake (eds.), *Language typology and historical contingency: In honor of Johanna Nichols*, 153–170. Amsterdam: John Benjamins.
- Polinsky, Maria & Ezra Van Everbroeck. 2003. Development of gender classifications: Modeling the historical change from Latin to French. *Language* 79(2). 356–390.
- Quinlan, J. Ross. 1993. *C4.5: Programs for Machine Learning*. Burlington: Morgan Kaufmann Publishers.
- Quinlan, J. Ross. 1996. Improved use of continuous attributes in c4.5. *Journal of Artificial Intelligence Research* 4. 77–90.
- Rajabov, Ramazan. Undated. *Tsez Dictionary*. Unpublished MS (Los Angeles: University of Southern California).
- Senft, Gunter (ed.). 2000. *Systems of nominal classification*. Cambridge: Cambridge University Press.
- Sokolik, M. E. & Michael E. Smith. 1992. Assignment of gender to French nouns in primary and secondary language: A connectionist model. *Second Language Research* 8(1). 39–58.
- Tagliamonte, Sali A. & R. Harald Baayen. 2012. Models, forests and trees of York English: *Was/were* variation as a case study for statistical practice. *Language Variation and Change* 24(2). 135–178.
<https://doi.org/10.1017/S0954394512000129>
- Ting, Kai Ming. 2010. Precision and recall. In Claude Sammut & Geoffrey I. Webb (eds.), *Encyclopedia of Machine Learning*, 781–781. Boston: Springer.
https://doi.org/10.1007/978-0-387-30164-8_652
- Ulrich, Natalja & Marc Allasonnière-Tang & François Pellegrino & Dan Dediú. 2021. Identifying the Russian voiceless non-palatalized fricatives /f/, /s/ and /ʃ/ from acoustic cues using Machine Learning. *Journal of the Acoustical Society of America* 150(3). 1806–1820. <https://doi.org/10.1121/10.0005950>
- Wichers Schreur, Jesse. 2021. Nominal borrowings in Tsova-Tush (Nakh-Daghestanian, Georgia) and their gender assignment. In Diana Forker & Lore A. Grenoble (eds.), *Language contact in the territory of the former Soviet Union*, 15–33. Amsterdam: John Benjamins.

Wurm, S. A. & I. Heyward & Unesco. 2001. *Atlas of the world's languages in danger of disappearing*. Paris: Unesco Pub. Website <http://www.unesco.org/languages-atlas/> consulted on 7-12-2021.

CONTACT

jesse.jessews@gmail.com

marc.allasonniere-tang@mnhn.fr

k.r.bellamy@hum.leidenuniv.nl

neige.rochant@sorbonne-nouvelle.fr

Cross-linguistic sources of anticausative markers

GUGLIELMO INGLESE

UNIVERSITÀ DI TORINO

Submitted: 13/01/2022 Revised version: 11/10/2022

Accepted: 11/10/2022 Published: 22/12/2022

Abstract

The (anti)causative alternation, that is, the alternation whereby languages contrast intransitive verbs expressing spontaneous events with transitive ones expressing externally caused events, has been the object of extensive language-specific and cross-linguistic studies. Within this type of alternation, marking on the intransitive member goes under the name of anticausative marking, while marking on the transitive member is causative marking. Historical research has mostly focused on causatives, while the diachrony of anticausative markers has largely been neglected. In the literature, only two possible cross-linguistic sources of anticausatives are mentioned: reflexive and passive markers. In this paper, I explore the sources of anticausative markers in a sample of 98 languages and show that they are much more varied than what is currently reported in the literature. Taking this richer diachronic evidence into account also sheds light on some yet controversial aspects concerning the relationship between anticausativization and reflexivity.

Keywords: diachronic typology, anticausative marking, reflexive, grammaticalization

1. Introduction

With the term *anticausative alternation* linguists refer to the way in which languages express events that are construed as coming about spontaneously as opposed to those that are construed as caused by an external entity.¹ In this paper, I follow Haspelmath's (2016: 37) proposal and refer to the two members of this alternation as NONCAUSAL vs. CAUSAL verb forms, respectively. Decades of research on this topic have shown that languages resorts to a wide array of morphosyntactic strategies to express

¹ By alternation, I refer to the possibility of individual verbs to occur in multiple argument structure constructions or valency frames (see Levin 1993; Malchukov 2015).

this alternation. Overt marking on the causal member of the alternation (e.g. Hittite *ze-* ‘cook (intr.)’ vs. *zai-nu-* ‘cook (tr.)’) goes under the name of *causative marking*, whereas overt marking on the noncausal verb (e.g. Russian *serdit* ‘make angry’ vs. *serdit’-sja* ‘be/get angry’) is referred to as *anticausative marking* (Nedjalkov & Silnitsky 1973: 2).

Diachronic studies have mostly focused on causative markers, both in terms of their possible (lexical) sources (Song 1996: Chap. 3; Zúñiga & Kittilä 2019: 220–221) and in their connections to other valency operations, such as passives and applicatives (Bahrt 2021: Chap. 7). Less attention has been paid to the diachronic typology of anticausative markers (henceforth, AMs), perhaps also due to the well-known fact that anticausatives are cross-linguistically less frequent than causatives (Nichols et al. 2004: 162; Zúñiga & Kittilä 2019: 53; Bahrt 2021: 147). In reference works such as Zúñiga & Kittilä (2019: 233) and Kuteva et al. (2019: 363) only two possible sources of AMs are mentioned: reflexive and passive markers. In fact, while extensive work has been carried out on the REFLEXIVE > ANTICAUSATIVE development, the lexical sources of AMs remain largely underexplored. Moreover, even with respect to the relationship between AMs and other voice markers there is evidence for developments that go against the directionalities commonly discussed in the literature (Bahrt 2021: Chap. 7).

Taking stock of these premises, this paper aims to fill this gap and offer an overview of the diachrony of AMs in the languages of the world. The work is couched in the framework of diachronic typology, understood here as the sub-field of linguistic typology where cross-linguistic research on linguistic phenomena meets historical linguistics and grammaticalization studies (e.g. Sansò 2017, 2020; Cristofaro 2021). While cross-linguistic data indeed points towards a robust connection between reflexives and anticausatives, working with a convenience sample of 98 languages I show that reflexives are by no means the only available source of AMs. This richer diachronic evidence can also contribute to clarifying long-lasting debates on the relationship between anticausatives and reflexives.

The paper is structured as follows. In Section 2, I briefly review the typological literature on the anticausative alternation and illustrate its morphosyntactic realizations. Section 3 illustrates the diachronic development of reflexives into anticausatives. Section 4 is devoted to non-reflexive sources of AMs. After a presentation of the sample (Section 4.1), I turn to discussing each type of source in some detail in Sections 4.2 to 4.7. Section 5 discusses the findings of the previous

section and offers an overview of how AMs come about (5.1-5.2) and their relationship to reflexives (5.3). Section 6 features the conclusions of this work.

2. The (anti)causative alternation

A decades-long body of research on the (anti)causative alternation has unveiled much of the morphosyntactic and semantic variation that exists in this domain within and across languages (see Tubino-Blanco 2020 for an overview), both in formal (e.g. Schäfer 2008, 2009; Alexiadou 2010; Alexiadou et al. 2015) and in functional/typological frameworks (e.g. Nedjalkov & Silnitsky 1973; Haspelmath 1987, 1993b, 2016; Levin 1993; Levin & Rappaport Hovav 1995; Nichols et al. 2004). As Schäfer (2009: 641) puts it:

the causative alternation is characterized by verbs that have an intransitive as well as a transitive use, where the intransitive use typically denotes a change-of-state event undergone by some entity and the transitive use denotes that this change-of-state event has been brought about or caused by some different entity.

Syntactically, this alternation involves a transitivity shift, as noncausal verbs are typically intransitive whereas causal ones are transitive (see also Alexiadou *et al.* 2015).

Cross-linguistic research has shown that languages display a variety of formal strategies to encode this alternation. These can be subsumed under a few general types based on the locus of marking (Nedjalkov & Silnitsky 1973; Nichols et al. 2004). The main types are the *anticausative pattern*, the *causative pattern* and the *equipollent pattern*.² In the anticausative pattern, an unmarked causal verb is opposed to a marked noncausal counterpart, as in (1).³ By contrast, in the causative pattern the causal verb form carries overt marking while the noncausal verb is morphologically simple, as in (2). Finally, in the equipollent pattern both the causal and the noncausal verb forms are equally marked, as in (3).

² Other patterns include labiality, as in the case of English *break* (tr./intr.), and suppletion, as e.g. English *kill* vs. *die* (see Nichols et al. 2004). Since neither of these offers evidence for overt AMs, I will exclude them from this study.

³ Glosses and translations generally reproduce those of the sources, with a few adjustments. In all examples, the AM is in bold and is consistently glossed as ANTC, irrespective of the original glossing in the source.

(1) ANTICAUSATIVE: Kammu (Austroasiatic; Zúñiga & Kittilä 2019: 49)

- a. *ʔòʔ p̄ir tóʔ* CAUSAL
 I shake table
 ‘I shake the table.’
- b. *tóʔ hm-p̄ir* NONCAUSAL
 table ANTC-shake
 ‘The table shakes.’

(2) CAUSATIVE: Turkish (Turkic; Zúñiga & Kittilä 2019: 16)

- a. *Hasan öl-dü*
 H.(NOM) die-PST
 ‘Hasan died.’
- b. *Ali Hasan-i öl-dür-dü*
 A.(NOM) H.-ACC die-CAUS-PST
 ‘Ali killed Hasan.’

(3) EQUIPOLLENT: Yaqui (Uto-Aztecan; Tubino-Blanco 2020: 19)

- a. *U kari bee-te*
 DET.NOM house.NOM burn-ANTC
 ‘The house is burning.’
- b. *Joan kari-ta bee-ta*
 J. house-ACC burn-CAUS
 ‘Juan is burning the house.’

In this paper, I will focus on AMs, that is, on markers that occur on the noncausal member in patterns such as (1) and (3). Anticausative marking on the verb can also be described as *anticausative voice* and anticausativization can be defined as the diathesis in which the Agent participant is removed from both the semantic and the syntactic valency of bivalent verbs and the Patient is encoded as subject (Zúñiga & Kittilä 2019: 41–53; Bahrt 2021: 37). The main semantic effect of anticausativization is that, the Agent being removed from the semantic valency of the verb, the event is construed as coming about spontaneously (Haspelmath 1993b: 90). Agent removal is

also the key difference between anticausatives and passives, as the latter still include the Agent in the event frame, as shown by the fact that some languages allow its expression via oblique phrases (Zúñiga & Kittilä 2019: 83).

AMs show notable restrictions with respect to the verb bases that they may apply to (see Cennamo et al. 2015: 680–681; Alexiadou et al. 2015: 20–23, 52–56; Tubino-Blanco 2020). First, anticausativization mostly concerns change-of-state predicates (Schäfer 2009; Alexiadou et al. 2015: 53), while other aspectual classes, such as atelic activity verbs, are only marginally included (e.g. Cennamo 2012; Cennamo et al. 2015). Most importantly, transitive verbs that lexicalize what Haspelmath (1987: 12) calls *agent-oriented meaning components*, that is, that have a lexically specified manner and/or causer, such as *cut* or *murder*, are excluded from the anticausative alternation (Koontz-Garboden 2009; Rappaport Hovav & Levin 2010).

Anticausative verbs can semantically be split up in two main classes: *decausative* and *autocausatives* or *endoreflexive* verbs (Geniušienė 1987: 86–89, 98–104; Haspelmath 1987: 27; Greissels 2006: 10). The main semantic difference between the two concerns control: decausative verbs involve (inanimate) participants that undergo an uncontrolled change of state (e.g. *melt*, *fall sick*) (see also Haspelmath 2016), whereas autocausatives involve (animate) participants that are conceived as partly controlling (at least the onset of) the event. In this respect, autocausative are semantically closer to reflexives, and typically include verbs of self-induced motion such as *mobilize* and *rise* (Geniušienė 1987: 87; Cennamo et al. 2015: 680). Crucially, as already noted by Geniušienė (1987: 108), some verbs allow both a decausative and an autocausative reading, as is the case of Lithuanian *kelia* ‘rise’ in (4a) and (4b), respectively:

(4) Lithuanian (Indo-European; Geniušienė 1987: 108)

- a. *kelia-si rūk-as*
lifts-ANTC fog-NOM
‘The fog lifts.’
- b. *zmones kelia-si*
people lift-ANTC
‘People get up.’

Finally, it has repeatedly been pointed out that AMs are often polyfunctional and typically also express other valency changing functions such as passive, reflexive, reciprocal and antipassive (thus already Nedjalkov & Silnitsky 1973: 22–24). In particular, drawing from a sample of 222 languages, Bahrt (2021: 147, 150) found that 48 languages feature anticausative syncretic markers whereas only 32 display dedicated AMs. Among syncretic patterns, anticausatives are most often co-expressed with reflexives and less so with passives (Bahrt 2021: 162; Inglese 2022a: 508).

3. AMs from reflexives

The reflexive and the anticausative diatheses are frequently co-expressed in the world's languages and this recurrent polyfunctionality pattern has diachronically been explained as the result of the grammaticalization of reflexive markers into AMs. A well-studied case is that of the development of Proto-Indo-European reflexive **swe* into the reflexive-anticausative marker of several Romance, Germanic, and Balto-Slavic languages (this is actually part of a wider development of reflexives into middle markers, see Geniušienė 1987; Kemmer 1993: 151–193; Cennamo 1993; Holvoet 2020 for discussion). For example, while Latin *se* essentially functioned as a reflexive marker, as in (5), its French continuant *se* also acquired an anticausative function (Heidinger 2010), as in (6), in which *se fondre* 'melt (intr.)' is the anticausative counterpart of transitive *fondre* 'melt (tr.)'.⁴

(5) Latin (Indo-European; Pinkster 2015: 262)

occido 'kill' → *se occidere* 'kill oneself'

(6) French (Indo-European; Heidinger 2010: 80)

ils se fondent aux rayons du soleil [...]

3PL ANTC melt.PRS.3PL in.DEF.3PL ray.PL of sun

'They [= hearts] melt in the rays of the sun'.

⁴ Anticausative usages of *se* occur already in Latin, so that the REFLEXIVE > ANTICAUSATIVE shift must have happened at an early date (Cennamo et al. 2015; Martínez Rojas et al. 2021). I refer to Cennamo (2020a; 2020b) for a more extensive discussion of the development of reflexives from Latin to Romance languages.

Reflexives are only compatible with agentive subjects, whereas anticausatives typically involve non-agentive subjects. This is why the development from reflexive to anticausative is semantically described as entailing a loss of agency and control restrictions (Haspelmath 1993b: 44; Heine 2002: 89; Heidinger 2010: 55–65). In this shift, a key role is arguably played by motion verbs, which, as discussed above for example (4), are compatible with both autocausative and decausative interpretations (Holvoet 2020: 118–119; Inglese 2020: 236–237). Once reflexives extend to motion verbs, they develop into autocausative markers, and they can subsequently be generalized as AMs with all verb bases, including decausatives. This means that the shift REFLEXIVE > ANTICAUSATIVE can more accurately be described as REFLEXIVE > AUTOCAUSATIVE > DECAUSATIVE (Haspelmath 1987: 29). A sketch of this development can be found in (7).

- (7) STAGE I: reflexive marking with typically two-participant events, e.g. *hit oneself* [+ control, + animacy]
STAGE II: reflexive marking extended to verbs of motion, which are also compatible with inanimate subjects, e.g. *move oneself/itself* [\pm control, \pm animacy]
STAGE III: reflexive used in decausative contexts proper, e.g. *melt (itself)* [-control, \pm animacy]

4. Non-reflexive sources of AMs

The development proposed in (7) is supported by abundant cross-linguistic evidence (Bahrt 2021: 173–175), to the extent that the REFLEXIVE > ANTICAUSATIVE shift is often described as a typologically frequent and unidirectional grammaticalization pathway (e.g. Kuteva et al. 2019: 363).

Non-reflexive sources of AMs have occasionally been mentioned in language-specific studies, but these findings have not yet made their way into the typological literature (but see Haspelmath 1987 for an early discussion). In fact, besides the possibility of anticausatives deriving from passives, no source other than reflexive is mentioned by Zúñiga & Kittilä (2019: 233) and Kuteva et al. (2019: 363). Only recently has Bahrt (2021: Chap. 7) brought together some evidence suggesting the possibility of AMs deriving from non-reflexive voice markers. In the remainder of this section, I discuss evidence for non-reflexive sources of AMs.

4.1. Data and methods

For this paper, I investigated the sources of AMs in a sample of 98 languages, for a total of 112 markers (languages may feature more than one AM). This is a convenience sample that includes data from 47 language families plus 9 isolates (see Appendix). Since this is not a variety sample, I will refrain from making quantitative-based generalizations from it.

I have included in my sample only AMs that comply with the following criteria: (i) they occur on the noncausal member in either anticausativizing or equipollent verb pairs; (ii) if they are syncretic, they do not encode reflexivity; (iii) they must be used with at least decausative verbs (i.e., with inanimate non-controlling Patients).

Let me briefly comment upon the choice of these criteria. According to some scholars (e.g. Haspelmath 2016: 39 fn. 5), the label AM should be restricted to markers on noncausal verbs in anticausative patterns proper, as in (1). In this paper, I rather follow Bahrt (2021: 38–39) and extend the definition of AM so as to include markers of noncausal verbs in equipollent pairs, as in (3). The reason to do so is that there is little evidence that equipollent noncausal markers are different in nature than those found in anticausative patterns proper. In fact, one finds languages in which the same marker can be used in both equipollent and anticausative patterns, so that the choice of the pattern is ultimately independent from the marker itself (see examples in Sections 4.5.2 and 4.7.4).

Since I am interested in exploring non-reflexive sources of AMs, criterion (ii) is meant to exclude those markers that synchronically also function as reflexives, because with these the widespread assumption is that the reflexive function must be historically prior (Section 3; but see Section 4.7.4 for a critical discussion). Note that I am aware that lack of a synchronic reflexive function does not necessarily exclude a reflexive origin, as this original function might simply have been lost in the course of time. A case in point is the suffix *-əm* in the Musqueam dialect of Halkomelem. This suffix productively occurs, among other things, in both anticausative and antipassive function, as in (8a-b), but never in prototypical reflexive contexts.

(8) Musqueam, Halkomelem (Salishan; Suttles 2004: 230–231)

- a. *hánək^w-əm* ‘get warm’ vs. *hánək^w-t* ‘warm something up’
 b. *sól-əm* ‘spin (wool)’ vs. *sól-ət* ‘spin something’

However, as discussed by Gerdts & Hukari (2006), there are good reasons to believe that the Halkomelem $-(ə)m$ ultimately goes back to a reflexive source. Once the original meaning of the suffix was progressively bleached as it extended to new functions, including anticausative and antipassive, it was replaced in reflexive function by a newly created reflexive $-θət$ (on this type of reflexive renewal see Kemmer 1993: Chap 5).⁵ For this reason, I have excluded from my sample also AMs which synchronically do not also function as reflexive markers, but for which a reflexive origin is nonetheless likely.

Criterion (iii) is meant to exclude constructions that do not encode prototypical anticausativization (see Zúñiga & Kittilä 2019: 43–48). For example, a construction that I have excluded from my sample is the Yakkha marker $-siʔ$. This suffix does function as an intransitivizer, but it is consistently associated with involuntary or unintentional actions and is restricted to animate subjects (Schackow 2015: 307–309), as in (9). Similarly, I have also excluded intransitivizers that only derive resultative/stative verbs, e.g. Ket (Yeniseian) $-jə-$ (Vajda 2015: 664).

(9) Yakkha (Sino-Tibetan; Schackow 2015: 307)

mendhwak = ci phaps-a-sy-a-ci

goat = NSG entangle-PST-INTR-PST-[3]DU

‘The two goats lost their way.’

AMs that I have identified thanks to the criteria (i) – (iii) fall into morphosyntactic types already well-known from previous cross-linguistic studies (e.g. Haspelmath 1990; Nichols et al. 2004). These include both analytic (auxiliaries, clitics) and synthetic strategies (affixation, morphophonological alternations, conjugation class change). In total, I have collected 112 AMs: the mismatch between the number of languages and that of markers is due to the fact that some languages feature more than one AM. This is the case of Huarjio (Uto-Aztecan), where both the suffixes $-i$ and $-pa$ function as AMs (Félix Armendáriz 2005: 222–228).

As is it often the case in studies of diachronic typology, for most of the languages under analysis one cannot rely on the necessary historical documentation to track down the actual source of AMs. Instead, one must operate with indirect evidence

⁵ The reflexive suffix $-θət$ also extended to autocausative contexts and to the derivation of change-of-state verbs from stative roots, e.g. *nás* ‘fat’ → *nás-θət* ‘get fat’, thus intruding into the anticausative domain (Suttles 2004: 244).

based on comparative and/or internal reconstruction (Sansò 2020: 407–408). Even then, for several AMs a specific lexical source cannot be pinned down with certainty. For example, Siar-Lak (Austronesian) features an anticausative prefix *ta(k)-* (Frowein 2011: 274–279). Comparative evidence shows that this prefix goes back to a reconstructable Proto-Oceanic prefix **ta-*, for which however no further lexical source can be reconstructed (Pawley 1972: 39). Similarly, Zenzontepec Chatino (Otomanguean) has an anticausative suffix *-y* (Campbell 2015: 1408–1409). This suffix is a continuant of a reconstructable Proto-Zapotecan intransitivizing suffix **-i*, which however lacks further etymology (Campbell 2011: 238). I have treated these and similar cases as having an unknown source, on par with cases in which no etymology for a given AM is given.

The overview of the sources of AMs attested in my sample is given in Table 1. For the sake of exposition, sources are grouped in six main groups: lexical verbs, spatial markers, spontaneous events, aspectual markers, nominalizations and verbalizations and non-reflexive voice markers (plus a seventh group for unknown sources). In the next sections, I shall discuss each group in more detail. In particular, I focus on the semantic features of the source construction that might have triggered the emergence of the anticausative function and discuss possible bridging contexts for such shifts (see Heine 2002; Sansò 2017).

Type of source	Frequency
Lexical verbs	32
Spatial markers	2
Spontaneous events markers	2
Aspectual markers	6
Nominalizers and verbalizers	3
Non-reflexive voice markers	4
Unknown source	63

Table 1: Sources of AMs

4.2. Lexical verbs > AMs

Lexical verbs constitute a well-known source of voice marking (see Kuteva et al. 2019, e.g. EAT > PASSIVE, DO > CAUSATIVE), so that it comes as no surprise that they may also turn into AMs. This shift follows the well-known grammaticalization path whereby

lexical verbs evolve into grammatical markers (e.g. Lehmann 2015: 35–39). In our case, lexical verbs first develop a grammatical meaning when used in various types of analytic constructions, either combined with other verbs in auxiliary or serial verb constructions or combined with nouns or ideophones in light verb constructions, and out of these contexts they may eventually develop into bound anticausative morphology.⁶

Elaborating upon a distinction proposed by Haspelmath (1990) for sources of passive markers, verbal sources of AMs can further be distinguished into intransitive inactive verbs ('be', 'become', 'happen', 'suffer', 'fall'), intransitive active verbs ('go'), and transitive verbs ('do', 'say', 'get', 'give', 'hit').⁷

4.2.1. Intransitive inactive verbs

The AMs discussed in this section go back to intransitive inactive verbs, that is, monovalent verbs featuring a non-controlling/non-agentive subject (Haspelmath 1990: 38). Let us begin by considering analytic anticausative constructions. Quite unsurprisingly, verbs meaning 'be' and 'become', which constitute frequent sources of various types of auxiliaries cross-linguistically (Anderson 2006: 359), may also be involved in anticausative analytic constructions. Two examples are Lezgian *ʃun* 'be(come)' in (10) and Hindi *honā* 'be' in (11). Note that the two patterns are different. Lezgian *ʃun* can rightfully be described as an auxiliary, as it combines with a verb stem. By contrast, Hindi *honā* is involved in a light verb construction (Shamin 2018) in combination with nominal and adjectival bases, giving rise to an equipollent opposition with its causative counterpart *karnā* 'do'.

(10) Lezgian (Nakh-Daghestanian; Haspelmath 1993a: 165–166)

bašlamišun 'finish (tr.)' → *bašlamiš ʃun* 'finish (intr.)'

⁶ According to Haspelmath (1990: 40), in the case of periphrastic passive constructions "it is misleading to attribute the passive function to the auxiliary" because the passive meaning component usually comes from the verbal form with which the auxiliary combines, typically a passive participle. This does not entirely hold for the sources of anticausatives surveyed in Sections 4.2.1 to 4.2.3, given that both inactive and active intransitive verbs actively contribute with an inherent change-of-state lexical semantics to the analytic constructions in which they are involved (Section 5.1).

⁷ In yet other cases, AMs go back to verbal elements whose precise lexical nature remains unknown. For example, the Creek (Muskogean) suffix *-k-* (Martin 2011: 216–218) is reconstructed as going back to a Pre-Proto-Muskogean auxiliary verb **-ka* (Haas 1977: 528–529), whose semantics cannot however be reconstructed.

(11) Hindi (Indo-European; Koul 2008: 102)

<i>daraāzā</i>	<i>band</i>	<i>karo</i>	vs.	<i>daraāzā</i>	<i>band</i>	<i>huā</i>
door	close	CAUS		door	close	ANTC
'Close the door.'				'The door (was) closed.'		

Other languages offer evidence for later stages of this grammaticalization path (Haspelmath 1990: 38), as in these languages verbs originally meaning 'be(come)' have yielded derivational bound affixes. This is the case of the Huaylas Ancash Quechua (Quechuan) suffix *-ka*, historically connected with the verb *ka* 'be' (Parker 1976: 116), as well as of AMs in a number of Pama-Nyungan languages, such as the Ngaanyatjarra suffix *-ri*, which is likely cognate with the Warlpiri (Pama-Nyungan) verb *jarri* 'become' (McGregor 2002: 144; this might possibly be further traced back to a verb 'fall', McGregor 2013: 120; see below).

Other intransitive inactive sources of AMs are 'happen', 'suffer' and 'fall'. The Rama (Chibchan) verb *ting* 'happen' is involved in compound verb forms with anticausative meaning in equipollent contrast with *uung* 'do, make', e.g. *tup-ting* 'sink (intr.)' vs. *tup-uung* 'sink (tr.)' (Grinevald 1990: Chap. 5, 21). The Vietnamese (Austroasiatic) auxiliary verb *bị* 'suffer' has given rise to analytic constructions often described as passive in the literature (Haspelmath 1990: 41; Kuteva et al. 2019: 414) but it can also occur in anticausative equipollent verb pairs in contrast with causative *làm* 'do, make', e.g. *làm/bị ốm* 'make/get sick' (see discussion in Simpson & Tâm 2013; Bruening & Tran 2015).

A verb *ti* 'fall, disappear' is the source of the Korean AM *-eci* (Ahn & Yap 2017: 444; Yap & Ahn 2019: 19–20). Already in 15th century texts, the verb *ti* is used either with its full lexical meaning, as in (12a), or as the second member of serial verb constructions (linked to preceding verb by means of a linking vowel), in (12b). Ahn & Yap (2017) argue that in these serial verb constructions *ti* underwent semantic bleaching, lost its spatial 'downwards' component and fused with the preceding linker to yield a new AM *-eti-*, attested already in the 17th century, as in (12c).

(12) Korean (Koreanic; Ahn & Yap 2017: 444–446)

- a. *apa-nim* *po-si-ko* *stah-ay* *ti-ye*
 father-HON see-HON-and earth-LOC fall-LNK
 'Father saw (him) and fell to the ground.'

- b. *sot-a-ti-ye* → *sot-ati-ye*
 pour-LNK-fall-LNK pour-ANTC-LNK
 ‘pour and fall’ ‘pour (intr.)’
- c. *elAm-i* *muntuk* *phul-ety-e*
 ice-NOM suddenly undo-ANTC-because
 ‘And because the ice suddenly broke.’

Further evidence for the FALL > ANTICAUSATIVE shift comes from AMs in Pama-Nyungan languages, e.g. Warlpiri *-wanti* and Martuthunira *-npa*, both originally meaning ‘fall’ (McGregor 2002: 140; McGregor 2013: 120). More generally, ‘fall’ verbs have been reported as sources of markers expressing sudden/unexpected events (Anderson 2006: 348), which bear semantic similarities to AMs.

4.2.2. Intransitive active verbs

The label *intransitive active verbs* refers to the motion verbs ‘go’ and ‘come’, which, unlike inactive verbs in Section 4.2.1, feature a controlling and volitional subject. Deictic motion verbs such as ‘go’ have repeatedly pointed out as sources of auxiliaries and voice markers (see Anderson 2006: 345–352; Devos & van der Wal 2014; Kuteva et al. 2019) and also constitute sources of AMs.

For example, in Jaminjung the verb *ijga-* ‘go’, besides other various functions (Schultze-Berndt 2000: 258–266), marks the noncausal member of equipollent alternations in contrast with the causative verb *ma-* ‘hit’, as in (13).⁸

(13) Jaminjung (Mirndi; McGregor 2002: 124)

bag-ijga- ‘break (intr., lit. break-go)’ vs. *bag-ma-* ‘break (tr., lit. break-hit)’

A less clear example comes from Mosestén(-Chimané). In this language, anticausativization may be encoded by the polyfunctional suffix *-ki*, as in (14a) (Sakel 2011: 306–312). In addition, a comparable form is found in associated motion

⁸ Another possible example is offered by Italian, where the combination of the verb *andare* ‘go’ plus past participle may encode spontaneous events and does not allow Agent NPs. However, the construction does not constitute a widespread anticausativization strategy, as it is virtually restricted to a subset of negative verbs, typically verbs of destruction, e.g. *la conoscenza andò perduta* ‘knowledge got (lit. went) lost’ (see Mocciaro 2014 for details).

constructions with the meaning ‘go there (to do something)’ (Sakel 2011: 273–275), as in (14b), and both forms are formally similar to the lexical verb *ka-* ‘bring there’. A plausible scenario is that the verb *ka-* ‘bring/go there’ was first used in combination with other verbs in coverb constructions expressing associated motion, as in (14b), and later extended to the anticausative use. Interestingly, a spontaneous change-of-state semantics is found in lexicalized usages of associated motion *-ki*, as in (14c).

(14) Mosestén (isolate, South America; Sakel 2007: 328; Sakel 2004: 307, 275)

a. *jofor’yi-* ‘open (tr.)’ → *jofor’ya-ki-* ‘open (intr.)’

b. *phan’-ye-ki-*’

feather-do-go-[3]F[SG]

‘She goes there to ask for feathers.’

c. *ö-yä-ye-ki-*’

F-AD-do-go-[3]F[SG]

‘She is getting better (lit. she is going there).’

Evidence for a ‘come’ origin of AMs also comes from a variety of languages. In Italian, the verb *venire* ‘come’ can be used as a passive auxiliary. As shown by Giacalone-Ramat & Sansò (2014: 31–34), passive *venire* originated out of an earlier anticausative state, already attested in Old Italian, as shown in example (15), in which the periphrasis *vennero smarriti* ‘came to be lost’ constitutes the anticausative counterpart of *smarrirono* ‘lost (tr.)’ (Squartini 2003).

(15) Italian (Indo-European; Squartini 2003: 25)

<i>e</i>	<i>allora</i>	<i>gli</i>	<i>cavalieri</i>	<i>tutti</i>	<i>vennero</i>	<i>smarriti</i>
and	then	DEF.PL	knight.PL	all.PL	come.PST.3PL	lose.PPP.PL

‘And then all knights got lost.’

Semantically, the development of verbs ‘go’ and ‘come’ into AMs can be linked to a well-known GO/COME > CHANGE-OF-STATE semantic shift (Schultze-Berndt 2000: 262). This shift is based on a conceptual metaphor whereby EVENTS/STATES ARE LOCATIONS and change of state can consequently be conceptualized as a change of location (e.g., Eng. *go mad*; see also Kuteva et al. 2019: 101–102, 204–205). In these metaphorical contexts, the agency and animacy restrictions on the subject of the verb can

progressively be lost, favoring its extension to decausative verbs (e.g., Eng. *go cold*). Note that a comparable metaphorical shift, whereby entry into/exit from a location stands for beginning of an event, also underlies the development of the verbs ‘go’ and ‘come’ into ingressive markers ‘begin to’ (Anderson 2006: 347; Kuteva et al. 2019: 101), and ingressives constitute another source of AMs (Section 4.5.2).

4.2.3. Transitive verbs

Transitive verbs like ‘do’ and ‘give’ typically give rise to causative markers (Zúñiga & Kittilä 2019: 220–221), but, perhaps surprisingly, they may also serve as sources of AMs. Transitive verbs found as sources of AMs in the sample are ‘do’, ‘say’, ‘get’, ‘give’ and ‘hit’.

Ainu features a suffix *-ke* that can be used in anticausative function when opposed to causative verbs in *-V*, as in (16a). The same suffix also has a verbalizing function, as it forms either agent-oriented causative verbs, as in (16b), or change-of-state intransitive verbs from ideophones, as in (16c). In addition, *-ke* is sporadically used in causative contexts, as in (16d) (Bugueva 2015: 473–474).

(16) Ainu (isolate, Eurasia; Bugueva 2015: 449, 473, 451)

- a. *mak-ke* ‘open (intr.)’ vs. *mak-a* ‘open (tr.)’
- b. *su* ‘pan’ → *su-ke* ‘cook (tr.)’
- c. *pat* IDEO → *pat-ke* ‘explode’
- d. *ray* ‘die’ → *ray-ke* ‘kill’

The perhaps puzzling anticausative/causative syncretism of Ainu *-ke* can diachronically be explained as follows. As argued by Bugueva (2015: 473–474), *-ke* can etymologically be traced back to a full verb **ki* ‘do’. As a lexical verb, **ki* ‘do’ could combine with various non-verbal elements, and these analytic constructions served as basis for its grammaticalization into a verbalizer. Depending on the nature of the element it combines with, verbalizations with **ki* shows different semantics (this is a typical feature of verbalizers, see e.g. Karaj & Sansò forth.). When combined with nominal roots the result is a [N **ki*] construction ‘do (with) N’ indicating an agent-oriented activity involving the nominal element as instrument/theme, as in (16b), while when combined with ideophones the result is an emission verb ‘do the

IDEO sound’, as in (16c). These are two typical contexts in which ‘do’ verbs occur cross-linguistically: the use as *activity verb* and the use as *verbalizer with sound-symbolic expressions* (Schultze-Berndt 2008: 190–191, 193–194). From the former, the suffix further extended to direct causative contexts, as in (16d), while the latter served as the basis for the development of the anticausative function, as in (16a), based on the fact that emission verbs can easily be interpreted as spontaneous situations involving a non-controlling inanimate participant. Thus, the anticausative and the causative functions of *-ke* independently go back to the verbalizing function of the lexical verb **ki* ‘do’ (on the connections between ‘do’ and ‘become’ verbs see also more generally Gil 2017). I return to the connection between verbalizers and AMs in Section 4.6.

The verbalizing function with ideophones also plays a key role in the development of ‘say’ verbs into AMs. This development underlies the anticausative use of the suffix *-me* in Bininj Kun-Wok (Gunwinyguan), which according to Alpher et al. (2003: 332–333) goes back to a verb **me* ‘do, say’, the anticausative function of the *in-* prefix in Semitic languages such as Arabic (Afro-Asiatic; Roset 2018: 247–248), which can be reconstructed as going back to a generic action verb **n-* ‘say, do, become’ (Kouwenberg 2010: 314–317), and the Yuracare AM *-tA*, likely from the verb *ta-* ‘say’ (Section 4.7.4). Comparable developments have also been discussed for a number of languages from East Africa by Cohen et al. (2002).

Acquisitive ‘get’ verbs are a known source of both anticausative (and also passive) morphology (Lenz & Rawoens 2012; Kuteva et al. 2019: 187–189).⁹ According to Gronemeyer (1999), historical corpus evidence from English shows that the development GET > CHANGE-OF-STATE (Kuteva et al. 2019: 186) possibly originated in the use of *get* in locative construction with a Goal, e.g. *get to the shop*. In these constructions, the slot of the Goal was progressively filled first by adverbs and then by adjectives/participles, while the construction kept the general meaning of ‘reaching a goal’. When combined with adjectives/participles, this resulted into an anticausative reading, e.g. *get burned*, partly based on the CHANGE OF STATE IS CHANGE OF LOCATION metaphor discussed in Section 4.2.2. The passive reading of English *get* eventually developed from the anticausative one (Gronemeyer 1999: 29).

⁹ Interestingly, a connection between possession and change-of-state is also documented for several denominal verbalizers that may variously be translated as ‘have N’ or ‘get N’ (Karaj & Sansò forth.).

Another compelling case for a ‘get’ origin of an AM is made by Frellesvig & Withman (2016: 296–306) for the Japanese *-e* suffix.¹⁰ Already in Old Japanese, one finds alternations between stem in *-e* (bigrade) opposed to consonantal stems (quadrigrade). As shown in (17), the *e*-suffixed verbs occurs both in causative, in (17a), and in anticausative function, as in (17b).

(17) Old Japanese (Japonic; Frellesvig & Withman 2016: 291)

- a. *tat-* ‘rise, set out’ → *tat-e-* ‘raise’
b. *yak-* ‘burn (tr.)’ → *yak-e-* ‘burn (intr.)’

The anticausative/causative suffix *-e* derives from the combination of basic stems with the verb *e-* ‘get’, still attested as a full lexical verb in Old Japanese. Both its valency-related functions arose in constructions where the verb *e-* ‘get’ was used with secondary predicates (though probably the causative was formed earlier and remained the predominant pattern). In particular, as argued by Frellesvig & Withman (2016), the anticausative usage likely emerged as the reinterpretation of a secondary predicate transitive construction ‘A gets P to X’ as an intransitive construction denoting a spontaneous event ‘P gets (to) X’. A context in which this reinterpretation might have taken place are occurrences in which the transitive construction features an omitted non-agentive experiencer/goal subject participant, as in (18).

(18) Old Japanese (Japonic; Frellesvig & Withman 2016: 299)

kari wo tukapi ni e-tesika mo
goose ACC messenger be.INF get-OPT even

‘Would that (I) had gotten the wild geese as messengers! > Would the wild geese had become messengers.’

The development of transfer/contact verbs ‘give’ and ‘hit’ into AMs appears to be less common. In the sample, I have only found one possible example of each. The Ingush

¹⁰ Still, as remarked by Frellesvig & Withman (2016: 308), the English and the Japanese cases are not fully equivalent. The anticausative meaning of the English *get*-constructions is also due to the combination with resultative participles (which inherently indicate a change of state), while in Japanese the verb *e-* ‘get’ is the solely responsible of the change-of-state semantics, as it combines with basic stems.

anticausative marker *-lu*, shown in (19), is homophonous with the verb ‘give’ (Nichols 2011: 491), thus possibly deriving from the latter. In Yagaria, causative transitive verbs can be turned into anticausatives by compounding them with the verb *ei-* ‘hit’, as in (20).¹¹

(19) Ingush (Nakh-Daghestanian; Nichols 2011: 751)

d.iell ‘open (tr.)’ → *d.iella-lu* ‘open (intr.)’

(20) Yagaria (Nuclear Trans New Guinea; Renck 1975: 154)

lo’ao- ‘break (tr.)’ → *ei-lo’ao-* ‘break (intr.)’

Also due to the rarity of these shifts, it is unclear how ‘hit’ and ‘give’ verbs might have grammaticalized as AMs. In the case of ‘hit’, perhaps a role could have been played by constructions in which *hit* denotes a motion event, e.g. *he hit the ground (with his body)*, thus linking the development of *hit* to that of ‘go’ (Section 4.2.1). I return to the development of ‘give’ in Section 4.7.2.

5.2. Spatial markers > AMs

In section 4.2.2, I have discussed the developmental path connecting motion verbs to the anticausative domain. In my sample, there are two more cases of AMs that go back to sources with spatial semantics.

The first example is that of the so-called ‘separative’ extension in Bantu languages. Consider the suffix *-uw-* in Chuwabu, which occurs in passive and anticausative contexts, as in (21).

(21) Chuwabu (Atlantic-Congo; Guérois & Bostoen 2018: 212, 219)

a. *mí-ri dhí-ni-ó-j-uw-á na nyenyéle*
 4-tree SM4-IPFV.DJ-15-eat-PASS-FV by 10a.ant

‘The trees are being eaten by the ants.’

b. *gob-ól-a* ‘break (tr.)’ → *gob-ów-a* ‘break (intr.)’

c. *fúga* ‘shut’ → *fúg-uw-a* ‘open (intr.)’

¹¹ Renck (1975: 154) also mention that *ei-* ‘hit’ can be compounded with intransitive verbs to add a causative meaning, but no data is given to illustrate this pattern.

Chuwabu *-uw-* goes back to the Proto-Bantu intransitive separative suffix **-uk-* (Guérois & Bostoen 2018: 218–222), which in origin indicated “movement out of some original position” (Schadeberg & Bostoen 2019: 185–186).¹² This semantics is especially visible in intransitive ablative motion verbs such as Proto-Bantu **-tá-uk-* ‘come from’ (Schadeberg 1982: 61–65). Guérois & Bostoen (2018: 219) link the anticausative function of Chuwabu *-uw-* directly to the reconstructed separative semantics, but how this shift actually took place is not straightforward.

One can hypothesize that the potential bridging context between the separative/ablative and the anticausative functions was the reversive function displayed by outcomes of **-uk-* in several Bantu languages, such as Fwe in (21c). The ABLATIVE > REVERSIVE shift can be motivated by a metaphorical extension whereby inverting the state resulting from an event can be conceived as exit from a location (Gibert-Sotelo 2018). In some contexts, reversive verbs in fact encode spontaneous change-of-state events, possibly only physical in origin, as in (21c), but secondarily also involving more abstract situations, as in (22). It is out of the latter that **-uk-* was possibly reinterpreted as an AM proper.

(22) Fwe (Atlantic-Congo; Gunnik 2018: 234)

-rwârà ‘become sick’ → *-rwárùkà* ‘become better’

The second, albeit less assured, case comes from Moskona (East Bird’s Head). In this language, anticausativization is expressed by adding to transitive verbs an enclitic element =*ef*, which elsewhere serves as a proximal demonstrative enclitic ‘near’, e.g. *og* ‘bend (tr.)’ → *og=i-ef* ‘bend (intr.)’ (Gravelle 2010: 123–125, 196). The historical relationship between the deictic and the AM functions of Moskona =*ef* remains unclear, but it is possible that the deictic function is prior, as it is the only one attested for the cognate demonstrative *if* ‘this’ in the closely related language Meyah (Gravelle 2002: 149–150). Perhaps the development of Moskona =*ef* can be linked to the development of deictic motion verbs into anticausatives (Section 4.2.2).

¹² On the polyfunctionality of outcomes of Proto-Bantu **-uk-* in individual Bantu languages see Dom et al. (2016: 140–143).

4.4. Spontaneous events > AM

AMs may emerge from markers that encode of spontaneous events. In fact, a historical link between the two is not surprising, given that both encode uncontrolled events (Fauconnier 2011: 323–327).

Sino-Tibetan languages offer a case in point. As shown in Table 2, in languages such as Galo and Northern Pumi anticausativization is expressed by an alternation between voiceless and voiced onsets of verbal roots, with the intransitive member being associated with the voiced variants.

Language	‘break (tr.)’	‘break (intr.)’	Source
Galo (Macro-Tani)	<i>tíř-</i>	<i>díř-</i>	(Post 2007: 97)
Northern Pumi (Burmo-Qlangic)	<i>tʰwǎ</i>	<i>ɖwǎ</i>	(Daudey 2014: 295)
Japhug (Burmo-Qlangic)	<i>prɣt</i>	<i>mbrɣt</i>	(Jacques 2021: 917)

Table 2: The anticausative alternation in Sino-Tibetan languages

The origin of this alternation is a much-discussed topic in Sino-Tibetan linguistics: some scholars argue that the direction of the derivation is from transitive to intransitive, with voicing reflecting an intransitivizing prefix *N-, while others argue for de-voicing of transitive verbs due to a causative prefix *s- (see Handel 2012 for an overview). As argued by Jacques (2015; 2021: 918–922) and Gates et al. (2022), decisive evidence in favor of an intransitivizing origin of the pattern comes from Gyalrongic languages. This is particularly clear in Japhug, in which anticausativization is expressed by pre-nasalization of transitive roots. For example, pre-nasalization of the transitive verb *prɣt* ‘break’ yields anticausative *mbrɣt* ‘break (intr.)’. This pre-nasalization historically underlies the pre-voicing pattern in Table 2.

Jacques goes one step further and argues that anticausative pre-nasalization in Japhug is historically connected with the ‘autive’ prefix *nu-* (Jacques 2021: 967–982). This prefix is used either to encode self-benefactive events of various types (but never in reflexive contexts proper) or it can be added to intransitive verbs to indicate that the event takes place spontaneously or accidentally. The prefix is never connected with a change in transitivity, and can likewise occur with intransitive and transitive verbs, as in (23a-b).

(23) Japhug (Sino-Tibetan; Jacques 2021: 974, 975)

a. *ɲu-ku-nu-βze*

IPFV-SUBJ:PTCP-AUTO-grow

‘It grows by itself.’

b. *k^hutsa pu-nu-qru-t-a*

bowl AOR-AUTO-break-PST:TR-1SG

‘I broke the bowl (by mistake).’

Jacques hypothesizes that both anticausative pre-nasalization and the ‘spontaneous’ use of *nu-* ultimately derive from a common source which he reconstructs as a “nasal prefix expressing spontaneous/non-volitive actions” (2015: 18), and which was not originally connected with intransitivization (see Fauconnier 2011 on the link between involuntary agent constructions and transitivity). Evidence for this comes also from rare voicing pairs where both members are syntactically transitive, e.g. Khaling (Himalayish) *plum-* ‘rinse in water’ vs. *blum-* ‘sink in water’. It is only later that outcomes of this prefix become associated with transitivity change, hence turning into AMs.

A connection with spontaneous events has also been proposed for two anticausativization strategies found in ancient Indo-European languages. The first example is that of the Old Indo-Aryan suffix *-yá-*, which is continued by AMs/passive markers of modern Indo-Aryan languages, as is the case of Palula *-ĩj-* in (24a-b) (Liljegren 2016: 240–241).

(24) Palula (Indo-European; Liljegren 2016: 241)

a. *bilá-* ‘melt (tr.)’ → *bil-ĩj-* ‘melt (intr.)’

b. *de-* ‘give’ → *da-ĩj-* ‘be given’

The origin of passive *-yá-* in Indo-Aryan still constitutes a disputed topic among specialists (see Lazzeroni 2004; Kulikov 2012; Luraghi et al. 2021 with references). There is a consensus that the passive function of *-yá-* is likely secondary and must be connected with the Sanskrit 4th class presents in *-ya-*, which were also associated with the anticausative alternation, especially when opposed to causative suffixes such as *-áya-*, e.g. *nás-ya-ti* ‘perish’ vs. *nās-áya-ti* ‘make disappear’ (Kulikov 2012: 727–729).

According to Lazzeroni (2004), intransitivization was however not the original function of *-ya-*, which was instead connected with the characterization of spontaneous (unaccusative) change-of-state events.¹³

A similar scenario has also been proposed for the origin of the Active vs. Middle voice alternation in ancient Indo-European languages such as Hittite, Ancient Greek and Latin. For reasons of space, I will not discuss the details here (see Lazzeroni 1990; Luraghi 2012; Inglese 2020: Chap 3).¹⁴ In short, scholars have argued that in Proto-Indo-European verbal voice originally followed a lexical distribution, with the Middle inflection specifically being confined to verbs indicating uncontrolled change-of-state events or states, e.g., Lat. *morior* and Hitt. *kištari* ‘die’. Out of this original middle-only group, voice alternation with anticausative function first arose, e.g. Hitt. *zinnizi* ‘finish.ACT (tr.)’ vs. *zinnatari* ‘finish.MID (intr.)’, and was later extended to other functions such as the passive.

4.5. Aspectual markers > AMs

In a number of languages, AMs go back to aspectual-like markers (here broadly understood as per Croft 2012). The three sources that I have detected in the sample are resultative, ingressive and stative markers.

4.5.1. Resultative markers > AMs

Resultative markers indicate “a state that was brought about by some action in the past” (Bybee et al. 1994: 63) and bear notorious resemblances to anticausatives (Zúñiga & Kittilä 2019: 43), thus making the RESULTATIVE > AM shift unsurprising.¹⁵ Recall for example that anticausative *get*-constructions in English are built by combining *get* with resultative participles. Another language that instantiates the RESULTATIVE > AM shift is possibly Hausa. The Hausa suffix *-u* shows a variety of

¹³ The ultimate source of Indo-Aryan *-ya-* remains highly contested. For a discussion of possible etymologies see Kulikov (2012: 748–759) and Willi (2018: 582 fn. 60).

¹⁴ None of these languages has been included in my sample because their Middle inflection synchronically also shows a reflexive function, which is however historically secondary with respect to the anticausative one (see Inglese 2020: 235–237).

¹⁵ Vajda (2015: 657–658) discusses the existence of anticausative verbs in Ket featuring a suffix *-j-* etymologically connected with resultative *-jə-*. However, since from the available data it is not clear whether *-j-* complies with the criteria laid out in Section 4.1, I have excluded it from my sample.

functions (Jaggar 2001: 260–267), including proper resultative and anticausative, as in (25a-b). Jaggar (1988: 405–408) reconstructs Hausa *-u* as the outcome of the Proto-Chadic resultative suffix **-k^wo*, so that the resultative function is likely to be historically prior.

(25) Hausa (Afro-Asiatic; Jaggar 2001: 263, 264, 267)

- a. *fas-à* ‘smash’ vs. *fàs-u* ‘(the glass) is smashed’
b. *kad-à* ‘shake (tr.)’ vs. *kàd-u* ‘shake (intr.)’

4.5.2. Ingressives > AMs

That AMs may also be associated with ingressive semantics has already been noted by Haspelmath (1987: 34). An ingressive source might be proposed for the Filomeno Mata Totonac prefix *ta-*. Among its various functions, the prefix can be used in anticausative verb pairs following two patterns (McFarland 2009: 182–186). First, it may stand in an equipollent opposition to causative *maa-*, as in (26a). This pattern is limited to roots that never occur in isolation. Second, *ta-* can also be used as an AMs attached to unmarked transitive verbs, as in (26b). In addition, the prefix shows an ingressive, or ‘inceptive’ (McFarland 2009: 144), use, by creating dynamic motion verbs when added to a close set of so-called positional verbs, i.e. stative roots indicating location, as in (26c).

(26) Filomeno Mata Totonac (Totonacan; McFarland 2009: 185, 184, 109)

- a. *maa-čaw-ii* ‘open (tr.)’ vs. *ta-čawa* ‘open (intr.)’
b. *lak-ponqa* ‘knock down (tr.)’ → *ta-lak-ponqa* ‘fall down’
c. *-xuu-* ‘inside’ → *ta-xuu-maa* ‘he’s getting inside’

Even though the etymology of Filomeno Mata Totonac *ta-* remains unknown, one can tentatively sketch the following diachronic scenario.¹⁶ A comparable prefix occurs in

¹⁶ McFarland (2009: 96) mentions a possible connection with the prefix *ta-* found in result nominalizations, e.g. *pink* ‘split (tr.)’ → *ta-pínk* ‘split (intr.)’ and *ta-pínk* ‘the crack’. This connection remains speculative, and might be due to chance, but note that a NOMINALIZATION > ANTICAUSATIVE development is perfectly conceivable (Section 4.6).

virtually every Totonacan language, both of the Totonac and the Tepehua branches (Beck 2012: 593), and its range of functions greatly varies: in some languages, such as Upper Naxaca, it also occurs with autocausative and grooming verbs (Beck 2011: 15), while in other it also has a passive/resultative meaning, as in e.g. Yecuatla (or Misantla) Totonac (MacKay 1999: 257). Nevertheless, the ingressive function seems to be widespread among the family and indeed the prefix is only compatible with stative verbs in the languages of the Tepehua branch (Kung 2007: 287). In particular, other Totonacan languages attest to the existence of a pattern similar to (26c), but in which a stative root co-exists alongside a *ta*-form with change-of-state semantics and a causative *ma*-form, as in the Upper Naxaca example in (27):

(27) Upper Naxaca Totonac (Totonacan; Beck 2011: 35)

lakí: ‘be open’ vs. *ta-lakí*: ‘open (intr.)’ vs. *ma-lakí*: ‘open (tr.)’

One may speculate that the ingressive function with stative verbs shown in (26c) and (27) is older and served as the source for the development of *ta*- into an AM. This was possibly favored by the loss of the original stative root, as in Filomeno Mata (26a), so that both the change-of-state and the spontaneous semantic components were reinterpreted as being expressed by *ta*- alone. Once reinterpreted as an AM, *ta*- could be extended to transitive verbs in an anticausativization pattern proper, as in (26b).

Further comparative work on *ta*- in Totonacan might shed light on the likelihood of this reconstruction, but a good typological parallel comes from Latin. As discussed in Inglese (2021: 148–153 with references), Latin features a peculiar pattern of anticausativization for some stative roots, whereby alongside a bare stative root one finds a telic spontaneous counterpart with a suffix *-sc-* and a causative counterpart with *-facio* ‘make’, as in (28a). This pattern is reminiscent of that found in Upper Naxaca Totonac in (27).

(28) Latin (Inglese 2021: 149)

a. *pateo* ‘be open’ vs. *patesco* ‘open up’ vs. *patefacio* ‘open (tr.)’

b. *tremo* ‘tremble’ vs. *tremesco* ‘start shaking’ vs. *tremefacio* ‘make tremble’

c. *morbus* ‘illness’ → *morbescio* ‘fall ill’ vs. *morbificio* ‘make ill’

In pairs such as (28a), the suffix *-sc-* arguably only adds a telic meaning component to stative roots. This aspectual meaning is particularly clear when the verb applies to atelic but dynamic roots such as *tremo* ‘shake’ in (28b), with which it gives an ingressive reading ‘start shaking’. However, *-sc-* alone may also function as an equipollent AM, especially in the case of verbs derived from nominal roots for which a corresponding stative verb is lacking, as in (28c). This anticausativization function of Latin *-sc-* is secondary. In fact, comparative evidence robustly points towards an original aspectual use of the Proto-Indo-European suffix **-ské/o-*, connected with imperfectivity, atelicity and pluractionality (Berrettoni 1971; Inglese & Mattiola 2020: 286–291; Inglese 2021: 151). Out of this original meaning, the ingressive semantics only developed in combination with atelic roots (as evidenced already by Hittite; Inglese & Mattiola 2020), and from there it was further extended to stative roots, thus leading to the change-of-state and eventually anticausative semantics.

4.5.3. Stative markers > AMs

A connection between stativity and anticausativization has traditionally been reconstructed for Ancient Greek *-ē-*. This suffix is described in reference grammars as a dedicated passive affix in the Aorist system, as in (29a), but in fact it also functions as an AM (Allan 2003: Chap. 3; Romagno 2014), as shown in (29b).

(29) Ancient Greek (Indo-European; Allan 2003: 95, 99)

- a. *éplēksa* ‘hit’ vs. *eplég-ē* ‘was hit’
b. *ékausa* ‘burned (tr.)’ vs. *éka-ē* ‘burned (intr.)’

Much has been written about the origin of the *ē*-aorist in Greek (see Luraghi et al. 2021: 14–16 for an overview), but the general consensus is that it derives from the Proto-Indo-European stative suffix **-eh₁-* (e.g. Willi 2018: 15). Outside of Greek, this suffix predominantly occurs with stative verbs, e.g. Latin *rub-ē-o* ‘be red’, Hittite *marš-e-* ‘be false, corrupted’, and this is taken as evidence to reconstruct an original intransitive stative semantics for Proto-Indo-European **-eh₁-* (thus Ruijth 2004).¹⁷ The

¹⁷ Ancient Greek features another ‘passive’ Aorist suffix *-thē-* which also functions as AM. This suffix is reconstructed as a Greek combination of stative **-eh₁-* with a suffix **-dh-*, the latter possibly resultative in origin (Luraghi et al. 2021: 14–16 with references). A comparable suffix *-th-* is also found in the

anticausative function of Greek *-ē-* likely emerged in its combination with the Aorist and was facilitated by the semantic proximity between stative markers and AMs, both typically denoting uncontrolled situations (thus Haspelmath 1990: 51–52).¹⁸

A stative origin has also been proposed for AMs in Tibeto-Burman languages. In Bunan, the suffix *-s* performs an anticausative function, among its various usages, as in (30a), and it also occurs with some inherently stative verbs, as in (30b) (Widmer 2018: 361–382). Based on comparable suffixes showing also a reflexive function in other Tibeto-Burman languages, LaPolla (1996: 3) reconstructs a reflexive origin for Proto-Tibeto-Burman **-si*, from which both the stative and the anticausative use emerged.

(30) Bunan (Sino-Tibetan; Widmer 2018: 363)

a. *al-t̥-um* ‘open (tr.)’ vs. *al-s-̣-um* ‘open (intr.)’

b. *noŋs-men* ‘be spoilt’, *t^hos-men* ‘be high’

However, an alternative reconstruction has been proposed by Matisoff (2003: 471–472), who connects **-si* with the Sino-Tibetan nominalizing stative suffix **-s*, as attested in e.g. Tibetan *za* ‘eat’ → *zas* ‘food’ (Jacques 2016). How a (stative) nominalizing suffix can turn into an AM remains a matter of speculation, and I return to this point in Section 4.6.

4.6. Nominalizers and verbalizers > AMs

Nominalizers and verbalizers constitute a well-known sources of voice morphology, including passives and antipassives (Sansò 2016; 2017), and there is evidence that

formation of present stem verbs with anticausative function, e.g. *phlég-ō* ‘burn (tr.)’ vs. *phlegé-th-ō* ‘burn (intr.)’ (see Magni 2010). If so, Greek *-thē-* provides further evidence for the RESULTATIVE > ANTICAUSATIVE shift discussed in Section 4.5.3.

¹⁸ This is admittedly not the only possible scenario. In fact, to account for the shift from stative to change-of-state semantics, Ruijth (2004: 59–61) himself propose that the anticausative/passive meaning actually arose only in combination with the Aorist formant **-s-* and that Ancient Greek *-ē-* thus goes back to **-eh₁-s-*. Other scholars instead have argued that the Proto-Indo-European suffix **-eh₁-* was not simply stative but that a change-of-state component must be reconstructed for the proto-language as well (e.g., Hardarson 1998), so that one cannot state with certainty which semantic component was the original one. Finally, particularly interesting is a proposal that links stative **-eh₁-* to the nominal instrumental case ending (Jasanoff 2004): in this view, the suffix was originally used in stative nominal predications ‘X is endowed with N-*eh₁-*’ and was later reinterpreted as a denominative verbal suffix ‘X is/becomes V-*eh₁-*’.

they can also give rise to AMs. This is not entirely surprising and a synchronic association between AMs and denominal verbalizers has already been noted in the literature (Haspelmath 1987: 33; Aikhenvald 2011: 244–245; Grestenberger 2016: 105).

In Section 4.5.3, I have mentioned the stative nominalizer *-s as a possible source of the Tibeto-Burman AM -s-. A potential parallel is discussed by Authier (2012) for the AM -aR- of Kryz and other related Nakh-Daghestanian languages. In Kryz, the suffix -aR- can be used for anticausativization (as well as other voice operations), as in (31a), and a comparable suffix is found in nominalizations, as in (31b-c).

(31) Kryz (Nakh-Daghestanian; Authier 2012: 149, 160)

- a. ^ʔuf-a- ‘open (tr.)’ vs. ^ʔuf-**ar**- ‘open (intr.)’
- b. ke-xh-r-ic ‘move’ → xh-**ar** ‘wind (lit. the moving)’
- c. x-irayc ‘weave’ → x-**al** ‘roof; cobweb (lit. the woven)’

Based on comparative evidence, Authier argues that the nominalizing function of -aR- in (31b-c) is historically older. He further hypothesizes that these deverbal nouns with S or P orientation ‘V-ing, V-ed’ could have been used in predicative function ‘X (is a) V-ing/ed’ and that out of these contexts, deverbal nouns could have been reinterpreted as intransitive V-aR- verbs. This is possibly how -aR- developed into a voice marker encoding anticausativization.

Verbalizers may also be sources of AMs. In Section 4.2.3, I have discussed how in their development into AMs ‘do, say’ verbs first go through a stage in which the function as verbalizers with ideophones (see also Section 4.7.4). Karaj & Sansò (forth.) propose a verbalizing origin also for AMs in Malayic languages. For example, in Malay (Austronesian) the prefix *bər-* also occurs in anticausative contexts, e.g. *tolak* ‘push someone away’ → *bər-tolak* ‘shove off’. As argued by Karaj & Sansò (forth.), this prefix goes back to Proto-Malayic *(*mb*)AR-, which can be reconstructed as a verbalizer deriving either activities/states or change-of-state verbs from nouns. Out of the latter, the prefix further developed a full-fledged anticausative function.

A brief digression is in order on the status of inchoative verbalizers as AMs. Verbalizers that create (spontaneous) change-of-state events ‘become X’ from adjectives and nouns have been reported for several languages, and these often go back to lexical verbs meaning ‘(be)come’ or ‘do’ (Aikhenvald 2011: 232, 237; Mattioli

& Sansò 2021). However, inchoative verbalizers can rightfully be described as AMs only when they participate in the anticausative alternation (criterion (i) in Section 4.1). An example from Misantra Totonac will serve to illustrate this point. Misantra Totonac features two synthetic strategies to create denominal verbs meaning ‘become X’: the suffixes *-nan* and *-la* (MacKay 1999: 339–342).¹⁹ While both suffixes derive inchoative verbs from nominals, as in (32a-b), only *-nan* also occurs as an intransitivizer with causative (denominal) stems, as in (32c). This means that while *-nan* qualifies as an AM, *-la* does not.²⁰

(32) Misantra Totonac (Totonacan; MacKay 1999: 339, 341)

- a. *haks-ta* ‘smelly place’ → *haks-ta-nan* ‘become smelly’
 b. *siski* ‘sweet’ → *siski-la* ‘become sweet’
 c. *papaks-nV[?]-ii* ‘make old’ → *papaks-nV[?]-ii-nan* ‘grow old’

4.7. Non-reflexive voice markers > AMs

As already mentioned, reflexives are commonly held to be the main valency-related source of AMs (Section 3). However, there is an increasing body of evidence suggesting that also voice markers originally dedicated to other valency operations may also extend to the anticausative domain (Bahrt 2021: Chap. 7).

4.7.1. Passive markers > AMs

While there is ample evidence that anticausatives may develop into passives (e.g. Haspelmath 1990), the reverse shift, that is PASSIVE > ANTICAUSATIVE, is held to be

¹⁹ The latter can in fact either occur bound to nominal roots or in isolation taking its own inflection and is cognate to the Upper Necaxa Totonac verb *la* ‘do, make, become’ (Beck 2011: 204).

²⁰ MacKay (1999: 339) does not discuss the etymology of *-nan*, but it is strikingly similar to the antipassive suffix *-nan*, which is widely attested in the Totonacan family (Beck forthc.) and possibly derives from a suffix for agent nominalizations (Sansò 2017: 180–181). A connection between anticausative and antipassive *-nan* can tentatively be sought in the use of *-nan* as a verbalizer: as discussed by Beck (2008: 17), in Upper Necaxa Totonac *-nan* can be used to derive activity verbs from nouns, e.g. *tfanáx* ‘coa’ → *tfanax-nán* ‘work with a coa’, or from ideophones, e.g. *ftayayá* ‘object gliding’ → *ftayayá-nán* ‘glide’. Notably, the latter may also encode non-volitional events, and thus may have served as bridge towards the inchoative and anticausative functions (see also Section 4.2.3).

much rarer. This shift has been proposed by Kulikov (2011) for various passive formations in Vedic Sanskrit (Indo-European). In his view, this development was mostly confined to experiencer verbs and the bridging contexts was provided by impersonalized passives: when these verbs were used with generic human Agents, they were reinterpreted as encoding spontaneous events, e.g. *śru-* ‘hear’ > *śrū-yá-te* ‘be heard (by someone)’ > ‘be audible’ (on the etymology of *-yá-* see Section 4.4).

Evidence for comparable developments remains rather scanty. In my sample, one possible instance comes from Tzeltal. In this language, the infix <*j*> can be used in both anticausative and passive function, as in (33a-b), but the latter is probably original, as the infix is reconstructed as a passive infix * <*h*> in Proto-Mayan (Polian 2013: 292). Note however that Tzeltal <*j*> has a much wider anticausative usage than the one discussed for Vedic by Kulikov (2011), so that it is doubtful whether the PASSIVE > ANTICAUSATIVE shift in Mayan took place following the same path.

(33) Tzeltal (Mayan; Polian 2013: 291, 293)

- a. *tsak* ‘take’ → *tsa<j>k* ‘be taken’
b. *puk* ‘melt (tr.)’ → *pu<j>k* ‘melt (intr.)’

4.7.2. Causative markers > AMs

Causative markers have mostly been discussed as sources of passive markers (e.g. Haspelmath 1990: 46–49, Bahrt 2021: 216–218), but they are also connected to AMs (cf. Kittilä 2000). This connection should sound less surprising in light of the discussion in Section 4.2.3 on the development of transitive verbs into AMs. In the case of syncretic causative-anticausative(-passive) markers, it is commonly believed that it is the causative meaning that gives rise to the anticausative one, as compelling evidence of the reverse development has not yet been found (cf. Bahrt 2021: 107–109, 214–215). Besides the examples discussed in Section 4.2.3, further possible candidates for the CAUSATIVE > AM shift comes from Tungusic languages and from Korean.

Tungusic languages feature cognate suffixes with either anticausative or passive function, as Even *-b* in (34) and Evenki *-v/-p* in (35). The situation of Manchu is slightly different: here the suffix *-bu* shows either passive or causative functions, as in (36), but it never functions as an AM.

(34) Even (Tungusic; Malchukov & Nadjalkov 2015: 598)

aŋa- ‘open (tr.)’ → *aŋa-b-* ‘open (intr.)’

(35) Evenki (Tungusic; Malchukov & Nadjalkov 2015: 609)

ula- ‘make wet’ → *ula-v-/ula-p-* ‘become wet’

(36) Manchu (Tungusic; Malchukov & Nadjalkov 2015: 610)

va- ‘kill’ → *va-bu-* ‘make kill, be killed’

Comparative evidence shows that Even *-b-*, Evenki *-b-/v-* and Manchu *-bu-* historically all derive from a lexical verb **-bu-* ‘give’ (Nadjalkov 1993). According to Malchukov & Nadjalkov (2015: 608–612), the most likely diachronic scenario is that Tungusic **-bu-* ‘give’ first developed into a causative marker, subsequently shifted to a passive function, as shown in Manchu, and in languages such as Even and Evenki the passive served as basis for the anticausative function. Both developments are compatible with the known grammaticalization path that links ‘give’ verbs with causatives and passives (Kuteva et al. 2019: 195–196, 198–199). In other words, in Malchukov & Nadjalkov’s account, the passive is an intermediate bridging context in the CAUSATIVE > PASSIVE > AM shift.

While this might be true for Tungusic languages, a direct CAUSATIVE > AM shift has been hypothesized for Korean by Yap & Ahn (2019). Besides the anticausative marker *-eci* discussed in (12), Korean also feature a syncretic voice marker *-i* with (at least) causative, anticausative and passive functions, as in (37a-c):

(37) Korean (Koreanic; Yap & Ahn 2019: 2, 6, 11)

a. *emma-ka aki-eykey cec-ul mek-y-ess-ta*
 mother-NOM baby-DAT breast-ACC eat-CAUS-PST-DEC
 ‘Mother breast-fed her baby.’

b. *pang-mwun-i cecello camk-y-ess-ta*
 room-door-NOM by_itself lock-ANTC-PST-DEC
 ‘The door of (my) room locked by itself.’

c. *manhun mwulkoki-tul-i sange-eykey capamek-hy-ess-ta*
 lots_of fish-PL-NOM shark-DAT eat-PASS-PST-DEC
 ‘A lot of fish got eaten by the shark.’

Unfortunately, the lexical source of Korean *-i* is yet unknown (Yap & Ahn 2019: 20). Taken at face value, the polyfunctionality of Korean *-i* is compatible with the development CAUSATIVE > PASSIVE > ANTICAUSATIVE proposed for Tungusic **-bu* by Malchukov & Nedjalkov (2015). However, the available historical data does not fully support this scenario: while the causative usage of *-i* is clearly primary, since it is attested already in Old Korean documents dating to the 10th century, the anticausative and the passive functions both first appear in Middle Korean texts of the 15th century (Yap & Ahn 2019: 15). To account for this chronological distribution, Yap & Ahn (2019: 16–19) argue in favor of the existence of two independent developments CAUSATIVE > PASSIVE and CAUSATIVE > ANTICAUSATIVE. Concerning the latter, they propose that a crucial bridging might have been reflexive-causatives constructions, as in (38), which, in absence of an overt Causee Agent and of a body-part expression (omitted for e.g., politeness reasons), were liable to be reinterpreted as intransitive constructions. Once *-i* started being used in intransitive constructions, it could have well been extended to anticausative contexts proper featuring a omission of the causee ‘(someone) closed the gate’ > ‘the gate closed’.

(38) Korean (Koreanic; Yap & Ahn 2019: 17)

namwu tok-(k)uy . . . (ne-eykey) koloī kulk-hi-ketun
tree poison-NOM 2SG-DAT painfully scratch-CAUS-if

‘If the poison of a tree causes (you) to scratch (yourself) badly’ → ‘If the poison of a tree scratches badly’.

4.7.3. Reciprocal markers > AMs

Reciprocals remain a lesser explored source of AMs. A possible example comes from Bantu languages.²¹ In Orungu, the suffix *-an* can be used in both reciprocal and anticausative function, as shown in (39a-b).²²

²¹ Other putative examples of the RECIPROCAL > ANTICAUSATIVE shift are discussed by Bahrt (2021: 192–195). However, since the markers discussed by Bahrt also display a reflexive function, I have excluded them from my sample.

²² See Bostoen & Nzang-Bie (2010: 1276–1283) and Bostoen et al. (2015) for an exhaustive discussion of the polysemy of *-an* in various Bantu languages.

(39) Orungu (Atlantic-Congo; Ambouroue 2007: 191)

- a. *-r̀̀nd̀̀à* ‘love’ → *-r̀̀nd̀̀à̀̀* ‘love each other’
 b. *-β̀̀r̀̀à* ‘bend (tr.)’ → *-β̀̀r̀̀à̀̀* ‘bend (intr.)’

Comparative evidence suggests that the reciprocal meaning is primary and the suffix is reconstructable to Proto-Bantu as an associative/reciprocal suffix **-an*, etymologically related to the comitative preposition *n(a)-* ‘with’ (Schadeberg & Bostoen 2019: 174, 182–184). Bostoen et al. (2015) suggest that the link between the reciprocal and anticausative use of **-an* should be sought in the fact that both situations are associated with low elaboration of events (Kemmer 1993; Lichtenberk 2000). This view is partly questionable, in that it is not clear why anticausatives should feature low degree of elaboration of events/participants, as they are simply one-participant events (see Inglese 2022a: 521). More importantly, it does not explain how and out of which contexts the anticausative meaning actually arose. Based on data from Hittite, Inglese (2020: 238–239) argues that a possible bridging context between reciprocals and anticausatives is offered by the class of lexical spatial reciprocals of the type ‘gather (intr.)’. These, much in the same vein as autocausatives, may also license non-agentive subjects, e.g. ‘the leaves gathered (because of the wind)’, thus giving rise to a spontaneous reading which can be then extended to decausative verbs proper. More research is needed to assess the likelihood of this diachronic scenario for Bantu languages.

4.7.4. AM > reflexive markers?

As discussed in Section 4.1, I have excluded from my sample those markers that besides anticausatives synchronically also function as reflexives, based on the widespread assumption that with these the reflexive function must be the original one. Nevertheless, there is evidence that syncretic anticausative/reflexive markers can also originate from non-reflexive sources and develop a reflexive function only after the anticausative one. If this is correct, then the numbers in Table 1 underestimate the actual frequency of non-reflexive sources of anticausatives. For reasons of space, I will only discuss the case of Yuracare and refer to Inglese (2022b) for other possible examples.

Yuracare features a suffix *-tA* that functions, among other things, as an anticausative and reflexive marker (van Gijn 2010). Notably, while in reflexive usage the suffix *-tA* is opposed to unmarked transitive verbs, as in (40a), in its anticausative function it typically stands in an equipollent alternation with verbs marked by causative suffixes such as *-pi* or *-che*, as in (40b).

(40) Yuracare (isolate, South America; van Gijn 2010: 277, 278)

a. *chumë* ‘cut (tr.)’ → *chumë-të* ‘cut oneself’

b. *pishij-pi* ‘break (tr.)’ vs. *pishij-ta* ‘break (intr.)’

c. *dürrüm ta-ø = ya*

IDEO.SI say-3 = REP

“Broom” it went.’

Van Gijn (2010) convincingly shows that *-tA* is etymologically connected to the verb *ta-* ‘say’. As a full verb, *ta-* could be involved in combinations with sound-imitating ideophones ‘say X’, as in (40a). Following the developmental trajectory discussed in Section 4.2.3, *ta-* was progressively extended to other visual and more abstract ideophones, thus undergoing semantic bleaching, and also became morphologically bound to the preceding ideophone. As a result, it effectively grammaticalized into a new suffix *-ta* ‘say, be, become, do X’ which essentially functioned as a verbalizer creating intransitive change-of-state verbs (Section 4.6). At this point, IDEO-*ta* verbs started being paired with the independently created causative constructions IDEO-*che/pi* (or other suffixes), thereby giving rise to an equipollent anticausative alternation. Confirmation for this reconstruction comes from the fact that most roots that take part in the pattern (40b) are in fact ideophones. The anticausative use of *-tA* served as basis for the extension to other valency-reducing functions, including the reflexive, which must therefore be secondary.

How the ANTICAUSATIVE > REFLEXIVE shift actually takes place remains unclear. Inglese (2020: 236–237) argues that a key role is played by autocasative motion verbs, which, due to their intermediate semantic status, may serve as bridging contexts between reflexives and anticausatives in either direction (see Section 2). A crucial piece of evidence in support of the scenario proposed in Inglese (2020) comes from the fact that markers that originate out of non-agentive sources may partially extend to the autocasative domain without ever fully expanding to reflexive contexts

proper (in some cases, the only reflexive-like verb discussed in the sources is ‘hide’, whose reflexive status is however questionable). For example, the Teltzal infix <*j*>, which derives from a passive source (Section 4.7.1), is also attested with the anticausative verb *lijk* ‘rise (oneself) up’ (Polian 2013: 293), or the Hausa suffix *-u* of resultative origin (Section 4.5.1), is also used with the anticausative reciprocal verb *târ-u-* ‘(the doctors) gather’ (Jaggar 2001: 265). This evidence suggests that anticausatives are not necessarily an extension of reflexive verbs as an intermediate step in the REFLEXIVE > ANTICAUSATIVE shift (Section 3), but that they may actually arise as a direct extension of decausative markers proper to contexts with more agent-like subjects.

5. Discussion

5.1. *Where do AMs come from?*

In Section 4, I have shown that AMs cross-linguistically derive from a much more varied pool of sources than what is reported in the typological literature. Overall, the changes whereby the source constructions surveyed in Sections 4.2 to 4.7 develop into AMs boil down to two general paths.

Sources that already lexicalize an uncontrolled change-of-state semantics are particularly suitable to turn into AMs. This is the case of verbs ‘become’ and of spontaneous event markers. Other sources first undergo some intermediate, often metaphorical, semantic shifts in order to acquire change-of-state semantics. For example, intransitive inactive verbs all inherently denote uncontrolled change-of-state events undergone by a Patient participant, and they all first develop into ‘be(come)’ verbs and subsequently into anticausatives. The verb ‘go’ acquires a change-of-state component metaphorically through its change-of-location semantics, and this might be the case of other deictic spatial sources. Acquisitive verbs ‘get’ also develop into anticausative markers through a stage in which they function as motion verbs. Ablative sources possibly first acquire a reversive function, which out of some contexts may give rise to the anticausative semantics. Finally, aspectual sources share different semantic components with anticausatives: ingressesives and resultatives already lexicalize a change-of-state component, while stative markers feature the necessary uncontrolled semantics. All these cases can be subsumed under a more general (SOURCE) > BECOME > ANTICAUSATIVE shift.

Transitive verbs do not obviously share any semantic component with AMs. Their development into anticausatives follows two trajectories. Generic action verbs ‘do, say’ first develop into inchoative verbalizers (especially when combined with ideophones) and subsequently into anticausatives. Other transitive verbs, e.g. ‘give’ (and partly ‘get’), likely go through an intermediate stage as causative markers.

For other sources, the connection with anticausatives lies with intransitivization, as these sources lack a spontaneous change-of-state component. Instead, all these constructions are first reinterpreted as (agentless) intransitive verb constructions and eventually extend to anticausative contexts. This is the case of (predicative) nominalizers and of non-reflexive voice markers. Concerning the latter, the specific mechanisms are quite varied. Causatives either first develop into passives, or they directly develop into anticausatives due to the reinterpretation of agentless constructions as intransitive ones. Passives possibly develop into anticausatives via impersonalization. Finally, reciprocals license an anticausative-like reading only with verbs that lexically encode symmetric spatial relations (e.g., *gather*).

Note also that the same source can turn into an AM following multiple pathways, as is the case of verbs ‘get’, which either follow the trajectory of motion verbs or that of causatives (Section 4.2.3). The pathways of development of AMs can be schematized in the network in Figure 1.²³

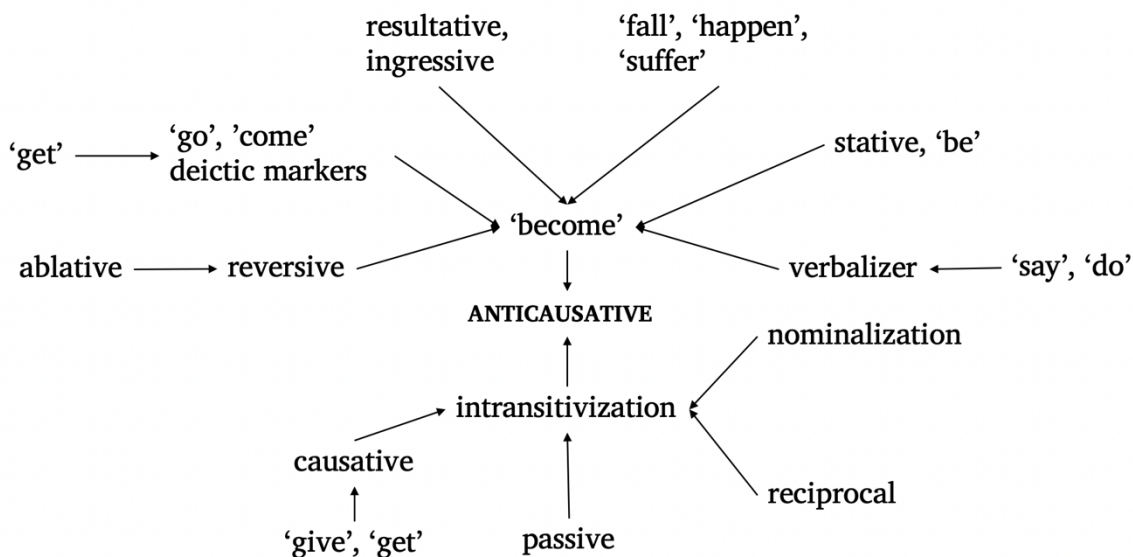


Figure 1: The origin of AMs

²³ The network merely visualizes how the various sources relate to the anticausative function, but by no means constitute a semantic map of anticausativization in the technical sense of Georgakopoulos & Polis (2018). A semantic map of anticausativization has been proposed by Haspelmath (1987: 35).

5.2. *How does the anticausative alternation come about?*

In Section 4, I have investigated the (lexical) sources of AMs in the languages of the world. A distinct question concerns how the anticausative alternation emerges in the first place. Addressing this question falls out of the scope of this paper, so that I will limit myself to some preliminary considerations here.

The developmental pathways of AMs discussed in Section 5.1 partly account for how individual AMs start being involved in the anticausative alternation. As hinted in Section 4.6, in the case of BECOME-sources the anticausative alternation only emerges once noncausal BECOME-verbs start being paired with corresponding causal verbs, as discussed for the Yurakare suffix *-tA* (Section 4.7.4). Notably, in this scenario the rise of the anticausative alternation constitutes a logically distinct and secondary phenomenon with respect to the diachronic processes that lead individual sources to develop a spontaneous change-of-state semantics. By contrast, INTRANSITIVIZING-source constructions are already involved in a transitivity alternation to begin with (with the likely exception of nominalizers), so that with these the anticausative alternation constitutes the result of individual sources extending to the expression of spontaneous change-of-state events.

Another question concerns the origin of anticausative and equipollent patterns. In principle, one may speculate that specific source types will preferably give rise to either pattern. For example, AMs that go back to light verb constructions with ‘do, say’ verbs might be expected to give rise to equipollent patterns, as their causal counterpart will likely be marked by a corresponding causative light verb, as is the case of Hindi in (11) and Jaminjung in (13) (as similar scenario may hold for several BECOME-sources). By contrast, sources that already combine with verbs in auxiliary verb constructions, coverb constructions or serial verb constructions might be connected with the anticausative pattern, as shown by Lezgian in (10), Mose-tén in (14) and Korean in (12), respectively.²⁴ Unfortunately, in most cases we lack the necessary historical evidence to assess the syntactic pattern in which sources of AMs originally occurred, so that this hypothesis cannot be thoroughly tested at present.

²⁴ I owe this observation to one of the anonymous reviewers.

Source	Anticausative pattern	Equipollent pattern
Lexical verbs	14	17
‘be(come)’	5	4
‘come’	1	0
‘do’	0	2
‘fall’	1	2
‘get’	2	0
‘give’	1	1
‘go’	2	2
‘happen’	0	1
‘hit’	1	0
‘say, do’	1	4
‘suffer’	0	1
Spatial markers	2	1
Spontaneous events markers	1	1
Aspectual markers	4	3
Nominalizers and verbalizers	3	-
Non-reflexive voice markers	4	1

Table 3: Sources of AMs and types of alternation

At any rate, the data from my sample, shown in Table 3, only partially supports the idea that the nature of the source construction determines the preference for one pattern over the other.²⁵ For example, as expected voice markers seem to preferably give rise to anticausative patterns and so do verbalizers and nominalizers. For the other source types no clear pattern can be detected, even if one considers the more fine-grained etymology of lexical verb sources, in part because the numbers are so small that they do not allow compelling generalizations. Nevertheless, the fact that there appears to be a weak link between type of sources and types of patterns further supports to the idea that markers that occur in the two patterns are not different in nature (see Section 4.1).

5.3. *Are anticausative a type of reflexives?*

Especially in formally-oriented research, which has mostly focused on syncretic reflexive-anticausative markers in Germanic and Romance languages, the relationship between reflexives and anticausatives has been the object of much discussion focusing

²⁵ AMs that are in both detransitivizing and equipollent pairs have been counted twice in Table 3.

on whether anticausatives constitute a type of reflexives or not (see Tubino-Blanco 2020: 22–29). According to some authors, for example Koontz-Garboden (2009), anticausativization is in fact a type of reflexivization. This view has been challenged by e.g. Horvath & Siloni (2011), who instead argue that anticausativization must be kept distinct from reflexivity proper, despite the fact that the two may be co-expressed. Indeed, Koontz-Garboden (2009: 92) maintains that “a compelling argument for the reflexivization approach to anticausativization comes from the fact that anticausativization and reflexivization are very commonly marked identically to one another cross-linguistically.” I do not wish to enter here into the debate concerning the status of reflexives and anticausatives, but I would like to point out that the syncretism argument made by Koontz-Garboden (2009) is a rather weak one, since, as demonstrated in this paper, languages often feature non-reflexive anticausativization strategies. In particular, the diachronic evidence discussed in this paper shows that AMs can arise out of source constructions that have little or nothing to do with reflexivity, so that it is not clear how a reflexive analysis may be viable for these.

6. Conclusions

In this paper, I have offered a detailed discussion of the sources and processes that lead to the rise of AMs in a sample of 98 languages. Against the *communis opinio* that AMs invariably grammaticalize out of reflexive and passive markers, I have shown that AMs in fact emerge out of a much wider pool of sources in the languages of the world. I have also discussed how the variety of sources observed in fact boils down to two main developmental paths: individual sources develop into AMs either because they are prone to developing a change-of-state semantics or because they are connected with intransitivity. Another interesting finding is that there does not appear to exist a strong correlation between types of sources and the type of morphosyntactic pattern in which individual AMs are involved in terms of the anticausative vs. equipollent distinction. This speaks to the fact that AMs can rightfully be considered so irrespective of the pattern in which they occur. More research is needed to better understand how and why the two patterns come about. Finally, the diachronic evidence presented in this paper on non-reflexive sources of AMs also compels us to rethink the purported connection between reflexivity and anticausativization, and suggests that anticausativization, at least historically, is a phenomenon of its own.

Acknowledgements

I would like to thank Simone Mattioli for reading through an earlier draft of this paper as well as two anonymous reviewers for their insightful observations, which have contributed to improving this paper. All remaining shortcomings are my own.

Abbreviations

1 = 1 st person	DJ = disjoint	OPT = optative
2 = 2 nd person	DU = dual	PASS = passive
3 = 3 rd person	F = feminine	PTCP = participle
ACC = accusative	FV = final vowel	PL = plural
AD = adessive relation	HON = honorific	PRS = present
ANTC = anticausative	IDEO = ideophone	PPP = past passive participle
AOR = aorist	INF = infinitive	PST = past
AUTO = autive	INTR = intransitive	REP = reportative
CAUS = causative	IPFV = imperfective	SUBJ = subject
DAT = dative	LKN = linking element	SG = singular
DEC = declarative	LOC = locative	SI = sound imitation
DEF = definite article	NOM = nominative	TR = transitive
DET = determiner	NSG = non-singular	

References

- Ahn, Mikyung & Foong Ha Yap. 2017. From middle to passive: A diachronic analysis of Korean *-eci* constructions. *Diachronica* 34(4). 437–469.
- Aikhenvald, Alexandra. 2011. Word-class changing derivation in typological perspective. In Alexandra Aikhenvald & R. M. W. Dixon (eds.), *Language at Large: Essays on Syntax and Semantics*, 221–289. Leiden: Brill.
- Alexiadou, Artemis, Elena Anagnostopoulou & Florian Schäfer. 2015. *External Arguments in Transitivity Alternations: A Layering Approach*. Oxford: Oxford University Press.
- Alexiadou, Artemis. 2010. On the morpho-syntax of (anti-)causative verbs. In Malka Rappaport Hovav, Edit Doron & Ivy Sichel (eds.), *Lexical Semantics, Syntax, and Event Structure*, 177–203. Oxford: Oxford University Press.
- Alhoniemi, Alho. 1993. *Grammatik des Tscheremissischen (Mari)*. Hamburg: Buske.
- Allan, Rutger J. 2003. *The middle voice in ancient Greek: a study in polysemy*. Amsterdam: J.C. Gieben.

- Alpher, Barry, Nicholas Evans & Mark Harvey. 2003. Proto-Gunwinyguan verb suffixes. In Nicholas Evans (ed.), *The Non-Pama-Nyungan languages of Northern Australia*, 305–352. Canberra: Pacific Linguistics.
- Ambouroué, Odette. 2007. *Eléments de description de l'orungu. Langue bantu du Gabon (B11b)*. PhD Dissertation, Université Libre de Brussels.
- Anderson, Gregory D. S. 2006. *Auxiliary Verb Constructions*. Oxford: Oxford University Press.
- Armendáriz, Rolando Félix. 2005. *A Grammar of River Warih'io*. PhD Dissertation, Rice University.
- Authier, Gilles. 2012. The detransitive voice in Kryz. In Gilles Authier & Katharina Haude (eds.), *Ergativity, Valency and Voice*, 133–164. Berlin / New York: de Gruyter.
- Bahrt, Nicklas N. 2021. *Voice syncretism*. Berlin: Language Science Press.
- Barlow, Russel. 2018. *A Grammar of Ulwa*. PhD Dissertation, University of Hawai'i at Mānoa.
- Beck, David. 2008. Ideophones, Adverbs, and Predicate Qualification in Upper Necaxa Totonac. *International Journal of American Linguistics* 74(1). 1–46.
- Beck, David. 2011. *Upper Necaxa Totonac Dictionary*. Berlin / New York: de Gruyter.
- Beck, David. 2012. Apéndice: Tablas de morfología comparativa. In Paulette Levy & David Beck (eds.), *Las lenguas totonacas y tepehuas: textos y otros materiales para su estudio*, 587–596. Mexico City: UNAM Press.
- Beck, David. Forthcoming. Totonacan languages. In Søren Wichmann (ed.), *The Languages and Linguistics of Middle and Central America: A Comprehensive Guide*. Berlin / New York: de Gruyter.
- Berrettoni, Pierangiolo. 1971. Considerazioni sui verbi latini in *-scō*. *Studi e Saggi Linguistici* 11. 89–169.
- Bostoen, Koen, Sebastian Dom & Guillaume Segerer. 2015. The antipassive in Bantu. *Linguistics* 53(4). 731–772.
- Bostoen, Koen & Yolande Nzang-Bie. 2010. On how “middle” plus “associative/reciprocal” became “passive” in the Bantu A70 languages. *Linguistics* 48(6). 1255–1307.
- Bruening, Benjamin & Thuan Tran. 2015. The nature of the passive, with an analysis of Vietnamese. *Lingua* 165. 133–172.
- Bugaeva, Anna. 2015. Causative constructions in Ainu: A typological perspective with remarks on the diachrony. *STUF - Language Typology and Universals* 68(4). 439–

484.

- Bybee, Joan, Revere Perkins & William Pagliuca. 1994. *The Evolution of Grammar: Tense, Aspect, and Modality in the Languages of the World*. Chicago: University of Chicago Press.
- Campbell, Eric. 2011. Zenzontepec Chatino Aspect Morphology and Zapotecan Verb Classes. *International Journal of American Linguistics* 77(2). 219–246.
- Campbell, Eric. 2015. Valency classes in Zenzontepec Chatino. In Andrej Malchukov & Bernard Comrie (eds.), *Valency Classes in the World's Languages*, vol. 2, 1391–1426. Berlin / New York: de Gruyter.
- Castro Alves, Flávia de. 2004. *O Timbira Falado Pelos Canela Apãniekrá: Uma Contribuição aos Estudos da Morfossintaxe de uma Língua Jê*. PhD Dissertation, Universidad Estadual de Campinas.
- Cennamo, Michela. 1993. *The reanalysis of reflexives: a diachronic perspective*. Napoli: Liguori.
- Cennamo, Michela. 2012. Aspectual Constraints on the (Anti)Causative Alternation in Old Italian. *Transactions of the Philological Society* 110(3). 394–421.
- Cennamo, Michela. 2020a. The actualization of new voice patterns in Romance. Persistence in diversity. In Bridget Drinka (ed.), *Historical Linguistics 2017*, 109–142. Amsterdam / Philadelphia: John Benjamins.
- Cennamo, Michela. 2020b. Mechanisms and paths of grammaticalization and reanalysis in Romance. In Walter Bisang & Andrej Malchukov (eds.), *Grammaticalization scenarios: cross-linguistic variation and universal tendencies. Volume 1 Grammaticalization Scenarios from Europe and Asia*, 165–248. Berlin / New York: de Gruyter.
- Cennamo, Michela, Thórhallur Eythórsson & Jóhanna Barðdal. 2015. Semantic and (morpho)syntactic constraints on anticausativization: Evidence from Latin and Old Norse-Icelandic. *Linguistics* 53(4). 677–729.
- Cohen, David, Marie-Claude Simeone-Senelle & Martine Vanhove. 2002. The grammaticalization of ‘say’ and ‘do’: An areal phenomenon in East Africa. In Tom Güldemann & Manfred von Roncador (eds.), *Reported Discourse*, 227–251. Amsterdam / Philadelphia: John Benjamins.
- Constenla Umaña, Adolfo. 1998. *Gramática de lengua Guatusa*. Heredia (Costa Rica): EUNA.
- Creissels, Denis. 2006. *Syntaxe générale: une introduction typologique*. Paris: Lavoisier.

- Creissels, Denis & Séckou Biaye. 2016. *Le balant ganja*. Dakar: IFAN Cheikh Anta Diop.
- Cristofaro, Sonia. 2021. Towards a source-oriented approach to typological universals. In Peter Arkadiev, Jurgis Pakerys, Inesa Šeškauskienė & Žeimantienė Žeimantienė (eds.), *Studies in Baltic and Other Languages*, 97–117. Vilnius: Vilniaus Universiteto Leidykla.
- Croft, William. 2012. *Verbs: Aspect and Causal Structure*. Oxford: Oxford University Press.
- Cunha de Oliveira, Christiane. 2005. *The Language of the Apinajé People of Central Brazil*. Eugene: PhD Dissertation, University of Oregon (Eugene).
- Daudey, Gerdine Henriëtte. 2014. *A grammar of Wadu Pumi*. PhD Dissertation, La Trobe University.
- Dench, Alan. 1995. *Martuthunira: A Language of the Pilbara Region of Western Australia*. Canberra: Research School of Pacific and Asian Studies, ANU.
- Devos, Maud & Jenneke van der Wal (eds.). 2014. “COME” and “GO” off the Beaten Grammaticalization Path. Berlin / New York: de Gruyter.
- Dom, Sebastian, Leonid Kulikov & Koen Bostoen. 2016. The middle as a voice category in Bantu: Setting the stage for further research. *Lingua Posnaniensis* 58(2). 129–149.
- Dunn, Michael John. 1999. *A Grammar of Chukchi*. PhD Dissertation, Australian National University.
- Estrada Fernández, Zarina, Mercedes Tubino & Jesús Villalpando. 2015. Valency classes in Yaqui. In Andrej Malchukov & Bernard Comrie (eds.), *Valency Classes in the World's Languages*, vol. 2, 1359–1389. Berlin / New York: de Gruyter.
- Evans, Nicholas. 2003. *Bininj Gun-Wok: A Pan-dialectal Grammar of Mayali, Kunwinjku and Kune*. Canberra: Research School of Pacific and Asian Studies, ANU.
- Fauconnier, Stefanie. 2011. Involuntary agent constructions are not directly linked to reduced transitivity. *Studies in Language* 35(2). 311–336.
- Félix Armendáriz, Rolando. 2005. *A Grammar of River Warihio (Mexico)*. PhD Dissertation, Rice University.
- Filchenko, Andrey Yury. 2007. *A grammar of Eastern Khanty*. PhD Dissertation, Rice University.

- Foley, William A. 1991. *The Yimas Language of New Guinea*. Stanford: Stanford University Press.
- Forker, Diana. 2013. *A Grammar of Hinuq*. Berlin / New York: de Gruyter.
- Frachtenberg, Leo Joachim. 1922. Coos. In Franz Boas (ed.), *Handbook of American Indian Languages* 2, 297–429. Washington: Government Printing Office.
- Frellesvig, Bjarke & John Whitman. 2016. The historical source of the bigrade transitivity alternations in Japanese. In Taro Kageyama & Wesley M. Jacobsen (eds.), *Transitivity and Valency Alternations: Studies on Japanese and Beyond*, 289–310. Berlin / New York: de Gruyter.
- Frowein, Friedel Martin. 2011. *A grammar of Siar, an Oceanic language of New Ireland Province, Papua New Guinea*. PhD Dissertation, La Trobe University.
- Frowein, Friedel Martin. 2011. *A grammar of Siar, an Oceanic language of New Ireland Province, Papua New Guinea*. PhD Dissertation, La Trobe University.
- Gates, Jesse P., Sami Honkasalo & Yunfan Lai. 2022. From transitive to intransitive and voiceless to voiced in Proto-Sino-Tibetan. *Language & Linguistics* 23(2). 212–239.
- Geniušienė, Emma Š. 1987. *The Typology of Reflexives*. Berlin / New York: de Gruyter.
- Georgakopoulos, Thanasis & Stéphane Polis. 2018. The semantic map model: State of the art and future avenues for linguistic research. *Language and Linguistics Compass* 12(2). e12270.
- Gerds, Donna B. & Thomas E. Hukari. 2006. The Halkomelem Middle: A Complex Network of Constructions. *Anthropological Linguistics* 48(1). 44–81.
- Giacalone-Ramat, Anna & Andrea Sansò. 2014. *Venire* 'come' as a passive auxiliary in Italian. In Maud Devos & Jenneke van der Wal (eds.), *“COME” and “GO” off the Beaten Grammaticalization Path*, 21–44. Berlin / New York: de Gruyter.
- Gibert-Sotelo, Elisabeth. 2018. Deriving ablative, privative, and reversative meanings in Catalan and Spanish. *Borealis* 7(2). 161–185.
- Gijn, Rik van. 2010. Middle voice and ideophones, a diachronic connection: The case of Yurakaré. *Studies in Language* 34(2). 273–297.
- Gil, David. 2017. Roon *ve*, DO/GIVE coexpression, and language contact in Northwest New Guinea. *NUSA* 62. 41-100.
- Gravelle, Gilles. 2002. Morphosyntactic properties of Meyah word classes. In Ger P. Reesink (ed.), *Languages of the eastern Bird's Head*, 109–180. Canberra: Research School of Pacific and Asian Studies.
- Gravelle, Gloria Jean. 2010. *A Grammar of Moskona: an East Bird's Head Language of*

- West Papua, Indonesia*. PhD Dissertation, Vrije Universiteit Amsterdam.
- Grestenberger, Laura. 2016. Reconstructing Proto-Indo-European Deponents. *Indo-European Linguistics* 4(1). 98–149.
- Grinevald, Colette. 1990. *A grammar of Rama*. Report to National Science Foundation.
- Gronemeyer, Claire. 1999. On deriving complex polysemy: the grammaticalization of *get*. *English Language & Linguistics* 3(1). 1–39.
- Guérois, Rozenn. 2015. *A grammar of Cuwabo (Mozambique, Bantu P34)*. PhD Dissertation, Université Lumière Lyon 2.
- Guérois, Rozenn & Koen Bostoen. 2018. On the origins of passive allomorphy in Cuwabo (Bantu P34). *Southern African Linguistics and Applied Language Studies* 36(3). 211–233.
- Gunnik, Hilde. 2018. *A grammar of Fwe. A Bantu language of Zambia and Namibia*. PhD Dissertation, Ghent University.
- Haas, Mary. 1977. From auxiliary verb to inflectional suffix. In Charles Li (ed.), *Mechanisms of Syntactic Change*, 525–537. Austin: University of Texas Press.
- Haiman, John. 2011. *Cambodian. loall.16*. Amsterdam / Philadelphia: John Benjamins.
- Handel, Zev. 2012. Valence-Changing Prefixes and Voicing Alternation in Old Chinese and Proto-Sino-Tibetan: Reconstructing *s- and *N- Prefixes. *Language and Linguistics* 13(1). 61–82.
- Hardarson, Jón A. 1998. Mit dem Suffix *-eh1- bzw. *(e)h1-je/o- gebildete Verbalstämme im Indogermanischen. In Wolfgang Meid (ed.), *Sprache und Kultur der Indogermanen: Akten der X. Fachtagung der Indogermanischen Gesellschaft Innsbruck, 22.-28. September 1996*, 323–339. Innsbruck: Institut für Sprachwissenschaft.
- Hardy, Donald E. 1994. Middle Voice in Creek. *International Journal of American Linguistics* 60(1). 39–68.
- Hardy, Heather K. & Timothy Montler. 1991. The formation of the Alabama middle voice. *Lingua* 85(1). 1–15.
- Haspelmath, Martin. 1987. *Transitivity alternations of the anticausative type*. Köln: Institut für Sprachwissenschaft.
- Haspelmath, Martin. 1990. The Grammaticization of Passive Morphology. *Studies in Language* 14(1). 25–72.
- Haspelmath, Martin. 1993a. *A Grammar of Lezgian*. Berlin / New York: de Gruyter.
- Haspelmath, Martin. 1993b. More on the typology of the inchoative/causative verb

- alternations. In Bernard Comrie & Maria Polinsky (eds.), *Causatives and Transitivity*, 87–120. Amsterdam / Philadelphia: John Benjamins.
- Haspelmath, Martin. 2016. Universals of causative and anticausative verb formation and the spontaneity scale. *Lingua Posnaniensis* 58(2). 33–63.
- Healey, Phyllis M. 1965. *Levels, constituent strings, and agreement in Telefol syntax*. PhD Dissertation, Australian National University.
- Heath, Jeffrey. 2005. *A Grammar of Tamashek (Tuareg of Mali)*. Berlin / New York: de Gruyter.
- Heath, Jeffrey. 2014. *Grammar of Humburi Senni (Songhay of Hombori, Mali)*. Ms.
- Heidinger, Steffen. 2010. *French anticausatives: A diachronic perspective*. Berlin / New York: de Gruyter.
- Heine, Bernd. 2002. On the role of context in grammaticalization. In Ilse Wischer & Gabriele Diewald (eds.), *New Reflections on Grammaticalization*, 83–102. Amsterdam / Philadelphia: John Benjamins.
- Hill, Jane H. 2005. *A Grammar of Cupeño*. Berkeley & Los Angeles: University of California Press.
- Hill, Jane H. & Kenneth C. Hill. 2019. *Comparative Takic Grammar*. Survey Reports, Survey of California and Other Indian Languages. Berkeley.
<https://escholarship.org/uc/item/6tr732gg>.
- Holvoet, Axel. 2020. *The Middle Voice in Baltic*. Amsterdam / Philadelphia: John Benjamins.
- Honeyman, Thomas. 2017. *A grammar of Momu, a language of Papua New Guinea*. PhD Dissertation, Australian National University.
- Horvath, Julia & Tal Siloni. 2011. Anticausatives: Against reflexivization. *Lingua* 121(15). 2176–2186.
- Hualde, José Ignacio & Jon Ortiz de Urbina (eds.). 2003. *A Grammar of Basque*. Berlin / New York: de Gruyter.
- Inglese, Guglielmo & Simone Mattioli. 2020. Pluractionality in Hittite: A new look at the suffix -ške/a-. *STUF - Language Typology and Universals* 73(2). 261–303.
- Inglese, Guglielmo. 2020. *The Hittite middle voice: synchrony, diachrony, typology*. Leiden: Brill.
- Inglese, Guglielmo. 2021. Anticausativization and basic valency orientation in Latin. In Silvia Luraghi & Elisa Roma (eds.), *Valency over time*, 133–168. Berlin / New York: de Gruyter.
- Inglese, Guglielmo. 2022a. Towards a typology of middle voice systems. *Linguistic*

- Typology* 26(3). 489–531.
- Inglese, Guglielmo. 2022b. The rise of middle voice systems: a study in diachronic typology. *Diachronica* aop.
- Jacques, Guillaume. 2015. The spontaneous-autobenefactive prefix in Japhug Rgyalrong. *Linguistics of the Tibeto-Burman Area* 38(2). 271–291.
- Jacques, Guillaume. 2016. How Many *-s Suffixes in Old Chinese? *Bulletin of Chinese Linguistics* 9(2). 205–217.
- Jacques, Guillaume. 2021. *A grammar of Japhug*. Language Science Press. Berlin: Language Science Press.
- Jaggar, Philip. 1988. Affected-subject ('grade 7') verbs in Hausa: What are they and where do they come from. In Masayoshi Shibatani (ed.), *Passive and Voice*, 387–416. Amsterdam / Philadelphia: John Benjamins.
- Jaggar, Philip. 2001. *Hausa*. Amsterdam / Philadelphia: John Benjamins.
- Jakobi, Angelika, Joachim Crass & Bakhit Seby Abdoulaye. 2004. *Grammaire du beria (langue saharienne)*. Köln: Köppe.
- Janhunen, Juha A. 2012. *Mongolian*. Amsterdam / Philadelphia: John Benjamins.
- Jasanoff, Jay H. 2004. 'Stative' *-ē- revisited. *Die Sprache* 43. 127–170.
- Joswig, Andreas. 2019. *The Majang Language*. Utrecht: LOT.
- Karaj, Dawid & Andrea Sansò. Forthcoming. From agent-oriented verbalizer to middle marker: The diachrony of the middle voice in Malayo-Sumbawan. *STUF - Language Typology and Universals*.
- Kawachi, Kazuhiro. 2007. *A Grammar of Sidaama (Sidamo), a Cushitic Language of Ethiopia*. PhD Dissertation, State University of New York.
- Kemmer, Suzanne. 1993. *The Middle Voice*. Amsterdam / Philadelphia: John Benjamins.
- Kim, Yuni. 2005. *Topics in the Phonology and Morphology of San Francisco del Mar Huave*. PhD Dissertation, University of California (Berkeley).
- Kittilä, Seppo. 2000. Causative morphemes as a de-transitivizing device: what do non-canonical instances reveal about causation and causativization? *Folia Linguistica* 47(1). 113–137.
- Koontz-Garboden, Andrew. 2009. Anticausativization. *Natural Language & Linguistic Theory* 27(1). 77–138.
- Koul, Omkar Nath. 2008. *Modern Hindi grammar*. Springfield: Dunwoody Press.
- Kouwenberg, Norbertus J. C. 2010. *The Akkadian Verb and Its Semitic Background*. Winona Lake (Indiana): Eisenbrauns.

- Kulikov, Leonid. 2011. Passive to anticausative through impersonalization. The case of Vedic and Indo-European. In Andrej Malchukov & Anna Siewierska (eds.), *Impersonal Constructions: A Cross-linguistic Perspective*, 229–254. Amsterdam / Philadelphia: John Benjamins.
- Kulikov, Leonid. 2012. *The Vedic -ya-presents: Passives and Intransitivity in Old Indo-Aryan*. Amsterdam: Rodopi.
- Kullmann, Rita & D. Tserenpil. 2008. *Mongolian Grammar*. Ulaanbaatar: Academy of Sciences.
- Kung, Susan Smythe. 2007. *A Descriptive Grammar of Huehuetla Tepehua*. PhD Dissertation, University of Texas (Austin).
- Kuteva, Tania, Bernd Heine, Bo Hong, Haiping Long, Heiko Narrog & Seongha Rhee. 2019. *World Lexicon of Grammaticalization*. Cambridge: Cambridge University Press.
- LaPolla, Randy. 1996. Middle voice marking in Tibeto-Burman. *Proceedings of the Fourth International Symposium on Languages and Linguistics: Pan-Asiatic Linguistics (January 8-10, 1996)* 5. 1940–1954.
- Lazzeroni, Romano. 1990. La diatesi come categoria linguistica. Studi sul medio indoeuropeo. *Studi e Saggi Linguistici* 33(2). 11–23.
- Lazzeroni, Romano. 2004. Inaccusatività indoeuropea e alternanza causativa vedica. *Archivio Glottologico Italiano* 89(2). 139–164.
- Lehmann, Christian. 2015. *Thoughts on grammaticalization (3rd edn.)*. Berlin: Language Science Press.
- Lenz, Alexandra N. & Gudrun Rawoens. 2012. The art of getting: GET verbs in European languages from a synchronic and diachronic point of view: Introduction. *Linguistics* 50(6). 1075–1078.
- Levin, Beth. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago: University of Chicago Press.
- Lichtenberk, Frantisek. 2000. Reciprocals without reflexives. In Zygmunt Frajzyngier & Traci S. Curl (eds.), *Reciprocals: Forms and Functions*, 32–62. Amsterdam / Philadelphia: John Benjamins.
- Liljegren, Henrik. 2016. *A grammar of Palula*. Berlin: Language Science Press.
- López Nicolás, Óscar. 2016. *Estudios de la fonología y gramática del zapoteco de Zochina*. PhD Dissertation, CIESAS (Ciudad de México).
- Luraghi, Silvia. 2012. Basic valency orientation and the middle voice in Hittite. *Studies in Language* 36(1). 1–32.

- Luraghi, Silvia, Guglielmo Inglese & Daniel Kölligan. 2021. The passive voice in ancient Indo-European languages: inflection, derivation, periphrastic verb forms. *Folia Linguistica Historica* 42(2). 339–391.
- MacKay, Carolyn Joyce. 1999. *A Grammar of Misantra Totonac*. Salt Lake City: University of Utah Press.
- Magni, Elisabetta. 2010. L'evoluzione semantico-funzionale dell'elemento -θ- nella morfologia verbale del greco. In Ignazio Putzu, Giulio Paulis, Gianfranco Nieddu & Pierluigi Cuzzolin (eds.), *La morfologia del greco tra tipologia e diacronia*, 266–287. Milano: FrancoAngeli.
- Malchukov, Andrej & Igor V. Nedjalkov. 2015. Valency classes in Even (North Tungusic) in a comparative Tungusic perspective. In Andrej Malchukov & Bernard Comrie (eds.), *Valency Classes in the World's Languages*, vol. 1, 571–627. Berlin / New York: de Gruyter.
- Malchukov, Andrej. 2015. Valency classes and alternation: parameters of variation. In Andrej Malchukov & Bernard Comrie (eds.), *Valency Classes in the World's Languages*, vol. 1, 73–130. Berlin / New York: de Gruyter.
- Martin, Jack B. 2011. *A Grammar of Creek (Muskogee)*. Lincoln / London: University of Nebraska Press.
- Martínez Corripio, Israel & Ricardo Maldonado. 2010. Middles and reflexives in Yucatec Maya: Trusting speaker intuition. In Daisy Rosenblum, Jean Gail Mulder & Andrea L. Berez (eds.), *Fieldwork and linguistic analysis in indigenous languages of the Americas*, 147–171. Honolulu: University of Hawai'i Press.
- Martínez Rojas, Lucía, Verónica Orqueda, Francisca Toro Varela, & Berta González Saavedra. 2021. Desarrollo diacrónico de las funciones de *se* y *sibi* del latín arcaico al latín clásico. *Boletín de Filología* 56(2). 483–517.
- Martins, Silvana Andrade. 2004. *Fonologia e Gramática Dâw*. Utrecht: LOT.
- Matisoff, James. 2003. *Handbook of Proto-Tibeto-Burman: System and Philosophy of Sino-Tibetan Reconstruction*. Los Angeles: University of California Press.
- Mattiola, Simone & Andrea Sansò. 2021. A typology of denominal verb formation. Paper presented at the 54th SLE Meeting, online.
- McFarland, Teresa Ann. 2009. *The phonology and morphology of Filomeno Mata Totonac*. PhD Dissertation, University of California (Berkeley).
- McGregor, William B. 2002. *Verb Classification in Australian Languages*. Berlin / New York: de Gruyter.
- McGregor, William B. 2013. Grammaticalisation of verbs into temporal and modal

- markers in Australian languages. In Folke Josephson & Ingmar Söhrman (eds.), *Diachronic and typological perspectives on verbs*, 107–132. Amsterdam / Philadelphia: John Benjamins.
- Mithun, Marianne. 2000. Valency-changing derivation in Central Alaskan Yup'ik. In R. M. W. Dixon & Alexandra Aikhenvald (eds.), *Changing Valency: Case Studies in Transitivity*, 84–114. Cambridge: Cambridge University Press.
- Mocciaro, Egle. 2014. Passive in motion: the Early Italian auxiliary *andare* ('to go'). In Maud Devos & Jenneke van der Wal (eds.), *"COME" and "GO" off the Beaten Grammaticalization Path*, 45–68. Berlin / New York: de Gruyter.
- Morgan, Lawrence. 1991. *A Description of the Kutenai Language*. PhD Dissertation, University of California (Berkeley).
- Mushin, Ilana. 2012. *A Grammar of (Western) Garrwa*. Berlin / New York: de Gruyter.
- Narrog, Heiko. 2016. Japanese transitivity pairs through time—a historical and typological perspective. In Taro Kageyama & Wesley M. Jacobsen (eds.), *Transitivity and Valency Alternations: Studies on Japanese and Beyond*, 249–268. Berlin / New York: de Gruyter.
- Nava, Fernando & Ricardo Maldonado. 2004. Basic Voice Patterns in Tarascan (P'orhepecha). In Michael Achard & Suzanne Kemmer (eds.), *Language, Culture, and Mind*, 461–477. CSLI Publications. Stanford.
- Nedjalkov, Igor V. 1993. Causative-passive polysemy of the Manchu-Tungusic *-bu/-v(u)-*. Antwerpen: Hoger Instituut voor Vertalers en Tolken 27. 193–202.
- Nedjalkov, Vladimir P. & Georgij G. Silnitsky. 1973. The Typology of Morphological and Lexical Causatives. In Ferenc Kiefer (ed.), *Trends in Soviet Theoretical Linguistics*, 1–32. Dordrecht: Reidel Publishing.
- Nichols, Johanna. 2011. *Ingush Grammar*. Los Angeles: University of California Press.
- Nichols, Johanna, David A. Peterson & Jonathan Barnes. 2004. Transitivity and detransitivizing languages. *Linguistic Typology* 8(2). 149–211.
- Nikolaeva, Irina & Maria Tolskaya. 2001. *A Grammar of Udihe*. Berlin / New York: de Gruyter.
- Okrand, Marc. 1977. *Mutsun Grammar*. PhD Dissertation, University of California (Berkeley).
- Pacchiarotti, Sara & Leonid Kulikov. 2022. Bribri media tantum verbs and the rise of labile syntax. *Linguistics*. 60(2). 617–657.
- Park, Indrek. 2012. *A grammar of Hidatsa*. PhD Dissertation, Indiana University.

- Parker, Gary J. 1976. *Gramática del quechua Áncash-Huaylas*. Lima: IEP.
- Pawley, Andrew K. 1972. On the internal relationship of eastern Oceanic languages. In Roger C. Green & Marian Kelly (eds.), *Studies in Oceanic culture history*, 1–142. Honolulu: Bernice Pauahi Bishop Museum.
- Peña, Jaime G. 2015. *A grammar of Wampis*. PhD Dissertation, University of Oregon.
- Petrollino, Sara. 2016. *A Grammar of Hamar: A South Omotic Language of Ethiopia*. PhD Dissertation, Leiden University.
- Pharris, Nicholas J. 2006. *Winuunsi Tm Talapaas: a grammar of the Molalla language*. PhD Dissertation, University of Michigan.
- Pinkster, Harm. 2015. *The Oxford Latin Syntax*. Oxford: Oxford University Press.
- Polian, Gilles. 2013. *Gramática del tseltal de Oxchuc*. Ciudad de Mexico: CIESAS.
- Post, Mark. 2007. *A Grammar of Galo*. PhD Dissertation, La Trobe University.
- Pucilowski, Anna. 2013. *Topics in Ho morphophonology and morphosyntax*. PhD Dissertation, University of Oregon.
- Rappaport Hovav, Malka & Beth Levin. 2010. Reflections on manner/result complementarity. In Malka Rappaport Hovav, Edit Doron & Ivy Sichel (eds.), *Lexical Semantics, Syntax, and Event Structure*, 21–38. Oxford: Oxford University Press.
- Reh, Mechthild. 1985. *Die Krongo-Sprache (N̄inò Mó-Dì)*. Berlin: Dietrich Reimer.
- Renck, Günther L. 1975. *A Grammar of Yagaria*. Canberra: Research School of Pacific and Asian Studies, ANU.
- Riese, Timothy. 2001. *Vogul*. Munich: Lincom Europa.
- Robinson, Stuart. 2011. *Split intransitivity in Rotokas, a Papuan language of Bougainville*. PhD Dissertation, Radboud Universiteit Nijmegen.
- Rojas Berscia, Luis Miguel. 2013. *La sintaxis y semántica de las construcciones causativas en el chayahuilla de Balsapuerto*. MA. Thesis, Pontificia Universidad Católica del Perú.
- Romagno, Domenica. 2014. The aorist in “-en” in Homeric Greek: at the morphosyntax-semantics interface: a thorough analysis of Iliad and Odyssey. *Archivio glottologico italiano* 99(2). 155–186.
- Rombandeeva, E. I. 1973. *Mansijskij (vogul'skij) jazyk*. Moscow: Nauka.
- Roset, Caroline. 2018. *A grammar of Darfur Arabic*. Utrecht: LOT.
- Ruijh, Cornelius J. 2004. The Stative Value of the PIE Verbal Suffix *-eh1-. In John Penney (ed.), *Indo-European Perspectives: Studies In Honour of Anna Morpurgo Davies*, 48–64. Oxford: Oxford University Press.

- Sakel, Jeanette. 2007. The verbness markers of Mosestén from a typological perspective. In Matti Miestamo & Bernhard Wälchli (eds.), *New Challenges in Typology: Broadening the Horizons and Redefining the Foundations*, 315–335. Berlin / New York: de Gruyter.
- Sakel, Jeanette. 2011. *A Grammar of Mosestén*. Berlin / New York: de Gruyter.
- Salminen, Mikko. 2016. *A Grammar of Umbeyajts as spoken by the Ikojts people of San Dionisio del Mar, Oaxaca, Mexico*. PhD Dissertation, James Cook University.
- Sansò, Andrea. 2016. Agent-defocusing constructions from nominalized VPs: A cross-linguistic type? *Studies in Language* 40(4). 894–954.
- Sansò, Andrea. 2017. Where do antipassive constructions come from?: A study in diachronic typology. *Diachronica* 34(2). 175–218.
- Sansò, Andrea. 2020. Routes towards the irrealis. *Transactions of the Philological Society* 118(3). 401–446.
- Sasaki, Kan. 2016. Anticausativization in the northern dialects of Japanese. In Taro Kageyama & Wesley M. Jacobsen (eds.), *Transitivity and Valency Alternations: Studies on Japanese and Beyond*, 183–214. Berlin / New York: de Gruyter.
- Schackow, Diana. 2015. *A grammar of Yakkha*. Berlin: Language Science Press.
- Schadeberg, Thilo. 1982. Les suffixes verbaux séparatifs en bantou. *Sprache und Geschichte in Afrika* 4. 55–66.
- Schadeberg, Thilo & Koen Bostoen. 2019. Word formation. In Mark Van de Velde, Koen Bostoen, Derek Nurse & Gérard Philippson (eds.), *The Bantu Languages (2nd edn.)*, 172–203. London: Routledge.
- Schäfer, Florian. 2008. *The Syntax of (anti-)causatives: External Arguments in Change-of-state Contexts*. Amsterdam / Philadelphia: John Benjamins.
- Schäfer, Florian. 2009. The Causative Alternation. *Language and Linguistics Compass* 3(2). 641–681.
- Schrock, Terrill. 2017. *The Ik language*. Language Science Press. Berlin: Language Science Press.
- Schultze-Berndt, Eva. 2000. *Simple and Complex Verbs in Jaminjung: A Study of Event Categorisation in an Australian Language*. PhD Dissertation, Katholieke Universiteit Nijmegen.
- Schultze-Berndt, Eva. 2008. What do “do” verbs do? The semantic diversity of generalised action verbs. In Elisabeth Verhoeven, Stavros Skopetas, Yong-Min Shin, Yoko Nishina & Johannes Helmbrecht (eds.), *Studies on Grammaticalization*, 185–208. Berlin / New York: de Gruyter.

- Segerer, Guillaume. 2002. *La langue bijogo de Bubaque (Guinée Bissau)*. Louvain: Peeters.
- Seidel, Frank. 2008. *A Grammar of Yeyi: A Bantu Language of Southern Africa*. Köln: Köppe.
- Shamin, Fatma. 2018. Conjunct verbs in Hindi. In Ghanshyam Sharma & Rajesh Bhatt (eds.), *Trends in Hindi Linguistics*, 214–244. Berlin / New York: de Gruyter.
- Simpson, Andrew & Hồ Hảo Tâm. 2013. Vietnamese and the typology of passive constructions. In Daniel Hole & Elisabeth Löbel (eds.), *Linguistics of Vietnamese: An International Survey*, 155–184. Berlin / New York: de Gruyter.
- Song, Jae Jung. 1996. *Causatives and Causation: A Universal -typological perspective*. London: Longman.
- Squartini, Mario. 2003. La grammaticalizzazione di <venire + participio> in Italiano. In Claus D. Pusch & Andreas Wesch (eds.), *Verbalperiphrasen in den (ibero-)romanischen Sprachen*, 23–34. Hamburg: Helmut Buske Verlag.
- Steeman, Sander. 2012. *A grammar of Sandawe a Khoisan language of Tanzania*. Utrecht: LOT.
- Suttles, Wayne P. 2004. *Musqueam Reference Grammar*. Vancouver: UBC Press.
- Troiani, Duna. 2007. *Fonología y morfosintaxis de la lengua totonaca: Municipio de Huehuetla, Sierra Norte de Puebla*. Mexico, D.F.: Instituto Nacional de Antropología e Historia.
- Tubino-Blanco, Mercedes. 2020. Causative-Inchoative in Morphology. *Oxford Research Encyclopedia of Linguistics*. Oxford: Oxford University Press.
- Turner, Katherine. 1987. *Aspects of Salinan Grammar*. PhD Dissertation, University of California (Berkeley).
- Vajda, Edward J. 2015. Valency properties of the Ket verb clause. In Andrej Malchukov & Bernard Comrie (eds.), *Valency Classes in the World's Languages*, vol. 1, 629–668. Berlin / New York: de Gruyter.
- Wagner, Peter. 2012. *A Grammar of North West Lovari Romani*. PhD Dissertation, Univerzita Karlova v Praze.
- Wakasa, Motomichi. 2008. *A Descriptive Study of the Modern Wolaytta Language*. PhD Dissertation, University of Tokyo.
- Watters, David. 2006. Notes on Kusunda Grammar: A language isolate of Nepal. *Himalayan Linguistics Archive* 3. 1–182.
- Wegener, Claudia. 2012. *A Grammar of Savosavo. A Grammar of Savosavo*. Berlin / New York: de Gruyter.

- Widmer, Manuel. 2018. *A Grammar of Bunan*. Berlin / New York: de Gruyter.
- Willi, Andreas. 2018. *Origins of the Greek Verb*. Cambridge: Cambridge University Press.
- Yap, Foong Ha & Mikyung Ahn. 2019. Development of grammatical voice marking in Korean: On the causative, middle and passive uses of suffix *-i*. *Lingua* 219. 1–23.
- Yarapea, Apoi Mason. 2006. *Morphosyntax of Kewapi*. PhD Dissertation, Australian National University.
- Yeon, Jaehoon & Lucien Brown. 2011. *Korean: A Comprehensive Grammar*. London: Routledge.
- Zúñiga, Fernando & Seppo Kittilä. 2019. *Grammatical Voice*. Cambridge: Cambridge University Press.

CONTACT

guglielmo.inglese@unito.it

Appendix: language sample

The sample used in this study derives from the combination of various existing samples of AMs and voice markers more generally (Nichols et al. 2004; Hartmann et al. 2013; Muysken et al. 2016; Bahrt 2021; Inglese 2022a) to which I have also added languages from other sources. Language names and genealogical classification are taken from Glottolog, consulted on 30/08/2022.

Language	Glottocode	Family	Marker	Origin	References
Afar	afar1241	Afro-Asiatic	Aux. <i>edhe</i>	verb ‘say, do’	(Cohen et al. 2002)
Ainu	ainu1251	Ainu	Suff. <i>-ke</i>	verb ‘do’	(Bugueva 2015)
Alabama	alab1237	Muskogean	Infix <i>-li</i>	unknown	(Hardy & Montler 1991)
Ancient Greek	anci1242	Indo-European	Suff. <i>-th(ē)</i>	resultative	(Luraghi et al. 2021)
Ancient Greek	anci1242	Indo-European	Suff. <i>-ē</i>	stative	(Luraghi et al. 2021)
Apinayé	apin1244	Nuclear-Macro-Je	<i>a</i> -series prefixes	unknown	(Cunha de Oliveira 2005)
Arabic, Sudanese	suda1236	Afro-Asiatic	Pref. <i>in-</i>	verb ‘say, do’	(Kouwenberg 2010; Roset 2018)
Balanta (Ganja)	bala1302	Atlantic-Congo	Suff. <i>-le</i>	unknown	(Creissels & Biaye 2016)
Basque	basq1248	Isolate	Aux. <i>da</i>	verb ‘be’	(Hualde & de Urbina 2003)
Bidyogo	bidy1244	Atlantic-Congo	Suff. <i>-ok</i>	verb ‘be’	(Seeger 2002)
Bininj Kun-Wok	gunw1252	Gunwinyguan	Suff. <i>-me</i>	verb ‘say, do’	(Evans 2003)
Bribri	brib1243	Chibchan	Suff. <i>-r</i>	unknown	(Pacchiarotti & Kulikov 2022)

Bunan	gahr1239	Sino-Tibetan	Suff. <i>-s</i>	stative	(Widmer 2018)
Canela	cane1242	Nuclear-Macro-Je	Pref. <i>-pi</i>	unknown	(de Castro Alves 2004)
Central Alaskan Yupik	cent2127	Eskimo-Aleut	Intransitive inflection	unknown	(Mithun 2000)
Chukchi	chuk1273	Chukotko-Kamchatkan	Suff. <i>-et</i>	unknown	(Dunn 1999)
Chuwabu	chuw1238	Atlantic-Congo	Suff. <i>-uw</i>	separative	(Guérois 2015)
Creek	cree1270	Muskogean	Suff. <i>-k</i>	auxiliary	(Hardy 1994; Martin 2011)
Cupeño	cupe1243	Uto-Aztecan	Suff. <i>-yax</i>	unknown	(Hill 2005; Hill & Hill 2019)
Dâw	daww1239	Naduhup	Tone shift	unknown	(Martins 2004)
East Kewa	east2516	Nuclear Trans New Guinea	Suff. <i>-ba/bi</i>	unknown	(Yarapea 2006)
East Khanty	east2774	Uralic	Suff. <i>-uj</i>	unknown	(Filchenko 2007)
Eastern Mari	east2328	Uralic	Conj. <i>-am</i>	unknown	(Alhoniemi 1993)
English	stan1293	Indo-European	Verb <i>get</i>	verb 'get'	(Gronemeyer 1999)
Even	even1260	Tungusic	Suff. <i>-b</i>	causative	(Malchukov & Nedjalkov 2015)
Even	even1260	Tungusic	Suff. <i>-rga</i>	unknown	(Malchukov & Nedjalkov 2015)
Even	even1260	Tungusic	Suff. <i>-lbe</i>	unknown	(Malchukov & Nedjalkov 2015)
Fang	fang1247	Atlantic-Congo	Suff. <i>-əbə</i>	unknown	(Bostoen & Nzang-Bie 2010)

Fwe	fwee1238	Atlantic-Congo	Suff. <i>-ahar</i>	unknown	(Gunnik 2018)
Galo	galo1242	Sino-Tibetan	Prevoicing	spontaneous	(Post 2007)
Garrwa	gara1269	Garrwan	Suff. <i>-j</i> (Class 2 verbs)	unknown	(Mushin 2012)
Hamer-Banna	hame1242	Afro-Asiatic	Suff. <i>-(a)?</i>	unknown	(Petrollino 2016)
Hanis	coos1249	Coosan	Suff. <i>-ē</i>	unknown	(Frachtenberg 1922)
Hausa	haus1257	Afro-Asiatic	Grade 7	resultative	(Jaggar 1988; Jaggar 2001)
Hidatsa	hida1246	Siouan	Middle inflection	unknown	(Park 2012)
Highland Totonac	high1243	Totonacan	Suff. <i>-kan</i>	unknown	(Troiani 2007)
Hindi	hind1269	Indo-European	Aux. <i>hona</i> ‘be’	verb ‘be’	(Koul 2008)
Hinukh	hinu1240	Nakh-Daghestanian	Suff. <i>-t</i>	unknown	(Forker 2013)
Ho	hooo1248	Austroasiatic	Suff. <i>-en</i>	unknown	(Pucilowski 2013)
Hokkaido Japanese	hokk1249	Japonic	Suff. <i>-rasar</i>	unknown	(Sasaki 2016)
Huambisa	huam1247	Chicham	Suff. <i>-na</i>	unknown	(Peña 2015)
Huarijio	huar1255	Uto-Aztecan	Suff. <i>-i</i>	unknown	(Armendáriz 2005)
Huarijio	huar1255	Uto-Aztecan	Suff. <i>-pa</i>	unknown	(Armendáriz 2005)
Huave, San Dionisio-San Mateo	sand1278	Huavean	Suff. <i>-j</i>	unknown	(Kim 2005; Salminen 2016)
Huaylas Ancash Quechua	huay1240	Quechuan	Suff. <i>-ka</i>	verb ‘be’	(Parker 1976)
Humburi Senni Songhay	humb1243	Songhay	Tone shift	unknown	(Heath 2014)

Ik	ikkk1242	Kuliak	Suff. <i>-Vm</i>	unknown	(Schrock 2017)
Ingush	ingu1240	Nakh-Daghestanian	Suff. <i>-lu</i>	verb ‘give’	(Nichols 2011)
Ingush	ingu1240	Nakh-Daghestanian	Aux. <i>d.uoda</i>	verb ‘go’	(Nichols 2011)
Italian	ital1282	Indo-European	Aux. <i>venire</i>	verb ‘come’	(Squartini 2003)
Italian	ital1282	Indo-European	Aux. <i>andare</i>	verb ‘go’	(Mocciaro 2014)
Jaminjung	jami1236	Mirndi	Suff. <i>-ijga</i>	verb ‘go’	(McGregor 2013)
Jaminjung	jami1236	Mirndi	Verb <i>-yu(nngu)</i>	verb ‘say, do’	(Schultze-Berndt 2000)
Japanese	nucl1643	Japonic	Suff. <i>-ar</i>	unknown	(Narrog 2016; Frellesvig & Whitman 2016)
Japanese	nucl1643	Japonic	Suff. <i>-e</i>	verb ‘get’	(Narrog 2016; Frellesvig & Whitman 2016)
Khmer (Central)	cent1989	Austroasiatic	Pref. <i>ra-</i>	unknown	(Haiman 2011)
Korean	kore1280	Koreanic	Suff. <i>-i</i>	causative	(Yeon & Brown 2011; Yap & Ahn 2019)
Korean	kore1280	Koreanic	Suff. <i>-eci</i>	verb ‘fall’	(Yeon & Brown 2011; Ahn & Yap 2017)
Krongo	kron1241	Kadugli-Krongo	Suff. <i>-ani</i>	unknown	(Reh 1985)
Krongo	kron1241	Kadugli-Krongo	Suff. <i>-i</i>	unknown	(Reh 1985)
Kryz	kryt1240	Nakh-Daghestanian	Suff. <i>-aR</i>	nominalizer	(Authier 2012)
Kusunda	kusu1250	Isolate	Suff. <i>-q</i>	unknown	(Watters 2006)
Kutenai	kute1249	Isolate	Suff. <i>-p</i>	unknown	(Morgan 1991)
Latin	lati1261	Indo-European	Suff. <i>-sc</i>	ingressive	(Inglese 2021)
Lezgian	lezg1247	Nakh-Daghestanian	Aux. <i>χun</i> ‘be’	verb ‘become’	(Haspelmath 1993)
Majang	maja1242	Surmic	Suff. <i>-(d)i^L</i>	unknown	(Joswig 2019)

Malay	mala1479	Austronesian	Pref. <i>bər</i>	verbalizer	(Karaj & Sansò forth.)
Maléku Jaíka	male1297	Chibchan	Suff. <i>-ti</i>	unknown	(Constenla Umaña 1998)
Mansi (Northern)	mans1258	Uralic	Suff. <i>-l</i>	unknown	(Rombandeeva 1973; Riese 2001)
Martuthunira	mart1255	Pama-Nyungan	Suff. <i>-npa</i>	verb ‘fall’	(Dench 1995)
Molale	mola1238	Isolate	Pref. <i>-taŋ</i>	verb ‘do’	(Pharris 2006)
Momu-Fas	fass1245	Balbai-Fas	Suff. <i>-ni/-nu</i>	unknown	(Honeyman 2017)
Mongolian	mong1331	Mongolic-Khitán	Suff. <i>-r</i>	unknown	(Kullmann & Tserenpil 2008; Janhunen 2012)
Mosetén-Chimané	mose1249	Isolate	Suff. <i>-ki</i>	verb ‘go’	(Sakel 2007, 2011)
Moskona	mosk1236	East Bird’s Head	Clitic = <i>ef</i>	spatial	(Gravelle 2010)
Ngaanyatjarra	ngaa1240	Pama-Nyungan	Suff. <i>-ri</i>	verb ‘become’	(McGregor 2013)
Ohlone, Southern	sout2986	Miwok-Costanoan	Suff. <i>-n(i)</i>	unknown	(Okrand 1977)
Orungu	orun1242	Atlantic-Congo	Suff. <i>-an</i>	reciprocal	(Ambouroué 2007)
Palula	phal1254	Indo-European	Suff. <i>-ŋ</i>	spontaneous	(Liljegren 2016)
Purepecha	pure1242	Tarascan	Suff. <i>-ra</i>	unknown	(Nava & Maldonado 2004)
Purepecha	pure1242	Tarascan	Suff. <i>-ku</i>	unknown	(Nava & Maldonado 2004)
Rama	rama1270	Chibchan	Suff. <i>-ting</i>	verb ‘happen’	(Grinevald 1990)
Romani, Vlax	vlax1238	Indo-European	Suff. <i>-uv</i>	verb ‘become’	(Wagner 2012)
Rotokas	roto1249	North Bougainville	Class A verbs	unknown	(Robinson 2011)
Salinan	sali1253	Isolate	Pref. <i>k-</i>	unknown	(Turner 1987)
Sandawe	sand1237	Isolate	Suff. <i>-ts̺</i>	unknown	(Steeman 2012)
Sandawe	sand1237	Isolate	Suff. <i>-ts̺</i>	unknown	(Steeman 2012)

Savosavo	savo1255	Isolate	Suff. <i>-za</i>	unknown	(Wegener 2012)
Shawi	chay1248	Cahuapanan	Pref. <i>-ya</i>	unknown	(Rojas Berscia 2013)
Siar-Lak	siar1238	Austronesian	Pref. <i>ta(k)-</i>	unknown	(Frowein 2011)
Sidaama	sida1246	Afro-Asiatic	Suff. <i>-am</i>	unknown	(Kawachi 2007)
Tamasheq	tama1365	Afro-Asiatic	Pref. <i>m-/n- /nvy-</i>	unknown	(Heath 2005)
Telefol	tele1256	Nuclear Trans New Guinea	Aux. <i>tébemin</i>	verb ‘become’	(Healey 1965)
Tigre	tigr1270	Afro-Asiatic	Aux. <i>bala</i>	verb ‘say, do’	(Cohen et al. 2002)
Totonac (Filomeno Mata)	filo1235	Totonacan	Pref. <i>ta-</i>	ingressive	(McFarland 2009)
Tzeltal	tzel1254	Mayan	Inf. <i><j></i>	passive	(Polian 2013)
Udihe	udih1248	Tungusic	Suff. <i>-ptA/-ktA</i>	unknown	(Nikolaeva & Tolskaya 2001)
Udihe	udih1248	Tungusic	Suff. <i>-kpi</i>	unknown	(Nikolaeva & Tolskaya 2001)
Vietnamese	viet1252	Austroasiatic	Aux. <i>bị</i>	verb ‘suffer’	(Simpson & Tâm 2013; Bruening & Tran 2015)
Warlpiri	warl1254	Pama-Nyungan	Suff. <i>-wanti</i>	verb ‘fall’	(McGregor 2013)
Wolaytta	wola1242	Ta-Ne-Omotiic	Suff. <i>-t</i>	unknown	(Wakasa 2008)
Yagaria	yaga1260	Nuclear Trans New Guinea	Aux. <i>ei-</i>	verb ‘hit’	(Renck 1975)
Yaqui	yaqu1251	Uto-Aztecan	Suff. <i>-te</i>	unknown	(Estrada Fernández et al. 2015)

Yaqui	yaqu1251	Uto-Aztecan	Suff. <i>-tu</i>	verb 'become'	(Estrada Fernández et al. 2015)
Yaul	maru1253	Keram	Pref. <i>na-</i>	unknown	(Barlow 2018)
Yecuatla Totonac	yecu1235	Totonacan	Suff. <i>-nan</i>	verbalizer	(MacKay 1999)
Yeyi	yeyi1239	Atlantic-Congo	Suff. <i>-aak</i>	unknown	(Seidel 2008)
Yimas	yima1243	Lower Sepik-Ramu	Suff. <i>-ara</i>	unknown	(Foley 1991)
Yucatec Maya	yuca1254	Mayan	CVVC pattern	unknown	(Martínez Corripio & Maldonado 2010)
Zaghawa	zagh1240	Saharan	Verb Class 1	unknown	(Jakobi et al. 2004)
Zapotec	zapo1437	Otomanguean	Pref. <i>d-</i>	auxiliary	(López Nicolás 2016)
Zenzontepec Chatino	zenz1235	Otomanguean	Pref. <i>y-</i>	unknown	(Campbell 2015)